

# Analyses of Synteny Between *Arabidopsis thaliana* and Species in the Asteraceae Reveal a Complex Network of Small Syntenic Segments and Major Chromosomal Rearrangements

Lee Timms,\* Rosmery Jimenez,\* Mike Chase,\* Dean Lavelle,<sup>†</sup> Leah McHale,<sup>†</sup> Alexander Kozik,<sup>†</sup> Zhao Lai,<sup>‡</sup> Adam Heesacker,<sup>§</sup> Steven Knapp,<sup>§</sup> Loren Rieseberg,<sup>‡</sup> Richard Michelmore<sup>†</sup> and Rick Kesseli<sup>\*,1</sup>

\*Biology Department, University of Massachusetts, Boston, Massachusetts 02125, <sup>†</sup>Genome Center and Department of Plant Sciences, University of California, Davis, California 95616, <sup>‡</sup>Biology Department, Indiana University, Bloomington, Indiana 47405 and <sup>§</sup>Center for Applied Genetic Technologies, University of Georgia, Athens, Georgia 30602

Manuscript received August 10, 2005  
Accepted for publication June 7, 2006

## ABSTRACT

Comparative genomic studies among highly divergent species have been problematic because reduced gene similarities make orthologous gene pairs difficult to identify and because colinearity is expected to be low with greater time since divergence from the last common ancestor. Nevertheless, synteny between divergent taxa in several lineages has been detected over short chromosomal segments. We have examined the level of synteny between the model species *Arabidopsis thaliana* and species in the Compositae, one of the largest and most diverse plant families. While macrosyntenic patterns covering large segments of the chromosomes are not evident, significant levels of local synteny are detected at a fine scale covering segments of 1-Mb regions of *A. thaliana* and regions of <5 cM in lettuce and sunflower. These syntenic patches are often not colinear, however, and form a network of regions that have likely evolved by duplications followed by differential gene loss.

**P**LANT nuclear genomes are extremely dynamic in nature and vary considerably in structure and size. Variation in genome size is largely attributable to repetitive DNA content (SANMIGUEL and BENNETZEN 1998) and ploidy changes. Ploidy changes have long been accepted to be a driving force in the evolution of plants (STEBBINS 1950, 1971) and upward of 70% of all angiosperms have likely undergone at least one polyploidization event (MASTERSON 1994, LEITCH and BENNETT 1997, BOWERS *et al.* 2003) and subsequent chromosomal rearrangements as well as gene loss and functional diversification (ARABIDOPSIS GENOME INITIATIVE 2000; KU *et al.* 2000; ADAMS *et al.* 2003; BLANC and WOLFE 2004; ADAMS and WENDEL 2005). *Arabidopsis thaliana*, with  $n = 5$  chromosomes and a small genome, was once thought likely to be a simple diploid but now appears to have passed through multiple whole genome duplications (ARABIDOPSIS GENOME INITIATIVE 2000; VISION *et al.* 2000; SIMILLION *et al.* 2002; BOWERS *et al.* 2003). Despite this activity, comparative genetic mapping has revealed conservation of gene content, order, and function among closely related taxa (BENNETZEN and FREELING 1993; PATERSON *et al.* 1996; BENNETZEN 2000, MCCOUCH 2001). The most striking

case of genome conservation is in the Poaceae, where detailed genetic maps can be used to infer gene content and order in related grass species (BENNETZEN *et al.* 1998; GALE and DEVOS 1998; FREELING 2001). Extended macrosynteny has also been observed within the Solanaceae (TANKSLEY *et al.* 1992; PRINCE *et al.* 1993) and the Pinaceae (KRUTOVSKY *et al.* 2004). Synteny is not, however, always evident. Comparative genome analyses of members of the Brassicaceae have shown complex syntenic relationships; gene content is often conserved, but the genomes are highly duplicated and rearranged (LAGERCRANTZ *et al.* 1996; LAGERCRANTZ 1998; QUIROS *et al.* 2001). Close relatives of *A. thaliana*, for example, have clear syntenic blocks but often differ by numerous chromosomal rearrangements (BOIVIN *et al.* 2004; KUITTINEN *et al.* 2004).

Comparative studies among distantly related species are more problematic due to reduced gene similarities, making orthologous relationships difficult to identify and mapping with common markers almost impossible. Reduced genomic conservation is also expected because of the increased time since the divergence of the taxa, and colinearity is expected only over small genetic distances (PATERSON *et al.* 1996; VISION *et al.* 2000; SALSE *et al.* 2002). Substantial progress unraveling the evolutionary history of eukaryotes has been made, however, precisely because of efforts to compare distantly related species (KELLOGG 2003). These studies have also aided

<sup>1</sup>Corresponding author: Biology Department, University of Massachusetts, 100 Morrissey Blvd., Boston, MA 02125. E-mail: rick.kesseli@umb.edu

the characterization of ploidy changes and the fate of duplicated genes in these species (BLANC and WOLFE 2004). Comparisons between Arabidopsis and tomato, belonging to Brassicaceae and Solanaceae, respectively, that diverged more than 90 million years ago, revealed a complex microstructure resulting from genome duplications followed by extensive gene loss (KU *et al.* 2000). Conserved synteny was observed between tomato chromosome 2 and four homologous regions in Arabidopsis. Comparisons between Arabidopsis and rice, the model dicot and monocot species, respectively, showed only very small intervals of microsynteny in some regions (VAN DODEWEERD *et al.* 1999; LIU *et al.* 2001; SALSE *et al.* 2002), smaller than expected (PATERSON *et al.* 1996).

The Asteraceae, or the Compositae family, is one of the largest and most diverse families of angiosperms, making up one-tenth of all known flowering plants. The Compositae include many domesticated species with agricultural and economical value, such as lettuce and sunflower (KESSELI and MICHELMORE 1997). The wide range in chromosome numbers among species within genera and within the family suggests that polyploid changes and genome rearrangements have been critical factors in the evolution of the family (STEBBINS 1971; SOLBRIG 1977). In addition, there have been major hybridization events and introgression producing chimeric genomes in the evolutionary history of some taxa (RIESEBERG *et al.* 1996) and at least for some multigene families, such as disease resistance genes, there have been localized changes in gene number among species and with respect to *A. thaliana* (PLOCIK *et al.* 2004), all suggesting dynamic and flexible genomes. Lettuce (*Lactuca sativa*,  $n = 9$ ) and sunflower (*Helianthus annuus*,  $n = 17$ ), the two main targets of this study, are usually referred to as a diploid and an ancient tetraploid, respectively. The agronomic importance of this family provided incentive for comparative genomic analyses within the Compositae family and with the model *A. thaliana*. These comparisons allow us to exploit the knowledge being generated from the study of *A. thaliana* for crop improvement and to further our understanding of genome evolution.

We used three approaches to characterize the level of synteny between *A. thaliana* and species in the Compositae. First, using all genes mapped in the lettuce and sunflower, we identified those with homologous pairs in *A. thaliana* to assess the global level of synteny. We then examined macrosyntenic patterns using only low-copy markers. Finally we evaluated the local syntenic relationships of lettuce and Arabidopsis at five targeted regions: self-incompatibility (SI) locus (two regions), *Leafy*, *Ovate*, and a cluster of genes from the conserved orthologous set (COS). At the global level, our analyses demonstrated conserved genome organization only at a very narrow scale; less than a few centimorgans in the Compositae species and a few megabases in *A. thaliana*. The macrosyntenic analyses confirmed this, showing

that lettuce and *A. thaliana* have undergone numerous chromosomal changes since their last common ancestor. The local analyses revealed conserved microstructure between lettuce and *A. thaliana*; however the two genomes in the targeted regions have been highly duplicated and rearranged since their divergence from a common ancestor.

## MATERIALS AND METHODS

**Plant materials and databases:** The seed for parental accessions of *L. serriola* (wild lettuce) and *L. sativa* cv. *Salinas* (cultivated lettuce) are maintained at the University of California, Davis (<http://michelmorelab.ucdavis.edu/>). These accessions have been used to generate a mapping population and to develop the EST database by the Composite Genome Project (<http://cgpdb.ucdavis.edu/>). The mapping population consisted of F<sub>7,8</sub> recombinant inbred lines (RILs) derived by generating single-plant descent self-fertilized lineages from individuals of the F<sub>2</sub> population. Seed and mapping populations (LAI *et al.* 2005) for the sunflower accessions (RHA280 and RHA801) are maintained at both the University of Georgia (sjknapp@uga.edu) and Indiana University (<http://www.bio.indiana.edu/~rieseberglab/>).

The lettuce and sunflower EST information is stored and accessible through the Composite Genome Project database (CGPdb) (<http://cgpdb.ucdavis.edu/>). Libraries were constructed from two lettuce genotypes, cultivated (*L. sativa* cv. *Salinas*) and wild lettuce (*L. serriola*) and two sunflower genotypes (RHA280 and RHA801). More than 70,000 ESTs and nearly 20,000 unigenes derived from a variety of tissues have been identified for each genus.

**Global approach for the analysis of synteny:** For the first approach to characterize the level of synteny, we examined all pairs of genes mapped in lettuce and in sunflower and determined whether homologs of these genes were linked more often than expected in *A. thaliana*. All mapped genes in sunflower (>350) and in lettuce (>200) were used to query the Arabidopsis database. A BLAST search was used to identify all significant matches in the Arabidopsis genome (National Center for Biotechnology Information, NCBI, database). Scripts were written in Perl (<http://www.perl.com>) to filter the BLAST results. We selected best hits from the Arabidopsis genome for each EST from the CGPdb and recorded the coordinates of each match. Matches with expectations (*E*-values) of  $1 \times 10^{-25}$  or less were retained, but we deleted matches that were more than  $1 \times 10^{-25}$  from the best match considering these to be unlikely orthologs. We also removed tandem duplication in the Arabidopsis genome by selecting only the best match in a region if there were paralogs within 2 Mb of each other (data available upon request).

From this pared-down data set of lettuce or sunflower ESTs and their matches in the Arabidopsis genome (106,693 and 152,445 pairs, respectively) we selected sliding windows of various recombination distances (centimorgans) in the Compositae genomes and windows of various physical distance (megabase) in the Arabidopsis genome. Scripts were written to extract all pairs of loci that fit the specified criteria set for both genomes. For example, all pairs of ESTs within 6 cM of each other in the sunflower genome with matches within 5 Mb of each other in Arabidopsis could be extracted (data available upon request). We did this for sliding windows of 1–10, 15, and 20 cM for the lettuce and sunflower genomes and windows of 1–10 Mb for the Arabidopsis genome.

To determine if the number of pairs of genes landing within a given interval of the Arabidopsis genome was greater than

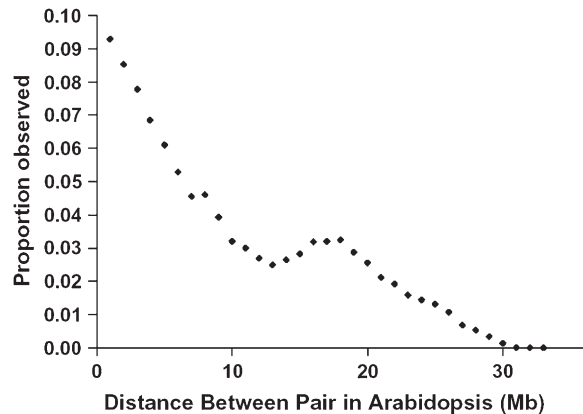


FIGURE 1.—Relative position in *Arabidopsis thaliana* of homologous, unlinked pairs of genes from sunflower ( $N = 23,735$  pairs). The frequency that homologs of a pair of unlinked genes in sunflower would land within 1-Mb intervals of each other in *A. thaliana* is plotted.

expected by chance alone, we simulated the probability distribution that random genes would have matches within an interval by chance. We did this by choosing all pairs of unlinked genes in the sunflower (23,735) and lettuce (16,610) maps and plotting the distribution of matches to these pairs within all 1-Mb intervals of the Arabidopsis genome; that is, what was the probability that two unlinked genes in sunflower have homologs 0–1 Mb, 1–2 Mb, 2–3 Mb, etc., apart in the Arabidopsis genome? The distributions based on sunflower (Figure 1) and lettuce were identical although the lettuce distribution (not shown) was less smooth due to the smaller sample size. These distributions were also markedly not uniform due to the nonrandom distribution of genes in the Arabidopsis genome and edge effects of chromosomes of finite lengths. With this distribution providing the expected values for markers landing in a given interval by chance, we calculated chi-squares from the observed and the expected data with the null hypothesis being that linked markers in sunflower or in lettuce are expected to be linked in Arabidopsis only at a level attributable to chance alone.

**Macrosyntentic approach:** Arabidopsis homologs were identified for all genes on the lettuce framework map by using lettuce singleton and contig sequences as queries for a tBLASTX search back into the Arabidopsis genome (<http://www.arabidopsis.org>). This allowed us to identify additional and possibly better matches for lettuce genes in the Arabidopsis genome and to examine macrosynteny. The lettuce framework markers that were part of the COS ([http://cgpdb.ucdavis.edu/COS\\_Arabidopsis/arabidopsis\\_single\\_copy\\_genes\\_2003.html/](http://cgpdb.ucdavis.edu/COS_Arabidopsis/arabidopsis_single_copy_genes_2003.html/)), which are defined as genes having one copy in most plant taxa (FULTON *et al.* 2002), were chosen. Because the genome of *A. thaliana* was likely duplicated once since its last common ancestor with the Compositae (BOWERS *et al.* 2003), we also retained markers with two distinct hits that were at least  $10^{10}$  better than all other matches to the Arabidopsis genome. We used the Arabidopsis genome matches to search for duplicated blocks (<http://wolfe.gen.tcd.ie/athal/dup/>) and identified linked markers from the lettuce map that match to known syntenic blocks in *A. thaliana*. Different markers in lettuce that match to common blocks in Arabidopsis were considered potentially syntenic even if the best matches for the pair of lettuce markers resided on the different chromosomes of *A. thaliana* that possess the block (see supplemental tables at <http://www.genetics.umb.edu/> and <http://cgpdb.ucdavis.edu/>). Finally we increased the number of framework markers on the

lettuce map by identifying a series of COS genes from one 5.5-Mb region on chromosome V of Arabidopsis that had a relatively high density of COS ([http://cgpdb.ucdavis.edu/COS\\_Arabidopsis/](http://cgpdb.ucdavis.edu/COS_Arabidopsis/)).

**Local syntenic approach and identification of candidate genes:** The third approach for the characterization of synteny involved targeting regions in species of the Brassicaceae and identifying all corresponding homologs in species of the Asteraceae. We focused on lettuce, the putative diploid, in this phase, and searched for single nucleotide polymorphisms (SNPs) in the set of candidate homologs and then mapped these in the lettuce RILs. Four regions in Arabidopsis defined by the key gene of interest, *Ark3*, *Sll2*, *Leafy*, *Ovate*, and one region, the S-locus (homologous to *Ark3/Sll2*), in *Brassica campestris* (CONNER *et al.* 1998), were chosen for comparative mapping studies with lettuce. Targets in Arabidopsis were identified using the NCBI database. The *Ark3*, *Sll2*, *Leafy*, and *Ovate* chromosomal positions were identified and all genes within 0.5 Mb of the targets were used to search for potential homologs in the CGPdb. Genes in any of these regions that are part of large multigene families were identified by a BLASTN search (ALTSCHUL *et al.* 1997) of the genes in our target regions against the entire Arabidopsis genome and excluded from further analyses. The remaining genes were entered into the tBLASTX program (ALTSCHUL *et al.* 1997) on the CGPdb with an *E*-value cutoff of  $10^{-5}$  and a list of candidate lettuce homologs was produced. The list was further refined by choosing lettuce contigs and unigenes that were low copy in the CGPdb and had the lowest *E*-values. In addition, the genes for the S-locus of *B. campestris* listed in NCBI were entered into tBLASTX search on the CGPdb lettuce EST database with an *E*-value cutoff of  $10^{-5}$ . All hits were used as candidate markers.

**DNA isolation, primer design, and PCR:** Primers were designed on the basis of consensus sequences obtained from the CGPdb using the program Primer3 (Whitehead Institute for Biomedical Research and Technical Report, Cambridge, MA) and required to have a GC cap at the 3' end and a maximum nucleotide repeat of three. Primers were obtained from Invitrogen (Carlsbad, CA) and Operon (San Diego). Genomic DNA (15 ng/ $\mu$ l) from each species was extracted (BERNATZKY and TANSKLEY 1986) and used as template in a 25- $\mu$ l PCR containing 1 mM  $MgCl_2$ , 0.1  $\mu$ M forward primer, 0.1  $\mu$ M reverse primer, 0.2 mM dNTPs, 1 $\times$  PCR buffer, and 1 unit *Taq* DNA polymerase (Promega) and standard cycling parameters. The amplification products were separated on a 1.5% TBE agarose and stained with ethidium bromide. If required, PCR products were treated with exonuclease and shrimp alkaline phosphatase (EXOSAP); 0.2  $\mu$ l EXO and 2  $\mu$ l of SAP were added to 15  $\mu$ l of PCR product. The mixture was incubated at 37° for 45 min and then the enzymes were inactivated at 80° for 15 min before sequencing.

**Polymorphism identification, genotyping, and linkage analysis:** Some polymorphisms were identified *in silico* by analyses of the EST database. Some codominant and dominant markers were identified directly from agarose gels of the parental PCR products as fragment size polymorphisms. In the absence of scorable polymorphisms, amplicons were sequenced to identify SNPs and short insertion/deletions (indels). DNA sequencing was performed in the University of Massachusetts (Boston) Environmental Genomics Center. Sequences were analyzed using Sequencher v3.0 (GeneCodes, Ann Arbor, MI).

Length polymorphisms and dominant markers were scored in the RILs directly from agarose gels. Several approaches, including direct sequencing, were used to score SNP genotypes in the RILs; the choice depended upon the length of the amplicon, presence of restriction sites, cost of the procedure, and other factors.



Single-stranded conformational polymorphism (SSCP) analysis was used for scoring many markers. Amplicons ranging from 200 to 450 bp were diluted 1:1 with amplification stop solution (AmpStop) of 40% formamide, 5 mM EDTA, 0.05% SDS, 0.25% bromophenol blue, 0.25% xylene cyanol, and 0.5× TBE (45 mM Tris-base, 45 mM boric acid, 1 mM disodium EDTA, pH 8.2). Amplicons >450 bp were modified with restriction enzymes. PCR products were heat denatured at 95° for 5 min and then cooled to 4° on ice. The resultant single-stranded DNA fragments were separated by gel electrophoresis through 6% polyacrylamide TBE gels. Gels were stained with the Silver Sequence Staining Kit (Promega, Madison WI).

For some markers with SNPs a 5' nuclease allele discrimination assay was used to score RILs according to the manufacturer's protocol (ABI TaqMan; Applied Biosystems, Foster City, CA). Primer and probe sequences for all genotyping assays are available (<http://www.genetics.umb.edu/>). A cleaved amplified polymorphic sequence (CAPS) protocol was used for other markers with SNPs. CAPS sites were identified using Sequencher. Each CAPS reaction consisted of 10 µl of unpurified PCR product digested with the appropriate restriction enzyme (NEB, Ipswich, MA) according to the manufacturer's protocols and visualized by agarose gel electrophoresis. The REVEAL mutation detection system (Spectrumedix, College Park, PA), which uses temperature gradient capillary electrophoresis, was used to screen additional SNP markers following manufacturer's protocols.

Framework maps were constructed with JoinMap (Kyazma, Wageningen, The Netherlands). The nine lettuce linkage groups and other mapping information are available from the CGPdb website. The relative positions of target markers to each other were identified with MapMaker/EXP, v3.0 (Whitehead Institute for Biomedical Research and Technical Report).

## RESULTS

**Global analysis of synteny:** Synteny between *A. thaliana* and species of the Compositae family is apparent only at the smallest of intervals. Pairs of markers that were <5 cM apart in lettuce or in sunflower had homologs within 1 Mb of each other in *A. thaliana* significantly more often than expected by chance alone (Figure 2). This was the only window that consistently showed this difference. For all pairs of markers in windows >5 cM we failed to reject the null hypothesis as their distribution matched the expected generated in the simulation (Figure 1). Because of the density of the maps in lettuce and sunflower, we did not have a sufficient number of markers to carefully examine smaller windows (<1 cM or <2 cM ranges). These would likely show stronger deviation from the expected.

**Macrosyntentic analysis:** To identify regions of macrosynteny between the lettuce and *Arabidopsis* genomes, *Arabidopsis* homologs were determined for genes on the lettuce framework map. The 124 *Arabidopsis*–lettuce gene pairs (<http://www.genetics.umb.edu/>) that fit our criteria for low copy number revealed no large syntenic blocks for the two genomes and presented a complex picture of the relationship between these two distantly related species. In addition, only 18 of 782 pairs of markers residing on a given linkage group of lettuce

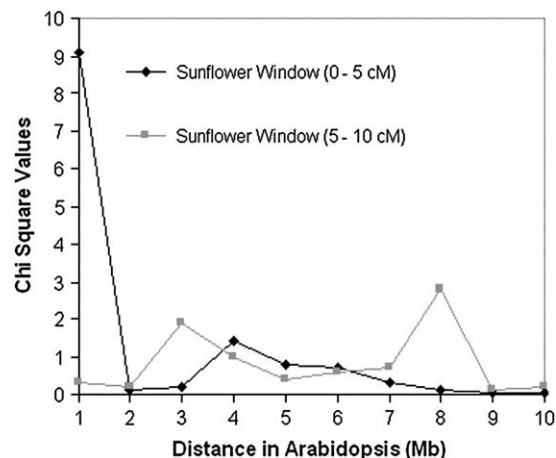


FIGURE 2.—Windows of synteny between sunflower and *Arabidopsis thaliana*. Chi-square values testing whether markers linked at specified recombination distances in sunflower (0–5 cM or 5–10 cM) are linked more often than expected by chance in *A. thaliana* at sliding 1-Mb windows. Chi-square values >3.84 are statistically significant.

had matches to common syntenic blocks in *A. thaliana*. To examine further the scale at which syntenic relationships may be detected, we increased the density of COS markers from one 5.5-Mb region of *A. thaliana* chromosome V and mapped them back into lettuce. Even with this increased density, macrosyntentic patterns were not evident. The 24 genes in this one region on chromosome V of *A. thaliana* mapped to all nine chromosomes in lettuce and only two pairs were tightly linked within 5 cM (Figure 3).

**Local analyses of synteny:** BLAST was used to identify ESTs in lettuce that showed homology to sequences from *Ark3*, *Sll2*, *Leafy*, and *Ovate* regions in *Arabidopsis* and the S-locus in *Brassica* sp. Primers were strategically chosen to flank introns to generate polymorphic primer pairs that could be easily scored in the lettuce mapping population. Potential homologs for between 36 and 75 genes (depending on the region) were identified for each of the target regions and one or two primer sets were designed for each gene (Table 1). Amplifications of single discrete products in both lettuce parents were obtained with most PCRs (53–88% depending on the target region) and of these ~74% were polymorphic. Polymorphisms varied as some, such as LK1504, showed length difference for the *L. sativa* and *L. serriola* parents while others, such as LK1452, were dominant, amplifying only in one parent. Most were codominant and alleles were distinguished as SNPs or short indels.

We scored genotypes for 103 EST markers (82% of the polymorphic markers) using a variety of methods in established lettuce mapping populations (Table 1; <http://www.genetics.umb.edu/>). Two markers were scored using the allele discrimination assay with real-time PCR. Many markers were scored by agarose gel electrophoresis: 9 as length polymorphisms, 3 as dominant,

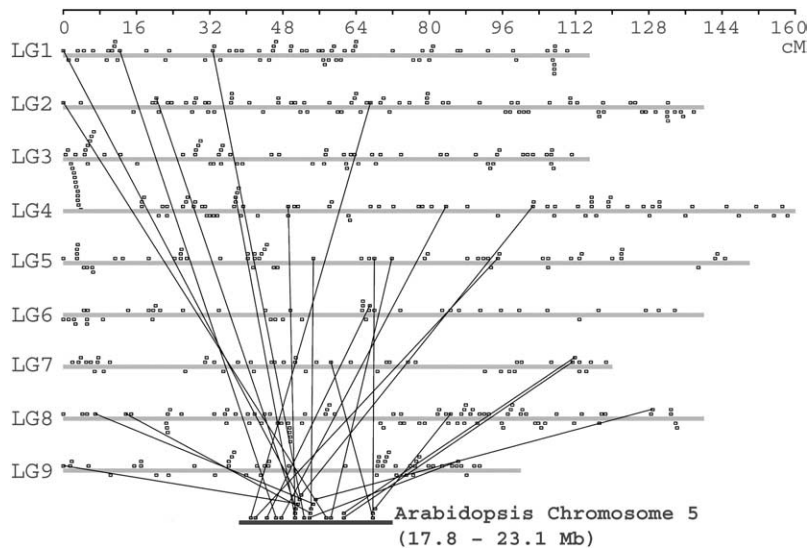


FIGURE 3.—Physical positions of conserved orthologous sequences (COS) in a 5.5-Mb region of chromosome V of *Arabidopsis thaliana* (bottom linkage group) and their corresponding mapped positions on the nine linkage groups of *Lactuca sativa* (top nine linkage groups, LG 1–9). The squares indicate the positions of genes in the lettuce map and the scale at the top for the lettuce maps is in centimorgans. The figure was created by GenomePixelizer (<http://www.atgc.org/GenomePixelizer/>).

and 10 scored as CAPS using various restriction enzymes to cut PCR products at polymorphic sites. The remaining markers involved SNPs with 47 scored as SSCPs, 21 by direct sequencing and 11 by the REVEAL mutational analysis system. The 103 polymorphic markers were mapped relative to each other using MapMaker v3.0 and 84 were oriented with Join Map within the nine linkage groups of lettuce with genotypic information available from the CGPdb ([http://cgpdb.ucdavis.edu/database/genome\\_viewer/viewer/](http://cgpdb.ucdavis.edu/database/genome_viewer/viewer/)).

Conservation of gene content and possibly gene order was detected between the genomes of lettuce and *Arabidopsis*. The precise gene orders in the narrow blocks of lettuce could not be evaluated, however, since they are based on recombination data in a small (<100) set of RILs. Ten regions on linkage groups (LG) 1, 2, 4, 5, and 7 of lettuce appear to be syntenic with local regions of the five chromosomes of *Arabidopsis* (Figures 4 and 5). Potential syntenic clusters with three or more genes were identified for the four small (1 Mb) target

regions, S-locus (*Ark3* and *Sll2*), *Leafy*, and *Ovate*. As mentioned above, only two pairs of genes from the larger COS region target were also tightly linked in lettuce.

Three tightly linked lettuce genes (identified by markers LK1403, LK1549, and LK1552) are 0.3 cM apart on LG 1 of lettuce and their homologs (At4g21390, At4g21540, and At4g21750, respectively) in *Arabidopsis* are within a 100-kb block of the *Ark3* region on AtIV (Figure 4c). An additional three loci (LK1540, LK1548, and LE0137) in a 1.6-cM interval of LG 7 in lettuce also have tentative orthologs (At4g20910, At4g21100, and At4g23560) in the *Ark3* region (Figure 4e). Two (At4g20910, At4g21100) are 70 kb apart; however, the third is ~1 Mb away. A third region on LG 4 of lettuce also may be syntenic with the *Ark3* region (Figure 4g). LK1503, LE1012, LK1527, and LK1406 are 13.1 cM apart and each has one hit (At4g22240, At4g21810, At4g21580, and At4g23230, respectively) to a 680-kb interval of the *Ark3* region. LK1503 and LE1012 also have significant hits to two other genes located on AtIV, At4g04020 and At4g04860. These two regions on AtIV are known to be part of a segmental duplication (ARABIDOPSIS GENOME INITIATIVE 2000).

Two regions in lettuce appear to be syntenic to the *Sll2* region on AtI. Three genes within a 4.1-cM interval on LG 2 (LK1481, LK1485, and LE0120) are homologous to three *Arabidopsis* genes (At1g65900, At1g66350, and At1g66250, respectively), ~800 kb apart (Figure 4d). A second cluster of three genes (At1g66670, At1g66680, and At1g66510) from a 50-kb block of *Sll2* map to a 10-cM block (LK1191, LK1164, LK1432) of lettuce (Figure 4f). A duplication of LK1432 (LK1447) is also present in this region.

A duplication of the *Leafy* region of AtV was found on AtIII and gene loss is apparent in both. Seven genes (LE0050, LK1333, LE3039, M2278, M275, LK1311, and LK1317) in a 32.6-cM segment of LG 1 of lettuce map to

TABLE 1

The number of EST markers screened and the number that were polymorphic and were mapped for the various targets in lettuce

Region	Gene targets	Primer pairs screened <sup>a</sup>	Targets amplified <sup>b</sup>	Polymorphic/mapped
COS	50	50	44	34/24
<i>Ovate</i>	43	44	23	14/11
<i>Leafy</i>	36	36	27	17/13
<i>ARK3/SLL2</i>	75	119	43	32/31
S-locus-like	48	106	31	28/28

<sup>a</sup> Multiple primer pairs were sometimes used for a given target gene.

<sup>b</sup> Gene targets with quality amplifications (single product in both parents).

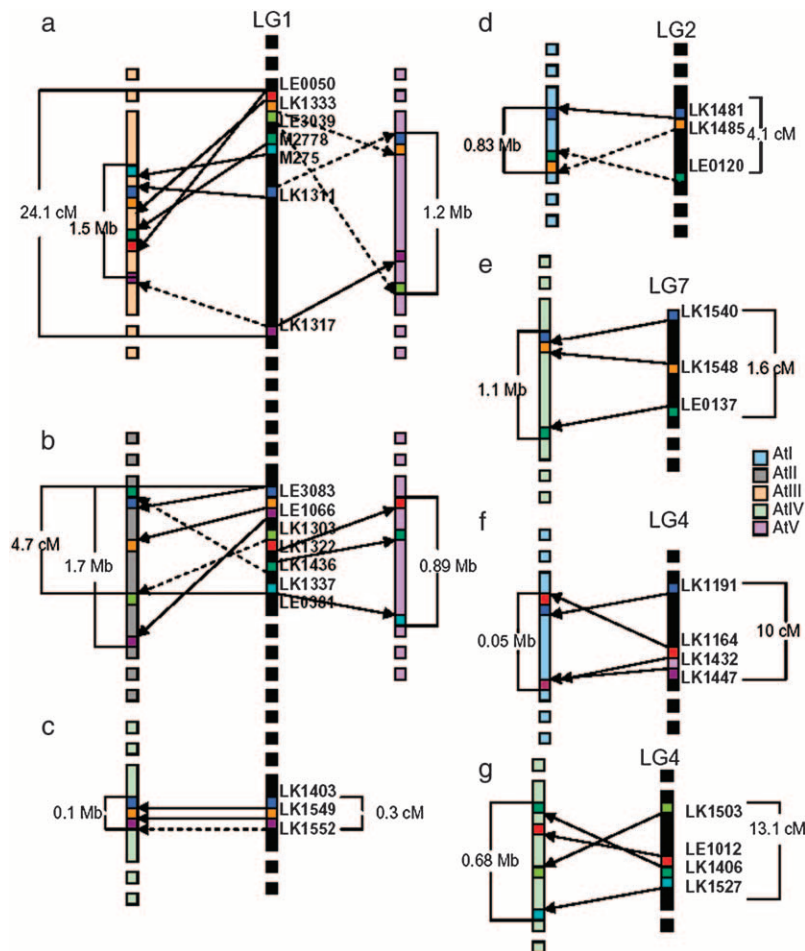


FIGURE 4.—Local analysis of synteny at the *Leafy* region (a and b) and the *S*-locus *Sl2* (d and f) and *Ark3* (c, e, and g). Solid and dashed arrows point to the best and second best match, respectively. Order of the markers in the linkage maps of lettuce is not precise.

these two regions of the Arabidopsis genome (Figure 4a). Six of these (all but LE3034) are in a 59-kb block of AtIII. Four genes are also in a 1.2-Mb block on chromosome AtV (LK1333/At5g61700, LE3039/At5g65060, LK1311/At5g61690, and LK1317/At5g63160). AtIII and AtV share orthologs to LK1333 (*E*-values for the AtIII and AtV homologs are  $1 \times 10^{-101}$  and  $6 \times 10^{-99}$ , respectively), LK1311 (*E*-values are  $1 \times 10^{-72}$  and  $1 \times 10^{-66}$ , respectively), and LK1317 (*E*-values are  $2 \times 10^{-71}$  and  $2 \times 10^{-77}$ , respectively). It appears that AtIII has lost the ortholog to LE3039 and AtV has lost the orthologs to LE0050, M2778, and M275. A second region on LG 1 also appears to be syntenic with *Leafy* on AtV and a different duplication on AtII (Figure 4b). This segment, which spans a 4.7-cM interval, contains seven tightly linked genes, five of which (LE3083, LE1066, LK1303, LK1322, and LE0381) align to a 1.7-Mb block on AtII (At2g27510, At2g28000, At2g31820, At2g28400, and At2g27500, respectively). Three loci (LK1322, LK1436, and LK1337) align with an 890-kb block in the *Leafy* region on AtV (At5g60680, At5g60900, and At5g62990, respectively).

The tomato *Ovate* region, previously described by Ku *et al.*, (2001), is syntenic to portions of AtII, AtIII, AtIV, and AtV. In lettuce, LG 2, 4, and two pieces of LG 5

possess syntenic regions to the same portions of the Arabidopsis chromosomes (Figure 5). The four regions in lettuce can be aligned to the four Arabidopsis chromosomes; however, none of the Arabidopsis segments contains a full complement of matching genes. Three of the four lettuce segments have homologs to the tomato bacterial artificial chromosome (GenBank accession no. AF273333) containing the *Ovate* region. LK1485 on LG 2 is homologous to open reading frame (ORF) 14, a Scarecrow-like protein of the tomato *Ovate* region (Ku *et al.* 2001). Markers for genes on LG 4, LK1344, and LE9003 are homologous to a UDP-glucose pyrophosphorylase and U2 snRNP auxiliary factor, ORF 10 and 11, respectively. The tomato and Arabidopsis *Ovate* genes (ORF6 and At2g18500, respectively) are homologous to an *Ovate*-like gene, LK1559b, on lettuce LG 5.

## DISCUSSION

Comparative analyses at multiple scales between the genomes of two species in the Compositae family and Arabidopsis were conducted. The global analysis using all mapped genes in lettuce and in sunflower to identify all homologs in *A. thaliana* indicated that syntenic relationships between the Compositae species and the



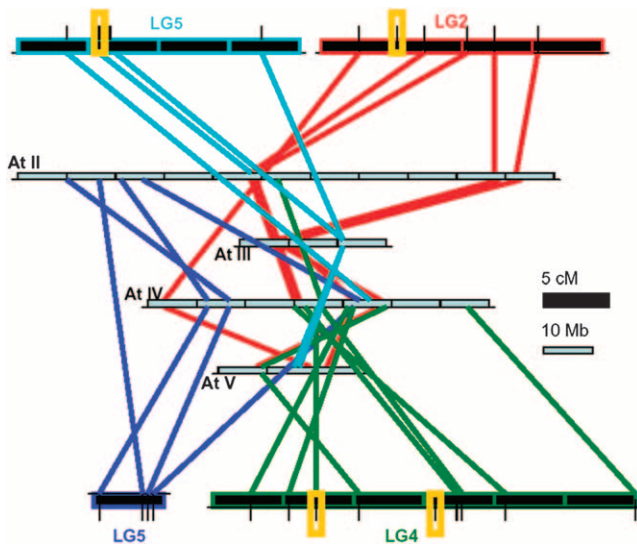


FIGURE 5.—Comparison of lettuce and Arabidopsis for the conserved *Ovale* region. Four syntenic blocks between lettuce (LG 2, 4, and 5) and Arabidopsis (AtII, AtIII, AtIV, and AtV) at the *Ovale* region. The scale of the lettuce blocks is in centimorgans and the scale of Arabidopsis blocks is in megabases. Gold boxes indicate markers that are putative homologs to genes in the tomato *Ovale* region.

distantly related *A. thaliana* would likely be detectable only in windows of a few centimorgans in the Compositae and a megabase or so in *A. thaliana*. This analysis does not try to distinguish orthologous *vs.* paralogous sequences between the genomes, but simply lets the statistical analysis determine if clustering is more often than expected for a given window.

**Macrosynteny not apparent between lettuce and Arabidopsis:** Macrosyntentic relationships are best displayed in closely related taxa (ECKARDT 2001) and have most often been characterized with restriction fragment length polymorphism markers (TARCHINI *et al.* 2000). Studies among families have generally found little evidence of macrosynteny (GRANT *et al.* 2000; KU *et al.* 2000). For example, ZHU *et al.* (2003), using 82 pairs of orthologous genes from *Medicago truncatula* and Arabidopsis, did not reveal a high level of macrosynteny between the two genomes. Our analyses with 124 low-copy genes from lettuce paralleled the majority of earlier studies and supported the finding of our global analysis as no large syntenic blocks between lettuce and *A. thaliana* were identified. Lettuce linkage groups included in the analysis could be aligned to multiple, if not all, of the *A. thaliana* chromosomes.

Duplications, rearrangements, and difficulties identifying orthologous genes may erode macrosynteny (LAGERCRANTZ 1998) or limit its detection among families. BLAST searches of the lettuce genes detected ~60% with multiple homologs in the Arabidopsis genome. These results were not surprising, given previous studies in Arabidopsis (BLANC *et al.* 2000; VISION *et al.* 2000; SIMILLION *et al.* 2002; BOWERS *et al.* 2003) and

studies comparing soybean (GRANT *et al.* 2000) or tomato (KU *et al.* 2000) to Arabidopsis, which suggested that the Arabidopsis genome had undergone multiple rounds of duplication. In addition, duplications in large-genome species such as lettuce are also likely and all copies will not be present in the EST databases. Given the large evolutionary distance and complexity of the genomes, it is not surprising that the degree of macrosynteny detected between lettuce and *A. thaliana* is reduced in contrast with that described in within-family comparisons.

**Local synteny over small chromosome segments:** Given the low level of macrosynteny for species of distantly related families and our lack of success identifying large syntenic blocks even when using a 5.5-Mb patch of COS genes from *A. thaliana* (Figure 3), we targeted four regions of *A. thaliana* for detailed local analyses. The S-locus in *Brassica* sp. and the two homologous *Ark3* and *Sll2* regions in Arabidopsis are known to control the self-incompatibility response in the Brassicaceae and were the first two regions targeted. For a species with this type of self-incompatibility, strong selection maintains a tight linkage relationship for the genes in the S-locus (CHARLESWORTH 2002). The region, however, has been broken apart in the self-compatible *A. thaliana*. The third region, surrounding the *Leafy* gene, which controls meristem identity in Arabidopsis (WEIGEL *et al.* 1992), was targeted with the objective of identifying tentative homologs in lettuce. The fourth targeted region flanks *Ovale*, a gene that controls fruit shape in tomato. This region in tomato has remained syntenic with several chromosome segments in Arabidopsis (KU *et al.* 2001).

Although there is a lack of evidence for extended synteny between the lettuce and Arabidopsis genomes, using criteria similar to those of DOMINGUEZ *et al.* (2003) we detected small syntenic blocks of genes in all Arabidopsis chromosomes and LG 1, 2, 4, 5, and 7 of lettuce (Figures 4 and 5). Other studies (KU *et al.* 2000; LEE *et al.* 2001) have detected a similar pattern in lineages separated for tens of millions of years and it has been suggested that these blocks may include co-adapted gene complexes that offer some sort of selective advantage to the organism (LEE *et al.* 2001). The SI region is one such region that forms a coadaptive complex and 17 low-copy homologs with only one to two hits for the genes of the SI regions of the Brassicaceae map to five short syntenic blocks on LG 1, 2, 4, and 7 of lettuce (Figure 4). Although the precise order of genes in the blocks is not always identical or determinable because of the limited population size of the RILs in lettuce, some blocks (Figure 4c) contain genes that have remained in the same order and tightly linked in both species since their divergence. Other blocks (*e.g.*, Figure 4f) contain potentially functional S-locus homologs with four genes that belong to the Arabidopsis *Sll2* region. These four genes in lettuce each have one hit to

three *Arabidopsis* genes, one to S-locus protein 2 (*SP2*) and S-locus linked protein 2 (*SLL2*), and two to S-locus protein 3 (*SP3*).

Despite the occurrence of sporophytic SI in both the Brassicaceae and Asteraceae and the presence of syntenic blocks common to the genomes of both families, our results do not suggest a common genetic basis. Domesticated lettuce has lost the ancestral SI found in most Compositae species and thus, like *Arabidopsis*, selection for retention of an SI block of genes may have been relaxed (CONNER *et al.* 1998). This could explain why the region appears broken into five small blocks. Since the level of synteny does not appear substantially greater in this region than others investigated (see below), a more likely scenario may be that sporophytic SI has evolved independently multiple times in the angiosperms and involves different sets of genes in different families (UYENOYAMA 1995).

Syntenic blocks were detected between two regions of lettuce LG 1 and *Arabidopsis* at the *Leafy* region (Figure 4, a and b). These two syntenic regions comprise fourteen lettuce genes, all of which had one or two significant matches to genes in the *Arabidopsis* genome, and all of which are rearranged in comparison to the *Arabidopsis* segment. Synteny at the *Leafy* region has not been previously reported; however, our data suggest that multiple rounds of duplication followed by selective gene loss have occurred in the *Arabidopsis* genome. Similar syntenic networks have been obtained in comparisons of *Arabidopsis* with tomato (KU *et al.* 2000), soybean (GRANT *et al.* 2000), and *M. truncatula* (ZHU *et al.* 2003). Furthermore we suspect that the *Leafy* segment of lettuce resembles more closely that of the ancestral dicot as the two *Arabidopsis* regions are overlapping subsets, suggesting duplication with subsequent redundant gene loss or functional divergence in *A. thaliana*.

The most striking region of synteny identified was the *Ovate* region; four blocks containing genes that were in close proximity in lettuce are syntenic to four chromosomes in *Arabidopsis* (Figure 5). Furthermore, these are the same regions in *Arabidopsis* that are syntenic to the tomato *Ovate* region (KU *et al.* 2001). BLAST searches using the tomato *Ovate* BAC revealed that four genes, one being the tomato *Ovate* gene, are homologous to four lettuce genes found on three of the four lettuce blocks. All of the lettuce genes had homologs on the several *Arabidopsis* *Ovate* segments; however, *Arabidopsis* segments did not contain the full complement of genes. These results were also obtained in tomato–*Arabidopsis* genome comparisons (KU *et al.* 2001).

These findings suggest that the *Ovate* region has remained partially conserved since the last common ancestor of the Asteraceae, Brassicaceae, and Solanaceae. Since the *Ovate* region has been broken in both lettuce and *Arabidopsis* there is likely no evolutionary significance to these blocks in these species. While the region may have remained intact in tomato, there

appears to have been gene loss in this species similar to what has been noted in *A. thaliana* (KU *et al.* 2001) and what we see in lettuce. LK1374, located in LG 4 of lettuce, is homologous to genes in two of the *Arabidopsis* clusters (At2g18330 and At4g36580) but absent from the tomato BAC. Given the common position in lettuce and *Arabidopsis*, this gene was likely present in the last common ancestor and lost in the tomato BAC region, but possibly present in other duplicated regions of that species.

Polyploidization, segmental duplications, chromosomal rearrangements, and gene losses have likely led to the erosion of macrosynteny between *A. thaliana* and species in the Asteraceae. While these events have not totally obscured our detection of short regions with conserved gene repertoire and occasionally gene order predicted by PATERSON *et al.* (1996), elucidating gene homologies or positional cloning of target genes using *Arabidopsis* as a model for the structure of distantly related genomes will require the careful dissection of these complex syntenic networks and comparisons among multiple species (BOWERS *et al.* 2003). Comparative mapping between *Arabidopsis* and more divergent genomes will serve, however, to improve our understanding of the dynamic nature of plant genomes and the mechanisms that have shaped angiosperm diversification.

We thank Trevor Morin, Deidre Morgan, Patty Szczys, and Alex Plocik for technical support. This work was supported in part by grants from United States Department of Agriculture, Initiative for Future Agriculture and Food Systems (2000-04292), National Science Foundation, Plant Genome (DBI-0421630), National Science Foundation Undergraduate Mentoring in Environmental Biology (DEB-0080286) and National Science Foundation Research Experience for Undergraduates (DBI-0354125).

## LITERATURE CITED

- ADAMS, K., R. CRONN, R. PERCIFIELD and J. WENDEL, 2003 Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. USA* **100**: 4649–4654.
- ADAMS, K. L., and J. F. WENDEL, 2005 Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**: 135–141.
- ARABIDOPSIS GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.
- ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFER, 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.
- BENNETZEN, J., 2000 Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12**: 1021–1029.
- BENNETZEN, J., and M. FREELING, 1993 Grasses as a single genetic system: genome composition, colinearity, and compatibility. *Trends Genet.* **9**: 259–261.
- BENNETZEN, J. L., P. SANMIGUEL, M. CHEN, A. TIKHONOV, M. FRANCKI *et al.*, 1998 Grass genomes. *Proc. Natl. Acad. Sci. USA* **95**: 1975–1978.
- BERNATZKY, R., and S. D. TANSKLEY, 1986 Genetics of actin-related sequences in tomato. *Theor. Appl. Genet.* **72**: 314–321.
- BLANC, G., and K. WOLFE, 2004 Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**: 1679–1691.



- BLANC, G., A. BARAKAT, R. GUYOT, R. COOKE and M. DELSENY, 2000 Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* **12**: 1093–1101.
- BOIVIN, K., A. ACARKAN, R. MBULU, O. CLARENZ and R. SCHMIDT, 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.
- BOWERS, J., B. CHAPMAN, J. RONG and A. PATERSON, 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosome duplication events. *Nature* **422**: 433–438.
- CHARLESWORTH, D., 2002 Self-incompatibility: how to stay incompatible. *Curr. Biol.* **12**: R424–R426.
- CONNER, J. A., P. CONNER, M. E. NASRALLAH and J. B. NASRALLAH, 1998 Comparative mapping of the *Brassica* S locus region and its homeolog in Arabidopsis: implications for the evolution of mating systems in the Brassicaceae. *Plant Cell* **10**: 801–812.
- DOMINGUEZ, I., E. GRAZIANO, E. GEBHARDT, A. BARAKAT, S. BERRY *et al.*, 2003 Plant genome archaeology: evidence for conserved ancestral chromosome segments in dicotyledonous plant species. *Plant Biotechnol.* **1**: 91–99.
- ECKARDT, N. A., 2001 Everything in its place: conservation of gene order among distantly related plant species. *Plant Cell* **13**: 723–725.
- FULTON, T., R. V. D. HOEVEN, N. EANNETTA and S. TANKSLEY, 2002 Identification, analysis and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* **14**: 1457–1467.
- FREELING, M., 2001 Grasses as a single genetic system. Reassessment. *Plant Physiol.* **125**: 1191–1197.
- GALE, M. D., and K. M. DEVOS, 1998 Plant comparative genomics after 10 years. *Science* **282**: 656–659.
- GRANT, D., P. CREGAN and C. SHOEMAKER, 2000 Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proc. Natl. Acad. Sci. USA* **97**: 4168–4173.
- KELLOGG, E., 2003 It's all relative. *Nature* **422**: 383–384.
- KESSELI, R. V., and R. W. MICHELMORE, 1997 The Compositae: systematically fascinating but specifically neglected, pp. 179–191 in *Genome Mapping of Plants*, edited by A. H. PATERSON R.G. Landes, Georgetown, TX.
- KRUTOVSKY, K. V., M. TROGGIO, G. R. BROWN, K. D. JERMSTAD and D. B. NEALE, 2004 Comparative mapping in the Pinaceae. *Genetics* **168**: 447–461.
- KU, H.-M., T. VISION, J. LIU and S. D. TANKSLEY, 2000 Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA* **97**: 9121–9126.
- KU, H.-M., J. LIU, S. DOGANLAR and S. D. TANKSLEY, 2001 Exploitation of Arabidopsis-tomato synteny to construct a high resolution map of the ovate-containing region in tomato chromosome 2. *Genome* **44**: 470–475.
- KUITTINEN, H., A. HAAN, C. VOGL, S. OIKARINEN, J. LEPPALA *et al.*, 2004 Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **168**: 1575–1584.
- LAGERCRANTZ, U., 1998 Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes have evolved through extensive genome replication accompanied by chromosomal fusions and frequent rearrangements. *Genetics* **150**: 1217–1228.
- LAGERCRANTZ, U., J. PUTTERILL, G. COUPLAND and D. LYDIATE, 1996 Comparative mapping in Arabidopsis and Brassica, fine scale genome colinearity and congruence of genes controlling flowering time. *Plant J.* **9**: 13–20.
- LAI, Z., T. NAKAZATO, M. SALMASO, J. M. BURKE, S. TANG *et al.*, 2005 Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species. *Genetics* **171**: 291–303.
- LEE, J. M., D. GRANT, C. E. VALLEJOS and R. C. SHOEMAKER, 2001 Genome organization in dicots. II. Arabidopsis as a 'bridging species' to resolve genome evolution events among legumes. *Theor. Appl. Genet.* **103**: 765–773.
- LEITCH, L., and M. BENNETT, 1997 Polyploidy in angiosperms. *Trends Plant Sci.* **2**: 470–476.
- LIU, H., R. SACHIDANANDAM and L. STEIN, 2001 Comparative genomics between rice and Arabidopsis shows scant colinearity in gene order. *Genome Res.* **11**: 2020–2026.
- MASTERSON, J., 1994 Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**: 421–424.
- MCCOUCH, S. R., 2001 Genomics and synteny. *Plant Physiol.* **125**: 152–155.
- PATERSON, A. H., T. H. LAN, K. P. REISCHMANN, C. CHANG, Y.-R. LIN *et al.*, 1996 Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nat. Genet.* **14**: 380–381.
- PLOCIK, A., J. LAYDEN and R. KESSELI, 2004 Comparative analysis of NBS domain sequences of NBS-LRR disease resistance genes in sunflower, lettuce and chicory. *Mol. Phylogenet. Evol.* **31**: 153–163.
- PRINCE, J., E. POCHARD and S. TANKSLEY, 1993 Construction of a molecular linkage map of pepper and a comparison of synteny with tomato. *Genome* **36**: 404–417.
- QUIROS, C., F. GRELLER, 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.
- RIESEBERG, L., B. SINERVO, C. R. LINDER, M. UNGERER and D. ARIAS, 1996 Role of gene interactions in hybrid speciation: Evidence from ancient and experimental hybrids. *Science* **272**: 741–745.
- SALSE, J., B. PIEGU, R. COOKE and M. DELSENY, 2002 Synteny between Arabidopsis thaliana and rice at the genome level: a tool to identify conservation in ongoing rice genome sequencing project. *Nucleic Acids Res.* **30**: 2316–2328.
- SANMIGUEL, P., and J. L. BENNETZEN, 1998 Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* **82**: 37–44.
- SIMILLION, C., K. VANDEPOELE, M. VAN MONTAGU, M. ZABEAU and Y. VAN DE PEER, 2002 The hidden duplication past of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **99**: 13627–13632.
- SOLBRIG, O., 1977 Chromosomal cytology and evolution in the family Compositae, pp. 267–282 in *The Biology and Chemistry of the Compositae*, edited by V. HEYWOOD, J. HARBORNE and B. TURNER. Academic Press, London.
- STEBBINS, G. L., 1950 *Variation and Evolution in Plants*. Columbia University Press, New York.
- STEBBINS, G. L., 1971 *Chromosomal Evolution in Higher Plants*. Addison-Wesley Publishing, Reading, MA.
- TANKSLEY, S. D., M. W. GANAL, J. P. PRINCE, M. C. DE VICENTE, M. W. BONIERBALE *et al.*, 1992 High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**: 1141–1160.
- TARCHINI, R., P. BIDDLE, R. WINELAND, S. TINGEY and A. RAFALSKI, 2000 The complete sequence of 340 kb of DNA around the rice Adh1-Adh2 regions reveals interrupted colinearity with maize chromosome 4. *Plant Cell* **12**: 381–391.
- UYENOYAMA, M. K., 1995 A generalized least-squares estimate for the origin of sporophytic self-incompatibility. *Genetics* **139**: 975–992.
- VAN DODEWEERD, A. M., C. R. HALL, E. G. BENT, S. J. JOHNSON, M. W. BEVAN *et al.*, 1999 Identification and analysis of homoeologous segments of the genomes of rice and Arabidopsis thaliana. *Genome* **42**: 887–892.
- VISION, T. J., D. G. BROWN and S. D. TANKSLEY, 2000 The origins of genomic duplications in Arabidopsis. *Science* **290**: 2114–2117.
- WEIGEL, D., J. ALVAREZ, D. R. SMYTH, M. F. YANOFKY and E. M. MEYEROWITZ, 1992 LEAFY controls floral meristem identity in Arabidopsis. *Cell* **69**: 843–859.
- ZHU, H., D.-J. KIM, J.-M. BAEK, H.-K. CHOI, L. C. ELLIS *et al.*, 2003 Syntenic relationships between *Medicago truncatula* and Arabidopsis reveal extensive divergence of genome organization. *Plant Physiol.* **131**: 1018–1026.