The Plant Cell, Vol. 12, 1093–1101, July 2000, www.plantcell.org © 2000 American Society of Plant Physiologists

Extensive Duplication and Reshuffling in the Arabidopsis Genome

Guillaume Blanc, Abdelali Barakat, Romain Guyot, Richard Cooke,[1] and Michel Delseny

Laboratoire Génome et Développement des Plantes, Unité Mixte de Recherche 5096, Centre National de la Recherche Scientifique, University of Perpignan, 66860 Perpignan Cédex, France

Systematic analysis of the Arabidopsis genome provides a basis for detailed studies of genome structure and evolution. Members of multigene families were mapped, and random sequence alignment was used to identify regions of extended similarity in the Arabidopsis genome. Detailed analysis showed that the number, order, and orientation of genes were conserved over large regions of the genome, revealing extensive duplication covering the majority of the known genomic sequence. Fine mapping analysis showed much rearrangement, resulting in a patchwork of duplicated regions that indicated deletion, insertion, tandem duplication, inversion, and reciprocal translocation. The implications of these observations for evolution of the Arabidopsis genome as well as their usefulness for analysis and annotation of the genomic sequence and in comparative genomics are discussed.

INTRODUCTION

Since the decision to adopt Arabidopsis as a model for plant genome studies ~10 years ago, a concerted international effort has led to the accumulation of a vast amount of information. Generating and analyzing expressed sequence tags (ESTs) led the way in this effort (Höfte et al., 1993; Newman et al., 1994; Cooke et al., 1996), followed by genome sequencing as the next step in the systematic study of the Arabidopsis genome (Bevan et al., 1998; Kaneko et al., 1999; Lin et al., 1999; Mayer et al., 1999; Terryn et al., 1999). The complete genome sequence should be available in the year 2000, and the rapidly accumulating data have already begun to provide information on genome structure and evolution.

A surprising observation based largely on EST studies was that despite consisting of only ~140 Mb, the Arabidopsis genome contains many small gene families (Höfte et al., 1993; Newman et al., 1994; Cooke et al., 1996; Bevan et al., 1998). This observation led to the question of how these multiple copies, derived from a single ancestor, arose during evolution. Previous work had shown that various copies are dispersed within the genome (van Lijsebettens et al., 1994; Rounsley et al., 1995; Membre et al., 1997; Romero et al., 1998) or are duplicated in tandem (Krebbers et al., 1988; Axelos et al., 1989; Kurkela and Borg-Franck, 1992; Terryn et al., 1999). One possible mechanism for the former type of distribution was that individual genes or groups of genes had been duplicated, giving rise to different members of the gene families.

[1] To whom correspondence should be addressed. E-mail cooke@univ-perp.fr; fax 33-468668499.

Large-scale duplication in the Arabidopsis genome was proposed on the basis of comparative mapping of molecular markers in Arabidopsis and *Brassica oleracea*. Kowalski et al. (1994) found that 14% of these markers mapped to duplicate locations in the Arabidopsis genome and identified a region of Arabidopsis chromosome 1 that appeared to be homeologous with a region of chromosome 5. Study of a larger number of plant species revealed short regions of synteny and a further possible duplication between chromosomes 1 and 3 of Arabidopsis (Paterson et al., 1996). More recently, detailed analysis of a 400-kb sequence revealed that a 40-kb region near the *APETALA2* locus on chromosome 4 is duplicated on chromosome 2 (Terryn et al., 1999). The availability of whole-chromosome sequences has also suggested that large regions of the genome have been duplicated (Lin et al., 1999; Mayer et al., 1999), although detailed analysis is necessary to determine the extent of duplications and to attempt to elucidate the mechanisms involved.

One way to obtain information on the position and extent of duplications is to locate members of small gene families and determine other conserved sequences in the vicinity of the different copies. Cytoplasmic ribosomal proteins have been shown to be encoded by small gene families (van Lijsebettens et al., 1994; Williams and Sussex, 1995; Cooke et al., 1997). Because sequences of proteins from different species are generally highly conserved, amino acid sequences from organisms in which ribosomal proteins have been systematically studied can be used to identify members of corresponding families in Arabidopsis. In addition, nucleotide sequences of members of individual families are

usually highly conserved in Arabidopsis (Cooke et al., 1997). Therefore, we chose to locate individual members of these families on the Arabidopsis genome (A. Barakat, R. Guyot, G. Blanc, R. Cooke, and M. Delseny, manuscript in preparation) and use them as a framework for detailed studies on genome structure and evolution. Using a combination of dot plot (Sonnhammer and Durbin, 1995) and BLAST (Altschul et al., 1990) analyses of sequences surrounding pairs of ribosomal protein genes, we have identified duplications covering more than half of the Arabidopsis genome.

RESULTS

Identification of a Large Duplicated Region

After locating individual members of cytoplasmic ribosomal protein gene families on the Arabidopsis genome by physical mapping or sequence analysis (A. Barakat, R. Guyot, G. Blanc, R. Cooke, and M. Delseny, manuscript in preparation), we used these genes as anchor points to identify flanking duplicated sequences. For example, members of the small subunit protein S25 (RS25) family are found on bacterial artificial chromosomes (BACs) F2G1 and T22F8 on chromosomes 2 and 4, respectively. After using the Dotter program (Sonnhammer and Durbin, 1995) to align the sequences, it became evident that large regions of the two BACs are very similar, suggesting that flanking genes are also conserved. Therefore, we constructed larger contigs from sequences of neighboring BACs on the two chromosomes to determine the extent of the conserved regions.

Figure 1A shows a dot plot of sequences covering nine BACs (657,655 bp) on chromosome 2 (from BACs F5H14 to F14M13) and seven BACs (550,140 bp) on chromosome 4 (from BACs F20M13 to T5J17) for which discontinuous nucleotide sequence conservation over large regions can be seen as a staggered diagonal on the dot plot. These data suggest that the two chromosome regions correspond to a single ancestral region that has been duplicated and has undergone limited rearrangement, including accumulation of point mutations and large-scale insertion or deletion, singly or in combination, of fragments. Detailed analysis of a smaller region (boxed in Figure 1A) and comparison with the GenBank annotations of the sequences (Figure 1B) revealed similarities covering regions of only a few kilobases, which apparently correspond to annotated genes. Of 12 annotated genes on chromosome 4 and 11 on chromosome 2, nine showed marked nucleotide sequence similarity. For the remaining genes, no similarity could be determined, suggesting divergent evolution of the sequences or further small-scale rearrangements since the original duplication.

Detailed Structure and Expression

Close examination of similar regions on the dot plot suggested that sequence conservation within individual genes

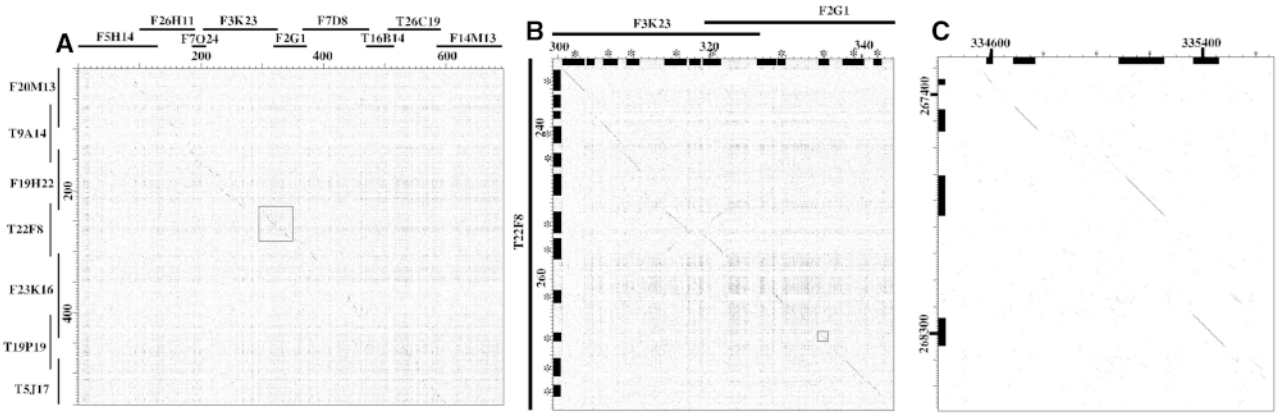

Figure 1. Dot Plot of BAC Contigs Containing *RS25* Sequences.

BAC contigs were constructed using the Sequencher program and aligned with the Dotter program.

(A) Dot plot of full-length contigs representing 657,655 bp on chromosome 2 (horizontal axis) and 550,140 bp on chromosome 4 (vertical axis). The positions of the BACs are indicated on the axes, and numbers show lengths in kilobases.

(B) Enlargement of the region highlighted in **(A)**. Numbers show positions on the complete contigs in kilobases. Black blocks on the axes show positions of predicted genes on the BACs. Genes for which sequences are conserved between the two chromosomes are shown by an asterisk. The square shows the region presented in **(C)**.

(C) Enlargement of *RS25* genes on both chromosomes. Black boxes show locations of exons in the two genes. The coding regions of the *RS25* genes cover bases 40,156 to 41,156 on BAC T22F8 and 38,478 to 39,359 on F2G1.

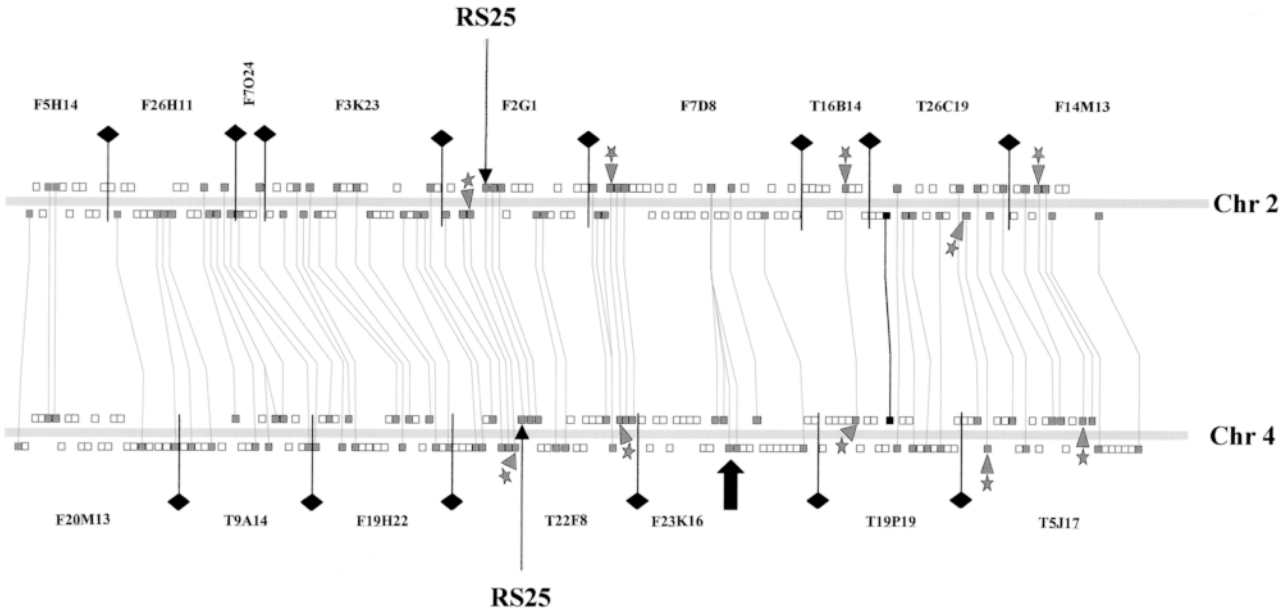

Figure 2. Schematic Representation of Regions Duplicated on Chromosomes 2 and 4.

BACs are given for each contig, with arrowheads indicating separations between clones. Predicted genes are shown by small blocks (not to scale) either above (forward strand) or below (reverse strand) the chromosome. Highly similar sequences identified by dot plot analysis and confirmed by BLAST alignment (see Methods) are shaded. Lines link the relative positions of these sequences on chromosomes 2 and 4 (Chr2 and Chr4). Genes encoding the RS25 proteins, which were used to detect the duplication, are labeled. Asterisks with arrowheads show the positions of conserved tRNA genes. The boldface line indicates a conserved gene that shows opposite polarity on the two chromosomes, and the large arrow shows the location of an inverted gene duplication.

is limited to short regions, which we presumed to correspond to exons. Figure 1C shows a typical dot plot for a single pair of genes, the *RS25* genes. On this plot, not only is sequence conservation clearly limited strictly to exon sequences, with no detectable similarities in the intron regions, but also deletions or insertions (or both) have led to considerable variation in intron size. Because dot plots are inefficient for nucleotide sequence comparisons, we used BLAST alignment to obtain more precise information on the structure and expression of genes in the regions duplicated between chromosomes 2 and 4.

By using the BLASTN program (Altschul et al., 1990) to align predicted coding sequences, which were obtained from GenBank annotations of BAC clones shown in Figure 1A, we confirmed that sequence conservation between corresponding genes on the two chromosomes is almost exclusively limited to exons. Only clearly important alignments were retained, as described in Methods, and the results of this analysis are shown schematically in Figure 2. The positions on the two contigs of all predicted genes are indicated, with lines linking pairs showing substantial sequence similarity. Of 151 pairs of genes, 59 (39%) show highly similar nucleotide sequences.

The order and distribution of the genes according to the Watson or Crick strand are conserved, as would be expected after duplication of a block of genes, with two notable exceptions. First, one conserved gene on BACs T26C19 and T19P19 (on chromosomes 2 and 4, respectively) shows different polarity. Second, the presence of four copies of a gene on chromosome 4, with two copies on each strand and only two copies, both on the same strand on chromosome 2, indicates that a single original gene was probably duplicated in tandem before duplication of the region and that this was followed by a duplication with an inversion on chromosome 4. Five conserved tRNA genes are also found within this region.

The presence of pairs of genes showing no nucleotide similarity in regions in which sequences of the majority of the duplicated genes have been conserved could arise either simply by sequence divergence or by more recent rearrangements. If rearrangements have occurred by insertion of genes from other chromosome locations, we would expect to detect nucleotide similarity between these nonconserved genes and sequences elsewhere in the genome. Therefore, we performed BLASTN alignments of all the corresponding predicted coding sequences with all known genomic sequences and found that in addition to the 59 genes from chromosome 2 duplicated on chromosome 4,

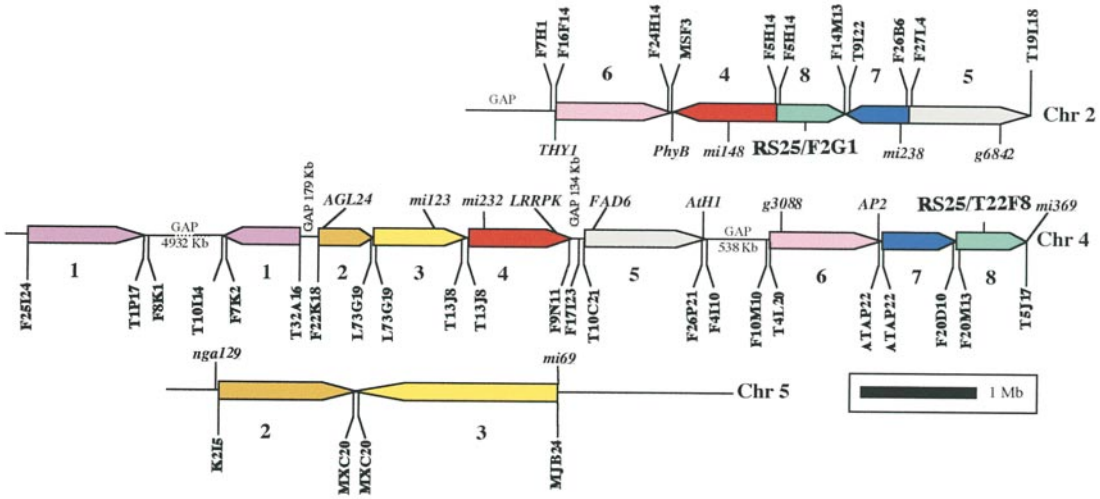

Figure 3. Schematic Representation of Regions Duplicated on Chromosome 4 and Chromosomes 2 and 5.

Highly similar sequences were identified by dot plot analysis and confirmed by BLAST alignment, as described in Methods. Colored blocks indicate the position and orientation of regions on the different chromosomes (Chr).The blocks presented here are identified on the chromosomes by diagonal stripes in Figure 4. Blocks are numbered sequentially on chromosome 4 and in duplicated regions to facilitate identification. The positions of the *RS25* genes shown in Figure 1 are indicated. The names of BACs (vertical orientation) are given only at the ends of individual regions. Selected genetic marker positions are indicated (horizontal orientation).

substantially similar sequences for an additional 47 could be found elsewhere in the genome. The remaining 45 predicted genes shared no sequence similarity with the genomic sequence that is currently available. Thus, sequences similar to at least 70% of all predicted genes on the region of chromosome 2 shown in Figure 1A are found elsewhere in the genome.

The identification of regions containing duplicate copies of many genes whose predicted protein products have highly similar or identical sequences raises the question of whether both copies are effectively expressed. Although expression data are not available for all genes, ESTs have been obtained for approximately half of the genes in Arabidopsis. BLASTN alignment of coding sequences with Arabidopsis ESTs in GenBank showed that for genes duplicated between chromosomes 2 and 4, 30% of those on chromosome 2 are tagged, compared with 45% on chromosome 4. For genes that are located on chromosomes 2 and 4 and for which copies are also found elsewhere in the genome, the percentages are roughly the same (26 and 51%, respectively), whereas of the genes on chromosome 2 for which no copy could be found, 43% are tagged compared with 37% on chromosome 4.

Patchwork Distribution of Duplications

Although we could detect no further sequence similarity by extending contigs of BACs shown in Figure 1A, alignment of sequences of BACs F20D10 and T28I19—which are adjacent to F20M13 on chromosome 4—showed marked similarity to sequences from BAC T9I22, which is immediately adjacent to BAC F14M13 on chromosome 2. The orientation of BAC T9I22, however, was inverted in relation to the other two sequences. Further extension using dot plots of sequences from neighboring BACs demonstrated clearly that a region of 664 kb on chromosome 4 is duplicated as a 585-kb region on chromosome 2, the separation between the two regions probably lying within BAC F14M13. Figure 3 presents the results of this analysis using dot plots followed by BLASTN alignments that allowed us to identify a 6.7-Mb region on chromosome 4. This region is duplicated on chromosomes 2 (4 Mb) and 5 (1.2 Mb) and within chromosome 4 itself (0.65 Mb) with three gaps. A portion of the 1.2-Mb region on chromosome 4 is inverted on chromosome 5. The duplication of 4 Mb is an extension of that described near the APETALA2 locus (Terryn et al., 1999). Several blocks on the chromosome 4 sequence are duplicated in the 4-Mb region of chromosome 2 but in a different order, suggesting multiple recombination events. Finally, two blocks, of 800 and 650 kb, are found as an inverted repeat separated by a gap of nearly 5 Mb on chromosome 4. The results of the BLAST analysis are summarized in Table 1. The percentage of conserved genes between the different regions varies, ranging from 18 to 45%. In fact, the actual percentage of genes showing conserved sequence is probably greater than this because we took into account only those BLASTN alignments that showed unambiguous

sequence similarity. In all cases, the order of duplicated genes is conserved.

The Majority of the Arabidopsis Genome Is Found in Duplications

The unexpected extent of duplications shown in Figure 3 led us to suspect that other large duplications could be found in the Arabidopsis genome. Therefore, we decided to adopt a random approach using arbitrarily selected BACs to identify regions in which sequences of small groups of genes were conserved at two loci and to determine whether these groups could then be extended. BLASTN analysis using the sequences of all BACs from the published genomic sequence allowed us to show that extensive duplication has effectively occurred. The distribution of duplications presented in Figure 4, based on analysis of ~80% of the complete sequence, shows that the majority of Arabidopsis genes are found in duplicated regions. In fact, the extent of duplication is certainly greater because information on coding sequences from chromosome 5 is not directly available in the databases and part of the genome remains to be sequenced. In addition, sequencing gaps remain in the pericentromeric regions on all chromosomes and in the nucleolar organizing regions on chromosomes 2 and 4 (Copenhaver and Pikaard, 1996). These regions have not been sequenced, and analysis would be impossible, given the high concentration of repetitive sequences. Figure 4 emphasizes the patchwork nature of duplications, which are similar to those we had already identified and analyzed in detail.

DISCUSSION

The results presented here show that the Arabidopsis genome contains megabase-sized blocks on pairs of chromosomes in which as many as 45% of the gene pairs show highly similar sequences. They also demonstrate that a large part of the genome results from duplication. This observation is surprising considering the small size of the genome but confirms and considerably extends previous observations based on mapping data (McGrath et al., 1993; Kowalski et al., 1994; Paterson et al., 1996) or sequence analysis (Lin et al., 1999; Mayer et al., 1999; Terryn et al., 1999).

The exact extent of this duplication will become clear only when the complete genome sequence has been established. For regions in which gaps remain to be sequenced, limited rearrangements possibly could be detected, although ongoing sequencing seems to confirm and extend our results. However, the detailed analysis presented here

Table 1. Conservation of Coding Sequences between the 6.7-Mb Region of Chromosome 4 and Regions on Chromosomes 2, 4, and 5

Regions[a]	Length (kb)	No. of Predicted Genes[b]	No. of Duplicated Genes	% of Duplicated Genes
4				
Chromosome 2	835	217	57	26
Chromosome 4	867	154	57	37
5				
Chromosome 2	1096	256	46	18
Chromosome 4	1008	257	44	18
6				
Chromosome 2	941	177	82	46
Chromosome 4	905	247	82	33
7				
Chromosome 2	482	140	38	27
Chromosome 4	664	125	38	30
8				
Chromosome 2	612	151	59	39
Chromosome 4	550	165	62	38
Duplicated on chromosome 5	NA[c]	310	138	44.5
1a				
Chromosome 4	803	151	61	40
1b				
Chromosome 4	647	170	80	44.7

[a] Numbers at left refer to regions shown in Figure 3.

[b] Numbers of predicted genes are based on annotations of BACs in the GenBank database after removal of duplicates from overlapping sequences. Precise details are not given for chromosome 5 because coding sequences are not cited in GenBank annotations.

[c] NA, not applicable.

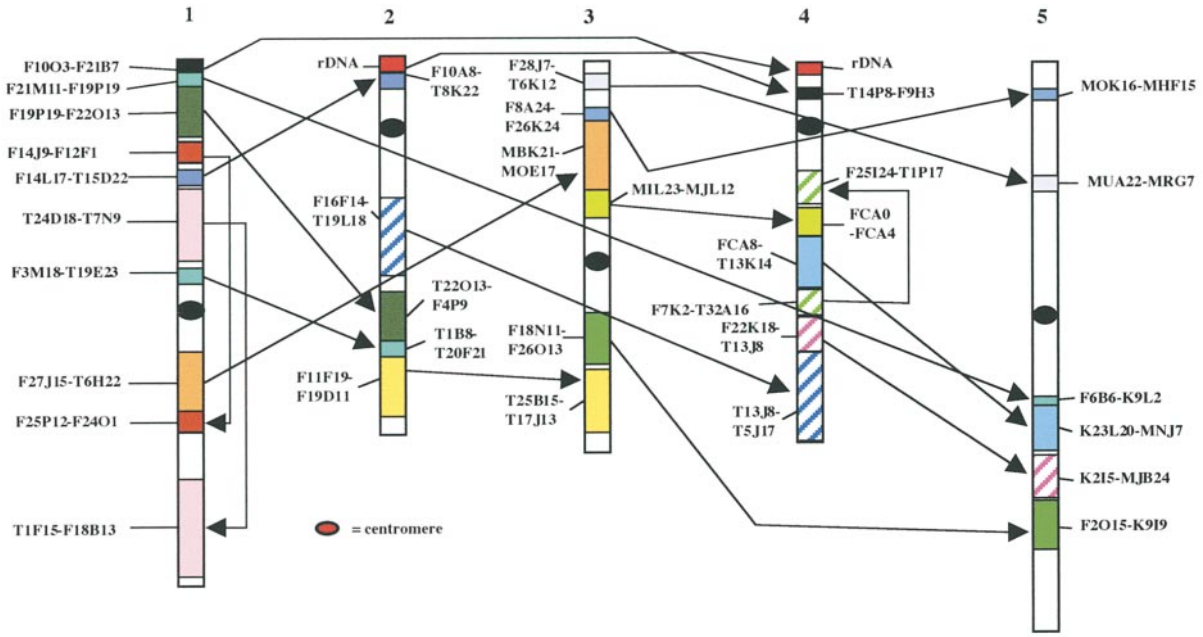

Figure 4. Schematic Representation of All Identified Duplications throughout the Arabidopsis Genome.

Duplicated regions were identified by BLASTN alignment of whole BAC sequences with all Arabidopsis genomic sequences, as described in Methods. Positions of centromeres and rDNA loci are indicated. Colored blocks identify similar regions on different chromosomes or within chromosomes. BAC clones at the ends of duplicated regions are shown. The regions shown on chromosomes 2, 4, and 5 by diagonal striping correspond to those presented in Figure 3.

shows that the duplication of large regions has been followed by extensive rearrangement and probably divergent evolution of the genes for which no sequence similarity can be detected elsewhere in the genome. In fact, our results, which indicate that >60% of the genome is found as duplications, provide only a minimum estimate. During these studies, we detected several short duplicated regions, containing only three or four genes, that are not shown in Figure 4. In addition, comparison of sequences of duplicated genes brought to light several obvious errors in annotation of the corresponding BAC sequences in international databases (G. Blanc, R. Guyot, R. Cooke, and M. Delseny, manuscript in preparation)—including erroneously annotated tRNA genes, additional or missing exons, and genes that have not been annotated in one of the copies. These errors certainly lead to an underestimation of the extent of gene sequence conservation when BLASTN alignment of predicted coding sequences is used.

Ab initio analysis of genomic sequence, based largely on computer-assisted prediction of exons, introns, and gene models, is still relatively inefficient in predicting whole-gene models (Rouze et al., 1999). In the absence of substantial alignment with protein or cDNA sequences, the structures of only ~20% of genes are correctly predicted. Comparison of conserved sequences between duplicated pairs of genes in which only the exon sequence has been conserved will provide a useful tool in the correct annotation of the complete genome sequence.

In light of these observations, the fact that the sequence of the genome is not yet complete, and given that the nucleolar organizing region and the pericentromeric and telomeric regions represent ~7 Mb, almost all of the "single copy" sequences of Arabidopsis appear to be found in regions resulting from ancient rearrangements. These results lead to the intriguing possibility that Arabidopsis could be a degenerate tetraploid. Ohno (1973) has previously suggested that whole-genome duplication is an important evolutionary mechanism, and evidence suggests that maize, yeast, and Xenopus could be degenerate polyploids (see Skrabanek and Wolfe, 1998). The pattern of duplication we found in Arabidopsis is similar to that observed in maize and yeast, that is, duplications are found as multiple regions, indicating considerable rearrangements, and not all genes in paired regions are conserved. The fact that two-thirds of the duplications presented in Figure 4 are found in the same orientation with respect to the centromeres is also in favor of a model of tetraploidy followed by translocation, as has been shown for yeast (Wolfe and Shields, 1997). Our results show that copies of some of the genes that are not paired within duplications can be found elsewhere in the genome and suggest

that a combination of divergent evolution, interchromosome recombination, and reciprocal transposition is probably responsible for the genome organization in Arabidopsis today.

Several observations suggest that these duplications are ancient events. First, the sequence of some genes has apparently diverged to the extent that no sequence similarity can be detected, although the positions of these genes in the duplicated regions strongly suggest that they are derived from a common ancestral sequence. Moreover, we have shown that some genes in duplicated regions have apparently been repositioned by transposition events since the original duplication occurred, but this is not the case for all of the genes, and the fact that many divergent regions are of similar lengths argues more favorably for divergent evolution of a common ancestral sequence than for replacement by transposition. Second, considerable sequence divergence has occurred in noncoding regions, to the extent that intron sequences, for example, vary greatly both in sequence and in length and in some cases are absent from one of the copies. This divergence is in striking contrast, for example, to the high degree of conservation of both exon and intron sequences for human and mouse (Ansari-Lari et al., 1998). Third, close inspection of sequences at the ends of duplicated regions shows no obvious sequence motifs to suggest the mechanisms involved.

In some cases, we observe considerable size differences between two duplicated regions. For example, the only duplicated regions between chromosomes 1 and 4 have lengths of 216 and 465 kb, respectively, and a 787-kb region of chromosome 4 is duplicated as a 1831-kb region on chromosome 5. Such extensions apparently have several origins. If we consider the former duplication, the gene number has increased (73 predicted genes in the 216-kb region of chromosome 1 and 108 in the 465-kb region of chromosome 4); however, intergenic regions have also probably increased because, assuming that most of the genes have been predicted, then one can calculate that the gene density is 1 per every 2.9 kb in the region on chromosome 1 and 1 per every 4.3 kb on chromosome 4. The increase in gene number also results from tandem duplication: only five genes are duplicated in tandem in the 216-kb region of chromosome 1 but 25 in the corresponding 465-kb region on chromosome 4.

An unexpected observation regarding genes in duplicated regions is the bias in expression between duplicated genes and apparently single-copy genes and also between the copies on different chromosomes. It is true that our analysis is based on EST sequences, which contain tags to no more than half of the estimated 20,000 to 25,000 genes. However, a comparison of gene pairs clearly shows that many more genes have been tagged on chromosome 4 than the corresponding genes on chromosome 2. This bias in expression could indicate that certain chromosomes or regions of chromosomes contain a greater density of pseudogenes, although little evidence is available to suggest the presence in the Arabidopsis genome of large numbers of pseudogenes. The highly conserved exon–intron structure of untagged genes is also an indication that these genes are in fact expressed. Another possibility is that the presence of at least two copies of a gene has allowed specialization of one of the two genes and that one is expressed only under conditions that have not yet been studied with ESTs. If this is the case, however, it is not clear why there should be a bias of expression in favor of genes on one chromosome over another.

This study sheds new light on Arabidopsis genome fluidity. It illustrates that during the evolution of this genome numerous rearrangements have occurred, including duplication, translocation, inversion, and deletion. All of these mechanisms were also probably at work in many species until heterologous chromosome pairing and recombination were prevented by specific mechanisms (Moore, 1998). An important consequence of duplication is that it should be considered in studies that use comparative mapping or sequencing based on the Arabidopsis genome. Comparative mapping with cultivated Brassica will show whether the Arabidopsis genome duplications occurred before or after the differentiation of the various species. Regions of collinearity between *Brassica* spp and Arabidopsis have already been identified (Osborn et al., 1997; Cavell et al., 1998). These regions may indicate that the duplications preceded speciation. More detailed studies are necessary to determine the exact extent of synteny, and careful comparison will reveal genes that have differentiated or disappeared during the evolution and domestication of crops. Fine-mapping has already shown the apparent deletion of self-recognition genes in Arabidopsis (Conner et al., 1998). In light of these observations, the expectation of finding large blocks of conserved regions of synteny between the Arabidopsis model genome (Paterson et al., 1996; Gale and Devos, 1998) and more distant major crop genomes is certainly limited.

METHODS

Bacterial artificial chromosome (BAC) contigs were constructed using Sequencher (Gene Codes Corp., Ann Arbor, MI). Dot plot analysis was conducted with the DOTTER program (Sonnhammer and Durbin, 1995). Dynamic zooming was used to focus on regions of interest, and gray-scale variation was determined to obtain clear plots. BLAST (Altschul et al., 1990) alignments were performed using either the BLAST network client or locally installed programs. For analysis of duplications in the whole genome, all predicted coding sequences were extracted from the databases by using the SRS program (Etzold et al., 1996) and aligned against all Arabidopsis genomic sequences by using the BLASTN program (Altschul et al., 1990). Only sequences giving a BLAST score >150 were considered for further analysis. Two regions were considered to be duplicated when (1) at least four different coding sequences, encoding four different proteins and located on the same BAC clone, matched contiguous coding sequences at another locus, and (2) the order of conserved genes and their orientation on the two DNA strands were identical. In cases in which BLASTN results were unclear (lower scores or shorter blocks

of similar sequences), alignments of derived amino acid sequences were inspected to confirm the results of the nucleotide alignments. Only sequences showing unambiguous similarity were considered to represent duplicated sequences.

Names and GenBank accession numbers of the BACs given in Figures 1 and 2 are as follows: F5H14, AC006234; F26H11, AC006264; F7O24, AC007142; F3K23, AC006841; F2G1, AC007119; F7D8, AC007019; T16B14, AC007232; T26C19, AC007168; and F14M13, AC006592 on chromosome 2; and F20M13, AL035540; T9A14, AL035656; F19H22, AL035679; T22F8, AL050351; F23K16, AL078620; T19P19, AL022605; and T5J17, AL035708 on chromosome 4.

Names and GenBank accession numbers of the BACs given in Figure 3 are as follows: F7H1, AC007134; F16F14, AC007047; F24H14, AC006135; MSF3, AC005724; F23N11, AC007048; F5H14, AC006234; T26C19, AC007168; F14M13, AC006592; T9I22, AC006340; F26B6, AC003040; F27L4, AC004482; and T19L18, AC004747 on chromosome 2; F25I24, AL049525; T1P17, AL049730; T20K18, AL049640; T10I14, AL021712; F7K2, AL033545; T32A16, AL078468; F22K18, AL035356; L73G19, AL050400; F14M19, AL049480; T27E11, AL049770; T13J8, AL035524; F9N11, AL109796; F17I23, AF160182; T10C21, AL109787; F26P21, AL031804; F4I10, AL035525; F10M10, AL035521; T4L20, AL023094; ATAP22, Z99708; F20D10, AL035538; F20M13, AL035540; and T5J17, AL035708 on chromosome 4; and K2I5, AB025613; MXC20, AB009055; and MJB24, AB019233 on chromosome 5.

Names and GenBank accession numbers of the BACs given in Figure 4 are as follows: F10O3, AC006550; F21B7, AC002560; F19P19, AC000104; F21M11, AC003027; F22O13, AC003981; F14J9, AC003970; F12F1, AC002131; F14L17, AC012188; T15D22, AC012189; T24D18, AC010924; T7N9, AC000348; F3M18, AC010155; T19E23, AC007654; F27J15, AC016041; T6H22, AC009894; F25P12, AC009323; F24O1, AC003113; T1F15, AC004393; and F18B13, AC009322 on chromosome 1; F10A8, AC006200; T8K22, AC004136; F16F14, AC007047; T19L18, AC004747; T22O13, AC007290; F4P9, AC002332; T1B8, U78721; T20F21, AC006068; F11F19, AC007017; and F19D11, AC005310 on chromosome 2; F28J7, AC010797; T6K12, AC016829; F8A24, AC015985; F26K24, AC016795; MBK21, AB024033; MOE17, AB025629; MIL23, AB019232; MJL12, AB026647; F18N11, AL132953; F26O13, AL133452; T25B15, AL132972; and T17J13, AL138651 on chromosome 3; T14P8, AF069298; F9H3, AF071527; F25I24, AL049525; T1P17, AL049730; FCA0, Z97335; FCA4, Z97339; FCA8, Z97343; T13K14, AL080282; F7K2, AL033545; T32A16, AL078468; F22K18, AL035356; T27E11, AL049770; T13J8, AL035524; and T5J17, AL035708 on chromosome 4; and MOK16, AB005240; MUA22, AB007650; F6B6, AP000368; K9L2, AB011475; K23L20, AB016874; MNJ7, AB025628; K2I5, AB025613; MJB24, AB019233; MRG7, AB012246; MHF15, AB006700; F2O15, AB025604; and K9I9, AB013390 on chromosome 5.

ACKNOWLEDGMENTS

This work strongly benefited from the public effort coordinated by the Arabidopsis Genome Initiative to make available Arabidopsis genomic sequences as soon as they were sequenced. We also acknowledge support of several European Union grants, which helped to make our research possible.

Received January 12, 2000; accepted May 17, 2000.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215,** 403–410.

Ansari-Lari, M.A., Oeltjen, J.C., Schwartz, S., Zhang, Z., Muzny, D.M., Lu, J., Gorrell, J.H., Chinault, A.C., Belmont, J.W., Miller, W., and Gibbs, R.A. (1998). Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. Genome Res. **8,** 29–40.

Axelos, M., Bardet, C., Liboz, T., Le Van Thai, A., Curie, C., and Lescure, B. (1989). The gene family encoding the *Arabidopsis thaliana* translation elongation factor EF-1α: Molecular cloning, characterization and expression. Mol. Gen. Genet. **219,** 106–112.

Bevan, M., et al. (1998). Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature **391,** 485–488.

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

Conner, J.A., Conner, P., Nasrallah, M.E., and Nasrallah, J.B. (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.

Cooke, R., et al. (1996). Further progress towards a catalogue of all *Arabidopsis* genes: Analysis of a set of 5000 non-redundant ESTs. Plant J. **9,** 101–124.

Cooke, R., Raynal, M., Laudie, M., and Delseny, M. (1997). Identification of members of gene families in *Arabidopsis thaliana* by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins. Plant J. **11,** 1127–1140.

Copenhaver, G.P., and Pikaard, C.S. (1996). Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. Plant J. **9,** 273–282.

Etzold, T., Ulyanov, U., and Argos, P. (1996). SRS: Information retrieval system for molecular biology data banks. Methods Enzymol. **266,** 114–128.

Gale, M.D., and Devos, K.M. (1998). Plant comparative genetics after 10 years. Science **282,** 656–659.

Höfte, H., et al. (1993). An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNA from *Arabidopsis thaliana*. Plant J. **4,** 1051–1061.

Kaneko, T., Katoh, T., Sato, S., Nakamura, Y., Asamizu, E., Kotani, H., Miyajima, N., and Tabata, S. (1999). Structural analysis of *Arabidopsis thaliana* chromosome 5. IX. Sequence features of the regions of 1,011,550 bp covered by seventeen P1 and TAC clones. DNA Res. **6,** 183–195.

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

Krebbers, E., Seurinck, J., Herdies, L., Cashmore, A.R., and Timko, M.P. (1988). Determination of the processing sites of an *Arabidopsis* 2S albumin and characterization of the complete gene family. Plant Physiol. **87,** 859–866.

Kurkela, S., and Borg-Franck, M. (1992). Structure and expression of *kin2*, one of two cold- and ABA-induced genes of *Arabidopsis thaliana*. Plant Mol. Biol. **19,** 689–692.

Lin, X., et al. (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402,** 761–768.

Mayer, K., et al. (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402,** 769–777.

McGrath, J.M., Jansco, M.M., and Pichersky, E. (1993). Duplicate sequences with similarity to expressed genes in the genome of *Arabidopsis thaliana*. Theor. Appl. Genet. **86,** 880–888.

Membre, N., Berna, A., Neutelings, G., David, A., David, H., Staiger, D., Saez-Vasquez, J., Raynal, M., Delseny, M., and Bernier, F. (1997). cDNA sequence, genomic organization and differential expression of three *Arabidopsis* genes for germin/oxalate oxidase-like proteins. Plant Mol. Biol. **35,** 459–469.

Moore, G. (1998). To pair or not to pair: Chromosome pairing and evolution. Curr. Opin. Plant Biol. **1,** 116–122.

Newman, T., et al. (1994). Genes galore: A summary of the methods for accessing the results of large-scale partial sequencing of anonymous *Arabidopsis thaliana* cDNA clones. Plant Physiol. **106,** 1241–1255.

Ohno, S. (1973). Ancient linkage groups and frozen accidents. Nature **244,** 259–262.

Osborn, T.C., Kole, C., Parkin, I.A., Sharpe, A.G., Kuiper, M., Lydiate, D.J., and Trick, M. (1997). Comparison of flowering time genes in *Brassica rapa*, *B. napus* and *Arabidopsis thaliana*. Genetics **146,** 1123–1129.

Paterson, A.A., et al. (1996). Towards a unified genetic map of higher plants transcending the monocot dicot divergence. Nat. Genet. **14,** 380–382.

Romero, I., Fuertes, A., Benito, M.J., Malpica, J.M., Leyva, A., and Paz-Ares, J. (1998). More than 80R2R3-MYB regulatory genes in the genome of *Arabidopsis thaliana*. Plant J. **14,** 273–284.

Rounsley, S.D., Ditta, G.S., and Yanofsky, M.F. (1995). Diverse roles for MADS box genes in Arabidopsis development. Plant Cell **7,** 1259–1269.

Rouze, P., Pavy, N., and Rombauts, S. (1999). Genome annotation: Which tools do we have for it? Curr. Opin. Plant Biol. **2,** 90–95.

Skrabanek, L., and Wolfe, K.H. (1998). Eukaryotic genome duplication—Where's the evidence? Curr. Opin. Genet. Dev. **8,** 694–700.

Sonnhammer, E.L.L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167,** 1–10.

Terryn N., et al. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the *APETALA2* locus on chromosome 4. FEBS Lett. **445,** 237–245.

van Lijsebettens, M., Vanderhaeghen, R., De Block, M., Bauw, G., Villarroel, R., and Van Montagu, M. (1994). An S18 ribosomal protein gene copy at the *Arabidopsis PFL* locus affects plant development by its specific expression in meristems. EMBO J. **13,** 3378–3388.

Williams, M.E., and Sussex, I.M. (1995). Developmental regulation of ribosomal protein L16 genes in *Arabidopsis thaliana*. Plant J. **8,** 65–76.

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