Comparison of a *Brassica oleracea* Genetic Map With the Genome of *Arabidopsis thaliana*

Lewis Lukens,*[,1] Fei Zou,[†] Derek Lydiate,[‡] Isobel Parkin[‡] and Tom Osborn*

**Department of Agronomy, University of Wisconsin, Madison, Wisconsin 53711, [†]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599 and [‡]Agriculture and Agri-Food Canada, Saskatoon Research Centre, Saskatoon, Saskatchewan S7N 0X2, Canada*

Manuscript received July 1, 2002
Accepted for publication January 25, 2003

ABSTRACT

Brassica oleracea is closely related to the model plant, *Arabidopsis thaliana*. Despite this relationship, it has been difficult to both identify the most closely related segments between the genomes and determine the degree of genome replication within *B. oleracea* relative to *A. thaliana*. These difficulties have arisen in part because both species have replicated genomes, and the criteria used to identify orthologous regions between the genomes are often ambiguous. In this report, we compare the positions of sequenced Brassica loci with a known position on a *B. oleracea* genetic map to the positions of their putative orthologs within the *A. thaliana* genome. We use explicit criteria to distinguish orthologous from paralogous loci. In addition, we develop a conservative algorithm to identify collinear loci between the genomes and a permutation test to evaluate the significance of these regions. The algorithm identified 34 significant *A. thaliana* regions that are collinear with >28% of the *B. oleracea* genetic map. These regions have a mean of 3.3 markers spanning 2.1 Mbp of the *A. thaliana* genome and 2.5 cM of the *B. oleracea* genetic map. Our findings are consistent with the hypothesis that the *B. oleracea* genome has been highly rearranged since divergence from *A. thaliana*, likely as a result of polyploidization.

ONE major goal of plant biologists is to compare the genomic information available from model species to other, nonmodel species for which genetic maps are available. If genome structures are highly conserved, candidate genes in the model species that correspond to loci mapped in the nonmodel species can be quickly identified. In addition, differences between the nonmodel genome and model genome can be used to infer the frequency of genome duplications and rearrangements over time.

The genus Brassica is an excellent system with which to develop tools for genome comparison and to examine the divergence of genome structure. Brassica species are closely related to the model plant species, *Arabidopsis thaliana*. Both Brassica and Arabidopsis are classified within the same family, the Brassicaceae, and diverged ~20 MYA (Koch *et al.* 2000). *Brassica oleracea* (including broccoli, cabbage, cauliflower, brussel sprouts, and kale) and its related crop species (including *B. napus* and *B. rapa*) have also been extensively studied genetically, and several molecular maps for *B. oleracea* and other species within the genera have been published (*e.g.*, Kianin and Quiros 1992; Camargo *et al.* 1997; Bohuon *et al.* 1998; Lan and Paterson 2000).

Despite the close relationship between Brassica species and *A. thaliana*, whole-genome mapping studies have found that the order of loci in Brassica genetic maps is only infrequently similar to the order of homologous loci in the *A. thaliana* genome (Kowalski *et al.* 1994; Lagercrantz 1998; Lan *et al.* 2000). Comparative studies of smaller genomic intervals have revealed more evidence for collinearity, but extensive deletions and genome rearrangements are still evident. Cavell *et al.* (1998) reported shared marker order and content between a 7.5-Mbp region of *A. thaliana* chromosome 4 with *B. napus*, and Parkin *et al.* (2002) and Schranz *et al.* (2002) have observed a high degree of collinearity between *A. thaliana* chromosome 5 and three chromosomal regions of diploid Brassica species. The order of loci within one 10-cM region within *B. oleracea* is well conserved in *A. thaliana* (Ryder *et al.* 2001), and many genes within a 222-kb interval of *A. thaliana* chromosome 4 hybridize to the same or contiguous bacterial artificial chromosomes (BACs) in *B. oleracea* (O'Neill and Bancroft 2000). However, Ryder *et al.* (2001) found that many regions of the *B. oleracea* genetic map did not have a clear relationship to the *A. thaliana* genome, and O'Neill and Bancroft (2000) found that several genes within the 222-kb *A. thaliana* interval were not found in the homologous region of *B. oleracea*.

The conserved and rearranged regions between Brassica and Arabidopsis genomes have been interpreted in different ways, leading to fundamental disagreements about Brassica genome structure. Lagercrantz (1998) suggested that the base diploid Brassica genome evolved

[1]*Corresponding author:* Department of Plant Agriculture, Crop Science Bldg., University of Guelph, Guelph, ON N1G 2W1, Canada. E-mail: llukens@uoguelph.ca

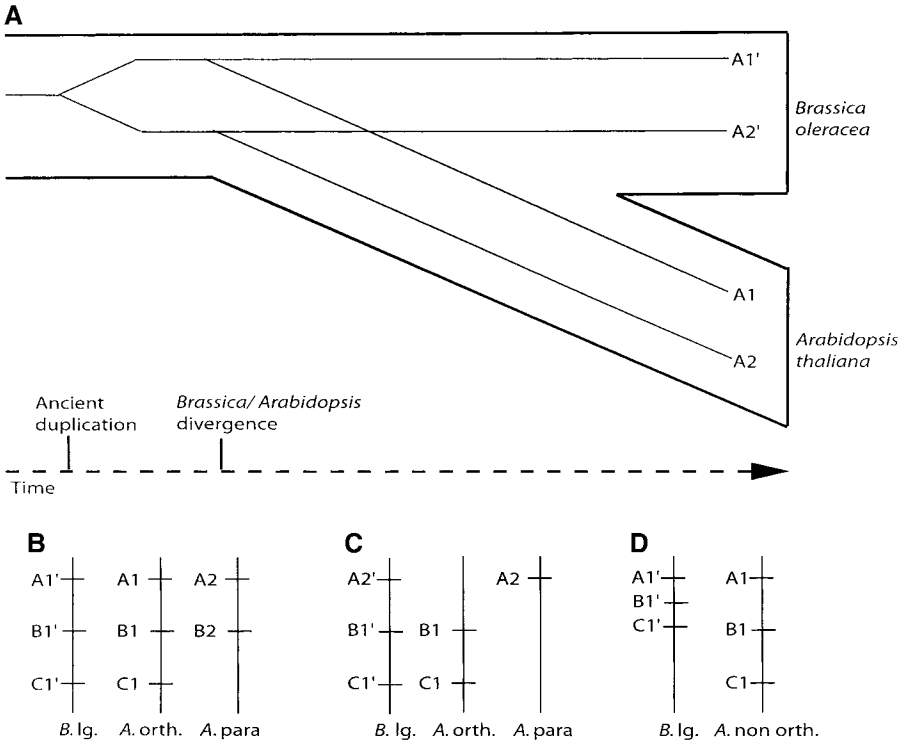

FIGURE 1.—Gene duplications and loosely linked markers complicate comparative mapping. (A) Hypothetical gene (thin line) and species (thick line) phylogeny for *Arabidopsis thaliana* and *Brassica oleracea*. In this scenario, a duplication event occurred prior to the divergence of both species (LYNCH and CONERY 2000; KOCH *et al.* 2001). (B) If Brassica sequences falsely detect *A. thaliana* sequences as orthologous, regions between the genomes (*i.e.*, A1′ and A2) can be falsely associated. (C) If intragenomic paralogous sequences are mapped in *B. oleracea*, a single linkage group may be incorrectly inferred to be orthologous with different regions of *A. thaliana*. (D) Brassica sequences that are closely linked may detect distantly linked sequences in *A. thaliana* that are not true orthologs.

from an ancient hexaploid with three highly rearranged *A. thaliana*-like genomes. LAN *et al.* (2000), supporting earlier cytogenetical studies (*e.g.*, HAGA 1938), found much stronger support for the hypothesis that the base Brassica genome is largely composed of duplicated regions.

Some of the difficulties in resolving intergenomic relationships and shared orthologous regions have arisen because genome replication confounds orthologous and paralogous relationships between loci (Figure 1A). The *A. thaliana* genome is partially duplicated (ARABIDOPSIS GENOME INITIATIVE 2000), and Brassica sequences can detect paralogous *A. thaliana* duplicates (PARKIN 2000). To identify the region from *A. thaliana* that most likely has the same gene content as a region of *B. oleracea*, one must identify the orthologous, not paralogous, loci within the *A. thaliana* genome (Figure 1B). Paralogy may further complicate genome comparisons because a probe may hybridize and be mapped to an ancient, paralogous locus within the Brassica genome. If one subsequently uses this probe for comparative mapping, this probe will likely identify its ortholog, and one may incorrectly associate two genome intervals. Although a single *B. oleracea* region may be orthologous to a single region within Arabidopsis, the probes that identified this region in *B. oleracea* may be orthologous to loci in different, duplicated regions within the *A. thaliana* genome, giving the appearance of (nonexistent) genome rearrangements (Figure 1C; I. PARKIN, unpublished data).

Finally, the lack of explicit criteria when evaluating putative orthologous regions and the potential for bias when making genome comparisons can lead to misclassification of intergenomic relationships (BENNETZEN 2000; GAUT 2001). In published comparisons between the Brassica and *A. thaliana* genomes, clear orthologous regions have been identified by the presence of several shared loci that are closely linked within both species. However, additional orthologous regions are inferred by a single marker or a small number of markers that are linked in both genomes but lie far from each other (Figure 1D; *e.g.*, RYDER *et al.* 2001). It is difficult to evaluate whether such associations are due to chance alone.

Here, we report on a comparison between a genetic map of *B. oleracea* and the *A. thaliana* genome using approaches that reduce the confounding effect of paralogous sequences. In addition, we developed an algorithm written in PERL that uses explicit criteria to identify orthologous regions and to establish their significance. Consistent with previous reports, we found evidence for substantial genomic replication in *B. oleracea* as compared to *A. thaliana* and found evidence that multiple chromosomal rearrangements have occurred since the species' divergence. However, we also found that the *B. oleracea* genetic map and the *A. thaliana* genome sequence share 34 significant, collinear regions. The average putative orthologous segment has 3.3 markers corresponding to 2.1 Mbp in *A. thaliana* and 7.1 cM in *B. oleracea*. In total, the significant regions identified in this study cover well over one-fourth of the *B. oleracea* genome. Of 22 previously published regions of predicted orthology, our algorithm identified 20, 17 of which were significant at $P < 0.05$.

Our data suggest three separate inferences. First, in general, published reports of collinear regions appear to have sampled highly conserved areas between the Brassica and *A. thaliana* genomes. Second, different interpretations of Brassica genome structure may have arisen because of different criteria used to define homologous regions between Brassica and *A. thaliana*. Finally, differences in the genomic arrangements between *A. thaliana* and *B. oleracea* appear to be due to the recent history of polyploidy in *B. oleracea*.

MATERIALS AND METHODS

***A. thaliana* sequence information source and *B. oleracea* map source:** The *B. oleracea* genetic map was developed by Bohuon *et al.* (1998) from a highly polymorphic cross between a double-haploid (DH) line of *B. oleracea* ssp. *italica* with a DH line of *B. oleracea* ssp. *alboglabra*.

The sequences of BACs used to assemble the *A. thaliana* genome sequence were downloaded from TIGR, http://www.tigr.org, on May 20, 2001. The number of nucleotides within all BACs totaled 132,101,284 bp. Dr. Eva Huala (Arabidopsis Information Resource) kindly provided the order of BACs and the estimated starting and ending position for each BAC within the *A. thaliana* genome on February 15, 2001. The nucleotide positions of BACs within the genome are estimates.

Sequencing and plasmid insert information: Brassica DNA mapping fragments were cloned into a pUC18-derived plasmid, pIJ2925, and two sequences were obtained for each clone. Most of these clones contain *Pst*I fragments of genomic DNA, although a few pW clones contain *Eco*RI genomic DNA fragments, and all have been used in mapping experiments. They are present in low-copy number within the Brassica genome with a mean of 1.8 polymorphic loci in *B. oleracea*. Sequencing reactions were performed using ABI Big-Dye Terminator cycle sequencing reagents. Reactions contained the ABI mix, ddH20, 500 ng of plasmid DNA, and 3.2 pmol of M13 forward or reverse primer to a final volume of 20 μl. The cycle-sequencing conditions were as follows: 25 cycles for 10 sec at 95°, 5 sec at 50°, and 4 min at 60°. Unincorporated nucleotides were removed by passing the reaction mixture through a Sephadex G-50 column. Sequence reactions were analyzed with an automated DNA sequencer (ABI model 377XL or 377-96) and base-pair calls were confirmed by visual inspection of chromatograms.

To identify molecular markers that had similar sequences but different names, we used blastn (Altschul *et al.* 1997) from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) to perform all pairwise comparisons between Brassica marker sequences. The following markers shared highly similar sequences: pW177 and pW148, pN3 and pW200, pN53 and pN96, and pW120 and pW101. If two or more similar markers were placed on the *B. oleracea* genetic map in the same or adjacent positions, one of the markers was removed from the analysis. Many of the marker sequences were similar to putative coding sequences within *A. thaliana*, and it is likely that a large portion of potentially orthologous noncoding sequences were omitted due to low blast scores (*e.g.*, Quiros *et al.* 2001). A rigorous classification of marker sequences as coding or noncoding from our preliminary sequence data was difficult because the *B. oleracea* genome likely contains a large number of pseudogenes relative to *A. thaliana* (*e.g.*, Quiros *et al.* 2001).

BLAST analysis: To identify BACs with nucleotide sequences similar to the Brassica query sequences, we used blastn. Low-complexity sequences were filtered in the blast analysis, and default values for cost (mismatch cost = −3.0), reward (match reward = 1.0), and word size (11 bp) were selected. The default gap opening penalty (5.0) and the gap extension penalty (2.0) were also selected. We recorded the bit score to evaluate sequence relationships. We did not align *A. thaliana* and Brassica sequences by eye in order to calculate additional sequence distances or other statistics. The number of marker sequences and the number of detected homologs in *A. thaliana* made such an approach impractical.

The results from the blastn analysis were parsed using a spreadsheet and short PERL scripts that we wrote for this purpose. From each "hit" to the *A. thaliana* BAC database by a Brassica query sequence, we retrieved the BAC name, the bit score, and the significance value. The nucleotide start position of the BAC was used as the approximate position of the Brassica query sequence in the *A. thaliana* genome. If both sequences from the same fragment detected the same BAC, only the highest scoring match was kept. Additional parsing was done to remove redundant data. If a query sequence had significant sequence similarity to BACs that overlapped (had overlapping base-pair intervals) or were immediately adjacent to each other (had shared beginning or ending nucleotide positions), a single BAC that was assigned the lower nucleotide position in the chromosome was recorded. This procedure would cause local/tandem duplications of a single gene within *A. thaliana* to be defined as a single locus. The raw and parsed data sets can be downloaded from http://www.plant.uoguelph.ca/faculty/llukens.

Collinearity analysis: "Conserved linkage" or "collinearity" (Ehrlich *et al.* 1997; Gaut 2001), the conservation of both synteny and order of orthologous loci between two species, can be used as a principle to compare the relatedness of genomes. We wrote PERL scripts to identify such collinear regions between *B. oleracea* and *A. thaliana* using two different definitions of collinearity. The first, "strict" definition defines two or more loci shared between genomes as collinear only if the loci are found on the same linkage group within each species and if the order of loci is the same in both species. The second, "general" definition was proposed by Gaut (2001). Under the general definition, collinear segments are defined as a series of uninterrupted markers within one chromosome [or linkage group (LG)], the "standard," that is found in a common orientation in the second chromosome (or LG), the "tester." In this case, the order of loci is not necessarily the same between the standard and the tester chromosome. In a comparison between *B. oleracea* LG 7 (designated O7) and *A. thaliana* chromosome (ch.) 5 (designated At5), the differences between the strict and general definitions can be seen (Figure 2A). Using the general definition, the identity of a collinear region may differ depending on which chromosome is the standard and which is the tester (compare Figure 2A to 2B).

Our method to assess the significance of an observed collinear region using both definitions is based on the approach outlined by Gaut (2001) but has major modifications (see discussion). If a collinear region with n shared markers was detected, it was scored. The scoring metric $S(n)$ is defined as the average distance (in centimorgans) between the loci within the collinear region: that is,

$$S(n) = (AR + B)/(2 \times (n - 1)).$$

A is the approximate distance (in kilobases) between the outermost loci detected within a collinear region in *A. thaliana*. The term R is the mean ratio of genetic distance (centimorgans) to physical distance (kilobases) for the *A. thaliana* chromosome on which the segment lies. R is 135/29,000, 97/17,463, 101/23,560, 125/22,140, 139/26,170 for chromosomes 1–5, respec-

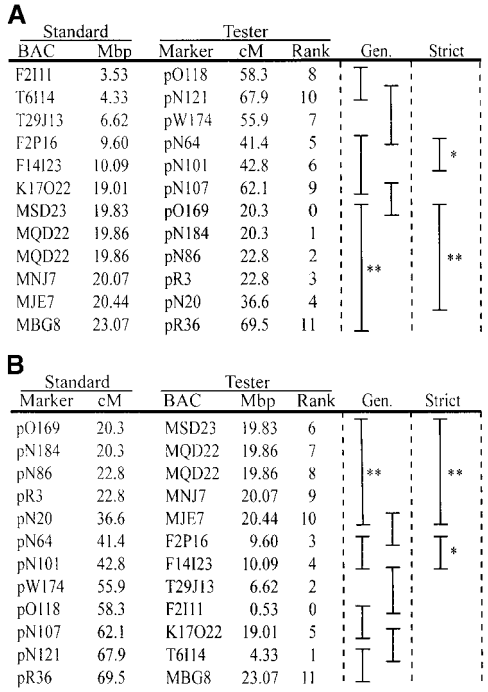

FIGURE 2.—Comparison of *Brassica oleracea* O7 and *Arabidopsis thaliana* At5 showing characteristics of the strict and general (Gen.) collinearity definitions. Collinear regions are identified by vertical lines. Nested collinear runs are identified and evaluated by the algorithm but for clarity are not shown here. (A) The physical order of the BACs within the "standard" *A. thaliana* chromosome is compared with the rank order of markers within the "tester" *B. oleracea* linkage group. (B) The genetic order of the markers within the standard *B. oleracea* linkage group is compared with the rank order of BACs within the tester *A. thaliana* chromosome. *, statistical significance at $P \leq 0.05$; **, statistical significance at $P \leq 0.01$.

tively (LISTER and DEAN 1993). B is the distance between markers in *B. oleracea* (in centimorgans).

To evaluate the probability that an observed collinear region arose by chance, an empirical permutation procedure was developed. In each permutation, markers from *B. oleracea* linkage groups were assigned randomly and uniformly to the *A. thaliana* genome. After the permutation, the scoring metric was calculated for each collinear region found in the permuted data set. If the permuted data set did not contain a collinear region of a given length, the metric was assigned a high value. The procedure was repeated 1000 times to obtain the expected distribution of scores for a particular length of a collinear region under the null hypothesis that collinear regions are due to chance association between loci. To obtain the P value for an observed collinear region, its distance metric was calculated and compared with the distribution of the metric scores for collinear regions of the same length between the same linkage group and chromosome generated from the permutations. The percentage of the scores from the permutations that were less than the observed score was defined as the P value for the observed collinear region.

The relative importance of the number of loci and the distance between loci in determining the significance of a collinear region is arbitrary. With our standards, $S(n)$ must be considerably lower than $S(n + 1)$ to be significant. For example, between O2 and At5, the 5% quantile of the scoring metric under the null distribution for collinear regions with two shared loci was 2.31 cM *vs.* 7.14 cM for three shared loci. Collinear regions with more than four shared loci were claimed to be significant at the 5% level.

The comparative data between *B. oleracea* and *A. thaliana* were parsed in two ways before testing for significant collinear regions. Both manipulations tended to increase the number of observed, collinear regions. First, if two or more markers had similarity to the same BAC in *A. thaliana* or shared the same centimorgan location in *B. oleracea*, we assigned the order of these markers or sequences relative to each other by eye. Second, in each pairwise comparison between a *B. oleracea* linkage group and *A. thaliana* chromosome, each locus on a *B. oleracea* linkage group was allowed a single position on the *A. thaliana* chromosome. Duplicates caused the algorithm to identify unlikely collinear regions. For example, if one linkage group of *B. oleracea* contained three closely linked loci of which two were recently duplicated, then this entire region would incorrectly be inferred to be collinear with a region defined by only two loci on an *A. thaliana* chromosome.

RESULTS

Identification of *A. thaliana* sequences putatively orthologous to *B. oleracea* marker sequences: The *A. thaliana* BAC database was queried with sequences from a total of 158 Brassica DNA probes using the nucleotide pattern-matching program blastn (ALTSCHUL *et al.* 1997), and 18,007 BACs had substantial similarity (bit score >32) to these query sequences. *A. thaliana* sequences detected within the initial database search may be derived from ancient, duplicated regions or ancient, paralogous gene family members within the *A. thaliana* genome (ARABIDOPSIS GENOME INITIATIVE 2000). To reduce the number of these matches, we first reasoned that Brassica query sequences would have low similarity to many *A. thaliana* sequences due to ancient events, but have high similarity to a much smaller number of *A. thaliana* sequences due to recent common ancestry. A truncated distribution of blastn scores >60 supports this hypothesis (Figure 3). At low blast scores (<82), the number of *A. thaliana* BACs similar to Brassica sequences begins to rise asymptotically. Second, the *A. thaliana* genome was duplicated ~45 million years prior to the divergence of *A. thaliana* and *B. oleracea* (KOCH *et al.* 2000; LYNCH and CONERY 2000). Thus, we also inferred that an *A. thaliana* sequence that is orthologous to a Brassica sequence should be more similar to the Brassica sequence than the mean duplicate *A. thaliana* sequences are to each other. Within the *A. thaliana* genome, the expected number of silent substitutions per silent site for duplicate coding sequences is 0.8 (LYNCH and CONERY 2000). We aligned our rough Brassica sequence with *A. thaliana* sequence, and we found that Brassica/*A. thaliana* coding sequence alignments with blast scores of 82 or above were highly similar at third position sites (data not shown). Thus, a Brassica query sequence and an *A. thaliana* BAC were considered putatively orthologous if they generated a blast score of 82 or higher in a pairwise comparison. If a BAC had a

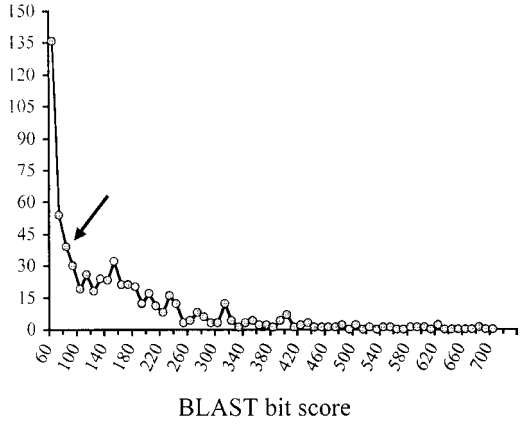

FIGURE 3.—Number of BACs detected by Brassica marker sequences at different BLAST bit scores. Only blastn scores ≥60 are plotted. The number of *Arabidopsis thaliana* BACs detected by markers rises asymptotically at lower BLAST scores. Sequence matches with scores <82 (identified by the arrow) were not included in our analysis. For plotting, one outlier with a bit score of 1104 was removed.

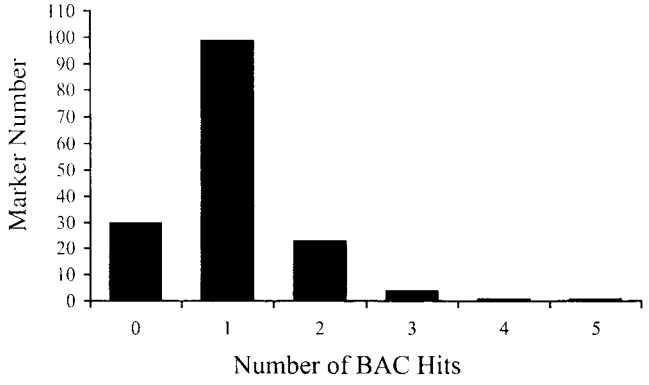

FIGURE 4.—Distribution of the number of *Arabidopsis thaliana* BACs detected by Brassica marker sequences with a blastn bit score >82.

score <82 when compared to a query sequence, it was removed from the analysis.

Of the 158 Brassica probes, sequences from 131 (83%) have significant similarity to one or more *A. thaliana* BACs. Four-fifths of the probes with significant similarity to *A. thaliana* have putative orthology to only a single BAC, and each probe has similarity to a mean of 1.4 loci within the *A. thaliana* genome. Only two probes have similarity to more than three BACs (Figure 4).

The majority of *B. oleracea* linkage groups are strongly associated with a single *A. thaliana* chromosome. Over one-half of the probes that mapped to five of the nine *B. oleracea* linkage groups were putatively orthologous to BACs within a single chromosome (O2 and At5; O4 and At2; O5 and At1; O8 and At1; O9 and At5; Table 1). Despite this general association, probes that map to each *B. oleracea* linkage group have putative orthologs throughout the *A. thaliana* genome, suggesting extensive chromosomal repatterning has occurred since the divergence of these two species (Table 1). In addition, sequences are not uniformly distributed across *A. thaliana* chromosomes, and the *P* value of the goodness-of-fit test was <0.05. On the basis of comparisons of total and expected numbers in Table 1, sequences with similarity to At5 are overrepresented in the *B. oleracea* genome, while sequences with similarity to At3 are underrepresented.

Significant collinear regions between *B. oleracea* and *A. thaliana*: Pairwise comparisons between each *B. oleracea* linkage group and *A. thaliana* chromosome show that several adjacent markers within a linkage group may correspond to adjacent markers within a single chromosome. The pairwise comparison between linkage groups and chromosomes reduces the confounding effect of markers mapped to paralogous regions within the *B. oleracea* genome (Figure 1C) because Brassica markers that have homology to different chromosomes are not simultaneously examined for collinearity.

Long collinear regions are shared between O1 and At4, O5 and At1, O3 and At5, as well as between several other linkage groups and chromosomes (Figure 5). Nonetheless, the relationship between many genomic regions remains ambiguous. Some markers may be closely linked in one genome but not in the other genome. For example, two loci >20 cM apart on O1 lie within 1 Mbp on At1 (Figure 5). Or, some markers closely linked in both genomes may clearly delineate a region of putative orthology, but another marker may not be closely linked to them. For example, two markers are tightly linked within O3 and At3 (both at 17.5 Mbp); a third marker is also tightly linked in *B. oleracea*, but it lies at 20 Mbp in *A. thaliana*.

The collinearity test identified all collinear, putatively orthologous regions with confidence. Using the strict definition or the general definition of collinearity with both species as testers (see MATERIALS AND METHODS), the algorithm aligned a total of 240 cM of the 872-cM *B. oleracea* genetic map, or 28% of the *B. oleracea* genome, to *A. thaliana* within 34 significant collinear regions at a $P < 0.05$ level of significance (Table 2). The algorithm identified regions that appeared collinear in the visual examination, regions such as the interval shared between O1 and At4. The largest segment of significant collinearity is the region in O5 that is putatively orthologous to At1. This region has 13 markers and spans 45.3 cM, corresponding to 7.4 Mbp within At1, and is significant at $P < 0.01$ (Table 2). Every *B. oleracea* linkage group has a significant collinear region on at least one *A. thaliana* chromosome (Table 2). Collinear regions contain an average of 3.3 markers corresponding to 2.1 Mbp in *A. thaliana* and 7.1 cM in *B. oleracea*. The distribution of the number of markers that define collinear regions is highly skewed because one-half of the regions have only two markers, while all but one of the remaining regions have three, four, or five markers. As a

TABLE 1

Number of *Arabidopsis thaliana* genome sequences detected from *Brassica oleracea* marker sequences

Brassica LG	No. markers[a]	Markers with hits[b]	Total hits[c]	At1	At2	At3	At4	At5	Mean loci/ marker
O1	24	19	28	3	5	4	9	7	1.5
O2	31	28	36	3	6	5	2	20	1.3
O3	54	42	56	8	11	12	9	16	1.3
O4	26	22	42	9	12	5	7	9	1.9
O5	28	26	31	17	2	8	2	2	1.2
O6	14	12	19	8	3	2	2	4	1.6
O7	33	27	39	7	7	4	8	13	1.4
O8	27	22	25	13	2	4	3	3	1.1
O9	36	31	44	9	2	3	6	24	1.4
Total	273	229	320	77	50	47	48	98	1.4
Expected[d]				80	54	64	48	74	

[a] Sequenced markers that were on the *B. oleracea* genetic map.
[b] Markers that were highly similar to an *A. thaliana* BAC (BLAST bit score >82).
[c] The number of BACs in the *A. thaliana* genome that are similar to the markers.
[d] Expected number of total hits if Brassica markers were distributed equally on separate *A. thaliana* chromosomes in proportion to their physical length.

result, the median collinear region length is 2.5 markers corresponding to 695,000 bp and 2.5 cM. Marker density may be an important factor in detecting collinear regions. O6 has the fewest number of markers and only a single significant collinear region.

The true length of each collinear region extends for some distance beyond the outermost loci detected here. An estimate of the true length of a particular segment can be made using the equation from Nadeau and Taylor (1984). With this correction, the mean estimated length of all significant collinear runs is 11.6 cM (median 5.1 cM) in *B. oleracea* and 3.8 Mbp (median

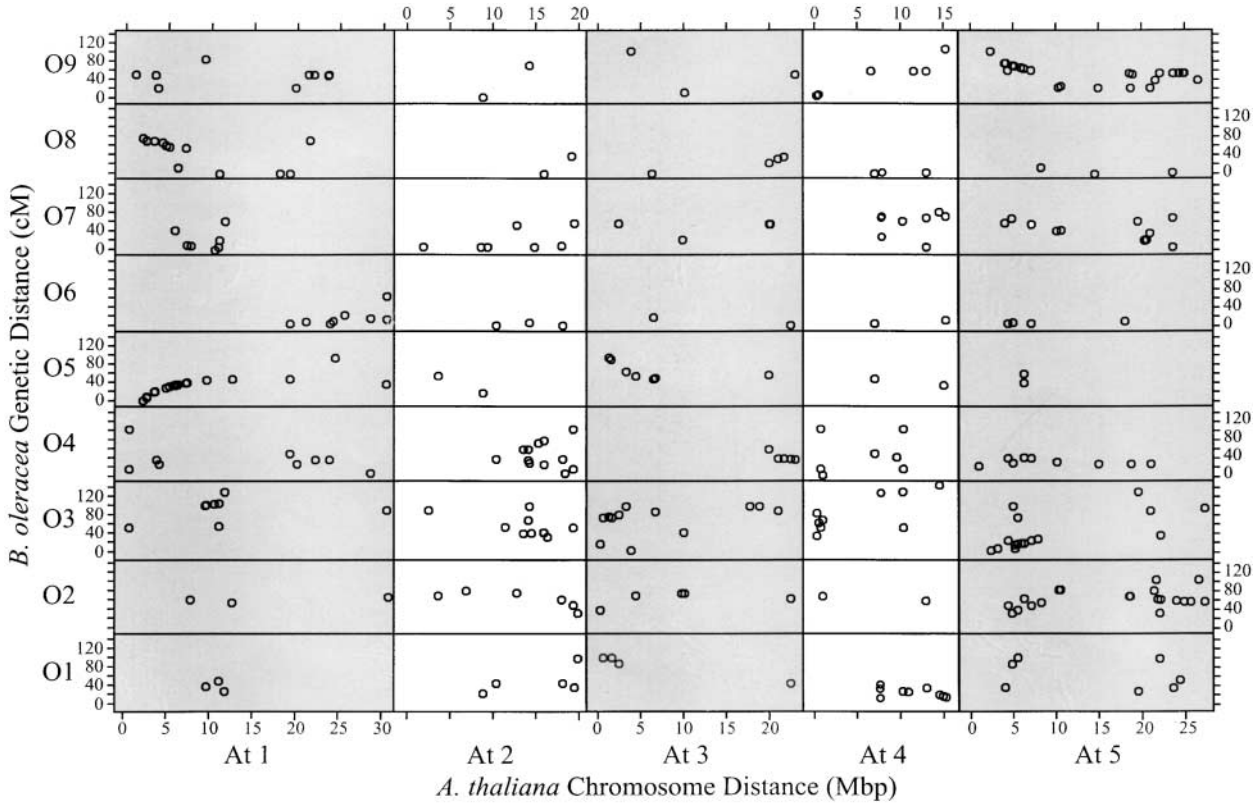

Figure 5.—Pairwise comparison between *Brassica oleracea* linkage groups and *Arabidopsis thaliana* chromosomes. The map position of *B. oleracea* probes (cM, *y*-axis) is plotted against the position of their putative orthologs within *A. thaliana* (Mbp, *x*-axis).

TABLE 2

Summary of significant collinear regions between *Brassica oleracea* and *Arabidopsis thaliana*

B.o. LG	*A.t.* Ch.	Run begin				Run end				S^a	N^b	Sig[c]
		Name	cM	bp	BAC	Name	cM	bp	BAC			
1	4	pW239	16.0	14.9	F4D11	pO43	21.3	14.2	T10C21	2.4	3	**
1	4	pW105	27.2	10.6	T6K22	pN107	28.5	9.9	T16H5	2.5	2	*
1	4	pN97	36.7	7.3	FCAALL	pN152	41.4	7.3	FCAALL	2.4	2	*
2	5	pW102	57.1	26.7	MXK3	pR34	61.7	21.6	MWD22	4.0	5	**
2	5	pW135	68.5	18.1	MMG4	pW218	68.5	18.0	MBD2	0.2	2	**
2	3	pR72	75.1	9.5	MPE11	pO125	75.1	9.8	MLJ15	0.7	2	*
2	5	pW167	82.7	9.8	F21A20	pO120	83.4	10.0	T1G16	0.9	2	*
3	5	pO111	16.3	4.7	T15N1	pN102	18.1	5.0	T20K14	1.8	2	*
3	5	pO160	19.9	5.4	MTG13	pW152	28.9	7.4	T6G21	3.2	4	**
3	2	pN22	41.9	14.2	F4P9	pN120	43.7	15.6	T2N18	1.7	4	**
3	3	pR85	75.4	0.3	F1C9	pO12	81.8	2.2	F17A9	3.6	3	**
3	3	pW188	90.1	20.7	T5N23	pO172	100.0	17.4	F12M12	6.1	3	*
3	1	pW146	102.0	9.0	F2J7	pN107	131.0	11.2	T19E23	5.0	5	**
4	1	pW137	25.3	19.6	F12M16	pO106	34.1	23.4	F2K11	6.6	3	*
4	3	pO145	34.1	22.7	F27H5	pN59	57.5	19.6	T18N14	4.6	5	**
4	2	pN66	57.5	13.8	T32F6	pO98	57.5	13.2	F7F1	1.6	2	*
5	1	pN21	0.0	1.7	T20M3	pN47	45.3	9.1	F28B23	3.1	13	**
5	3	pO128	47.0	6.2	MRC8	pN148	49.4	6.4	K24M9	1.6	2	*
5	3	pN215	53.6	4.1	MGH6	pO153	53.6	4.1	MGH6	0.0	2	**
6	5	pN180	4.8	3.9	MXC9	pR64	6.0	4.4	MXE10	2.1	2	*
7	2	pW186	6.6	8.3	F19F24	pW194	6.6	9.0	F5H14	2.1	2	*
7	5	pO169	20.3	19.8	MSD23	pN20	36.6	20.4	MJE7	2.4	5	**
7	5	pN64	41.4	9.6	F2P16	pN101	42.8	10.1	F14I23	2.0	2	*
7	3	pW104	55.8	19.7	ATEM1	pW174	55.9	19.8	T25B15	0.4	2	**
7	4	pO29	57.2	2.1	F3E22	pN97	57.9	7.3	FCAALL	0.4	2	**
8	4	pR54	0.7	6.7	T20K18	pR36	3.1	12.7	F10M23	9.0	3	*
8	3	pN168	1.3	6.0	MKP6	pW205	36.7	21.4	T5P19	16.9	4	*
8	1	pO159	14.0	5.7	F17F16	pW138	14.0	5.7	F6I1	0.1	2	**
8	1	pW123	59.0	4.8	F7A19	pO92	71.8	2.1	F10K1	18.7	5	**
9	4	pR116	0.0	0.0	F6N15	pN213	2.0	0.2	F6N23	1.6	2	*
9	1	pO106	49.2	23.4	F2K11	pN173	51.2	21.1	T8L23	3.2	3	**
9	5	pO168	54.8	23.9	MHM17	pW240	54.8	24.4	MQJ2	0.5	4	**
9	5	pN180	60.7	6.6	T20D1	pR64	71.3	4.4	MXE10	2.7	5	**
9	5	pO118	76.6	3.5	F2I11	pO7	77.2	3.6	F15N18	0.6	2	**

B.o., *B. oleracea*; *A.t.*, *Arabidopsis thaliana*.

[a] The estimated average distance in centimorgans between markers within a collinear region (the scoring metric).

[b] The number of markers shared between *A. thaliana* and *B. oleracea* within a collinear region.

[c] Sig, statistical significance. $*P < 0.05$; $**P < 0.01$.

1.2 Mbp) in *A. thaliana*, accounting for 45% of the *B. oleracea* genome.

Seven of the nine *B. oleracea* linkage groups have regions that are collinear with more than one *A. thaliana* chromosome (Figure 6), again suggesting that numerous translocations have occurred since the divergence of the two species. In addition, closely linked regions within a single *B. oleracea* linkage group may be collinear to different segments within the same *A. thaliana* chromosome, suggesting intrachromosomal rearrangements. On O7, for example, one collinear region (pN64–pN101) spans 9.6–10.1 Mbp within At5. This region lies adjacent to another region defined by pO169 and pN20 on the genetic map, but putative orthologs to these markers span 19.8–20.4 Mbp within At5 (Figure 6). Because neither interval corresponds to the known duplicated segment within At5 (Arabidopsis Genome Initiative 2000), this region likely defines an intrachromosomal rearrangement that has occurred since the divergence of *B. oleracea* and *A. thaliana*. Additional collinear segments that are contiguous on the genetic map and hybridize to the same *A. thaliana* chromosome are found on O3, O5, and O9. However, within these linkage groups, neighboring markers on the genetic map have similarity to the same general area within the *A. thaliana* genome, indicating either a local rearrangement of markers or that markers were misplaced on the *B. oleracea* genetic map.

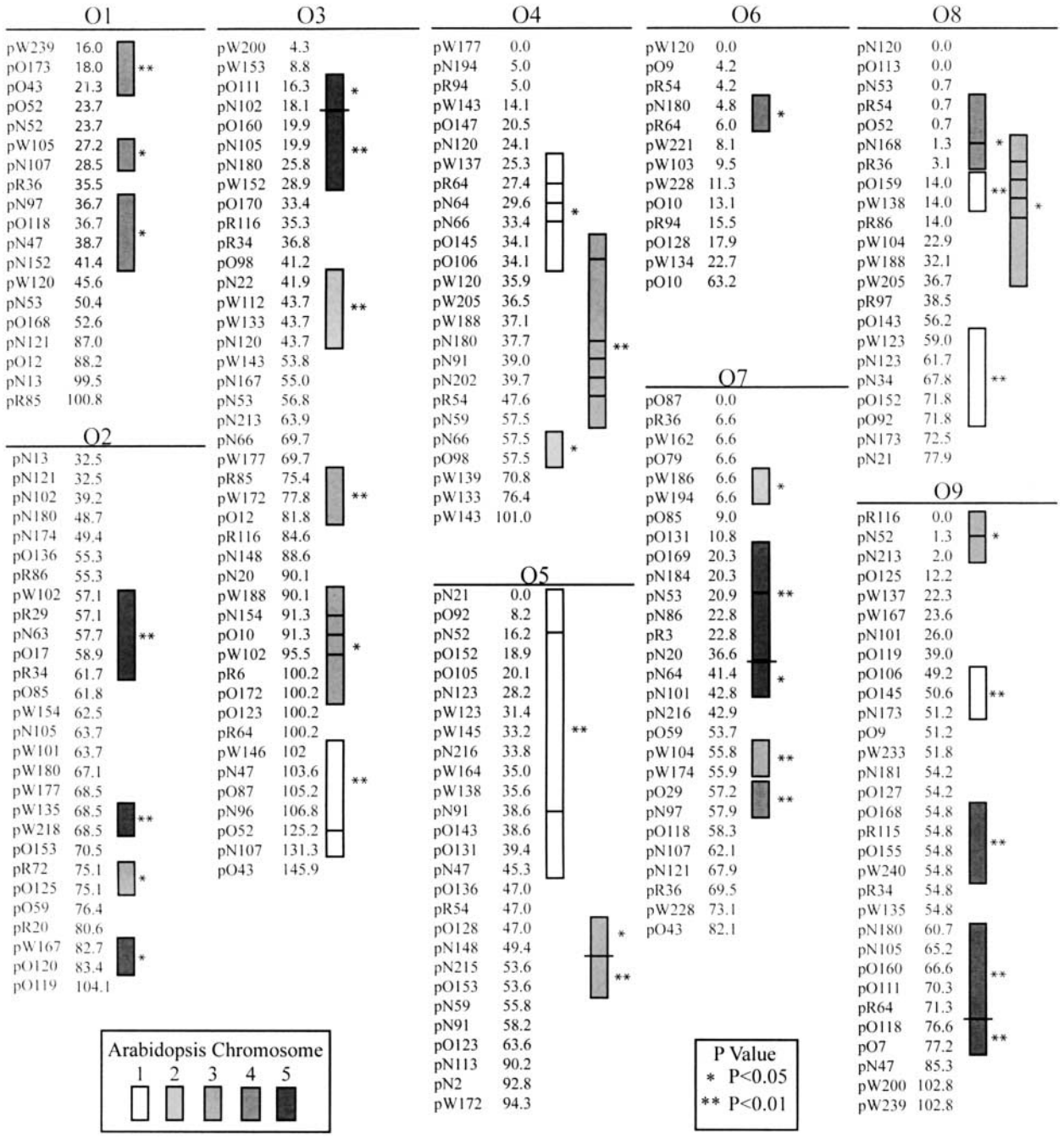

FIGURE 6.—Positions of significant collinear regions between the *Brassica oleracea* genetic map and the *Arabidopsis thaliana* genome sequence. Boxes enclose putative collinear runs. A line within the box indicates that the corresponding *B. oleracea* genetic marker was not found within the *A. thaliana* chromosome on which the collinear run was identified. A line through a box indicates that a new collinear run from the same chromosome was initiated. *, significance levels of $P < 0.05$; **, significance levels of $P < 0.01$.

The relationship between the *B. oleracea* genetic map distances and the *A. thaliana* genome sequence can be inferred from regions within both genomes with a large number of shared markers. In most such regions (*e.g.*, O4 and At3, O9 and At5), the genetic distance between each marker on the *B. oleracea* linkage group is roughly proportional to the nucleotide distance between each marker on the *A. thaliana* chromosome (Table 2, Figure 5), indicating that the relationship between the genetic distance in *B. oleracea* and physical distance in *A. thaliana* is linear for these regions. A notable exception is the longest region of putative homology shared between O5 and At1. As would be expected if the frequency of recombination per nucleotide is greater at chromosome ends than toward the centromere, the change in centimorgans *vs.* the change in base pairs is greater

between markers at the end of the linkage group than between markers in the center. At the top of the linkage group, markers are distantly linked in *B. oleracea* but lie fairly close together on the *A. thaliana* chromosome. The slope progressively declines toward the center of the *B. oleracea* linkage group as the markers approach the centromere of At1 located at ~14 Mbp (Arabidopsis Genome Initiative 2000). Markers that are closely linked in the center of O5 lie very far apart on At1.

DISCUSSION

Sequence similarity and orthology between *B. oleracea* and *A. thaliana*: Identification of orthologous regions between two genomes depends on correctly identifying orthologous sequences that are shared between the genomes (Figure 1). Orthology and the significance of a relationship between two sequences are very difficult to infer (Deken 1983; Yuan *et al.* 1998), especially when searching a partially duplicated genome such as *A. thaliana* (Arabidopsis Genome Initiative 2000). *A priori* assumptions about the extent of sequence similarity that indicates orthology are problematical for several reasons. For example, two sequences may be considered orthologous if they are matched with a high blast score or significance level (*e.g.*, Wolfe and Shields 1997). However, if the cutoff score is not high enough, sequence from one genome could have significant similarity to anciently duplicated paralogs within the second genome (see Figure 1B). In contrast, one can assume that a query sequence from one genome is orthologous only to the other genome's sequence to which it is most similar (*e.g.*, Grant *et al.* 2000). In this case, only a single sequence from a recent duplication would be detected although both duplicates have the same ancestral relationship to the query sequence.

In this study, we define criteria by which to accept or reject *A. thaliana* sequences as likely orthologs to Brassica sequences on the basis of the distribution of blastn scores and the *a priori* knowledge of expected nucleotide differences between paralogous sequences within the *A. thaliana* genome (Lynch and Conery 2000). We estimated that the point at which underlying similarity between two sequences is high enough to indicate recent common ancestry corresponds to a blastn score of 82. Over four-fifths of Brassica sequences with putative orthologs in the *A. thaliana* genome are similar to only one BAC, indicating that the criteria eliminated anciently duplicated sequences within the Arabidopsis genome from the analysis. A few Brassica markers did detect loci in duplicated regions of the *A. thaliana* genome. For example, pN102 lies on O2 at 39.2 cM and has similarity to a duplicated region between the top of At3 and At5 at ~5 Mbp (Arabidopsis Genome Initiative 2000). Without an outgroup, it is difficult to infer whether the small number of *A. thaliana* duplicated loci detected as orthologous to Brassica sequences arose before or after the divergence of *A. thaliana* from *B. oleracea*.

The degree of conservation inferred between genomes depends on the cutoff score used to define putative orthology. In this analysis, if sequence alignments with low blastn scores were considered orthologous, additional collinear regions would be identified. For example, the marker pW105 is positioned at 27 cM on *B. oleracea* LG 1 and is similar to a sequence at 10 Mbp in *A. thaliana* ch. 4 with a score of 119. Under less stringent criteria, pW105 could be considered collinear with pR36, a marker positioned at 35 cM on LG 1 and aligned to *A. thaliana* ch. 4 at 12 Mbp with a blastn score of 74. Similarly, if only sequence alignments with high blastn scores were considered orthologous, some collinear regions reported in this analysis, such as the region defined by markers pO125 and pR72 on LG 2, would not be identified. Marker pO125 is positioned at 75 cM and is aligned to a sequence at 9 Mbp on *A. thaliana* ch. 3 with a blastn score of 234. Marker pR72 is positioned at 75 cM and is aligned to a sequence at 9 Mbp on *A. thaliana* ch. 3 with a blastn score of 92. Finally, Brassica markers may be similar to a region of the *A. thaliana* genome to which other linked Brassica markers have no similarity. These singleton hits increase as the stringency for orthology is reduced.

Statistical test of collinearity: Probes may be closely linked in one genome while their orthologs are distantly linked in another (Figures 1D and 5). In such a case, it is difficult to determine if the intervening region between the probes is orthologous. This determination often depends on the researcher's judgment, and orthologous regions may be identified using very liberal criteria (Bennetzen 2000). The collinearity method presented here identifies collinear regions between two genomes and calculates the probability that these regions occurred by chance alone. Our method is based on the analysis outlined by Gaut (2001). Like Gaut's method, this method uses both the number of shared probes between two linkage groups and the distances between those probes as criteria to evaluate a metric of collinearity. In addition, this method, like Gaut's, does not define an *ad hoc* number of probes or an *ad hoc* distance between probes that is required for a collinear region to be significant. Rather, the metric of an observed collinear region is compared against the metrics of collinear regions that would be expected to occur by chance. Finally, in each pairwise comparison between linkage groups, only one copy of a locus and its putative ortholog may lie on each linkage group. The pairwise comparisons and the requirement of a single shared marker reduce the misleading effect of Brassica probes that map to paralogous loci within the *B. oleracea* genome (see Figure 1C).

Despite these similarities, this collinearity approach differs from Gaut's approach in several major respects. First, the metric to evaluate collinear regions integrates

physical distance from one genome with genetic distance from the other. Second, the collinearity program evaluates collinear regions nested within longer regions. This characteristic is important because a long collinear region shared between linkage groups or chromosomes may not be significant, but nested collinear regions within this long interval may be significant. Third, markers that lie at the end of one collinear region and at the beginning of another collinear region are evaluated in both positions (*i.e.*, pN121 in Figure 2). Fourth, although both models use empirical permutations to estimate statistical significance, in each permutation of our model, markers from each linkage group of one species are randomly and uniformly assigned a position within the second species' genome. Thus, to evaluate significance, the algorithm compares the metric of each collinear region with the expected distribution of the metric under the null hypothesis that collinear regions are due to chance association between loci across genomes. In contrast, in Gaut's permutation, markers from each linkage group are randomly and uniformly assigned to positions within every linkage group of the second species. Finally, the algorithm can evaluate genomes for collinear regions using both general and strict definitions of collinearity.

Under the general definition of collinearity, the identity of collinear regions may depend on which chromosome or linkage group is a tester and which is a standard in the comparison (Figure 2). In this study, if a collinear region was detected when a chromosome was used as both a standard and a tester, it was recorded. Collinear regions identified using the strict definition of collinearity were also recorded. Under the general definition of collinearity, many collinear regions that were significant when using one chromosome as the tester but not significant when using the same chromosome as the standard did not correspond to orthologous intervals. For example, the markers pO169 and pR36 define a significant, collinear region between O7 and At5 where At5 is the standard (Figure 2A). However, within this region, six markers lie between the last two markers (pN20 and pR36) within the *B. oleracea* genetic map, and all six markers have putative orthologs to different regions within At5 (Figure 2B).

Our method of scoring collinear regions, inferring collinearity, and establishing significance is based on several assumptions. First, when calculating the scoring metric, the ratio between genetic distance and physical distance is assumed to be constant over each *A. thaliana* chromosome. Nonetheless, because the genetic distance and physical distance are known to vary across *A. thaliana* chromosomes (Copenhaver *et al.* 1998), each chromosome is assigned a different centimorgan-to-base-pair conversion factor. Second, in the *B. oleracea* genetic map, probes are assumed to be in the correct order. Because mapping errors do occur, the collinearity algorithm is conservative and likely underestimates the number and length of collinear regions. Third, in each permutation, Brassica probes are uniformly distributed within each Arabidopsis chromosome. However, low-copy sequences tend not to be associated with genomic regions flanking the centromeres (Arabidopsis Genome Initiative 2000). Finally, we assumed, like Gaut (2001), that a large number of collinear probes shared between linkage groups is strong evidence for an orthologous region.

Several published reports of genome comparisons between *A. thaliana* and Brassica species were compared to our results to test the utility of the collinearity algorithm and the effect of these assumptions. If previously identified collinear regions were among the significant collinear regions identified by our analysis, then we judged that our algorithm has high utility. We used two distinct criteria to infer if previously reported homologous/orthologous regions were among those identified by the collinearity analysis. Under the first criterion, two requirements had to be met. The region of *A. thaliana* identified in our analysis must overlap with a region detected in the previous analysis. In addition, the *B. oleracea* region identified in our analysis must lie on a linkage group that was likely homologous or homeologous to the Brassica linkage group reported previously. A second, less stringent criterion was used if relationships between linkage groups could not be inferred because of experimental design (*i.e.*, O'Neill and Bancroft 2000). In this case, we inferred that an observed collinear region corresponded to previous reports if a region of *A. thaliana* within a significant collinear region in our analysis overlapped with the region of the *A. thaliana* genome previously reported as similar.

The collinearity test identified almost all putative homologous/orthologous regions reported in previous genetic and/or physical comparative mapping studies (Table 3). Out of 22 previously reported regions of similarity, the collinearity test identified 20. In addition, the test identified several regions that have not been reported previously and could be targets of future studies. Of the 22 previously characterized regions, 17 were collinear and significant at $P < 0.05$, and 3 additional regions were collinear but not significant, reflecting the conservative nature of the test. Two published regions were not identified by our analysis. Our collinearity analysis assigned somewhat more than one-quarter of the *B. oleracea* genetic map to putatively orthologous regions within *A. thaliana* (see below). The fact that our analysis identified over three-quarters of the previously reported collinear regions suggests both that comparative studies have not randomly sampled the Brassica genome for regions of collinearity and that long, conserved regions may be overrepresented in the literature.

Analysis of the ancestral Brassica genome: Several studies have provided evidence that the base Brassica genome is highly duplicated (Haga 1938; Röbbelen 1960; Truco *et al.* 1996; Lan *et al.* 2000). However,

TABLE 3

Comparison between published homologous regions and collinear regions

Reference species	Reference Brassica position	Reference *A.t.* (ch.)	Reference *A.t.* (Mbp)[a]	Citation	*B.o.* LG	*A.t.* position (Mbp)	Collinear
B. oleracea	C7	1	27.4	Quiros *et al.* (2001)	O3[b]	30	No
B. oleracea	C4	4	12.3	Quiros *et al.* (2001)	O9[b]	12.6	Yes[e]
B. oleracea	BAC contig	4	7.3	O'Neill and Bancroft (2000)	O1[c]	7.3	Yes
B. oleracea	BAC contig	4	7.3	O'Neill and Bancroft (2000)	O7[c]	7.3	Yes
B. oleracea	BAC contig	4	7.3	O'Neill and Bancroft (2000)	O3[c]	7.3	Yes[e]
B. oleracea	BAC contig	5	20.0	O'Neill and Bancroft (2000)	O7[c]	19.8	Yes
B. oleracea	BAC contig	5	20.0	O'Neill and Bancroft (2000)	O2[c]	18.1	Yes
B. oleracea	BAC contig	5	20.0	O'Neill and Bancroft (2000)	O9[c]	23.9	Yes
B. oleracea	C1	3	19.6	Sadowski *et al.* (1996)	O7[c]	19.8	Yes
B. oleracea	C6	3	19.6	Sadowski *et al.* (1996)	O7[c]	18.5	Yes
B. nigra	G5	5	7.0	Lagercrantz (1998)	O2[d]	9.8	Yes
B. nigra	G8	5	7.0	Lagercrantz (1998)	O3[d]	5.7	Yes
B. nigra	G2	5	7.0	Lagercrantz (1998)	O9[d]	6.6	Yes
B. nigra	G8	2	14.0	Lagercrantz (1998)	O3[d]	15.6	Yes
B. nigra	G1	2	14.0	Lagercrantz (1998)	O4[d]	13.2	Yes
B. nigra	G6	2	14.0	Lagercrantz (1998)	O4[d]	13.8	Yes
B. nigra	G7	3	1.6	Lagercrantz (1998)	O1[d]	1.3	No
B. nigra	G7	3	1.6	Lagercrantz (1998)	O3[d]	2.0	Yes
B. nigra	G1	3	8.1	Lagercrantz (1998)	O5[d]	6.4	Yes
B. napus	N1, N11	4	8.0	Cavell *et al.* (1998)	O1[d]	7.3	Yes
B. napus	N3, N17	4	8.0	Cavell *et al.* (1998)	O7[d]	7.3	Yes
B. napus	N8, N18	4	8.0	Cavell *et al.* (1998)	O3[c]	7.3	Yes[e]

A.t., *A. thaliana*; *B.o.*, *B. oleracea*.

[a] The position of a previously reported region within *A. thaliana* that was similar to Brassica was estimated one of two ways. An *A. thaliana* genetic marker(s) was identified within the region of Brassica-Arabidopsis homology, and the BAC location of the marker was acquired using Mapviewer (www.arabidopsis.org). Or, Brassica markers within the homologous region were sequenced and the homologous *A. thaliana* physical position was identified using blastn.

[b] Homology between the cited linkage group and our linkage group was determined using the relationships inferred by Hu *et al.* (1998).

[c] Our *B. oleracea* linkage group and a *B. oleracea* BAC were considered homologous if both had similarity to the same position within the *A. thaliana* genome.

[d] Homeology was determined using the relationships inferred by Lagercrantz and Lydiate (1996) and/or using markers shared in both studies.

[e] The region was detected by the collinearity algorithm as collinear, but it was not significant at $P < 0.05$.

Lagercrantz and Lydiate (1996) proposed that the Brassica genome is largely triplicated, and this hypothesis has been supported by subsequent studies (Lagercrantz 1998; O'Neill and Bancroft 2000). This discrepancy is likely due in large part to different criteria for inferring genome redundancy. In a visual examination of our comparative data, we found 15 segments within the *A. thaliana* genome that are similar but not necessarily collinear to more than one region of the *B. oleracea* genome (Table 4). Of the 15 regions, 5 are present in two copies, 8 are present in three copies, and 2 are present in more than three copies (Table 4). Because such a large number of *A. thaliana* regions are found in triplicate in *B. oleracea*, this visual inspection is consistent with the hypothesis that the base diploid Brassica species have evolved from an ancient hexaploid. However, the collinearity test offers little evidence for triplication. The test finds that only 3 of the 15 *A. thaliana* intervals are triplicated within *B. oleracea* (Table 4). Inferences about whole-genome relationships can greatly differ depending on the criteria used to infer those relationships. A future comparison between *A. thaliana* and a very high-density genetic or physical map of Brassica will be able to resolve conclusively the question of ancient hexaploidy in Brassica.

Replication and rearrangements within the *B. oleracea* genome: Our analysis does show that numerous chromosomal translocations, deletions, and duplications differentiate *A. thaliana* from *B. oleracea*. For example, different regions of O3 have high similarity to all five *A. thaliana* chromosomes (Figure 3). In addition, Brassica markers often have putative orthologs within a region of the *A. thaliana* genome to which other linked Brassica markers have no similarity (*i.e.*, between O5 and At4), suggesting that relatively short sequences have transferred between chromosomes. Such events have been identified in the recent evolutionary history of humans (O'Keefe and Eichler, 2000) and in plants (R. Schmidt,

TABLE 4

Putative duplicated regions of similarity shared between *A. thaliana* and *B. oleracea*

	Interval (Mbp)		*B. oleracea* linkage group								
A.t. ch.	Start	End	1	2	3	4	5	6	7	8	9
1	1	5				+	−			−	
1	7	12		+	−		−		+		
1	20	24				−		+			−
2	10	17			−	−			+		
3	1	5	+		−		−				
3	4	9		−			−				
3	17	21			−				−		
3	21	23				−				−	
4	0	1		+	+						−
4	6	10	−						−	−	
4	12	15	−						+		
5	1	3			+						−
5	3	7	+	+	−	+		−	+		−
5	10	18		−		+					+
5	20	29		−					−		−

− indicates that a region was identified using the collinearity test and by visual examination of the data. + indicates that a region was identified by visual examination alone.

personal communication). Finally, a putative intrachromosomal duplication can be seen on O4. Several Brassica probes hybridize to both the top and the bottom of O4 and have putative orthologs within the same position in *A. thaliana* At1, At2, and At4 (Figure 5). This duplication may be shared by all Brassica species; Lagercrantz and Lydiate (1996) identified an intrachromosomal duplication within the homeologous *B. nigra* G6 linkage group.

Although inferences about the rate of change within two genomes require a third, outgroup genome for comparison, we nonetheless suggest that most rearrangements reported here occurred within the Brassica lineage since its divergence from the Brassica-Arabidopsis common ancestor. If the *A. thaliana* genome has had many duplications since its divergence from the Brassica-Arabidopsis common ancestor, one would expect that more than one region of the *A. thaliana* genome would correspond to a single region within *B. oleracea.* This occurs infrequently (Figure 4), and only two regions of the *B. oleracea* genome were found by the collinearity test to be associated with more than one region in *A. thaliana* (Figure 6). Likewise, if many large deletions have occurred in the *A. thaliana* genome since its divergence from the Brassica-*Arabidopsis* common ancestor, large portions of the *B. oleracea* genome would not be similar to regions within the *A. thaliana* genome. With the possible exception of the top of O6, all regions of *B. oleracea* have similarity to *A. thaliana* regions (Table 1, Figure 5). The concept that genome change has occurred predominantly during Brassica evolution (as opposed to Arabidopsis) is also consistent with both linkage and microcollinearity studies. The genomes of the genus Brassica and that of *A. thaliana* likely diverged ~20–24 MYA (Koch *et al.* 2000, 2001). Although *Capsella rubella* and *A. thaliana* are <33% more similar at the nucleotide level than are *B. oleracea* and *A. thaliana* (Koch *et al.* 2000; Warwick and Black 1997), comparisons of several kilobases between *A. thaliana* and *C. rubella* have revealed perfect collinearity of genes (Rossberg *et al.* 2001). In contrast, comparisons of intervals between *A. thaliana* and Brassica species have been characterized by numerous gene rearrangements or deletions (*e.g.*, Grant *et al.* 1998; Quiros *et al.* 2001). In addition, restriction fragment length polymorphism markers spanning most of *A. thaliana* chromosome 4 are almost perfectly collinear with two *C. rubella* linkage groups (Acarkan *et al.* 2000). This conservation of marker order contrasts with the extensive rearrangements between the Brassica and Arabidopsis genomes. This study does not rigorously substantiate the hypothesis that extensive genome rearrangements occurred on the Brassica lineage because Capsella is more closely related to Arabidopsis than is Brassica. As a result, it is possible (although unlikely) that the common ancestor of Arabidopsis/Capsella and Brassica had a Brassica-like genome structure.

The large number of differences between the genomes may in part have been the result of the recent polyploidization of *B. oleracea.* Extensive repatterning of a genome subsequent to polyploidization has been suggested to account for the scattered, duplicate segments within the yeast genome (Wolfe and Shields 1997), and such repatterning has been observed within

the plant paleopolyploids *Zea mays* (Helentjaris *et al.* 1988; Gaut 2001) and *A. thaliana* (Arabidopsis Genome Initiative 2000). In both Brassica and wheat (Song *et al.* 1995; Liu *et al.* 1998; Shaked *et al.* 2001), polyploidization has been shown to be accompanied by rapid genome change. Nonetheless, genome evolution has been proposed to occur rapidly in diploids as well as polyploids (Brubaker *et al.* 1999). The structure of an outgroup genome such as *Aethionema grandiflora* (Galloway *et al.* 1998) could be used to estimate the rate of change on the Brassica and Arabidopsis lineages and test the correlation between polyploidy and genome change.

We thank the Arabidopsis Genome Initiative for open access to the sequence data. Financial support was provided by a National Science Foundation Postdoctoral Fellowship in Bioinformatics to L.L. and a National Science Foundation grant to T.O.

LITERATURE CITED

Acarkan, A., M. Rossberg, M. Koch and R. Schmidt, 2000 Comparative genome analysis reveals extensive conservation of genome organization for *Arabidopsis thaliana* and *Capsella rubella*. Plant J. **23:** 55–62.

Altschul, S., T. Madden, A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–812.

Bennetzen, J., 2000 Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell **12:** 1021–1029.

Bohuon, E. J. R., L. D. Ramsay, J. A. Craft, A. E. Arthur, D. F. Marshall *et al.*, 1998 The association of flowering time quantitative trait loci with duplicated regions and candidate loci in *Brassica oleracea*. Genetics **150:** 393–401.

Brubaker, C. L., A. H. Paterson and J. F. Wendel, 1999 Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome **42:** 184–203.

Camargo, L. E. A., L. Savides, G. Jung, J. Nienhuis and T. C. Osborn, 1997 Location of the self-incompatability locus in an RFLP and RAPD map of B. oleracea. J. Hered. **88:** 57–60.

Cavell, A. C., D. Lydiate, I. Parkin, C. Dean and M. Trick, 1998 Collinearity between a 30cM segment of *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome. Genome **41:** 62–69.

Copenhaver, G., W. Browne and D. Preuss, 1998 Assaying genome-wide recombination and centromere functions with *Arabidopsis* tetrads. Proc. Natl. Acad. Sci. USA **95:** 247–252.

Deken, J., 1983 Probabilistic behavior of longest common subsequence length, pp. 359–362 in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. Kruskal. Addison-Wesley, Reading, MA.

Ehrlich, J., D. Sankoff and J. Nadeau, 1997 Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics **147:** 289–296.

Galloway, G. L., R. L. Malmberg and R. A. Price, 1998 Phylogenetic utility of the nuclear gene arginine decarboxylase: an example from *Brassicaceae*. Mol. Biol. Evol. **15:** 1312–1320.

Gaut, B., 2001 Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Res. **11:** 55–66.

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

Grant, M., J. McDowell, A. Sharpe, M. de Torres Zabala, D. Lydiate *et al.*, 1998 Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **95:** 15843–15848.

Haga, T., 1938 Relationship of genome to secondary pairing in *Brassica*. Jpn. J. Genet. **13:** 277–284.

Helentjaris, T., D. Weber and S. Wright, 1988 Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphism. Genetics **118:** 353–356.

Hu, J., J. Sadowski, T. C. Osborn, B. S. Landry and C. F. Quiros, 1998 Linkage alignment from four independent *Brassica oleracea* RFLP maps. Genome **41:** 226–235.

Koch, M., B. Haubold, and T. Mitchell-Olds, 2000 Comparative evolutionary analysis of *chalcone synthase* and *alcohol dehydrogenase* loci in *Arabidopsis*, *Arabis*, and related genera. Mol. Biol. Evol. **17:** 1482–1498.

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

Kowalski, S., T. Lan, K. Feldmann and A. Paterson, 1994 Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveal 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 fustions and frequent rearrangements. Genetics **150:** 1217–1228.

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

Lan, T., and A. Paterson, 2000 Comparative mapping of quantitative trait loci sculpting the curd of *Brassica oleracea*. Genetics **155:** 1927–1954.

Lan, T., T. Delmonte, K. Reischmann, J. Hyman, S. Kowalski *et al.*, 2000 An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res. **10:** 776–788.

Lister, C., and C. Dean, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J. **4:** 745–750.

Liu, B., J. M. Vega and M. Feldman, 1998 Rapid genomic changes in newly synthesized amphiploids of *Triticum* and *Aegilops*. II. Changes in low-copy coding DNA sequences. Genome **41:** 535–542.

Lynch, M., and J. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science **290:** 1151–1155.

Nadeau, J., and B. Taylor, 1984 Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA **81:** 814–818.

O'Keefe, C., and E. Eichler, 2000 The pathological consequences and evolutionary implications of recent human genetic duplications, pp. 29–46 in *Comparative Genomics*, edited by D. Sankoff and J. H. Nadeau. Kluwer Academic Publishers, Dordrecht, The Netherlands.

O'Neill, C., and I. Bancroft, 2000 Comparative physical mapping of the genome of *Brassica oleracea* var. alloglabra that are homeologous to sequenced regions of chromosome 4 and 5 of *Arabidopsis thaliana*. Plant J. **23:** 233–243.

Parkin, I., 2000 Unraveling crucifer genomes through comparative mapping, pp. 425–437 in *Comparative Genomics*, edited by D. Sankoff and J. H. Nadeau. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Parkin, I., D. J. Lydiate and M. Trick, 2002 Assessing the level of collinearity between *Arabidopsis thaliana* and *Brassica napus* for *A. thaliana* chromosome 5. Genome **45:** 356–366.

Quiros, C. F., F. Grellet, J. Sadowski, T. Suzuki, G. Li *et al.*, 2001 Arabidopsis and Brassica comparative genomics: sequence, structure and gene content in the *ABI1-RPS2-Ck1* chromosomal segment and related regions. Genetics **157:** 1321–1330.

Röbbelen, G., 1960 Beiträge zur analyse des *Brassica* genomes. Chromosoma **11:** 205–228.

Rossberg, M., K. Theres, A. Acarkan, R. Herrero, T. Schmitt *et al.*, 2001 Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and *Capsella* genomes. Plant Cell **13:** 979–988.

Ryder, C. D., L. B. Smith, G. R. Teakle and G. J. King, 2001 Contrasting genome organisation: two regions of the *Brassica oleracea*

genome compared with collinear regions of the *Arabidopsis thaliana* genome. Genome **44:** 808–817.

Schranz, M. E., P. Quijada, S. Sung, L. Lukens, R. Amasino *et al.*, 2002 Characterization and effects of the replicated flowering time gene *FLC* in *Brassica rapa*. Genetics **162:** 1457–1468.

Shaked, H., K. Kashkush, H. Ozkan, M. Feldman and A. Levy, 2001 Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. Plant Cell **13:** 1749–1759.

Song, K., P. Lu, K. Tang and T. Osborn, 1995 Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc. Natl. Acad. Sci. USA **92:** 7719–7723.

Truco, M. J., J. Hu, J. Sadowski and C. F. Quiros, 1996 Inter- and intra-genomic homology of the *Brassica* genomes: implications for their origin and evolution. Theor. Appl. Genet. **93:** 1225–1233.

Warwick, S. I., and L. D. Black, 1997 Phylogenetic implications of chloroplast DNA restriction site variation in subtribes Raphaninae and Cakilinae (*Brassicaceae*, tribe Brassicieae). Can. J. Bot. **75:** 960–973.

Wolfe, K., and D. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387:** 708–713.

Yuan, Y., O. Eulenstein, M. Vingron and P. Bork, 1998 Toward detection of orthologues in sequence databases. Bioinformatics **14:** 285–289.

Communicating editor: O. Savolainen