

Partial Shotgun Sequencing of the *Boechera stricta* Genome Reveals Extensive Microsynteny and Promoter Conservation with *Arabidopsis*^{1[W]}

Aaron J. Windsor^{2*}, M. Eric Schranz², Nataša Formanová, Steffi Gebauer-Jung, John G. Bishop, Domenica Schnabelrauch, Juergen Kroymann, and Thomas Mitchell-Olds²

Max-Planck-Institut für chemische Ökologie, D-07745 Jena, Germany (A.J.W., M.E.S., N.F., S.G.-J., D.S., J.K., T.M.-O.); and Washington State University, School of Biological Sciences, Vancouver, Washington 98686 (J.G.B.)

Comparative genomics provides insight into the evolutionary dynamics that shape discrete sequences as well as whole genomes. To advance comparative genomics within the Brassicaceae, we have end sequenced 23,136 medium-sized insert clones from *Boechera stricta*, a wild relative of *Arabidopsis* (*Arabidopsis thaliana*). A significant proportion of these sequences, 18,797, are nonredundant and display highly significant similarity (BLASTn e-value $\leq 10^{-30}$) to low copy number *Arabidopsis* genomic regions, including more than 9,000 annotated coding sequences. We have used this dataset to identify orthologous gene pairs in the two species and to perform a global comparison of DNA regions 5' to annotated coding regions. On average, the 500 nucleotides upstream to coding sequences display 71.4% identity between the two species. In a similar analysis, 61.4% identity was observed between 5' noncoding sequences of *Brassica oleracea* and *Arabidopsis*, indicating that regulatory regions are not as diverged among these lineages as previously anticipated. By mapping the *B. stricta* end sequences onto the *Arabidopsis* genome, we have identified nearly 2,000 conserved blocks of microsynteny (bracketing 26% of the *Arabidopsis* genome). A comparison of fully sequenced *B. stricta* inserts to their homologous *Arabidopsis* genomic regions indicates that indel polymorphisms >5 kb contribute substantially to the genome size difference observed between the two species. Further, we demonstrate that microsynteny inferred from end-sequence data can be applied to the rapid identification and cloning of genomic regions of interest from nonmodel species. These results suggest that among diploid relatives of *Arabidopsis*, small- to medium-scale shotgun sequencing approaches can provide rapid and cost-effective benefits to evolutionary and/or functional comparative genomic frameworks.

The genomes of higher plants are dynamic entities and their evolutionary histories have been influenced by duplication, deletion, rearrangement, transposition, and changes in ploidy. As such it is not surprising that significant investments have been made toward the development of *Arabidopsis* (*Arabidopsis thaliana*), a species with a modest haploid genome composed of five chromosomes, approximately 125 Mb, and relatively little repetitive DNA, as a model genetic/genomic system. Beyond the superficial simplicity of the *Arabidopsis* genome, the species offers other characteristics that make it amenable to laboratory study, including self compatibility, rapid generation time, and responsiveness to culture conditions and transformation. These characteristics have translated themselves into an extensive history of both classical- and molecular-genetic application, the complete sequenc-

ing of this genome (*Arabidopsis* Genome Initiative, 2000), the accumulation of transcriptome data, and the development of a multitude of genomic resources. However, while highly informative, the resources built upon *Arabidopsis* cannot adequately address all aspects of plant biology.

In the postgenomic era, comparative analyses between related genomes have proved invaluable. Animal and fungal research communities have invested heavily in the establishment of comparative genomic resources (*Genomes OnLine Database v2.0* provides a comprehensive list of completed, proposed, and ongoing genome-sequencing projects; <http://www.genomesonline.org/>). The dividends of these investments include: better assessments of conserved microsynteny and colinearity; improved annotation and ab initio gene prediction; and the identification of novel genes, cis-regulatory sequences, and noncoding RNAs. Within the plant biology community, the development and application of comparative genomic approaches have also met with success in maize (*Zea mays*), rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), and other agronomically important members of Poaceae.

While many agriculturally important crop and weed species (Dietz et al., 1999; Meekins et al., 2001; Bleeker, 2003; Fumanal et al., 2004; Bleeker and Matthies, 2005; Durka et al., 2005) are found within the Brassicaceae, the Brassicaceae research community lags with regard to availability of interspecific comparative genomic

¹ This work was supported by the Max Planck Society.

² Present address: Duke University, Department of Biology, Box 91000, Durham, NC 27708-0338.

* Corresponding author; e-mail aaron.windsor@duke.edu; fax 919-613-8177.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Aaron J. Windsor (aaron.windsor@duke.edu).

^[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.105.073981.

vector and quality trimming (Fig. 2). In total, 33.5 Mb of high-quality *B. stricta* sequence were obtained. Analysis of the nucleotide content of the *B. stricta* sequences indicates that, like Arabidopsis (Arabidopsis Genome Initiative, 2000), the *B. stricta* genome is biased toward A-T (60%) versus G-C (40%) content. These end sequences serve as molecular identifiers for the inserts in the SAD12.4 sequence-indexed library (see "Materials and Methods") and, unless otherwise stated, are the primary dataset for subsequent analyses.

Identification of *B. stricta*-Arabidopsis Homologous Genomic Regions

We identified 25,879 *B. stricta* end sequences from the sequence-indexed library that display significant similarity (BLASTn e-value $\leq 10^{-30}$) to Arabidopsis genomic regions. Of these, 18,797 sequences were non-redundant within the context of the library and lacked significant similarity to annotated repetitive elements, rRNA genes, and organellar DNAs. We refer to this as the informative sequence set.

From the informative sequence set, 12,668 sequences corresponded to both sequencing reads from 6,334 inserts (paired end sequences; Fig. 2) and the other 6,129 sequences were classified as solo end sequences (Fig. 2). All informative sequences were mapped to Arabidopsis chromosomal pseudomolecules using the physical position of the Arabidopsis nucleotide homologous to the 5'-most *B. stricta* nucleotide in a given BLASTn high-scoring pair (HSP; Fig. 3; Supplemental Fig. 1). Approximately 15.5 Mb of nonredundant *B. stricta* sequence with highly significant similarity to Arabidopsis genomic sequences have been identified.

Nearly half (9,034) of the *B. stricta* end sequences in the informative set also have significant BLASTn hits (e-value $\leq 10^{-10}$) to annotated Arabidopsis coding

sequences (CDSs), with an average identity of 90.3% among all nonredundant HSPs (Supplemental Data 1). As expected, the number of CDS hits per Arabidopsis chromosome corresponds to the physical size of the chromosome (Table I; Supplemental Data 1).

Conservation of Microsynteny in *B. stricta* and Arabidopsis

The 12,668 paired end sequences (6,334 inserts) with similarity to Arabidopsis genomic sequences were analyzed to identify end sequences with conserved microsyntenic relationships relative to homologous Arabidopsis genomic regions. This analysis was performed with syntenyFinder.py that recognizes four syntenic categories (described in Fig. 4): syntenic, tight colinear, physically linked, and unlinked. From the starting collection of inserts, 4,691 (74.1%) qualified as having conserved synteny relative to Arabidopsis, while only 13.5% of inserts fell within the categories of tight colinear or linked (Fig. 4). The 1.2% of *B. stricta* end sequences scored as tight colinear (Fig. 4) indicated that the relevant Arabidopsis homologs show conserved colinearity and physical proximity, but diverge with regard to relative orientation. The physical sizes of the Arabidopsis chromosomal intervals bracketed by *B. stricta* end sequences with conserved microsynteny (Supplemental Data 2) are distributed as shown in Figure 5. In a one-tailed *t* test assuming unequal variances, the mean length of these Arabidopsis intervals (11,949 bp) differs significantly from the average *B. stricta* insert size (13,187 bp; $P < 0.005$), indicating that Arabidopsis genomic regions are, on average, smaller than homologous *B. stricta* genomic regions.

The *B. stricta* inserts displaying conserved microsynteny were subsequently organized into virtual

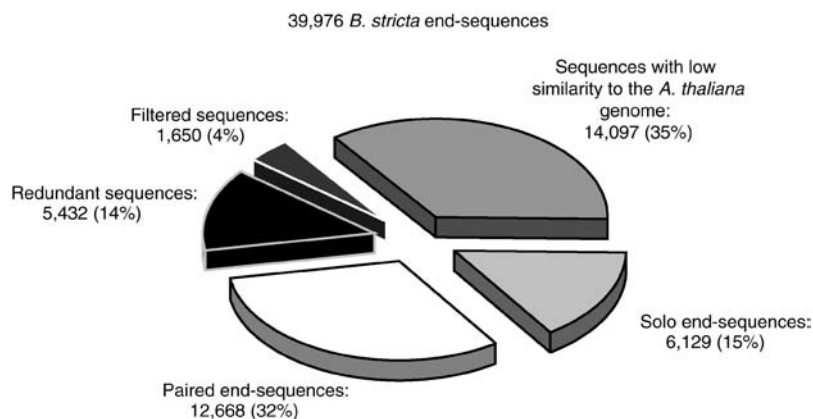


Figure 2. Summary of *B. stricta* sequence-indexed library end sequencing. Redundant sequences are end sequences that have been excluded from further analysis as they represent duplicated inserts (i.e. library amplification artifacts). Filtered sequences display significant similarity to repetitive DNA species or organellar genomes. Sequences with low similarity may be similar, but have failed to meet our significance threshold (e-value $\leq 10^{-30}$). The two remaining categories are comprised of *B. stricta* end sequences with highly significant similarity (e-value $\leq 10^{-30}$) to Arabidopsis genomic regions. The designation "paired end sequences" indicates that both T3 and T7 sequencing reads are available for a given insert; solo end sequences indicates that that only one sequencing read is available for a given insert, the second read having been placed into one of the initial four categories.

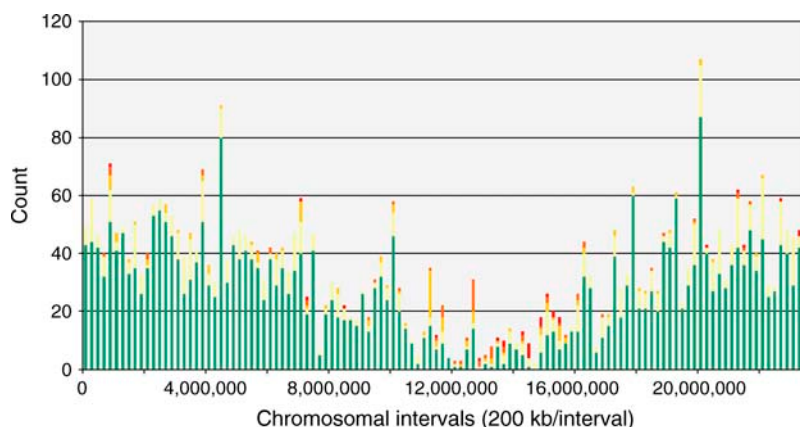


Figure 3. Distribution of nonredundant SAD12.4 sequence-indexed library end-sequence homologies across chromosome III (NC_003074.4) of Arabidopsis. The frequency of end sequences with homology to a given physical interval have been plotted along the 200 kb intervals comprising the Arabidopsis chromosome III pseudomolecule. *B. stricta* sequences are placed according to the physical position of the Arabidopsis nucleotide that is homologous to the 5'-most *B. stricta* nucleotide of a given blast HSP. While *B. stricta* end sequences with multiple Arabidopsis homologs are indicated, *B. stricta* end sequences are only mapped according to the most significant HSP. Color coding indicates the copy number of homologous sequences in the Arabidopsis genome: green, 1×; yellow, 2×; light orange, 3×; orange, 4×; and red, 5×.

microsyntenic blocks. These blocks are comprised of either singleton inserts or multiple, overlapping inserts as suggested by their similarity to the Arabidopsis genome. In total, 1,974 blocks have been identified, bracketing approximately 31.5 Mb of the Arabidopsis genome (Table II; Supplemental Data 3).

To verify the homology of the *B. stricta* sequence-indexed clones identified as displaying conserved microsynteny to their corresponding Arabidopsis genomic regions, two approaches were taken. In the first approach, eight *B. stricta* inserts were randomly selected from the distribution presented in Figure 5 and sequenced completely. Microsynteny and identity are conserved in all regions examined, with disruptions being attributed to indel polymorphisms between the two species (Fig. 6, A and C; Supplemental Fig. 2). In comparisons where there is a substantial difference in the sizes of the *B. stricta* genomic region versus the Arabidopsis homolog, the differences arise from large indel polymorphisms (>5.0 kb). Near-isometric regions contain multiple small indel polymorphisms that occur in a complementary fashion in both species.

For the second approach, 21 *B. stricta* sequence-indexed clones were specifically targeted for complete sequencing. End-sequence similarities for these clones suggested that their inserts should contain *B. stricta* genomic regions that are homologous to Arabidopsis loci involved in insect/pathogen resistance and flowering time. Every insert was correctly identified by syntenyFinder.py as a homolog of the relevant Arabidopsis genomic region (data not shown). As an example, the *B. stricta* class I chitinase region is presented in Figure 6B. The *B. stricta* region shown is composed of two overlapping, sequence-indexed inserts that were predicted by syntenyFinder.py to be members of a virtual microsyntenic block.

Orthology versus Paralogy

The *B. stricta* end sequences with significant similarity to annotated Arabidopsis CDSs were cross-referenced against the Arabidopsis duplication annotations of K. Wolfe (http://wolfe.gen.tcd.ie/athal/all_results/). In total, we identified 7,160 nonredundant *B. stricta* hits to the Arabidopsis CDSs duplicated by polyploidization within the last 24 to 40 million years (Blanc et al., 2003; Table III). Among these sequences, 72.3% had significant hits to both Arabidopsis homologs, 23.8% hit only one Arabidopsis homolog of a given gene pair, and 3.9% of the *B. stricta* end sequences showed significant similarity to members of extended gene families (Table III; Supplemental Data 4). The *B. stricta* end sequences of the latter group, where paralogs can arise through mechanisms other than genome-wide duplication, were excluded from further analysis.

To more closely investigate orthologous/paralogous relationships between *B. stricta* and Arabidopsis, we analyzed a subset of the above data (Table III, selected dataset; Supplemental Data 4). This dataset, which consists of 1,440 *B. stricta* end sequences, includes only

Table I. *B. stricta*/Arabidopsis CDS homologs

CHR, Chromosome.	
Total Arabidopsis CDS Homologs Identified ^a :	9,034 ^b
CHR I CDS sequences:	2,325
CHR II CDS sequences:	1,413
CHR III CDS sequences:	1,884
CHR IV CDS sequences:	1,340
CHR V CDS sequences:	2,072

^aFiltered *B. stricta* sequence-indexed clone dataset; e-value of BLASTn hits $\leq 10^{-10}$. ^bFor a given Arabidopsis CDS, only the most significant BLASTn hit is represented.

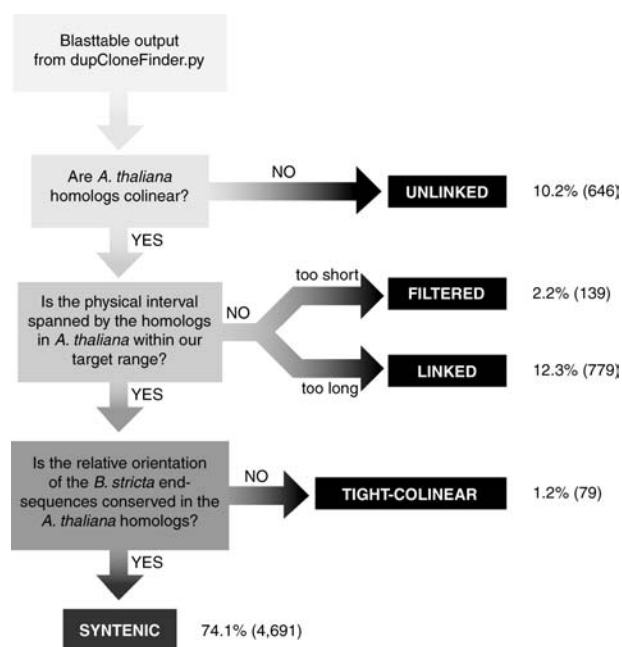


Figure 4. Flowchart of the algorithm used by syntenyFinder.py to identify *B. stricta* sequence-indexed inserts whose end sequences display microsynteny relative to the Arabidopsis genome. Output from dupCloneFinder.py for 6,334 inserts with paired end sequences is fed to syntenyFinder.py (top). End-sequence pairs then move down the vertical axis and are tested for the criteria indicated. If the end-sequence pair satisfies a given criterion, the pair continues down the vertical axis toward the designation SYNTENIC. If a given criterion is not satisfied, the end-sequence pair moves along the horizontal axis to the designation indicated. The designation FILTERED indicates that the physical length of a region in Arabidopsis identified by an end-sequence pair was less than twice the average read length of all *B. stricta* end sequences; LINKED indicates that the Arabidopsis region is greater than 50 kb. Percentage and total number (in parentheses) of *B. stricta* inserts placed in each designation are indicated. Descriptions of both dupCloneFinder.py and syntenyFinder.py can be found in Supplemental Text 1.

data from *B. stricta* sequences where similarity is detected to a least one member of a given Arabidopsis gene pair at an $e\text{-value} \leq 10^{-90}$. The calculation of $\Delta_{\log e\text{-value}}$ for a given *B. stricta* end sequence (see "Materials and Methods") yields an integer ranging from -170 to 0 . Lower values of $\Delta_{\log e\text{-value}}$ indicate greater differences in the $e\text{-values}$ of the BLASTn hits to the members of a given Arabidopsis gene pair and improve our ability to distinguish between candidates for orthology and close paralogs. The frequency distribution of $\Delta_{\log e\text{-value}}$ (Fig. 7) shows that a majority (98%) of Arabidopsis gene pairs have $\Delta_{\log e\text{-value}}$ scores < -5 , indicating more than five orders-of-magnitude difference in the significance of the *B. stricta* end-sequence BLASTn hits to the two Arabidopsis paralogs.

Comparison of Upstream Noncoding Sequences

B. stricta and Arabidopsis genomic regions upstream to homologous CDSs were analyzed using the

UntransID.py program. The analysis, which proceeds in two phases, was performed on *B. stricta* sequences from the sequence-indexed library dataset as well as sequences generated from a *B. stricta* small-insert library (see "Materials and Methods"). In the first phase, all *B. stricta* sequences with significant similarity to Arabidopsis CDSs were screened to identify sequences that traverse the translation initiation codon of the homologous Arabidopsis CDS and contain a minimum of 500 bp of sequence 5' to the presumptive initiation codon. Redundancies and *B. stricta* sequences with homology to Arabidopsis CDSs displaying complex paralogous relationships were purged during the screening process. At the conclusion of this phase of the analysis, 657 *B. stricta* sequences were deemed sufficiently informative for comparison to Arabidopsis upstream regions (Supplemental Data 5).

During the second phase of the analysis, the 657 *B. stricta* sequences identified in phase 1 were aligned to homologous Arabidopsis upstream regions (The Arabidopsis Information Resource, Arabidopsis sequence datasets; <ftp://ftp.arabidopsis.org/Sequences/>) using the algorithm of Needleman and Wunsch (1970). The portions of the *B. stricta* sequences with identity to Arabidopsis CDS were masked by UntransID.py in the alignments and did not contribute to the alignments or their quality scores. In 541 alignments (82.4%), the score of the alignment met or exceeded our alignment quality threshold; these alignments were categorized as informative. Among all alignments (noninformative alignments being treated as 0.0% identity), the mean identity of Arabidopsis upstream regions to their *B. stricta* homologs was 58.8%. When noninformative alignments are excluded from the calculation, this value rises to 71.4% (Supplemental Data 5). Breaks from identity are associated not only with mismatches, but also with the presence of short indel polymorphisms;

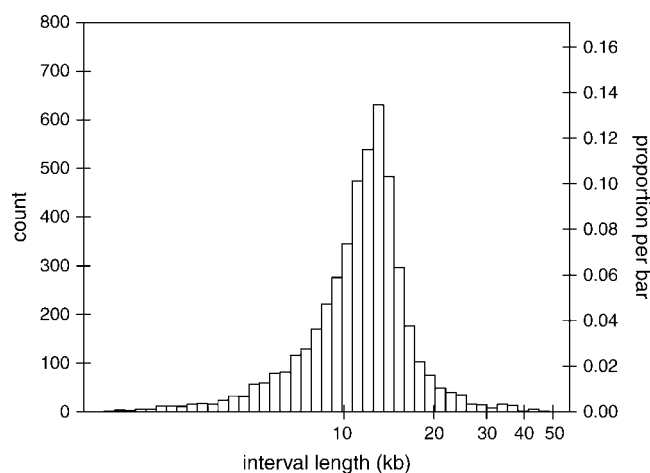


Figure 5. Frequency of *B. stricta* inserts versus the physical size of the corresponding microsyntenic regions in Arabidopsis. The mean physical size of the *B. stricta* λ -library inserts is 13,187 bp ($n = 27$; $sd = 1,716$ bp). The mean Arabidopsis interval length is 11,949 ($n = 4,691$; $sd = 4,637$ bp).

Table II. *Microsyntenic blocks shared between B. stricta and Arabidopsis*

CHR, Chromosome.

Arabidopsis Chromosome	Physical Size ^a	No. of Syntenic Blocks	Median Length of Syntenic Blocks	Range Length		Physical Coverage ^c	Percent Coverage ^b
				Shortest ^b	Longest ^b		
	<i>bp</i>		<i>bp</i>	<i>bp</i>		<i>bp</i>	
CHR I	30,432,563	514	13,145	1,675	114,255	8,169,580	26.8%
CHR II	19,705,359	278	12,480	2,323	61,260	4,175,476	21.2%
CHR III	23,470,805	424	14,223	1,728	73,324	7,235,484	30.8%
CHR IV	18,585,042	297	13,690	2,555	60,279	4,956,175	26.7%
CHR V	26,992,728	461	12,966	1,816	52,686	6,946,122	25.7%
Totals	119,186,497	1,974	–	–	–	31,482,837	26.4%

^aBased on the Arabidopsis chromosome pseudomolecules. ^bThe shortest blocks are comprised of individual *B. stricta* inserts whose homologous Arabidopsis regions are physically smaller in size; longer blocks are comprised of multiple, overlapping *B. stricta* inserts as suggested by their similarity to Arabidopsis genomic regions. ^cRelative to the Arabidopsis reference sequence.

B. stricta 5' noncoding regions are, generally, slightly larger than their Arabidopsis orthologs (data not shown).

As percent identity is a one-dimensional representation of sequence similarity, the proportions of Arabidopsis nucleotides scored as identity and noninformative were plotted against Arabidopsis nucleotide position for all 657 alignments (Fig. 8A). The proportion of identities is highest (approximately 0.70) at positions directly abutting the presumptive translation initiation codons and declines steadily at positions increasingly 5' to the start codon. This trend is linear and significant ($P < 0.005$; Supplemental Data 5).

DISCUSSION

In the postgenomic era, the utility of comparative genomics as applied to functional and evolutionary questions has become the subject of increasing interest. This work represents the first foray into establishing a whole-genome resource from a wild relative of Arabidopsis that provides immediate comparative genomic application for an established research community; literature searches demonstrate that *B. stricta* and the encompassing genus, *Boechera*, are among the most extensively studied wild relatives of Arabidopsis. Further, the relatedness of the two species, similar breeding systems, and comparable general genome organizations contrasted against different life histories, geographic distribution, and ecology suggested that a *B. stricta*-Arabidopsis comparison would be highly informative in both functional and evolutionary terms.

Global Comparison of Two Related Genomes

To establish a conservative estimate for coverage of the Arabidopsis genome by homologous *B. stricta* end sequences, a stringent significance threshold ($e\text{-value} \leq 10^{-30}$) was applied to BLASTn analyses. By this criterion, nearly half of the end sequences, 18,797, were identified as nonredundant and informative relative to Arabidopsis genomic sequences and were mapped to

physical positions along each Arabidopsis chromosome. For each of the five Arabidopsis chromosomes, the distribution of significant *B. stricta* end-sequence BLASTn hits was minimal in the Arabidopsis centromeric regions and higher at more distal positions of the chromosomal arms (Fig. 3; Supplemental Fig. 1). This is not surprising as *B. stricta* sequences corresponding to repetitive heterochromatic regions, such as centromeres, are likely to have been filtered during our analysis. In addition, centromeric regions are known to be underrepresented in the current version of the completed Arabidopsis genome sequence.

At the specified significance threshold, BLASTn identified a single, best Arabidopsis homolog for the vast majority of *B. stricta* end sequences (Fig. 3, green). Duplications, i.e. close paralogs, in Arabidopsis are also detected (Fig. 3, yellow), while regions with three or four paralogs in Arabidopsis (Fig. 3, light orange and orange) are observed less frequently. These results are consistent with what is known regarding segmental duplication in the lineage leading to Arabidopsis and related species (Arabidopsis Genome Initiative, 2000; Blanc et al., 2003). Further, the observation that the coverage of Arabidopsis chromosomes by *B. stricta* homologs is consistent between chromosomes (Supplemental Fig. 1), and that there are no obvious biases for one duplicated Arabidopsis region over another, suggests that the *B. stricta* genome has retained a similar repertoire of segmental duplications, and that orthologous relationships to duplicated Arabidopsis genomic regions, both coding and noncoding, can be established.

Several informative *B. stricta* sequences display similarity to Arabidopsis sequences with 5 times representation in the Arabidopsis genome (Fig. 3; Supplemental Fig. 1, red). These Arabidopsis homologs are associated with centromeres and pericentromeric regions, suggesting that they are heterochromatic repeats that were not identified by our filtering datasets. Further work is required to position these repeats in the *B. stricta* genome relative to centromeric and pericentromeric regions.

Nearly half of the informative *B. stricta* end sequences intersect unique, annotated Arabidopsis

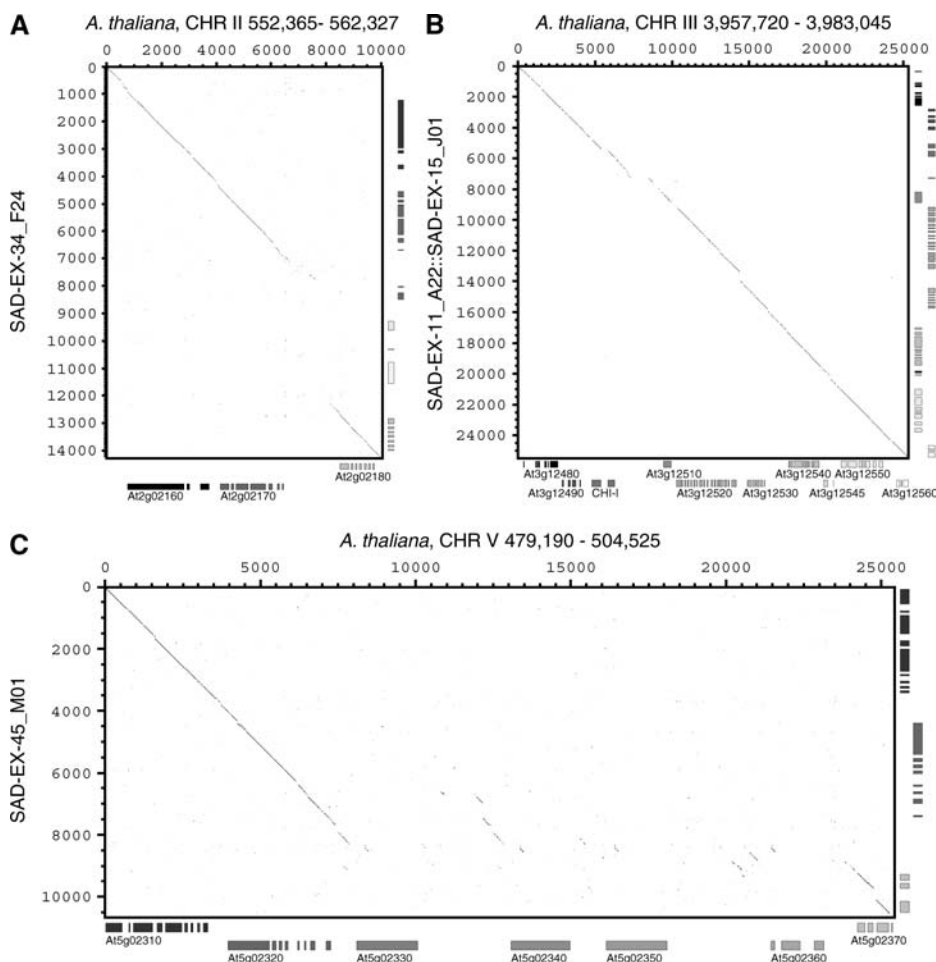


Figure 6. Dot-plot alignments of *B. stricta* inserts and the homologous Arabidopsis genomic regions identified by synteny-Finder.py. Annotated Arabidopsis CDSs with exon-intron structure are indicated along the horizontal axes. *B. stricta* CDSs, as predicted by Twinscan, are shown on the vertical axes. Watson-strand CDSs are positioned closer to the relevant axis. CDSs are presented in varying gray scale to highlight homologous relationships between *B. stricta* and Arabidopsis. A, *B. stricta* genomic inserts are larger than the homologous region(s) of Arabidopsis. This is generally attributable to one or more large indel polymorphisms. In this instance, two of the predicted CDSs are of unknown function; the predicted translation product of the 3'-most Watson-strand CDS, however, shares similarity to ping/pong/SNOOPY family transposases. B, Microsyntenic block comprised of two *B. stricta* inserts compared to the homologous, class I chitinase (CHI-I)-containing Arabidopsis region. This region is of similar size in the two species and contains many small indel polymorphisms. C, An example of a comparison where the Arabidopsis homolog is larger than the *B. stricta* region. The Arabidopsis indel in this region contains four tandem duplications (At5g02330–60) of a putative CDS encoding a DC-1 domain-containing protein.

CDSs (Supplemental Data 1). This accounts for a full third of the CDSs currently identified in the Arabidopsis genome (approximately 27,500). The average identity observed among HSPs (90.3%) coupled with the mean d_N/d_S observed in *B. stricta* by Arabidopsis comparisons (approximately 0.3; K. Schmid, personal communication) points toward functional constraints on the evolution and divergence of CDSs in these two species. On a per-site basis, d_N/d_S quantifies the level of non-synonymous to synonymous changes.

Our estimate of informative *B. stricta* end sequences relative to Arabidopsis is intentionally conservative in an attempt to make the conclusions drawn from this initial analysis as unambiguous as possible. Thus, the

approximately 14,000 *B. stricta* end sequences classified as having weak similarity to Arabidopsis genomic regions (Fig. 2) should not be interpreted as uninformative or as representing substantive interspecific differences, rather, they represent a resource for future analysis and interpretation.

Microsynteny and Resource Utility

From the informative *B. stricta* end-sequence dataset, we have identified 6,334 *B. stricta* sequence-indexed inserts in which both end sequences are anchored to homologous sequences in the Arabidopsis genome. The end sequences from three quarters of these inserts display conserved microsynteny or tight

Table III. Similarity analysis by BLASTn of *B. stricta* sequences with genes identified as duplicated by polyploidy in the *Arabidopsis* genome

Based on the analysis of K. Wolfe, and limited to the recent genome duplication in the Brassicaceae (24–40 million years ago).

Total <i>B. stricta</i> sequences with BLASTn hits to duplicated- <i>Arabidopsis</i> CDSs ^a :	7,160
BLASTn hits to both members of a gene pair in <i>Arabidopsis</i> :	5,180 (2,590 pairs)
BLASTn hits to a single member of a gene pair ^b in <i>Arabidopsis</i> :	1,704
BLASTn hits to extended gene families ^c in <i>Arabidopsis</i> :	276
Selected dataset: orthology/paralogy comparisons ^d	
BLASTn hits to both members of a gene pair ^e in <i>Arabidopsis</i> :	2,002 (1,001 pairs)
BLASTn hits to a single member of a gene pair ^b in <i>Arabidopsis</i> :	314
BLASTn hits to both members of a gene pair ^f with e-values $\leq 10^{-90}$:	250 (125 pairs)

^aNonredundant, excludes hits to multiple splice variants. ^bE-value of the BLASTn hit to the unrepresented *Arabidopsis* homolog is assigned an ad hoc value of 10^{-10} . ^cMultiple duplication events; no simple paralogous relationships (i.e. >2 paralogs). To limit the complexity of orthologous/paralogous relationships, this group has been excluded from further analysis. ^d*B. stricta* end sequences from the total dataset with at least one *Arabidopsis* homolog (e-value $\leq 10^{-90}$). ^ePutative ortholog BLASTn hit with an e-value $\leq 10^{-90}$; putative paralog BLASTn hit with an e-value $> 10^{-90}$. ^fSimilarity to both *Arabidopsis* homologs is highly significant; a clear a priori distinction between orthology and paralogy is not possible.

colinearity relative to *Arabidopsis* genomic regions (Fig. 4) and are distributed uniformly across each *Arabidopsis* chromosome (Table II; Supplemental Data 3). Microsyntenic relationships have been further confirmed by sequencing a subset of inserts to completion (Fig. 6; Supplemental Fig. 2). Taken together and in the absence of a high-resolution genetic map for *B. stricta*, the data demonstrate that the genomes of *B. stricta* and *Arabidopsis* are largely colinear. Further, we can infer conservatively that 26% of the *Arabidopsis* genome has been bracketed in our sequence-indexed clone collection (Table II) with the homologous *B. stricta* genomic regions comprising the sequence-indexed library.

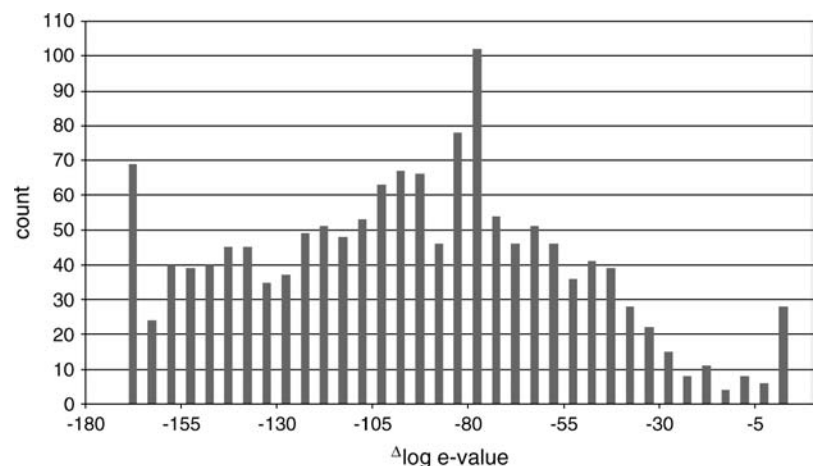
The informative *B. stricta* inserts with end sequences whose homologs in *Arabidopsis* have been categorized as linked or unlinked (Fig. 4) represent departures from conserved colinearity. These inserts may contain large indel polymorphisms, breakpoints of local rearrangements, and/or translocations in one

species relative to the other; however, further investigation is required to validate such conclusions.

Observations where only one end sequence for a given *B. stricta* insert displays similarity to *Arabidopsis* (Fig. 2) mostly arise as a result of poor sequence quality on one clone end, or where the significance of the BLASTn hit to a given *Arabidopsis* genomic region failed to meet our criterion, and cannot be meaningfully interpreted relative to conserved colinearity between the two species. However, a significant subset of solo end sequences results from transposon insertions or other indel polymorphisms that affect local but not global colinearity.

To further validate and demonstrate the utility of microsynteny as inferred from *B. stricta* end-sequence data, we have used the sequence-indexed library to clone several *B. stricta* genomic regions that contain homologs to *Arabidopsis* regions of interest. *B. stricta* inserts that were doubly anchored to *Arabidopsis* by similarity routinely contained the desired *B. stricta*

Figure 7. Differences in the log e-value for *B. stricta* BLASTn hits to *Arabidopsis* gene pairs ($n = 1,440$ gene pairs). E values were taken from the selected dataset (Table III; Supplemental Data 4). E values of 0.0 in the native BLASTn output have been adjusted to 10^{-180} ; e-values for unreported paralogs adjusted to 10^{-10} , the maximum e-value allowable for reporting in our analysis. As the $\Delta_{\log e\text{-value}}$ approaches 0, the ability to distinguish between orthologs and paralogs diminishes. $\Delta_{\log e\text{-value}}$ scores of -32 , -15 , and -5 correspond to the 95th, 97th, and 98th percentiles, respectively.



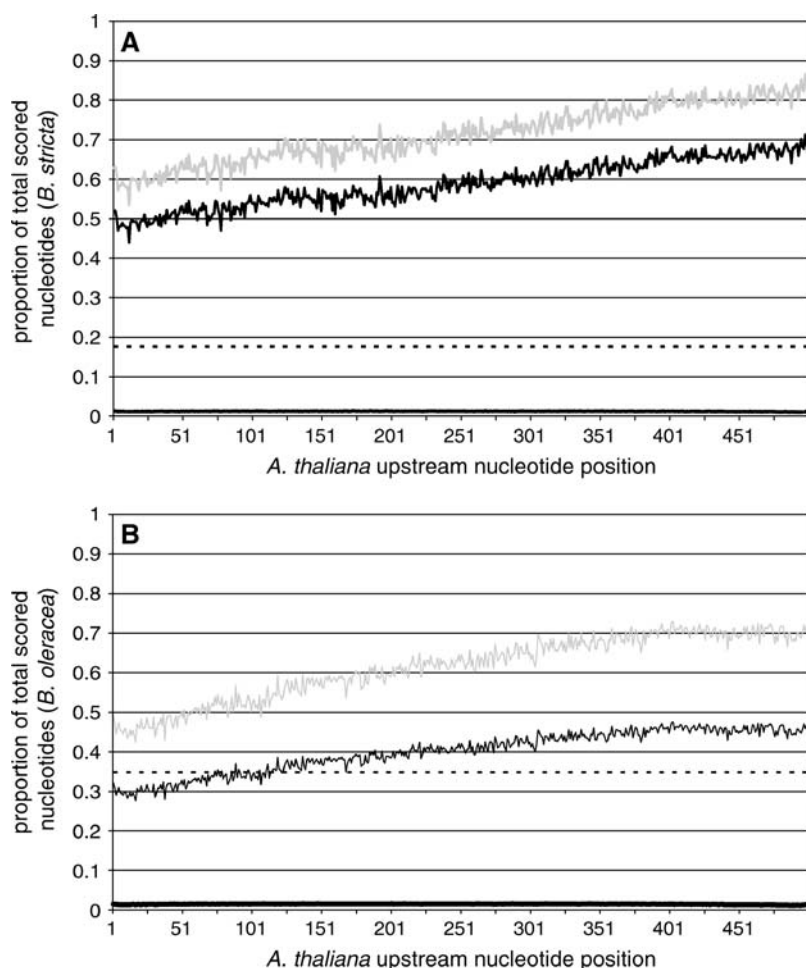


Figure 8. Summary of the per-nucleotide identity shared between homologous 5' regulatory regions. Values along the x axis correspond to Arabidopsis nucleotide positions 5' to CDSs; position 1 is the most 5' nucleotide, position 500 is the nucleotide adjacent to the translation initiation codon of a given CDS. For each position, the proportion of Arabidopsis nucleotides scored as identity is indicated as a jagged black line; the proportion of nucleotides from non-informative alignments (see Supplemental Text 1, UntransID.py, Test1 for details) at each position is designated by the dashed black line (17.7% and 34.8% of all nucleotides analyzed at every position for *B. stricta* and *B. oleracea*, respectively). The thick, black line at approximately 0.01 is the mean proportion of nucleotides scored as identities with error bars (SE of each mean) for each nucleotide position as determined via a permutation test (100 iterations; see Supplemental Text 1, UntransID.py, Test2 for details). The gray line indicates the proportion identity at each position when noninformative alignments are excluded from the calculation. A, $n = 657$ 5'-noncoding sequence alignments for the *B. stricta* by Arabidopsis comparison. B, $n = 1,208$ 5'-noncoding sequence alignments for the *B. oleracea* by Arabidopsis comparison.

CDSs (Fig. 6B; data not shown). Further, inserts anchored to the Arabidopsis genome at a single end and inserts anchored by identity to a fully sequenced *B. stricta* clone previously mapped to the Arabidopsis genome, i.e. walking out in silico, also proved successful in the identification of *B. stricta* inserts of interest (data not shown).

Genome Size

Genome size is heterogeneous both within the Brassicaceae (Johnston et al., 2005) and among accessions of Arabidopsis (Schmuths et al., 2004). While a consensus exists with regard to the relative contribution of polyploidy to variation in genome size, the role of directional bias in indel size (Gregory, 2003) during the genome size reduction in Arabidopsis is unclear. Recent studies involving interfamilial comparisons (Kirik et al., 2000; Orel and Puchta, 2003; Filkowski et al., 2004; Bennetzen et al., 2005) have identified the differentiation of DNA repair and recombination mechanisms as sources of genome size variation among plants. Thus, the availability of comparative sequence data from within the Brassicaceae and lineages that are closely

related to Arabidopsis provides an opportunity to address the biology of genome size more directly.

The genome of *B. stricta* is about 60% larger than that of Arabidopsis, but is similar in size to the diploids, *Arabidopsis lyrata* and *Capsella rubella* (Johnston et al., 2005). When the average insert size of sequence-indexed *B. stricta* inserts is compared to the mean size of the homologous Arabidopsis genomic intervals (Fig. 5), we find that that the *B. stricta* regions are, on average, 10% larger than the homologous Arabidopsis regions. Alignment of completely sequenced *B. stricta* inserts to their Arabidopsis homologs (Fig. 6; Supplemental Fig. 2) indicates that the observed difference in interval size is explained by the presence of large indel polymorphisms, generally ≥ 5 kb in length. It is likely that a greater proportion of the observed disparity in genome size is due to larger indel polymorphisms, but our methodology suffers from ascertainment bias in this respect; the maximal size of detectable *B. stricta* indel polymorphisms is limited by the size of the inserts themselves. Further, *B. stricta* sequence-indexed clones that fail to span insertions, where one end sequence corresponds to insertion-specific sequences, are likely to have been filtered during homology analyses, or

identified as either low similarity (Fig. 2) or as lacking conserved colinearity/synteny (Fig. 4). Nevertheless, our observation supports a coarse model of indel polymorphism as a major contributor to genome size variation among diploid relatives of Arabidopsis.

Identification of Orthologous Relationship in Close Relatives

A major obstacle in comparative genomics is inference of orthology versus paralogy for coding regions in interspecific comparisons. This is especially true in lineages, such as the Brassicaceae, that have experienced recurrent polyploidization and/or duplication events. In these lineages, the issue is complicated not only by the number of intraspecific paralogs, but also by the sub- or neofunctionalization of these paralogs (Lynch et al., 2001; Walsh, 2003). As the processes of sub- and neofunctionalization can involve alterations in the patterns of transcription for paralogous gene pairs (Haberer et al., 2004; Duarte et al., 2006), assessments of orthology rooted in direct genomic comparisons are preferable to those based on transcriptome data.

Throughout this study, the significance value attached to identified *B. stricta*-Arabidopsis similarity has been assumed to be the best indicator of orthology. To test this assertion more directly, we have contrasted the significance values of *B. stricta* BLASTn hits to known Arabidopsis paralog pairs. In a majority of instances (98%; Fig. 7), the best Arabidopsis candidate for orthology was readily identified by standard BLASTn analysis. Hence, computational approaches will prove highly informative in the establishment of orthology for comparisons within the genus Arabidopsis and among near diploid relatives (Fig. 1). These species, however, share a common history with regard to genome duplication and, as demonstrated by this study and related work (Acarkan et al., 2000; Yogeewaran et al., 2005), are likely to have retained similar repertoires of segmentally duplicated genomic regions. The establishment of orthology in wider comparisons involving lineages such as Brassica with divergent histories of polyploidization, will, however, require extensive ancillary data relating to global genomic organization and local context.

Global Analysis of 5' Regulatory Regions

Phylogenetic footprinting and shadowing have emerged as powerful techniques for the identification of candidates for functionally conserved domains in cis-transcriptional regulators of plant genes (Koch et al., 2001b; Colinas et al., 2002; Ayre et al., 2003; Guo and Moose, 2003; Hong et al., 2003; Bao et al., 2004; Buchanan et al., 2004; Lee et al., 2005). Sequence identity is not attributable solely to conservation of function; it is also retained as a stochastic consequence of common history. Similarly, differences in noncoding regions are not de facto indicators of functional divergence. In directed studies involving one to a few cis-

regulatory regions, these limitations can be partially overcome through multiple, interspecific comparisons. However, the logistics and expense of generating large comparative sequence datasets severely restricts the application of this strategy on a global scale.

To assess the utility of *B. stricta* in global comparisons of 5' noncoding regions, we have aligned 657 such *B. stricta* regions to their Arabidopsis orthologs. In informative alignments, the average identity displayed to the Arabidopsis regions was 71.4%, indicating that neutral divergence has not been saturated. These data indicate that *B. stricta* and Arabidopsis 5' regions are more diverged than identified orthologous coding regions, where the average identity among HSPs was observed to be 90.3%.

While more diverged than coding regions, orthologous 5' noncoding sequences still display a high degree of identity. A similar global comparison of 1,208 *B. oleracea* 5' regions to their Arabidopsis homologs (Fig. 8B) has indicated that while the 5' regions of Brassica are more diverged from Arabidopsis than those of *B. stricta*, conservation at the level of sequence is still high in these regions. On average, upstream regions displayed 61.4% identity (40.0% when non-informative alignments are included in the calculation) in *B. oleracea*-Arabidopsis comparisons. Interestingly, 34.8% of alignments in these comparisons were deemed noninformative (Fig. 8B). This may reflect complex ortholog-paralog relationships between *B. oleracea* and Arabidopsis, as evidenced by the observation that only 17.7% noninformative alignments were detected among *B. stricta* by Arabidopsis comparisons (Fig. 8A).

When the proportion of identity among all alignments is plotted against Arabidopsis nucleotide position (Fig. 8, A and B), comparisons to *B. stricta* and *B. oleracea* display declining identity 5' to translation initiation codons. Two processes may contribute to this pattern. First, insertions in either species may cause regulatory domains to fall outside our window of comparison, hence distal regulatory domains may not be included in some comparisons. We expect cumulative insertions (and the probability that regulatory domains fall outside our window of comparison) to show a linear increase toward distal promoter regions, as observed in *B. stricta*, but not *B. oleracea*. Second, natural selection to maintain regulatory function may be strongest near the initiation codon, and decline in distal promoter regions. Although it is difficult to disentangle these two processes using partial shotgun sequences, the nonlinear pattern of decrease in *B. oleracea* suggests that both factors contribute to observed patterns of promoter similarity, making our estimates of mean similarity conservative.

Our observations suggests that *B. stricta*, *C. rubella*, and even species as distant as *B. oleracea* are too similar to Arabidopsis to consistently identify highly conserved regulatory regions, since identity arising from functional constraint cannot be easily distinguished from that arising through common history. This highlights the necessity of multiple comparisons for effective

phylogenetic footprinting of regulatory regions, even at a global scale.

CONCLUSION

We have applied a random shotgun sequencing approach to a medium-insert size *B. stricta* genomic library. This approach was rapid, cost effective, and allowed the generation and analysis of a high-quality genomic sequence dataset with immediate application to comparative genomics within the Brassicaceae. Using this dataset, we have shown that analyses of *B. stricta* and, potentially, species of similar evolutionary distance from Arabidopsis will be informative relative to issues of genome size, coding regions, and gross genomic organization such as duplication, synteny, and orthology. Our data suggests that global comparisons to Arabidopsis have limited ability to identify conserved regulatory domains. Further, additional work is required to generalize our conclusions regarding genome evolution to lineages outside of Arabidopsis and its sister clade. As such, we suggest that similar, exploratory end-sequencing projects from a wide sampling of the Brassicaceae would be beneficial in establishing a comparative genomic strategy for the Brassicaceae research community. Such a sampling will help to inform the selection of candidate species and help to establish priorities for large-scale genome projects within the Brassicaceae in the future.

MATERIALS AND METHODS

Plant Materials and DNA Preparation

Seeds derived from a *Boechera stricta* individual, SAD12.4 (Taylor Creek, Colorado, population), were surface sterilized and grown in liquid culture as described previously for Arabidopsis (*Arabidopsis thaliana*; Windsor and Waddell, 2000). To reduce cytoplasmic DNA contamination and increase the yield of high-molecular-weight genomic DNA, genomic DNA was isolated from 7.3 g of SAD12 tissue using a nuclei-isolation protocol modified from Olszewski and coworkers (1988).

Library Construction

A SAD12 genomic library, λ 03, was constructed in the λ BlueSTAR λ -replacement vector system (Novagen). Approximately 8.0 μ g of SAD12 genomic DNA were partially digested with *Sau*3AI. The *Sau*3AI digestion was halted with a 20 min incubation at 65°C and size fractionated by means of pulse-field electrophoresis on a 1.0% agarose gel. DNA fragments in the size interval from 12 to 22 kb were excised in block and purified away from the encapsulating agarose using GELase (Epicentre Biotechnologies) according to the manufacturer's instructions. This DNA-insert size interval was selected to optimize subsequent packaging of the recombinant-phage genomes while taking advantage of the size-exclusion characteristics of the packaging reaction to eliminate chimeric clones. Microcon YM-100 columns (Millipore) were used to concentrate the size-selected SAD12 *Sau*3AI fragments to a volume of approximately 20 μ L in 10 mM Tris-HCl, pH 7.6. Six microliters of this preparation were ligated overnight at 16°C to *Bam*HI-digested λ BlueSTAR phage arms with T4 ligase in a total volume of 10 μ L. Two aliquots from the ligation reaction of 4 μ L each were packaged separately for *Escherichia coli* infection using Gigapack III Gold packaging extract (Stratagene) as instructed by the manufacturer. Subsequent *E. coli* ER1647 infection, titration, and amplification of the primary library were performed as described in the λ BlueSTAR manual. The primary λ 03 library contained 1.03e6 plaque forming

units (pfu) with an average insert size of 13.2 kb, as determined by completely sequencing 27 inserts (data not shown). This corresponds to approximately 52 times coverage of the *B. stricta* genome, based on a haploid genome size of approximately 260 Mb (highly replicated flow cytometry; data not shown). The final titer of the amplified λ 03 library is 5.0e8 pfu/mL.

A second, small-insert library was constructed by partially digesting 200 ng of SAD12 genomic DNA with *Sau*3AI. The digestion was halted with a 20 min incubation at 65°C followed by a 5 min incubation on ice. Digestion products were run on a standard 1.0% agarose gel and the fragments in the size interval of 1.5 to 5.0 kb were purified using NucleoSpin Extract (Macherey-Nagel) columns according to the manufacturer's instructions. Recovered SAD12 DNA fragments were incubated at 65°C for 10 min, snap cooled to 4°C for 5 min, and ligated to dephosphorylated, *Bam*HI digested pUC19 (Fermentas) as described for the production of λ 03. Ligation products were transformed into electrocompetent *E. coli* DH10b (Stratagene) using standard protocols and plated on Luria-Bertani (LB) media supplemented with 100 μ g/mL ampicillin and 10 μ g/mL 5-bromo-4-chloro-3-indolyl- β -D-galactoside (x-gal). From the resultant white colonies, 4,992 clones were isolated and grown for plasmid isolation and sequencing.

Production of the SAD12.4 Sequence-Indexed Library

λ 03 inserts were automatically subcloned into pBlueSTAR, a pUC19 derivative, using λ BlueSTAR's integrated plasmid excision system. *E. coli* BM25.8 cells were grown to an OD₆₀₀ of 1.0 in LB media supplemented with 34 μ g/mL chloramphenicol, 50 μ g/mL kanamycin, 0.2% maltose, and 10 mM MgSO₄. λ 03 phage particles, 10⁷ pfu from the amplified library, were combined with 500 μ L of BM25.8 cells and incubated at 37°C for 30 min with gentle agitation to facilitate phage adsorption. After adsorption, 4.5 mL of LB supplemented with 100 μ g/mL ampicillin was inoculated with the entire infection mixture and incubated at 30°C for 2 h followed by a second incubation at 37°C for 3 h. BM25.8 cells were subsequently centrifuged and pBlueSTAR clones were extracted from the pellet in a final volume of 100 μ L elution buffer (two, 50- μ L elution steps) using the NucleoSpin Plasmid Quick Pure columns (Macherey-Nagel) as instructed by the manufacturer.

Recovered SAD12 clones were transformed en masse into electrocompetent *E. coli* DH10b. From the resulting pool of DH10b transformants, DNA from 23,136 colonies was isolated, transferred to microtiter plates, and sequenced. As each clone has been cataloged to a unique plate and plate coordinate and each clone can, therefore, be accessed specifically on demand, we refer to this resource as the SAD12.4 sequence-indexed library.

Sequencing

All clones that comprise the SAD12.4 sequence-indexed library were end sequenced in both directions using T7 and T3 primers; the small-insert library clones were sequenced in both directions with M13 forward and reverse primers. Cycle-sequencing reactions were carried out in GeneAmp 2700 thermal cyclers (Applied Biosystems) using Big Dye terminator cycle-sequencing kits (Applied Biosystems) and read with an ABI PRISM 3730XL DNA sequencer (Applied Biosystems). The traces generated for both libraries were trimmed of vector and processed for quality using SeqMan 5.0 (DNASTar) at a quality threshold of 12. To augment sequence quality, SAD12.4 sequence-indexed reads with less than 200 bp of reliable sequence were excluded from further analysis. Small-insert clones were assembled into contigs with SeqMan and only clones with overlapping sequencing reads were considered further.

A subset of SAD12.4 sequence-indexed library inserts were sequenced to completion. Sequencing primer sites were introduced into inserts using the HyperMu <KAN-1> insertion kit (Epicentre Biotechnologies) via a scaled-down version of the manufacturer's protocol; sequencing reactions were performed as described earlier using the primers MUKAN-1 FP-1 and MUKAN-1 RP-1 supplied with the HyperMu system. Using SeqMan 5.0, traces were first quality trimmed followed by a fixed 5' trim of 60 bp to remove MuKan sequences. pBlueStar sequences were removed manually during contig assembly.

Sequence Analysis and Bioinformatics

The BLASTn program (Altschul et al., 1997) was used to identify similarity between *B. stricta* sequence-indexed end sequences and the Arabidopsis chromosome pseudomolecules (GenBank accession numbers: NC_003070.5,

NC_003071.3, NC_003074.4, NC_003075.3, and NC_003076.4). Genomic blasts were performed with default parameters with the exception of the gap-opening penalty, the gap-extension penalty, and the low-complexity filters that were set to 1, 1, and F, respectively. Both the sequence-indexed end sequences (filtered, see below) and the *B. stricta* small-insert library sequences were blasted against the Arabidopsis CDS set with low-complexity filters switched off, but all other parameters at the default settings. In comparison to Arabidopsis genomic sequences, BLAST hits with $e\text{-values} \leq 10^{-30}$ were scored homologous to the relevant Arabidopsis subject sequence(s). The $e\text{-value}$ threshold was increased to 10^{-10} for all other comparisons.

For the purpose of filtering the *B. stricta* sequence-indexed library dataset, BLASTn analyses with default parameters were used to compare *B. stricta* sequences to the sequences in the AtRepBase (Cold Spring Harbor, database and sequence set of repetitive DNAs identified in Arabidopsis; <http://nucleus.cshl.org/protarab/AtRepBase.htm/>) and the Arabidopsis mitochondrial and chloroplast genomes (<ftp://ftp.arabidopsis.org/Sequences/>). Further, BLASTx (default parameters) was used to identify *B. stricta* sequences with significant similarity to the transposable element translated coding-sequence dataset of Zhang and Wessler (Zhang and Wessler, 2004). *B. stricta* end sequences with matches having an $e\text{-value} \leq 10^{-30}$ were excluded from comparative analyses to the Arabidopsis genome (1,650; Fig. 2, light-gray wedge).

To analyze the *B. stricta* end-sequence data, a custom suite of software was developed. These scripts were written in the Python language using the Python core distribution (distribution site for the Python object-oriented programming language; <http://www.python.org/>), the BioPython extension (Chapman and Chang, 2000), and the stats.py and pstat.py modules of G. Strangman (statistical modules for Python developed by Gary Strangman; http://www.nmr.mgh.harvard.edu/Neural_Systems_Group/gary/python.html/). A description of each script is provided in Supplemental Text 1 and all code is available from either the sources specified earlier or from <http://boecheira.twibbit.net/>.

Duplicated SAD12.4 sequence-indexed inserts were identified by the dupCloneFinder.py script with an $e\text{-value}$ threshold 10^{-30} and a read_error setting of 25 bp.

Mapping of *B. stricta* sequence-indexed end sequences to physical intervals along the Arabidopsis chromosome pseudomolecules and filtering of end sequences with significant similarity to repetitive DNAs, rRNA genes, and organellar genomes was accomplished with the compGenomeFilterv2.py script and an $e\text{-value}$ threshold setting of 10^{-30} .

The syntenyFinder.py script was used to identify *B. stricta* sequence-indexed clones displaying microsynteny to Arabidopsis chromosomal regions with the following parameters: physical limit for a syntenic region in Arabidopsis, 50 kb; average read length, 826 bp; and the default duplicated region filtering mode.

B. stricta-Arabidopsis promoter region comparisons were performed using the UntransID.py program. UntransID.py was run with extra sequence quality measures; an $e\text{-value}$ threshold of 10^{-10} for Arabidopsis CDS screening, the Needleman-Wunsch global alignment algorithm, 10,000 random comparisons for the determination of the alignment quality score threshold, a quality score threshold of 300, and 100 iterations for the determination of the mean and se for the proportion of Arabidopsis nucleotides scored as identities in random comparisons. To implement the Needleman-Wunsch global alignment algorithm, UntransID.py invokes the needle program of the EMBOSS 3.0 sequence analysis suite (Rice et al., 2000). Needle alignments were performed with a gap-opening penalty of 12.0 and a gap-extension penalty of 2.0.

Brassica oleracea-Arabidopsis promoter region comparisons were performed as described for the *B. stricta*-Arabidopsis analysis with the exception of the quality score threshold parameter, which was set to 290. This analysis was based on an initial pool of 415,521 publicly available *B. oleracea* shotgun sequences (Ayele et al., 2005).

Annotation of fully sequenced *B. stricta* sequence-indexed inserts was performed using gene-prediction models generated with the Twinscan (Korf et al., 2001) online resource (TwinScan Web site; <http://genes.cs.wustl.edu/>). For these analyses, Arabidopsis was selected as the reference dicot.

Dot-plot alignments were performed with the Dotter program (Sonnhammer and Durbin, 1995) using the default sliding-window size of 25 bp and minimum and maximum visualization cutoffs of 90 and 100, respectively.

Ortholog-Paralog Analysis

B. stricta end sequences displaying significant similarity ($e\text{-value} \leq 10^{-10}$) to Arabidopsis CDSs with paralogs known to have been generated during the most recent polyploidization event in the Arabidopsis lineage (Blanc et al.,

2003) were identified by reference to the annotation of K. Wolfe. Only *B. stricta* end sequences with homology to distinct, Arabidopsis paralog pairs were considered further and sequences with BLASTn hits to members of extended gene families were excluded from further analysis (Supplemental Data 4, complete dataset). Both Arabidopsis paralogs of a given gene pair displayed significant similarity to the relevant *B. stricta* end sequence in a majority of comparisons. In cases where the $e\text{-value}$ of the BLASTn hit to the second Arabidopsis paralog was above our significance threshold, the $e\text{-value}$ for these comparisons was assumed to be 10^{-10} . BLASTn hits with $e\text{-values}$ of 0.0 were rescored to 10^{-180} , the value below which BLAST reports 0.0, for the purpose of log transformation.

Detailed analysis was performed on a more conservative dataset derived from the above (Supplemental Data 4, selected dataset). To be included in this dataset, the $e\text{-value}$ for a given *B. stricta* end-sequence BLASTn hit to at least one member of an Arabidopsis gene pair was $\leq 10^{-90}$. This stringent significance threshold was selected to enrich for *B. stricta* end sequences where a majority of the recovered *B. stricta* sequence intersects annotated Arabidopsis CDS, thus limiting the effect of secondary influences, such as sequence quality or partial CDS intersection, on the assessment of homology.

For a given *B. stricta* end sequence, $\Delta_{\log e\text{-value}}$ was calculated as: $\log(e\text{-value}_a) - \log(e\text{-value}_b)$; where $e\text{-value}_a$ is the $e\text{-value}$ from the Arabidopsis homolog with a more significant BLASTn hit and $e\text{-value}_b$ is the $e\text{-value}$ from the BLASTn hit to the second Arabidopsis homolog.

Statistical Analysis

All statistical analyses were performed with either the Excel spreadsheet package (Microsoft) or the stats.py and pstat.py modules of G. Strangman.

Sequence Availability

The sequence-indexed and small-insert datasets are available as GenBank accession numbers DU667459 to DU708532. The *B. stricta* class I chitinase genomic region is available as accession number DQ275145.

ACKNOWLEDGMENTS

The authors wish to thank X. Zhang and S. Wessler for supplying their *B. oleracea*/Arabidopsis transposable element CDS dataset, A. Heidel for data relating to the recovery of insect resistance loci cloned using inferred synteny and the sequence-indexed library, and U. Göbel for insightful discussions regarding promoter and 5'-untranslated region analyses. The authors also wish to thank C. Ortlepp and J. Haupt for technical support and R. Oyama, K. Schmid, E. Kellogg, and A. Navarro-Quezada for comments on the manuscript.

Received November 9, 2005; revised January 19, 2006; accepted February 4, 2006; published April 11, 2006.

LITERATURE CITED

- Acarkan A, Rossberg M, Koch M, Schmidt R (2000) Comparative genome analysis reveals extensive conservation of genome organisation for Arabidopsis thaliana and Capsella rubella. *Plant J* 23: 55–62
- Altschul SE, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (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: 768–815
- Ayele M, Haas BJ, Kumar N, Wu H, Xiao Y, Van Aken S, Utterback TR, Wortman JR, White OR, Town CD (2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in Arabidopsis. *Genome Res* 15: 487–495
- Ayre BG, Blair JE, Turgeon R (2003) Functional and phylogenetic analyses of a conserved regulatory program in the phloem of minor veins. *Plant Physiol* 133: 1229–1239
- Bao X, Franks RG, Levin JZ, Liu Z (2004) Repression of AGAMOUS by BELLINGER in floral and inflorescence meristems. *Plant Cell* 16: 1478–1489
- Beilstein MA, Al-Shehbaz IA, Kellogg EA (2006) Brassicaceae phylogeny and trichome evolution. *Am J Bot* 93: (in press)

- Bennetzen JL, Ma J, Devos KM** (2005) Mechanisms of recent genome size variation in flowering plants. *Ann Bot (Lond)* **95**: 127–132
- Blanc G, Hokamp K, Wolfe KH** (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* **13**: 137–144
- Blanc G, Wolfe KH** (2004a) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**: 1679–1691
- Blanc G, Wolfe KH** (2004b) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**: 1667–1678
- Bleeker W** (2003) Hybridization and *Rorippa austriaca* (Brassicaceae) invasion in Germany. *Mol Ecol* **12**: 1831–1841
- Bleeker W, Matthies A** (2005) Hybrid zones between invasive *Rorippa austriaca* and native *R-sylvestris* (Brassicaceae) in Germany: ploidy levels and patterns of fitness in the field. *Heredity* **94**: 664–670
- Boivin K, Acarkan A, Mbulu R-S, Clarenz O, Schmidt R** (2004) The Arabidopsis genome sequence as a tool for genome analysis in Brassicaceae: a comparison of the Arabidopsis and *Capsella rubella* genomes. *Plant Physiol* **135**: 735–744
- Buchanan CD, Klein PE, Mullet JE** (2004) Phylogenetic analysis of 5′-noncoding regions from the ABA-responsive *rab 16/17* gene family of sorghum, maize and rice provides insight into the composition, organization and function of cis-regulatory modules. *Genetics* **168**: 1639–1654
- Chapman BA, Chang J** (2000) Biopython: python tools for computational biology. *ACM SIGBIO Newsletter* **20**: 15–19
- Charlesworth D, Wright SI** (2001) Breeding systems and genome evolution. *Curr Opin Genet Dev* **11**: 685–690
- Clauss MJ, Mitchell-Olds T** (2003) Population genetics of tandem trypsin inhibitor genes in Arabidopsis species with contrasting ecology and life history. *Mol Ecol* **12**: 1287–1299
- Colinas J, Birnbaum K, Benfey PN** (2002) Using cauliflower to find conserved non-coding regions in Arabidopsis. *Plant Physiol* **129**: 451–454
- Dietz H, Fischer M, Schmid B** (1999) Demographic and genetic invasion history of a 9-year-old roadside population of *Bunias orientalis* L. (Brassicaceae). *Oecologia* **120**: 225–234
- Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW** (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Mol Biol Evol* **23**: 469–478
- Durka W, Bossdorf O, Prati D, Auge H** (2005) Molecular evidence for multiple introductions of garlic mustard (*Alliaria petiolata*, Brassicaceae) to North America. *Mol Ecol* **14**: 1697–1706
- Filkowski J, Kovalchuk O, Kovalchuk I** (2004) Dissimilar mutation and recombination rates in Arabidopsis and tobacco. *Plant Sci* **166**: 265–272
- Fumanal B, Martin J-F, Sobhian R, Blanchet A, Bon M-C** (2004) Host range of *Ceutorhynchus assimilis* (Coleoptera: Curculionidae), a candidate for biological control of *Lepidium draba* (Brassicaceae) in the USA. *Biol Control* **30**: 598–607
- Gao M, Li G, Yang B, McCombie WR, Quiros CF** (2004) Comparative analysis of a Brassica BAC clone containing several major aliphatic glucosinolate genes with its corresponding Arabidopsis sequence. *Genome* **47**: 666–679
- Gregory TR** (2003) Is small indel bias a determinant of genome size? *Trends Genet* **19**: 485–488
- Guo H, Moose SP** (2003) Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **15**: 1143–1158
- Haberer G, Hindemitt T, Meyers BC, Mayer KFX** (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. *Plant Physiol* **136**: 3009–3022
- Hong RL, Hamaguchi L, Busch MA, Weigel D** (2003) Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. *Plant Cell* **15**: 1296–1309
- Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ** (2005) Evolution of genome size in Brassicaceae. *Ann Bot (Lond)* **95**: 229–235
- Katari MS, Balija V, Wilson RK, Martienssen RA, McCombie WR** (2005) Comparing low coverage random shotgun sequence data from Brassica oleracea and *Oryza sativa* genome sequence for their ability to add to the annotation of Arabidopsis thaliana. *Genome Res* **15**: 496–504
- Kirik A, Salomon S, Puchta H** (2000) Species-specific double-strand break repair and genome evolution in plants. *EMBO J* **19**: 5562–5566
- Koch M, Al-Shehbaz IA, Mummenhoff K** (2003) Molecular systematics, evolution, and population biology in the mustard family (Brassicaceae). *Ann Mo Bot Gard* **90**: 151–171
- Koch M, Haubold B, Mitchell-Olds T** (2001a) Molecular systematics of the Brassicaceae: evidence from coding plastidic *matK* and nuclear *Chs* sequences. *Am J Bot* **88**: 534–544
- Koch MA, Kiefer M** (2005) Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species—*Capsella rubella*, *Arabidopsis lyrata* subsp. *petraea*, and *A. thaliana*. *Am J Bot* **92**: 761–767
- Koch MA, Weisshaar B, Kroymann J, Haubold B, Mitchell-Olds T** (2001b) Comparative genomics and regulatory evolution: conservation and function of the *Chs* and *Apeta13* promoters. *Mol Biol Evol* **18**: 1882–1891
- Korf I, Flicek P, Duan D, Brent MR** (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**: S140–S148
- Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppala J, Koch M, Mitchell-Olds T, Langley CH, Savolainen O** (2004) Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **168**: 1575–1584
- Lee J-Y, Baum SE, Alvarez J, Patel A, Chitwood DH, Bowman JL** (2005) Activation of CRABS CLAW in the nectaries and carpels of Arabidopsis. *Plant Cell* **17**: 25–36
- Li G, Gao M, Yang B, Quiros CF** (2003) Gene for gene alignment between the Brassica and Arabidopsis genomes by direct transcriptome mapping. *Theor Appl Genet* **107**: 168–180
- Li G, Quiros CF** (2003) In planta side-chain glucosinolate modification in Arabidopsis by introduction of dioxygenase Brassica homolog *BoGSL-ALK*. *Theor Appl Genet* **106**: 1116–1121
- Lukens L, Zou F, Lydiate D, Parkin I, Osborn T** (2003) Comparison of a Brassica oleracea genetic map with the genome of Arabidopsis thaliana. *Genetics* **164**: 359–372
- Lynch M, O’Hely M, Walsh B, Force A** (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**: 1789–1804
- Lysak MA, Koch MA, Pecinka A, Schubert I** (2005) Chromosome triplification found across the tribe Brassicaceae. *Genome Res* **15**: 516–525
- Ma XF, Gustafson JP** (2005) Genome evolution of allopolyploids: a process of cytological and genetic diploidization. *Cytogenet Genome Res* **109**: 236–249
- Meekins JF, Ballard HE Jr, McCarthy BC** (2001) Genetic variation and molecular biogeography of a North American invasive plant species (*Alliaria petiolata*, Brassicaceae). *Int J Plant Sci* **162**: 161–169
- Mitchell-Olds T, Al-Shehbaz IA, Koch M, Sharbel T** (2005) Crucifer evolution in the post-genomic era. In R Henry, ed, *Diversity and Evolution of Plants—Genotype and Phenotype Variation in Higher Plants*. CABI Press, Cambridge, MA, pp 119–138
- Needleman SB, Wunsch CD** (1970) A general method applicable to the search for similarities in the amino-acid sequence of 2 proteins. *J Mol Biol* **48**: 443–453
- Olszewski NE, Martin FB, Ausubel FM** (1988) Specialized binary vector for plant transformation expression of the Arabidopsis-thaliana *Ahas* gene in *Nicotiana-tabacum*. *Nucleic Acids Res* **16**: 10765–10782
- Orel N, Puchta H** (2003) Differences in the processing of DNA ends in Arabidopsis thaliana and tobacco: possible implications for genome evolution. *Plant Mol Biol* **51**: 523–531
- Osborn TC** (2004) The contribution of polyploidy to variation in Brassica species. *Physiol Plant* **121**: 531–536
- Pannell JR, Barrett SCH** (2001) Effects of population size and metapopulation dynamics on a mating-system polymorphism. *Theor Popul Biol* **59**: 145–155
- Rice P, Longden I, Bleasby A** (2000) EMBOSS: the european molecular biology open software suite. *Trends Genet* **16**: 276–277
- Roy BA** (1995) The breeding system of six species of *Arabis* (Brassicaceae). *Am J Bot* **82**: 869–877
- Schein M, Yang Z, Mitchell-Olds T, Schmid KJ** (2004) Rapid evolution of a pollen-specific oleosin-like gene family from Arabidopsis thaliana and closely related species. *Mol Biol Evol* **21**: 659–669
- Schmuths H, Meister A, Horres R, Bachmann K** (2004) Genome size variation among accessions of Arabidopsis thaliana. *Ann Bot (Lond)* **93**: 317–321

- Schranz ME, Dobes C, Koch MA, Mitchell-Olds T** (2005) Sexual reproduction, hybridization, apomixis, and polyploidization in the genus *Boechera* (BRASSICACEAE). *Am J Bot* **92**: 1797–1810
- Sharbel TE, Mitchell-Olds T** (2001) Recurrent polyploid origins and chloroplast phylogeography in the *Arabis holboellii* complex (Brassicaceae). *Heredity* **87**: 59–68
- Song B-H, Clauss MJ, Pepper A, Mitchell-Olds T** (2006) Geographic patterns of microsatellite variation in *Boechera stricta*, a close relative of *Arabidopsis*. *Mol Ecol* **15**: 357–369
- Sonnhammer EL, Durbin R** (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**: GC1–GC10
- Suzuki T, Grellet F, Potter D, Li G, Quiros CF** (2003) Structure, sequence, and phylogeny of the members of the Ck1 gene family in Brassica oleracea and *Arabidopsis thaliana* (Brassicaceae). *Plant Sci* **164**: 735–742
- Walsh B** (2003) Population-genetic models of the fates of duplicate genes. *Genetica* **118**: 279–294
- Weigel D, Bergelson JM, Borevitz JO, Clark R, Gaut BS, Hall AE, Langley CH, Nueffer B, Mayer KFX, Nordborg M, et al** (2005) Department of Energy Community Sequencing Program: *Arabidopsis lyrata* and *Capsella rubella* Whole-Genome Sequencing Proposal. Unpublished white paper. U.S. Department of Energy, Washington, DC
- Windsor AJ, Waddell CS** (2000) FARE, a new family of foldback transposons in *Arabidopsis*. *Genetics* **156**: 1983–1995
- Wright SI, Lauga B, Charlesworth D** (2002) Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Mol Biol Evol* **19**: 1407–1420
- Wright SI, Lauga B, Charlesworth D** (2003) Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. *Mol Ecol* **12**: 1247–1263
- Yang Y-W, Lai K-N, Tai P-Y, Ma D-P, Li W-H** (1999) Molecular phylogenetic studies of Brassica, Rorippa, *Arabidopsis* and allied genera based on the internal transcribed spacer region of 18S-25S rDNA. *Mol Phylogenet Evol* **13**: 455–462
- Yogeeswaran K, Frary A, York TL, Amenta A, Lesser AH, Nasrallah JB, Tanksley SD, Nasrallah ME** (2005) Comparative genome analyses of *Arabidopsis* spp.: inferring chromosomal rearrangement events in the evolutionary history of *A. thaliana*. *Genome Res* **15**: 505–515
- Zhang X, Wessler SR** (2004) Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and Brassica oleracea. *Proc Natl Acad Sci USA* **101**: 5589–5594