The Arabidopsis Genome Sequence as a Tool for Genome Analysis in Brassicaceae. A Comparison of the Arabidopsis and *Capsella rubella* Genomes[1][w]

Karine Boivin[2], Adile Acarkan[3], Rosa-Stella Mbulu[4], Oliver Clarenz[5], and Renate Schmidt[6]*

Max-Delbrück-Laboratorium in der Max-Planck-Gesellschaft, 50829 Cologne, Germany

The annotated Arabidopsis genome sequence was exploited as a tool for carrying out comparative analyses of the Arabidopsis and *Capsella rubella* genomes. Comparison of a set of random, short *C. rubella* sequences with the corresponding sequences in Arabidopsis revealed that aligned protein-coding exon sequences differ from aligned intron or intergenic sequences in respect to the degree of sequence identity and the frequency of small insertions/deletions. Molecular-mapped markers and expressed sequence tags derived from Arabidopsis were used for genetic mapping in a population derived from an interspecific cross between *Capsella grandiflora* and *C. rubella*. The resulting eight Capsella linkage groups were compared to the sequence maps of the five Arabidopsis chromosomes. Fourteen colinear segments spanning approximately 85% of the Arabidopsis chromosome sequence maps and 92% of the Capsella genetic linkage map were detected. Several fusions and fissions of chromosomal segments as well as large inversions account for the observed arrangement of the 14 colinear blocks in the analyzed genomes. In addition, evidence for small-scale deviations from genome colinearity was found. Colinearity between the Arabidopsis and Capsella genomes is more pronounced than has been previously reported for comparisons between Arabidopsis and different Brassica species.

Cross-hybridization and genetic mapping studies are a powerful combination when comparing the gross chromosomal organization of two or more species (for review, see Schmidt, 2000; Schmidt, 2002), but in order to draw firm conclusions about genome colinearity between genomes the analysis has to be restricted to orthologous loci. However, a large proportion of markers used for genetic mapping cross-hybridizes with several sequences in the species analyzed. Thus, unless all loci corresponding to a particular marker have been mapped in both species, it cannot be determined whether positions of orthologous or paralogous loci are being compared. Consequently, a deviation from colinearity can often not be discriminated from the mapping of a paralogous sequence (Bennetzen, 2000). Here we show that this shortcoming of comparative genetic mapping experiments can

be overcome by taking advantage of the Arabidopsis chromosome sequence maps (Arabidopsis Genome Initiative, 2000) for comparative genome analysis between species of the Brassicaceae. The Arabidopsis chromosome sequence maps offer unique opportunities for comparative studies. First, the annotated genome sequence provides detailed information about the gene content in any genomic segment of interest. Second, the chromosomal map position and copy number of any given sequence can be determined. Third, it can be assessed whether a particular locus is located in any of the duplicated chromosomal segments that have been identified in the Arabidopsis genome (Arabidopsis Genome Initiative, 2000; Blanc et al., 2000; Vision et al., 2000; Bowers et al., 2003; Ermolaeva et al., 2003; Raes et al., 2003).

Haploid chromosome numbers vary among Brassicaceous species. In Arabidopsis the haploid set consists of 5 chromosomes, whereas many close relatives such as *Capsella rubella* have $n = 8$ chromosomes. Phylogenetic analyses within the tribe Arabideae suggested that base chromosome numbers lower than $n = 8$ are derived because base chromosome number reduction from $n = 8$ to $n = 5$ to 7 occurred several times (Koch et al., 1999).

The progenitors of the lineage leading to Arabidopsis and *C. rubella* diverged approximately 10 million years ago (Koch et al., 2000, 2001) and the first comparative mapping studies revealed conserved gene repertoires among these species (Acarkan et al., 2000; Rossberg et al., 2001). Thus, these two closely related species that differ in respect to base chromosome number offer an excellent opportunity to reveal patterns of chromosome evolution in Brassicaceae.

All markers used for genetic mapping in a population derived from an interspecific cross of *Capsella grandiflora* and *C. rubella* were sequenced in order to enable the characterization of each marker with respect to copy number and chromosomal map position in Arabidopsis. This information was used for a comparison between the Arabidopsis chromosome sequence maps and the *Capsella* linkage map. Extensive colinearity of these genomes was apparent but large- and small-scale deviations from genome colinearity were also identified and characterized to provide insight into factors of importance to chromosome evolution.

The annotated Arabidopsis genome sequence also facilitates the study of sequence evolution. Comparative sequence analysis of selected orthologous regions of Arabidopsis and *C. rubella* revealed a high degree of sequence identity for protein-coding sequences. In contrast, intergenic regions and introns are differently sized in both species, and overall sequence identity is generally not found (Acarkan et al., 2000; Rossberg et al., 2001). In the study presented here, a set of random short *C. rubella* sequences was compared with the annotated Arabidopsis genome sequence to assess which kind of sequences show conservation between these closely related species. A detailed analysis of sequence alignments revealed that aligned protein-coding sequences do differ from aligned intron or intergenic sequences with respect to the degree of sequence identity and the frequency of small insertions/deletions (indels). Thus, significant sequence similarities between these closely related species are not restricted to protein-coding exon sequences.

RESULTS

Conservation of Sequence Repertoires in Arabidopsis and *C. rubella*

C. rubella DNA was restricted with *Mbo*I and cloned. A total of 137 different clones were sequenced (AJ581160-AJ581296). Insert sizes of the cloned *Mbo*I fragments ranged from 106 to 783 bp, with an average of 434 bp. A total of 113 (82.5%) of the fragments corresponded to Arabidopsis sequences, while 24 (17.5%) showed no significant identity to either Arabidopsis sequences or to any other sequences available in the databases. Sequence homology to single or low-copy regions in the Arabidopsis nuclear genome was established for 71 (51.8%) of the *C. rubella Mbo*I fragments. Twelve sequences (8.8%) were similar to repeated Arabidopsis DNA sequences of nuclear origin, and 30 sequences (21.9%) represented sequences of the organellar genomes (Table I).

Using the program BLAST 2 Sequences (Tatusova and Madden, 1999), the 71 *C. rubella Mbo*I fragments that corresponded to single or low-copy regions in the Arabidopsis genome were aligned with the Arabidopsis sequence showing the highest overall DNA se-

Table I. *Conservation of sequence repertoires in Arabidopsis and C. rubella: a comparison of 137 sequences of C. rubella MboI fragments with Arabidopsis DNA sequences from both nuclear and organellar genomes*

C. rubella Sequences Corresponding to	No. of C. rubella Sequences (of 137 MboI Fragments Analyzed)
Arabidopsis sequences	113 (82.5%)
Organellar sequences	30 (21.9%)
Chloroplast DNA sequences	26 (19.0%)
Mitochondrial DNA sequences	4 (2.9%)
Repetitive nuclear DNA sequences	12 (8.8%)
rDNA sequences	10 (7.3%)
18-5.8-25S rDNA	9 (6.6%)
5S rDNA	1 (0.7%)
Retrotransposon sequences	2 (1.5%)
Low-copy nuclear DNA sequences	71 (51.8%)
Predicted gene sequences	54 (39.4%)

quence similarity. With the parameters chosen, 23,200 of 30,856 bp of the *C. rubella* sequences (75.2%) were found in high scoring sequence pairs. This analysis was also performed with a 27-kbp long *C. rubella* region and the orthologous area of the Arabidopsis genome that is located on chromosome I (Rossberg et al., 2001; Table II).

A detailed analysis of the aligned sequences showed correspondence to annotated Arabidopsis protein-coding sequences for 54 out of the 71 *C. rubella Mbo*I fragments (76.1%; Table I). Based on the annotation of the Arabidopsis genome, 51.8% and 15.6% of the aligned sequences consisted of protein-coding exon and intron sequences, respectively. The remainder of the alignments, 32.5%, which included 5'- and 3'-untranslated regions of genes, was classified as intergenic sequences (data not shown). The comparison of the contiguous *C. rubella* genomic DNA region and its orthologous counterpart in Arabidopsis revealed that the alignments consisted of 55.7% protein-coding exon, 9.0% intron, and 35.3% intergenic sequences (data not shown). Thus, the fraction of aligned intron sequences in the contiguous orthologous regions was, at 9%, much lower than the fraction observed for the dataset of the aligned random *Mbo*I fragments (15.6%). This difference between the two datasets reflected that only about 20% of the Arabidopsis genic sequences corresponded to intron sequences in this particular region of Arabidopsis chromosome I (Rossberg et al., 2001), whereas the average value for the Arabidopsis genome amounts to approximately 35% (Arabidopsis Genome Initiative, 2000).

Sequence identity values of about 90% were found for aligned protein-coding exon sequences, whereas values of approximately 80% were observed for aligned sequences that consisted of intron or intergenic sequences regardless of whether the dataset of the random *Mbo*I fragments was analyzed or whether the contiguous orthologous regions were evaluated (Table II).

Table II. *Nature and characteristics of sequences conserved between the Arabidopsis and C. rubella genomes*

Features of High Scoring Sequence Pairs (C. rubella Compared to Arabidopsis Sequences)	Random *Mbo*I Fragments of *C. rubella* DNA (% of 30,856 bp Analyzed in Total)	Contiguous Region of *C. rubella* DNA (% of 27,056 bp Analyzed in Total)
Total length of high scoring sequence pairs	23,200 bp (75.2%)	19,468 bp (72.0%)
Average sequence identity of aligned sequences	86.0%	86.6%
Frequency of indels in aligned sequences	1 indel/80 bp	1 indel/77 bp
Total length of aligned exon sequences	12,024 bp (38.9%)	10,842 bp (40.1%)
Average sequence identity of aligned exon sequences	91.3%	90.0%
Frequency of indels in aligned exon sequences	1 indel/573 bp	1 indel/471 bp
Total length of aligned intron sequences	3,628 bp (11.8%)	1,759 bp (6.5%)
Average sequence identity of aligned intron sequences	78.9%	80.4%
Frequency of indels in aligned intron sequences	1 indel/39 bp	1 indel/46 bp
Total length of aligned intergenic sequences	7,548 bp (24.5%)	6,867 bp (25.4%)
Average sequence identity of aligned intergenic sequences	80.9%	81.5%
Frequency of indels in aligned intergenic sequences	1 indel/42 bp	1 indel/36 bp

In total, 291 small indels with an average length of 3.4 bp (1–21 bp) were found in the aligned sequences of the *Mbo*I fragments. In the alignments of the contiguous orthologous regions, 254 indels were found. These indels ranged in length from 1 to 18 bp and spanned on average 3.2 bp (Table II). Regardless of which of the two datasets was analyzed, indels were observed on average once every 80 bp, but they were much more frequent in aligned intron or intergenic sequences than in exon sequences (Table II). The alignments of intron and intergenic sequences showed indels on average once every 40 bp and more than one-half of the indels spanned 1 or 2 bp. In contrast, the sizes of indels in alignments of exon sequences corresponded to one or more codons. Treating all indels as small insertions showed that this type of sequence alteration was approximately twice as frequent in the *C. rubella* sequences when compared to the Arabidopsis sequences (data not shown).

Use of Arabidopsis Sequences as Markers in Genetic Mapping Experiments in Capsella

Fifty self-compatible F_2 plants that were derived from an interspecific cross of *C. grandiflora* and *C. rubella* made up the Capsella mapping population (Acarkan et al., 2000). Sixty-two Arabidopsis RFLP-markers and 36 Arabidopsis expressed sequence tags (ESTs) were chosen for genetic mapping. The prefixes "m" and "mi" designate RFLP markers developed by Fabri and Schäffner (1994) and Liu et al. (1996), respectively, whereas the prefix "E" denotes Arabidopsis ESTs (Höfte et al., 1993; Newman et al., 1994). Additionally, three *C. rubella* genomic DNA fragments were used for RFLP analysis (C1, C54, and Cos20).

A single codominant polymorphism was scored for each of 84 RFLP markers, 2 for 13 markers, and 3 for a single marker in Capsella. RFLP mapping of sequences derived from the 18S-25S rDNA loci in Arabidopsis identified 2 loci in Capsella, and a codominant polymorphism was scored for locus rDNAa, whereas a dominant polymorphism was evaluated for locus

rDNAb. Additionally, 19 loci were placed on the Capsella map using PCR-based methods. The resulting genetic map consisted of 133 codominant loci distributed over 8 linkage groups and spanned 582.1 cM (Fig. 1).

Three markers (E6, E9, and E20) harboring mitochondrial DNA sequences showed maternal inheritance; thus all F_2-plants carried the *C. grandiflora* allele of these markers. For a nuclear-encoded codominant locus, the expected segregation among the F_2 progeny is a 1:2:1 ratio of plants homozygous for the *C. grandiflora* allele, heterozygous, and homozygous for the *C. rubella* allele, respectively. The results of χ^2-tests ($P = 0.05$) revealed that the observed segregation ratios were significantly different from the expected distribution for 25 of the 133 loci (Fig. 1). All but 3 (E35b and mi353, linkage group B; E57, linkage group D) of these 25 loci map to linkage groups F and G. Markers CL5.1 and m326 delimit a chromosomal region on linkage group G that is characterized by a significant under-representation of homozygous *C. grandiflora* plants. In contrast, *C. grandiflora* alleles are significantly over-represented for 8 of the loci showing a distorted segregation (E31, mi219, mi90, mi433, mi138, mi438, mi74b, and mi174), which map to linkage group F. Segregation distortion has previously been noted for plant populations derived from interspecific crosses (e.g. Livingstone et al., 1999).

Comparison of the Capsella Linkage Maps with the Sequence Maps of the Arabidopsis Chromosomes

A comparison of the marker sequences with the Arabidopsis annotated gene sequences revealed that 113 out of the 117 markers (96.6%) harbored protein-coding sequences (Supplemental Table S1, which can be viewed at www.plantphysiol.org). Sixty-two of the markers corresponded to sequences mapping to a single locus (53.0%) in the Arabidopsis genome. Two loci were found for 33 markers, 3 or more loci were recorded for 21 markers, and 1 marker (C57) did not
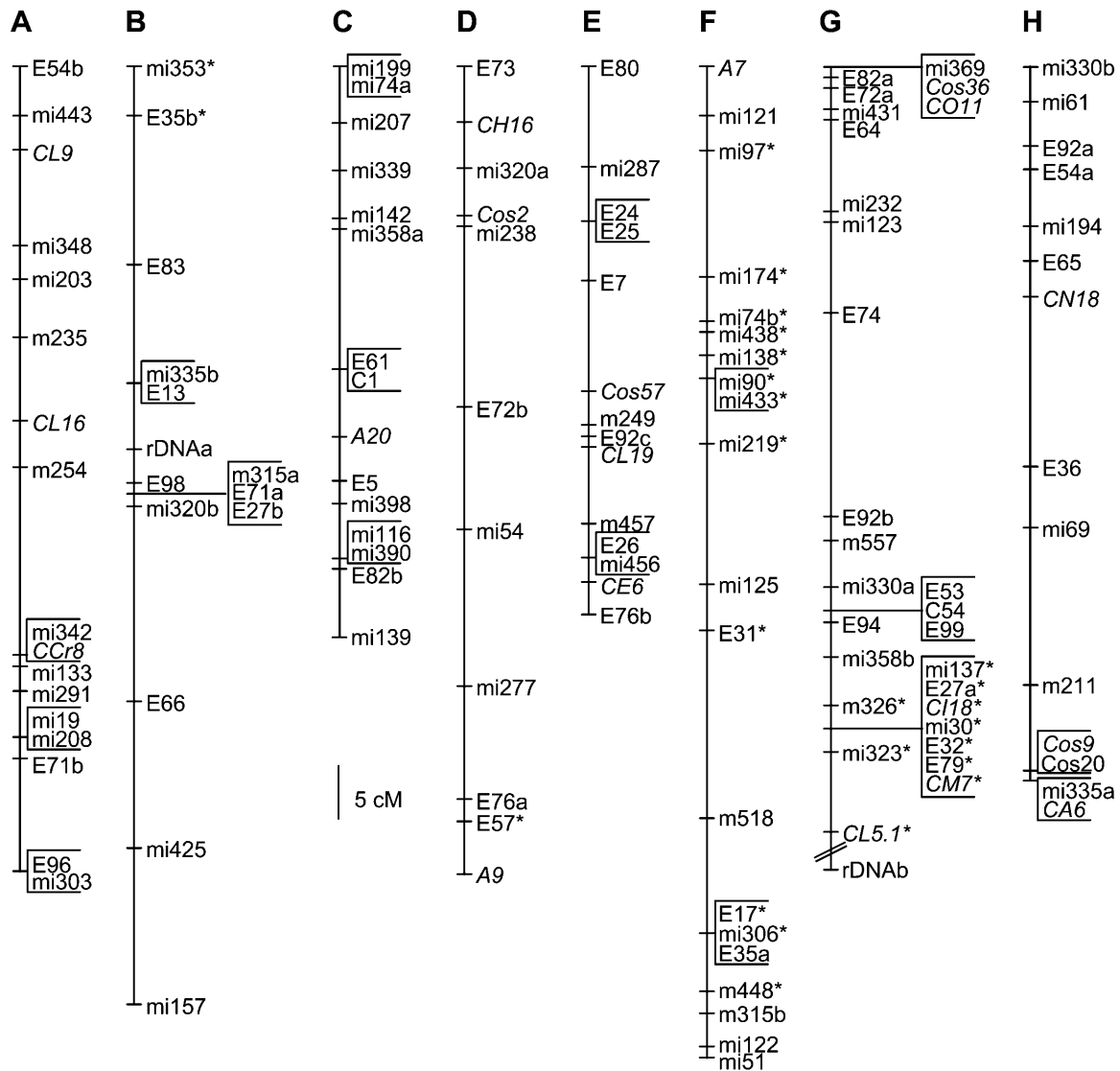
Figure 1. Genetic linkage map of Capsella. Linkage data for the Capsella mapping population derived from an interspecific cross were calculated with Map Manager QTX (Manly et al., 2001). Using a logarithm of odds ratio score of 5.0, 8 linkage groups labeled A to H resulted. Map distances are given in cM. PCR-based markers are shown in italics, and the remainder of the loci was mapped using RFLP. Nuclear codominant loci for which the observed ratios of plants homozygous for the *C. grandiflora* allele, heterozygous plants, and plants homozygous for the *C. rubella* allele that deviated significantly ($P = 0.05$) from the expected 1:2:1 ratio are marked by asterisks. Two loci corresponding to rDNA sequences were mapped: a codominant locus (rDNAa) was located on linkage group B and a dominant locus (rDNAb) was found linked to markers of linkage group G. An interrupted line indicates that the genetic distance separating rDNAb and CL5.1 has not been established.

show any significant sequence identity with Arabidopsis sequences (Supplemental Table S1).

Figure 2 shows a comparative map of the 8 Capsella linkage groups and the 5 Arabidopsis chromosomes. A comparison of the map positions of the 62 single-locus marker sequences with those of the corresponding loci in Capsella clearly indicated that 61 locus pairs reside in colinear positions on the Arabidopsis and Capsella maps. The mapping of marker E80, which represented a single-copy sequence in Arabidopsis as well as in Capsella, revealed a translocation, the extent of which is unknown.

For 48 out of the 54 markers for which 2 or more loci were found in the Arabidopsis genome, either the marker sequence itself or its closest Arabidopsis homolog resided in a colinear position with the genetically mapped Capsella locus. Colinearity between a Capsella locus and a homolog of a marker sequence was found for an additional 15 locus pairs.

In total, comparative mapping revealed 124 locus pairs distributed over 14 large colinear segments on the Arabidopsis and Capsella maps. Two of these segments may harbor small-scale inversions and/or translocations (Capsella linkage group B, Arabidopsis
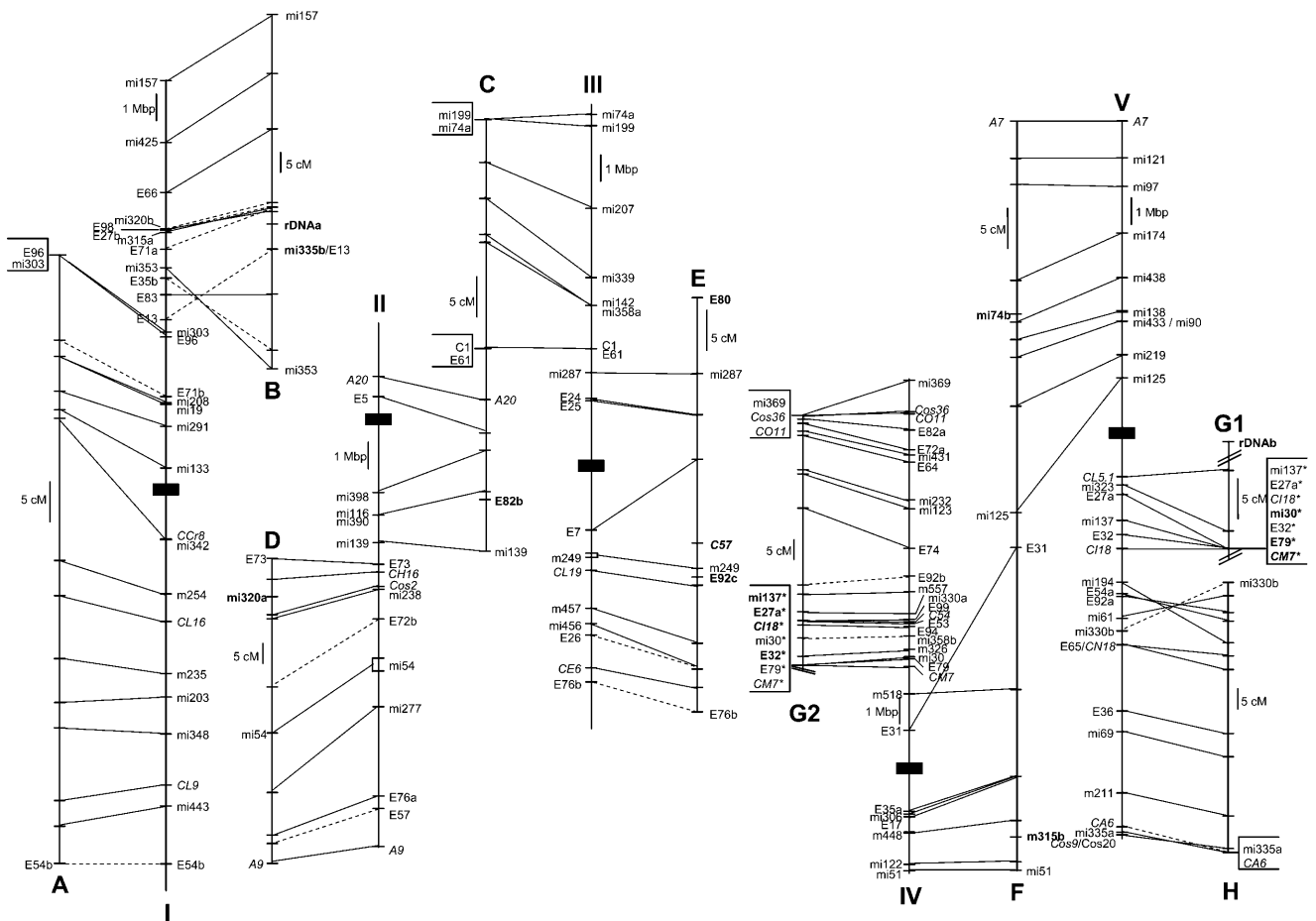
Figure 2. Comparative map between Arabidopsis and Capsella. The sequence maps of the five Arabidopsis chromosomes are shown. Horizontal bars mark the positions of the centromeres. The designations of the chromosomes (I to V) indicate the telomeric ends of the short arms. The eight Capsella linkage groups are labeled A to H. Linkage group G is shown in two parts, G1 and G2. Linkage groups B, D, G2, and H are shown at a smaller scale than the remainder of the genetic chromosome maps. Map positions of marker sequences are indicated as vertical lines on the chromosome maps. Two different loci in close physical proximity to each other (Supplemental Table S1) are shown for markers mi54 and m249. Lines connect all loci residing in colinear positions. A solid line indicates that the map position of the Capsella locus corresponds to the map position of the Arabidopsis marker sequence itself or the closest Arabidopsis homolog of a Capsella PCR-based marker. Colinearity between a Capsella locus and a homolog of a marker sequence is indicated by a dashed line. Capsella loci shown in bold do not have a corresponding Arabidopsis sequence in a colinear arrangement.

chromosome I; Capsella linkage group H, Arabidopsis chromosome V). Each of the five Arabidopsis chromosomes corresponds to two or three different Capsella linkage groups or segments thereof. Thus, translocations or fusions of large chromosome segments were an important factor in differentiating the genomes of the progenitors of Capsella and Arabidopsis since their divergence. Nuclear organizer regions (NORs) adjoin the telomeres on the short arms of Arabidopsis chromosomes II and IV (Copenhaver and Pikaard, 1996). RFLP mapping of the 18S rDNA sequences in Capsella also revealed two rDNA loci, but neither of these two loci map in a colinear position with the rDNA loci in Arabidopsis, indicating that the NORs had also been involved in translocations. Fur-

thermore, evidence for inversions of large chromosome segments was found (Capsella linkage group B, Arabidopsis chromosome I; Capsella linkage group F, Arabidopsis chromosome IV; Capsella linkage group H, Arabidopsis chromosome V).

For each of five markers that mapped to 2 loci in Arabidopsis (E27, E54, E72, E76, and mi358), 2 loci were also found on the Capsella genetic map. Colinear positions on the Arabidopsis and Capsella maps were found for all 10 locus pairs. With the exception of marker mi358, these markers resided in duplicated regions of the Arabidopsis genome. Likewise, the positions of loci E92a, E92b, mi330a, and mi330b in the Capsella genome were found to be colinear with the Arabidopsis loci, mapping to a duplicated region

between Arabidopsis chromosomes IV and V (Fig. 3A). These results are consistent with the occurrence of this segmental duplication in the Capsella genome.

In addition to large-scale rearrangements, evidence for small-scale deviations from colinearity was found. For example, Capsella loci C57, E92c, and mi320a did not have a corresponding Arabidopsis sequence in a colinear position. In three cases copy-number changes were detected; for Arabidopsis single-locus markers E82, m315A, and mi74, two loci each were mapped in Capsella. Despite the fact that markers
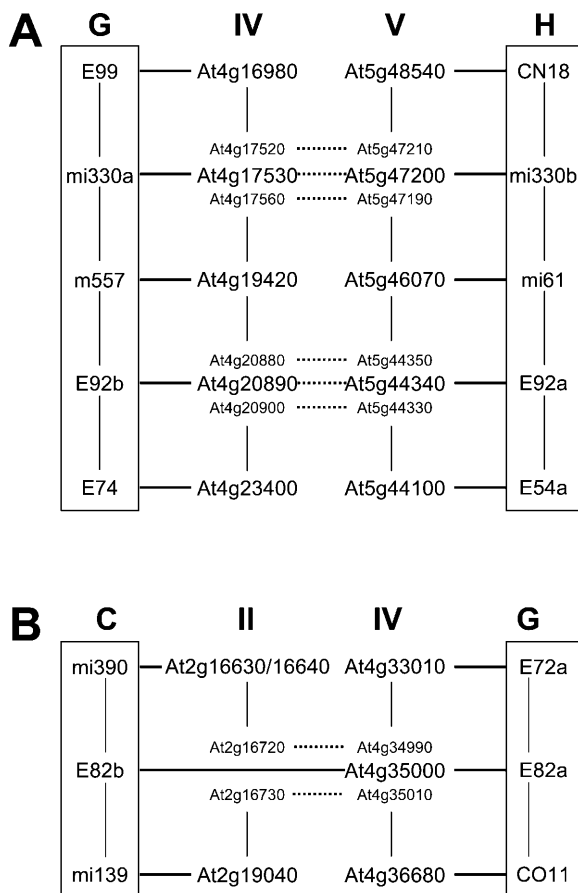
mi74 and E82 are single-locus sequences in Arabidopsis, they are located in segmentally duplicated regions of the Arabidopsis genome. The comparative mapping results indicate that two copies of markers mi74 and E82 each should have been present in the duplicated segments harbored by the common progenitor of Arabidopsis and Capsella, but one copy for each of the markers was lost in the lineage leading to Arabidopsis, whereas both copies of markers mi74 and E82 were retained in the lineage leading to Capsella (Fig. 3B and data not shown).

Two loci were found in Capsella for Arabidopsis single-locus marker mi335. Because marker mi335 harbors three different protein sequences (Supplemental Table S1), it cannot be discriminated whether this deviation from colinearity is due to a copy number change or a small-scale translocation involving part of the marker sequence.

Figure 3. Comparative mapping of segmental duplications. Arabidopsis markers mapping to duplicated regions of the Arabidopsis genome have been used for genetic mapping in Capsella. Selected parts of a segmental duplication harboring duplicated sequences of markers E92 and mi330 are shown in A. Genes mapping to a duplicated region between Arabidopsis chromosomes II and IV and their Capsella counterparts C and G are represented in B. In both sections, only selected parts of the Capsella genetic maps and Arabidopsis chromosome sequence maps are given schematically. The Capsella linkage groups are boxed. The representation of Arabidopsis gene codes mapping to the duplicated regions has been restricted to genes corresponding to genetically mapped markers in Capsella and those duplicated genes that are found immediately adjacent to the duplicated marker loci (shown at a smaller font size). Horizontal lines indicate colinear relationships between an Arabidopsis gene and a genetically mapped locus in Capsella. Dashed lines connect homologous Arabidopsis genes that are located in duplicated regions of the genome.

DISCUSSION

Earlier comparative sequence analyses of orthologous regions of Arabidopsis and *C. rubella* have revealed a high degree of sequence identity for protein-coding sequences. In contrast, intergenic regions and introns are differently sized in both species, and overall sequence identity is generally not found (Acarkan et al., 2000; Rossberg et al., 2001). The study of the alignments between the Arabidopsis and *C. rubella* sequences presented here corroborates these findings. The Arabidopsis genome consists of approximately 28.8% protein-coding exons, 15.7% introns, and 55.5% intergenic sequences (Arabidopsis Genome Initiative, 2000). Alignment of *C. rubella* DNA sequences with the corresponding Arabidopsis sequences revealed that protein-coding sequences represented approximately one-half of the aligned sequences, whereas the other one-half consisted of intron and intergenic sequences (Table II). Thus, significant levels of sequence conservation in the alignments are overrepresented in protein-coding exon sequences when compared to intron and intergenic sequences. Nevertheless, it is important to note that significant levels of sequence conservation between Arabidopsis and Capsella are not restricted to protein-coding exon sequences.

Regardless of whether sequences of random *C. rubella* Mbo*I fragments or of a contiguous 27-kbp region were compared to the corresponding sequences in Arabidopsis, the average degree of sequence identity was approximately 10% higher in aligned protein-coding sequences than in that of aligned intron or intergenic sequences. Comparative sequence analysis of eight orthologous genes in Arabidopsis and *C. rubella* revealed approximately 90% sequence identity at the nucleotide level, but for a rapidly evolving gene a much lower value of about 80% was observed (Rossberg et al., 2001). Thus, the observed 10% difference in average sequence identity levels in aligned

exon sequences compared to alignments of intron or intergenic sequences is not sufficient to distinguish aligned protein-coding sequences from alignments of intron or intergenic regions in an unambiguous manner. In contrast, both indel size and frequency are appropriate features to differentiate between aligned protein-coding exon sequences and alignments of intron or intergenic sequences of closely related species. Detailed inspection of the alignments revealed that these features are especially powerful to help determine the beginning and end of an open reading frame. Thus, the current weakness of gene structure prediction programs to identify coding region limits (Mathé et al., 2002) may be overcome by taking advantage of data concerning indel size and frequency in aligned sequences of closely related species.

The analysis of double strand break repair revealed that larger deletions were found more frequently in Arabidopsis than in *Nicotiana tabacum*. Whereas 40% of the deletions were accompanied by insertions in *N. tabacum*, this was not the case in Arabidopsis. Based on these results, Kirik et al. (2000) proposed that species-specific differences in double strand break repair might influence genome evolution. Interestingly, through analysis of indel frequencies in the alignments between the Arabidopsis and *C. rubella* sequences, another species-specific difference concerning sequence evolution was revealed. Treating all indels in the alignments of the Arabidopsis and *C. rubella* sequences as insertions, it emerged that small insertions occurred approximately twice as often in *C. rubella* sequences than in Arabidopsis sequences (data not shown). However, it was not possible to discriminate whether these results reflect an overall higher rate of small insertions in *C. rubella* and/or a generally higher rate of small deletions in Arabidopsis.

In comparative genetic mapping experiments, it is often not possible to distinguish deviations from colinearity from the mapping of paralogous sequences (Bennetzen, 2000). This shortcoming was overcome in the study of the Arabidopsis and Capsella genomes due to the availability of the Arabidopsis chromosome sequence maps (Arabidopsis Genome Initiative, 2000), which allow the establishment of copy numbers and positions for all markers used in the Arabidopsis genome. For example, marker E92 has 10 copies in the Arabidopsis genome (data not shown). In Capsella, three polymorphic loci and several monomorphic fragments were detected (data not shown). Loci flanking the three different loci on the Capsella linkage maps were found in colinear arrangements in Arabidopsis. In the colinear blocks defined by the flanking markers orthologous locus pairs were found for E92a and E92b, whereas none of the 10 copies corresponding to marker E92 was present in a colinear arrangement with E92c (Figs. 2 and 3A). Thus, mapping the E92c locus revealed a deviation from colinearity.

The availability of the Arabidopsis chromosome sequence maps was also exploited to ensure good coverage of the comparative map because it offers the opportunity to target any particular region of the genome for a comparative mapping study. This is particularly useful if mapping results indicate a deviation from colinearity. For example, the order of markers mi330b and mi194 on Capsella linkage group H was inverted when compared to their arrangement on the sequence map of Arabidopsis chromosome V. Mapping of additional markers (E54a, E92a, and mi61) located in this interval of the Arabidopsis genome unambiguously showed that this was due to a large chromosomal inversion and not by a translocation of a chromosomal segment harboring a marker sequence (Fig. 2).

Comparative physical mapping in Arabidopsis and *C. rubella* revealed that one particular region duplicated between Arabidopsis chromosomes I and III was also found in two copies in *C. rubella* (Rossberg et al., 2001). The comparative study of several markers located in duplicated regions of the Arabidopsis genome that were investigated here gave comparable results (Fig. 3 and data not shown), indicating that these segments should have been present in the common ancestor of Arabidopsis and Capsella (Fig. 3 and data not shown). Consistent with this finding the estimated values for the age of the duplication events in the Arabidopsis genome (Lynch and Conery, 2000; Vision et al., 2000; Bowers et al., 2003; Ermolaeva et al., 2003; Raes et al., 2003) are much higher than the divergence time of Arabidopsis and Capsella, which was calculated at 10 million years ago (Koch et al., 2000, 2001). Gene loss in duplicated regions of the genome has been proposed to be an important factor in shaping plant genomes (for review, see Schmidt, 2002); the comparative mapping results for markers

Table III. *Oligonucleotide combinations used for amplification of PCR-based markers from Capsella genomic DNA*

PCR Marker	Oligonucleotide Combination
A7	5′-gctgtatacatcacttctgg-3′/5′-cactcagtctatgcctggagc-3′
A9	5′-cactccatgtctcaacttcatc-3′/5′-ggtaatgtagcacactacgg-3′
CA6	5′-gcagcttctgtgaaaccettaacc-3′/5′-ccatcttctcacgctcaatcc-3′
CCr8	5′-ggagctctcaaacgactttg-3′/5′-ccaatgtcatgttcacttcatc-3′
CE6	5′-ccacgtccagcatcgtcgcatc-3′/5′-catacctcgcttgttatgtac-3′
CH16	5′-ggaagatacggtaactgg-3′/5′-gaatctggttcattcagcttcg-3′
CI18	5′-gatcttgcgttttgagctc-3′/5′-cttaatcactactggtcctgg-3′
CL5.1	5′-ggacaagcccttgctcaatac-3′/5′-gagaaccatatgatatcaactg-3′
CL9	5′-gcttgaccacaacacttatg-3′/5′-ggactgtgagcataatcctc-3′
CL16	5′-atcagcacgggtacggtcaaat-3′/5′-gcactccaagcttaatctgc-3′
CL19	5′-gctcgggtagagaatggagatg-3′/5′-cgttcctctgtttttcgctgc-3′
CM7	5′-gtcgaaatgtgtggtgccgagtc-3′/5′-gaccatttccttcagcttcatcc-3′
CN18	5′-gtctcggttacattagcaacg-3′/5′-ccaagtttccaagaacatcg-3′
CO11	5′-caacgccttgctcaacgcttgtc-3′/5′-cctttttgagcactcataagac-gaac-3′
Cos2	5′-gctaaggatatgtgatgataag-3′/5′-gctgccttgtaggcttacgag-3′
Cos9	5′-ggacgaggatgagctgctgagtc-3′/5′-ctaacttggaggtctaagagc-3′
Cos36	5′-cagcgttttcagcggttgccacgtg-3′/5′-cgtatatgtcacatgaacaa-caacgctg-3′
Cos57	5′-gagtggtattcattcaccatgac-3′/5′-ctttgagttgaacactacatat-gtag-3′

E82 and mi74 lend further support to this view (Fig. 3B and data not shown).

Comparing the organization of the colinear blocks and the NORs in the two genomes unveiled 14 large chromosomal rearrangements. In addition to these changes involving large chromosome segments, the genome arrangement of the two species is distinguished by numerous small rearrangements. Approximately 6% of the analyzed loci revealed such changes, which included deletions/insertions, duplications, and/or translocations of gene sequences. No attempt was made to map all loci corresponding to the different marker sequences in Capsella; thus it is reasonable to assume that such small-scale changes are far more frequent than indicated by the data presented here.

Koch et al. (1999) concluded from results of phylogenetic studies that base chromosome numbers lower than $n = 8$ are derived in the tribe Arabideae. The results of comparative mapping between Arabidopsis and Capsella showed that such a reduction of base chromosome number cannot be attributed exclusively to chromosome fusions. If the chromosome number $n = 8$ of Capsella represents the ancestral state, at least 6 chromosome fusion (involving linkage groups A/B, C/D, C/E, G/F, and F/G/H) and three chromosome fission events (involving linkage groups C, F, and G) must have taken place for the observed arrangement of the 14 colinear blocks in the 8 Capsella and the 5 Arabidopsis chromosomes to have occurred. Furthermore, comparative mapping revealed that the 2 genomes are distinguished by three large chromosomal inversions and 2 translocations involving the NORs.

Previous comparative genome analyses in the Brassicaceae have largely been focused on the different Brassica species or on comparisons of these genomes to that of Arabidopsis (for review, see Schmidt et al., 2001). Comparative mapping using Brassica oleracea, Brassica rapa, and Brassica nigra revealed an almost complete conservation of gene repertoire, but the genomes were distinguished by multiple rearrangements (Lagercrantz and Lydiate, 1996). Analyses between Arabidopsis and B. oleracea led to the detection of conserved linkage arrangements (Kowalski et al., 1994; Lan et al., 2000; Babula et al., 2003; Lukens et al., 2003), but these were much smaller than the sizes of the colinear blocks that were observed for the Arabidopsis and Capsella genomes. These encompassed 7.2 Mbp of the Arabidopsis genome and 38.2 cM of the Capsella genome. This finding may in part reflect the more ancient divergence of the Arabidopsis and Brassica lineages when compared to that of the progenitors of Arabidopsis and Capsella (Yang et al., 1999; Koch et al., 2000, 2001). However, it should also be noted that comparative mapping indicated the presence of many duplicated and triplicated segments in the B. oleracea, B. rapa, and B. nigra genomes. Detailed microcolinearity studies between selected genomic regions in Arabidopsis and B. oleracea showed that the separation of the lineages leading to Arabidopsis and Brassica predated the triplications seen in the Brassica genomes (O'Neill and Bancroft, 2000; Schmidt et al., 2003). Thus, it is conceivable that the high number of chromosomal rearrangements that distinguish the Arabidopsis from the B. nigra and B. oleracea genomes is at least in part due to the relatively recent polyploidization in the Brassica species (Lagercrantz, 1998; Lukens et al., 2003).

Probes derived from Arabidopsis chromosome IV were used for comparative chromosome painting in the closely related species Arabidopsis halleri, Arabidopsis lyrata, Cardaminopsis carpatica, and C. rubella, all of which share a base chromosome number of $n = 8$ (Lysak et al., 2003). Importantly, in all four species, the arrangement of colinear segments was consistent with the results of the comparative mapping between Arabidopsis and Capsella. Furthermore, the homeologs of Capsella linkage group G in these species carried a NOR in the same position as indicated by the RFLP mapping studies of 18S-25S rDNA sequences in Capsella (Figs. 1 and 2). These results clearly show that the genetic linkage map of Capsella based on Arabidopsis markers that we have established will prove to be an indispensable tool for chromosome mapping studies in close relatives and the study of chromosome evolution in Brassicaceae.

MATERIALS AND METHODS

Sequencing

DNA sequences were determined using PE/Applied Biosystems 377 and 3700 sequencers using BigDye-terminator chemistry (Perkin-Elmer, Überlingen, Germany) by the ADIS unit at the Max-Planck-Institut für Züchtungsforschung (Köln, Germany). Oligonucleotides were purchased from Metabion (Martinsried, Germany). Analysis of sequences was carried out using the Wisconsin Package (Version 10.0-UNIX, Genetics Computer Group, Madison, WI), BLAST (Altschul et al., 1990), and BLAST 2 sequences (Tatusova and Madden, 1999).

C. rubella Library of MboI Fragments

C. rubella DNA was digested with MboI, and fragments of a size range from 0.5 to 1.5 kbp were cloned into vector pGEM 7Zf+ (Promega GmbH, Mannheim, Germany). The inserts of 137 clones were sequenced and subjected to a BLASTN analysis (Altschul et al., 1990) to search for corresponding sequences in the Arabidopsis nuclear and organellar genomes. As a threshold for reporting a match, $E < 10^{-9}$ was chosen (parameters used for the alignments: nucleotide match 1, nucleotide mismatch -2, gap open penalty 5, and gap extension penalty 1). Sequences that did not match Arabidopsis sequences were then used to search the "nr," "est," and "gss" divisions of GenBank.

C. rubella sequences were aligned with the corresponding Arabidopsis sequences using the program BLAST 2 sequences (Tatusova and Madden, 1999) with the same parameters used for the BLASTN analysis.

RFLP Markers

Preparation of plant genomic DNA- and Southern-blot analyses were carried out as previously described (Schmidt et al., 1999). Genomic DNA of C. rubella and a pool of F_2-plants derived from an interspecific cross of Capsella grandiflora $\times$ C. rubella (Acarkan et al., 2000) were digested with restriction endonucleases (BglII, DraI, EcoRI, EcoRV, HindIIII, and XbaI). The resulting Southern blots were probed with Arabidopsis RFLP markers (Fabri and Schäffner, 1994; Liu et al., 1996), Arabidopsis ESTs (Höfte et al., 1993; Newman et al., 1994), and C. rubella genomic DNA fragments (C2, C20, and C54) to

reveal polymorphisms. A 420-bp fragment corresponding to the 18S rRNA gene was taken as a probe to establish RFLPs for the 18S-25S rDNA loci.

The Arabidopsis cDNA clones used as RFLP markers were denoted as follows: E5 (VBVEA05), E6 (FAFM25), E7 (VBVAH05), E9 (YAP234T7), E13 (139A22T7, AJ608275), E17 (172G2T7), E24 (102B12T7), E25 (198N17T7), E26 (104M9T7), E27 (92I17T7), E30 (YAY106), E31 (241C22T7), E32 (3H8TT), E35 (YAY337), E36 (85E6T7), E53 (113M1T7, AJ608276), E54 (G10A6T7), E57 (133K4T7), E61 (149C21T7), E64 (192P5T7), E65 (174E13T7), E66 (166D7T7), E71 (91P17T7), E72 (OAO172), E73 (206L7T7), E74 (109G22T7), E76 (176G19T7), E79 (VBV08–30792), E80 (TAP0180), E82 (VBVEB09), E83 (AJ608277), E92 (192F6T7), E94 (OAO217), E96 (G2F2T7), E98 (198N17T7), and E99 (c13.049, AJ299418).

PCR-Based Markers

PCR-Based markers were developed using sequence information derived from *C. rubella* genomic DNA fragments. For each of the PCR markers, the oligonucleotide combination that was used for amplification of a particular marker from Capsella genomic DNA is given in Table III.

For marker Cos2, a pronounced size difference between the *C. rubella* and *C. grandiflora* allele sequences was exploited for the mapping experiments. Single-strand conformational polymorphism analysis (Slabaugh et al., 1997) was carried out for loci A7, A9, CA6, CCr8, CE6, CH16, CI18, CL5.1, CL9, CL16, CM7, CO11, and Cos57. The PCR-amplified products were separated on gels prepared with MDE gel solution (FMC Bioproducts, Vallenbak Strand, Denmark) and the resulting gels were stained with silver (Sanguinetti et al., 1994). For loci CL19, CN18, Cos9, and Cos36 the PCR-amplified products were subjected to restriction with *Hind*III, *Hind*III/*Cla*I, *Hind*III/*Bam*HI, and *Rsa*I, respectively, prior to single-strand conformational polymorphism analysis.

An oligonucleotide combination that had been developed for the amplification of Arabidopsis sequences was used to amplify genomic DNA sequences of both Capsella species for locus A20 (5′-gcttccaaggctttgattctg-3′, 5′-ggcttagtctgaacaggttcg-3′). Restriction of the resulting PCR products with *Dra*I revealed a polymorphism between *C. rubella* and *C. grandiflora*.

Mapping Sequences on Arabidopsis Chromosome Sequence Maps

End sequences were determined for each of the Arabidopsis RFLP markers. These were used to position the markers on the sequence maps of the five Arabidopsis chromosomes using SeqViewer of The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org) and the Arabidopsis gene-mapping tool (http://signal.salk.edu/cgi-bin/tdnaexpress). The dataset of Arabidopsis genomic DNA sequences established by the Arabidopsis Genome Initiative was searched using BLASTN (parameters used for the alignments: nucleotide match 1, nucleotide mismatch −3, gap open penalty 5, and gap extension penalty 1) to establish the copy-number for each RFLP marker sequence in the Arabidopsis genome. Positions on the Arabidopsis chromosome maps were established for all homologous sequences sharing ≥80% sequence identity at the nucleotide level for a stretch of at least 100 bp. Sequences sharing lower nucleotide sequence identity values were disregarded because the conditions used for the hybridizations precluded the identification of poorly conserved sequences. The small size of the Capsella mapping population makes it unlikely that recombination events will be detected between closely linked genes, therefore homologous sequences mapping within very close physical proximity of each other (<100 kbp) were treated as a single locus.

To establish the copy-number of Capsella PCR-based markers, the following parameters were used for BLASTN: nucleotide match 1, nucleotide mismatch −2, gap open penalty 5, and gap extension penalty 1. All corresponding Arabidopsis sequences that showed $E < 10^{-9}$ were then mapped onto the sequence maps of the Arabidopsis chromosomes.

To reveal whether a particular marker was mapping to segmental duplications in the Arabidopsis genome the protein coding sequences harbored in both the marker sequence and its homolog(s) were assessed. The resulting gene codes and those of genes in the immediate vicinity of these genes were used to search the datasets describing the segmental duplications of the Arabidopsis genome (http://www.tigr.org/tdb/e2k1/ath1/Arabidopsis_genome_duplication.shtml; http://mips.gsf.de/proj/thal/db/gv/rv/rv_frame.html).

Genetic Analyses

The population used for genetic mapping experiments consisted of self-compatible F_2 plants derived from an interspecific cross of *C. grandiflora* and *C. rubella* (Acarkan et al., 2000). Genotypes of 44 to 50 F_2 plants were evaluated for the different markers. Genetic linkage analysis was carried out with the program Map manager QTX (Manly et al., 2001) and the Kosambi cM function. Using a minimum logarithm of the odds ratio score of 5.0, the 133 codominant loci mapped to eight linkage groups. Genetic distances were given in cM.

ACKNOWLEDGMENTS

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 SF, Gish W, Miller W, Myers EW, Lipman D (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Babula D, Kaczmarek M, Barakat A, Delseny M, Quiros CF, Sadowski J (2003) Chromosomal mapping of *Brassica oleracea* based on ESTs from *Arabidopsis thaliana*: complexity of the comparative map. Mol Genet Genomics **268:** 656–665

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

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12:** 1093–1101

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438

Copenhaver GP, Pikaard CS (1996) RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. Plant J **9:** 259–272

Ermolaeva MD, Wu M, Eisen JA, Salzberg SL (2003) The age of the *Arabidopsis thaliana* genome duplication. Plant Mol Biol **51:** 859–866

Fabri C, Schäffner A (1994) An *Arabidopsis thaliana* RFLP mapping set to localize mutations to chromosomal regions. Plant J **5:** 149–156

Höfte H, Desprez T, Amselem J, Chiapello H, Rouzé P, Caboche M, Moison A, Jourjon M-F, Charpenteau J-L, Berthomieu P, et al (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. Plant J **4:** 1051–1061

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, Bishop J, Mitchell-Olds T (1999) Molecular systematics and evolution of *Arabidopsis* and *Arabis*. Plant Biol **1:** 529–537

Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol **17:** 1483–1498

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

Kowalski SP, Lan T-H, Feldmann KA, Paterson AH (1994) Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization. Genetics **138:** 499–510

Lagercrantz U (1998) Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics **150:** 1217–1228

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

Lan TH, DelMonte TA, Reischmann KP, Hyman J, Kowalski SP, McFerson J, Kresovich S, Paterson AH (2000) An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res **10:** 776–788

Liu Y-G, Mitsukawa N, Lister C, Dean C, Whittier RF (1996) Isolation and mapping of a new set of 129 RFLP markers in *Arabidopsis thaliana* recombinant inbred lines. Plant J **10:** 733–736

Livingstone KD, Lackney VK, Blauth JR, van Wijk R, Jahn MK (1999) Genome mapping in Capsicum and the evolution of genome structure in the Solanaceae. Genetics **152:** 1183–1202

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, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science **290:** 1151–1155

Lysak MA, Pecinka A, Schubert I (2003) Recent progress in chromosome painting of Arabidopsis and related species. Chromosome Res **11:** 195–204

Manly KF, Cudmore RH, Jr., Meer JM (2001) Map Manager QTX, cross-platform software for genetic mapping. Mamm Genome **12:** 930–932

Mathé C, Sagot MF, Schiex T, Rouzé P (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res **30:** 4103–4117

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, et al (1994) Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol **106:** 1241–1255

O'Neill CM, Bancroft I (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J **23:** 233–243

Raes J, Vandepoele K, Simillion C, Saeys Y, Van de Peer Y (2003) Investigating ancient duplication events in the Arabidopsis genome. J Struct Funct Genomics **3:** 117–129

Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R (2001) Comparative sequence analysis reveals extensive microcolinearity in the *Lateral suppressor* regions of the tomato, Arabidopsis and Capsella genomes. Plant Cell **13:** 979–988

Sanguinetti CJ, Dias Neto E, Simpson AJ (1994) Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. Biotechniques **17:** 914–921

Schmidt R (2000) Synteny: recent advances and future prospects. Curr Opin Plant Biol **3:** 97–102

Schmidt R (2002) Plant genome evolution: lessons from comparative genomics at the DNA level. Plant Mol Biol **48:** 21–37

Schmidt R, Acarkan A, Boivin K (2001) Comparative structural genomics in the Brassicaceae family. Plant Physiol Biochem **39:** 253–262

Schmidt R, Acarkan A, Boivin K, Clarenz O, Rossberg M (2003) The sequence of the Arabidopsis genome as a tool for comparative structural genomics in Brassicaceae. *In* T Nagata, S Tabata, eds, Biotechnology in Agriculture and Forestry, Vol 52: Brassica and Legumes. Springer-Verlag, Berlin Heidelberg, pp 19–36

Schmidt R, Acarkan A, Koch M, Rossberg M (1999) A strategy for comparative physical mapping in cruciferous plants. *In* LWD van Raamsdonk, JCM den Nijs, eds, Plant Evolution in Man-Made Habitats. Proceedings of the VIIth Symposium of the International Organization of Plant Biosystematists. Hugo de Vries Laboratory, University of Amsterdam, Amsterdam, pp 183–196

Slabaugh MB, Huestis GM, Leonard J, Holloway JL, Rosato C, Hongtrakul V, Martini N, Toepfer R, Voetz M, Schell J, et al (1997) Sequence-based genetic markers for genes and gene families: single-strand conformational polymorphisms for the fatty acid synthesis genes of Cuphea. Theor Appl Genet **80:** 57–64

Tatusova TA, Madden TL (1999) BLAST 2 sequences - a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett **174:** 247–250

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science **290:** 2114–2117

Yang Y-W, Lai K-N, Tai P-Y, Li W-H (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. J Mol Evol **48:** 597–604