Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*

David Grant*†, Perry Cregan‡, and Randy C. Shoemaker*

*U.S. Department of Agriculture–Agricultural Research, Service Corn Insect and Crop Genetics Research Unit, Department of Agronomy, Iowa State University, Ames, IA 50011; and ‡U.S. Department of Agriculture–Agricultural Research Service, Soybean and Alfalfa Research Unit, Beltsville Agricultural Research Center West, Beltsville, MD 20705

Edited by Ronald R. Sederoff, North Carolina State University, Raleigh, NC, and approved February 1, 2000 (received for review October 6, 1999)

Synteny between soybean and *Arabidopsis* was studied by using conceptual translations of DNA sequences from loci that map to soybean linkage groups A2, J, and L. Synteny was found between these linkage groups and all four of the *Arabidopsis* chromosomes, where GenBank contained enough sequence for synteny to be identified confidently. Soybean linkage group A2 (soyA2) and *Arabidopsis* chromosome I showed significant synteny over almost their entire lengths, with only 2–3 chromosomal rearrangements required to bring the maps into substantial agreement. Smaller blocks of synteny were identified between soyA2 and *Arabidopsis* chromosomes IV and V (near the *RPP*5 and *RPP*8 genes) and between soyA2 and *Arabidopsis* chromosomes I and V (near the *Phy*A and *Phy*C genes). These subchromosomal syntenic regions were themselves homeologous, suggesting that *Arabidopsis* has undergone a number of segmental duplications or possibly a complete genome duplication during its evolution. Homologies between the homeologous soybean linkage groups J and L and *Arabidopsis* chromosomes II and IV also revealed evidence of segmental duplication in *Arabidopsis*. Further support for this hypothesis was provided by the observation of very close linkage in *Arabidopsis* of homologs of soybean *Vsp*27 and Bng181 (three locations) and purple acid phosphatase-like sequences and homologs of soybean A256 (five locations). Simulations show that the synteny and duplications we report are unlikely to have arisen by chance during our analysis of the homology reports.

The 145-Mbp genome of *Arabidopsis thaliana* is one of the smallest known among higher plants (1). Its low, interspersed, repetitive DNA content (2) makes it an ideal model for genomic studies. On the other hand, the soybean unreplicated haploid genome contains 1,115 Mbp (1). This almost 8-fold difference in genome size appears to be due to ancient polyploidization event(s) during the evolution of the Glycinea (3) and the high level of repetitive sequences in the soybean genome (4).

DNA hybridization under moderate stringency indicated that more than 90% of the nonrepetitive sequences in soybean are present in more than two copies, with the average chromosomal segment being duplicated approximately 2.55 times (3). *Arabidopsis* presents a different story. McGrath *et al.* (5) suggested that only about 15% of the *Arabidopsis* genes may be encoded by duplicate loci. A later study based on numbers of restriction fragment bands observed through hybridization with restriction fragment length polymorphism (RFLP) probes estimated that number to be 14% (6). Approximately 98% of *Arabidopsis* RFLP markers mapped to a single locus (7).

An analysis of approximately 25,000 *Arabidopsis* expressed sequence tags suggested that relatively few highly similar isoforms of genes are found in the *Arabidopsis* genome (8). Still, many examples of multigene families have been reported (9). A comparative mapping study between *A. thaliana* and *Brassica oleracea* revealed islands of conserved organization between the two chromosome complements (6) and identified a region of *Arabidopsis* chromosome I that seemed to be homeologous with a region on chromosome V. Short regions of synteny between a much broader sample of higher plant taxa have been reported, including a possible duplication between *Arabidopsis* chromosomes I and III (10). Recently, an analysis of a 400-kb contig from *Arabidopsis* chromosome IV uncovered a 45-kb segment that seemed to be duplicated on chromosome II (9). These isolated observations indicated that segmental duplications within the *Arabidopsis* genome may have occurred during its evolutionary past.

Comparative genome analyses between soybean and *Arabidopsis* could facilitate cross-utilization of genetic resources and tools of both species and could shed light on evolutionary events associated with the divergence of their seemingly disparate genomes. The public availability of data generated from various genomics programs makes possible the comparative analyses of plant genomes representing broadly divergent genera. However, the detection of duplicated genes by DNA hybridization is less effective than comparisons at the protein sequence level because of the degeneracy of the genetic code in directing amino acid sequence. Consequently, gene duplications that occurred long ago are not likely to be detected by hybridization techniques or direct DNA sequence comparisons although they may be inferred by comparisons of protein sequences.

The objectives of this project were to investigate the degree of synteny between *Arabidopsis* and soybean by using conceptual translations of newly available DNA sequences rather than hybridization techniques. During the course of this project we detected significant synteny between soybean and *Arabidopsis*. We also found compelling evidence for multiple segmental duplications or possibly whole genome duplication of the *Arabidopsis* genome during its evolutionary history.

Materials and Methods

Soybean RFLP probes were chosen from the composite molecular map described by Cregan *et al.* (11). Many of the probes have only a single reported map location, although upon hybridization to restriction enzyme-digested genomic DNA each probe produces an average of 2.55 RFLP bands (3). For this study, both the

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: RFLP, restriction fragment length polymorphism; SSR, simple sequence repeat; BAC, bacterial artificial chromosome; soyA2, soybean linkage group A2; arabI, *Arabidopsis* chromosome I; AP, acid phosphatase(s); PAP, purple AP(s).

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AZ044886–AF045004 and AF237389–AF237412).

†To whom reprint requests should be addressed at: U.S. Department of Agriculture–Agricultural Research Service and Iowa State University, G304 Agronomy Hall, Ames, IA 50011. E-mail: dgrant@iastate.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.070430597. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.070430597

23 clones from which simple sequence repeat (SSR) markers on linkage group J were derived and all 68 available RFLP probes that mapped to soybean linkage groups A2, L, and J were sequenced. These soybean linkage groups were chosen as being representative of a densely populated map (A2) and a pair of homeologous linkage groups (J and L).

Plasmid DNA of clones containing *Pst*I fragments of soybean genomic DNA (12) was prepared by using alkaline lysis minipreps and Qiagen columns. Single sequence runs were made from both ends of the cloned soybean DNA insert by using primers located in the cloning vector. Sequencing was performed by the DNA Sequencing Facility at Iowa State University. Reactions used the Applied Biosystems Prism BigDye terminator cycle sequencing kit with AmpliTaq DNA Polymerase FS and were electrophoresed on an Applied Biosystems Prism 377 DNA sequencer. DNA sequences for the SSR clones were obtained previously (11).

The acid phosphatase (*AP*) gene located on soybean linkage group A2 has not been isolated. In this case, ENTREZ searches of the *Arabidopsis* bacterial artificial chromosome (BAC) annotations at the National Center for Biotechnology Information were used to find acid phosphatase-related sequences in *Arabidopsis*.

Homology searches were performed by using BLAST programs (13–15) at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) and ATDB (Stanford University; http://genome-www.stanford.edu/Arabidopsis/). Default parameter values were used for all homology searches.

Because comparisons were between two evolutionarily distant species, we analyzed all matches to *Arabidopsis* BAC sequence conceptual translations whose region of homology to the soybean sequence was both a subset of, and in the same translation frame as, the most significant match, which had Expect values of less than 0.025 and which had map locations listed in the summaries at the Arabidopsis Genome Analysis Project (Cold Spring Harbor Laboratory; http://nucleus.cshl.org/protarab/). Other sources of *Arabidopsis* sequence and map information used in this analysis were the *Arabidopsis thaliana* BAC Sequencing Project (The Institute for Genomic Research; http://www.tigr.org/tigr_home/tdb/at/atgenome/atgenome.html), the Kazusa *Arabidopsis thaliana* Genome Project (http://www.kazusa.or.jp/arabi/), The Nottingham *Arabidopsis* Stock Centre (http://nasc.nott.ac.uk/), and the Munich Information Centre for Protein Sequences *Arabidopsis thaliana* Sequencing Project (http://www.mips.biochem.mpg.de/proj/thal/).

To assess the probability that the inter- and intragenomic synteny we report was detected by chance, a simulated *Arabidopsis* genome was divided into appropriately sized bins based on either the number of BACs in the putative homologous regions or the genetic size of the regions. We then randomly placed "homologies" in as many bins as there were sequence homologies detected for each soybean sequence. The order of the simulated homologies in each bin was not considered in the analysis. At least 10,000 simulated genomes were analyzed for each case of putative synteny or duplication. A simulated genome was considered a match to our results if the number of bins containing at least one copy of each soybean sequence homolog was at least equal to the number of such bins actually observed. Complete details of the soybean/*Arabidopsis* homologies we found, along with the simulation algorithms and the results of the simulations, can be found at http://soybase.agron.iastate.edu/publication_data/Grant/synteny1.

Results

Synteny Between Soybean and *Arabidopsis*. The gapped-TBLASTX program, which makes comparisons of predicted amino acid sequences by using each of the six reading frames (15), was used to compare DNA sequences from 68 soybean RFLP clones and 23 SSR-containing genomic DNA clones against all *Arabidopsis* sequences in GenBank. Conceptual translations of DNA sequences were used for the comparisons because they provide a more sensitive test of homology between evolutionarily widely separated species than do nucleotide sequence comparisons. The soybean RFLP probes used in this study were generated through the use of methylation-sensitive enzymes (12). This approach long has been thought to be a means to enrich for transcribed sequences (16). This is borne out by our finding that 72% of the RFLP clones showed significant homology to at least one *Arabidopsis* genomic or cDNA sequence. In contrast, only one of the sequences surrounding soybean SSRs had any detectable homology to *Arabidopsis*. In this case, the homology was to a putative exon and the relative position of the SSR was in an intron. The BLAST reports suggested that none of the homologies we detected were to known repetitive sequences. Many matches were to cDNAs or isolated genes that did not have a reported map location, although in some cases these sequences were contained in mapped BACs. The map location(s) of the matching BAC(s) was determined by using information provided at the Cold Spring Harbor web site.

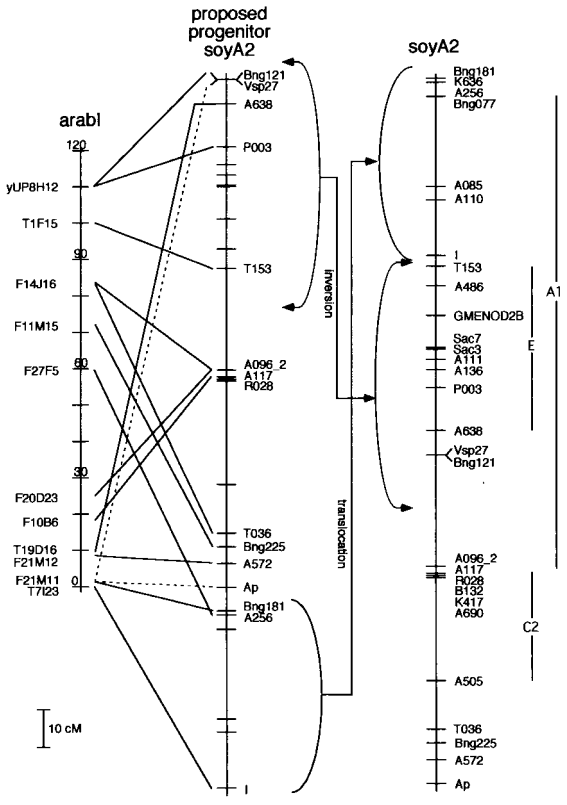

Fig. 1. Synteny between soyA2 and arabI. BACs showing homology to soybean sequences are indicated with lines connecting them to their soybean homolog(s). The soyA2 map at the right shows the modern linkage group, with each locus that was analyzed in this study indicated. The proposed progenitor soyA2 in the middle shows a rearranged soyA2 that maximizes the synteny with arabI. Soybean *Vsp*27 and AP cannot be distinguished at this level of analysis; this ambiguity is indicated by broken lines connecting *Arabidopsis* BAC F2 M11 and the two soybean loci. The *Arabidopsis* map is drawn inverted relative to the usual presentation. Tic marks and numbers indicate 10-cM intervals on arabI. The previously identified regions of homeology between soybean linkage groups (3) are identified and shown as vertical lines to the right of soyA2.

Comparisons of map positions of linked soybean RFLP probes and *Arabidopsis* BACs revealed many regions of synteny. Fig. 1 shows the homologies detected between sequences from soyA2

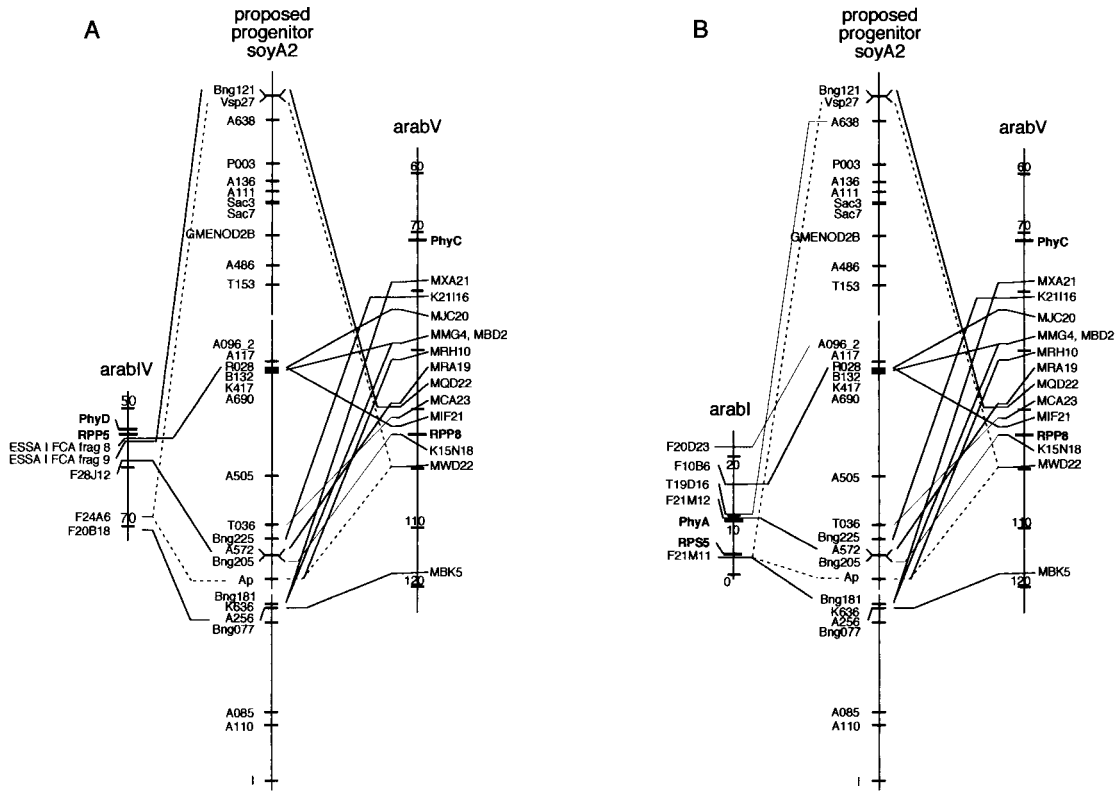

Fig. 2. SoyA2 is shown with the proposed progenitor locus order (see Fig. 1). Only those loci that had significant homology to *Arabidopsis* sequences on arabIV or arabV are connected by lines, although tic marks for every soybean sequence analyzed are shown on the proposed progenitor soyA2 map. Thin lines indicate soybean sequences that had homologs on only one *Arabidopsis* chromosome. Broken lines are used to indicate uncertainty in syntenic relationships because of duplicated loci in soybean. Known genes in *Arabidopsis* are shown in bold type. Tic marks and numbers indicate 10-cM intervals on the *Arabidopsis* chromosomes. (*A*) Synteny between loci on soybean linkage group A2 and duplicated segments of *Arabidopsis* chromosomes IV and V. (*B*) Synteny between soybean linkage group A2 and duplicated segments of *Arabidopsis* chromosomes I and V.

and *Arabidopsis* chromosome I (arabI), where soybean sequences distributed along the entire linkage group had homologs on arabI. Because the homologous sequences in soybean and *Arabidopsis* have been separated for approximately 90 million years (17) and, at least in soybean, there have been 1–2 rounds of genome duplication since their divergence (3), we did not expect *a priori* that there would be any correlation between a RFLP's map position in soybean and the location in *Arabidopsis* of the BAC that contained the most significant homology. However, for those 14 sequences on soyA2 that have homologs in BACs on arabI, seven (50%) had the lowest expect value returned by TBLASTX, whereas four were the second lowest. Eleven of the 14 soybean sequences that revealed synteny between soyA2 and arabI were homologous to *Arabidopsis* genes or cDNAs. The remaining three with no reported matches to expressed sequences (A096, A117, and T153) had expect values of 1.7e-7, 4e-9, and 7e-34, respectively. In three instances two soybean sequences were homologous to distinct sequences in the same *Arabidopsis* BAC (F21M11, yUP8H12, and F14J16). Simulations based on our data were conducted to test the likelihood that the synteny we observed was an artifact of analyzing a very large data set. This appears not to be the case because only 23 of 10,000 simulated *Arabidopsis* genomes had a chromosome that contained all of the loci we report. Our simulations did not consider order of loci on the chromosomes. If we had, the number of matches found would perforce have been much lower.

The soybean vegetative storage protein gene *Vsp*27 (18), which maps more than 50 cM from *AP* on linkage group A2 (11), has acid phosphatase activity (19), and its DNA sequence shows significant homology to many nonpurple AP. For this reason, we were unable to determine unambiguously the orthologous or paralogous relationships between these genes and homologous sequences in *Arabidopsis*. To indicate this ambiguity we use broken lines in Fig. 1 to show the most parsimonious syntenic relationships.

Although there is substantial synteny between soyA2 and arabI, the maps are not completely colinear. However, the magnitude of the differences are similar to those found in other interspecies comparisons in which synteny has been observed and a limited number of chromosomal rearrangements need be invoked to explain how the observed locus orders were derived from the ancestral ones (20, 21). Fig. 1 demonstrates how one translocation and one inversion in the evolution of soyA2 substantially explains the observed map order differences between soybean and *Arabidopsis*. Interestingly, the putatively rearranged blocks of soyA2 indicated in Fig. 1 are very similar to regions of homeology reported in the soybean genome (3).

Duplicated Segments Within the *Arabidopsis* Genome. Most soybean RFLP sequences had strong homologies to BACs on more than one *Arabidopsis* chromosome. Surprisingly, we found that these multiple homologies often identified homeologous regions in *Arabidopsis*.

Segmental duplication involving three Arabidopsis *chromosomes.* Fig. 2 shows the synteny between soyA2 and subchromosomal regions of arabI, arabIV, and arabV. To help in visualizing the relationships between soybean and *Arabidopsis* we have shown the proposed progenitor to soybean linkage group A2 in

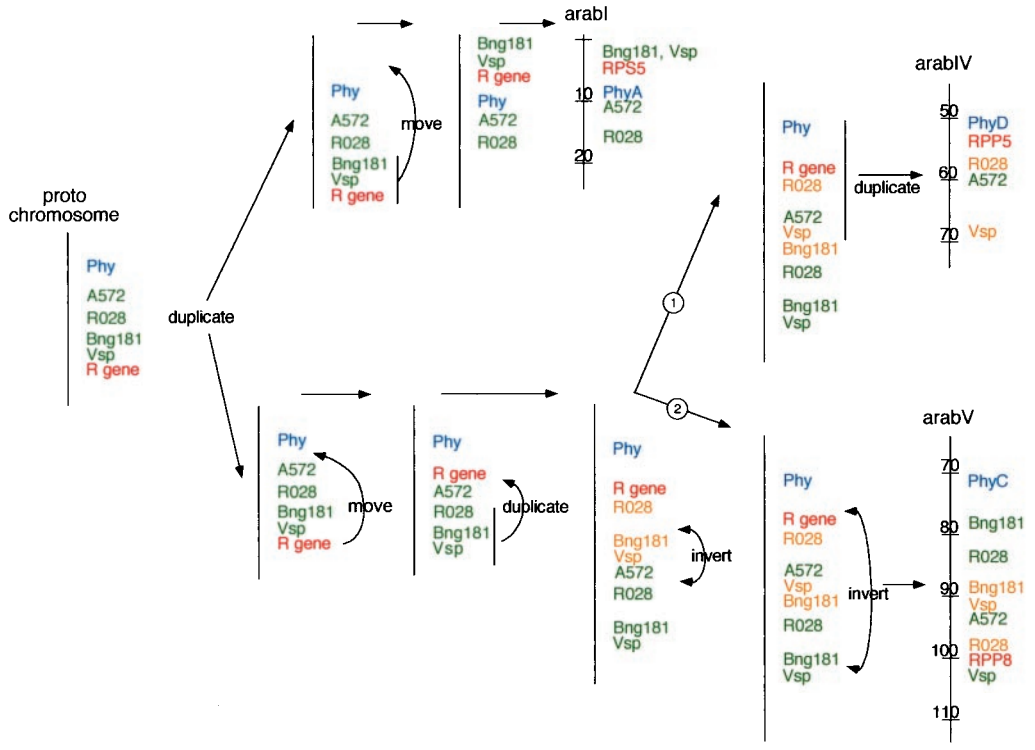

Fig. 3. Proposed evolutionary derivation of related regions of *Arabidopsis* chromosomes I, IV, and V. Green and orange are used to help track chromosomal segments only and do not necessarily indicate related functionality. In this model, an ancestral chromosome or chromosomal segment (protochromosome) was duplicated, producing lineages that culminated in parts of the modern *Arabidopsis* chromosomes I, IV, and V. In the arabIV/V lineage the path leading to arabIV branched off before the final inversion occurred. Vsp is used in the figure for both AP similarity and sequence homology to the soybean *Vsp*27 gene. Tic marks and numbers on the modern *Arabidopsis* chromosomes indicate 10-cM intervals.

Fig. 2 because this map likely is more similar to that of their last common ancestor than is the arrangement of the current soyA2. Nine soybean loci spanning the entire length of soyA2 had homologs in an approximately 20- to 30-cM region of arabV (Fig. 2). A subset of these loci also had homologs on arabIV (Fig. 2*A*), and a distinct but partially overlapping subset had homologs on arabI (Fig. 2*B*). Eight of 11 soybean sequences had matches to *Arabidopsis* genes or cDNAs. The remaining three had significant homology only to BACs (A096, E = 1.7e-7; Bng205, E = 2e-8; B132, E = 5e-22). In addition to the shared homologies to soybean RFLP sequences, and providing further support of their homeologous relationships, the three chromosomal segments also each contain at least one copy of a disease-resistance gene and a phytochrome gene. *RPP*5 and *RPP*8 (arabIV and arabV, respectively) both confer resistance to *Peronospora parasitica* (downy mildew). *RPP*5 is a member of the TIR-NBS-LRR R-gene subclass whereas *RPP*8 is an example of the LZ-NBS-LRR subclass (22, 23). *RPS*5 (arabI) conditions resistance to *Pseudomonas syringae* and is also a member of the LZ-NBS-LRR R-gene subclass (24). The five members of the phytochrome gene family in *Arabidopsis* can be grouped into three lineages: *Phy*A, *Phy*B/D/E, and *Phy*C (25). Each of these ancient lineages is represented on only one of the three evolutionarily related segments (*Phy*A, arabI; *Phy*D/E, arabIV; *Phy*C, arabV). Simulations showed that the probability of observing such apparently duplicated segments by chance is approximately 0.036.

A comparison of Fig. 2 A and *B* shows that some of the homologs to soybean sequences that define the arabIV/arabV homeologous regions also contribute to defining the arabI/arabIV regions. Despite this similarity, both pairs of homeologous regions have slightly different orders of soyA2 homologs between their members. This surprising overlap of homologous sequence composition suggests that all three of the modern *Arabidopsis* chromosomal regions could have been derived from a single progenitor chromosome. Fig. 3 shows how, starting with a single chromosomal segment, a relatively simple series of chromosomal duplications and rearrangements generate the order of homologous loci in all three regions of the modern *Arabidopsis* genome. In this model the progenitor chromosome contained single copies of each locus, which then diverged through a series of duplications and rearrangements to yield the chromosomal segments observed today.

Arabidopsis *chromosomes II and IV share a duplicated segment.* Based on duplicate RFLP loci, soybean chromosomes J and L (soyJ and soyL) have been proposed to be homeologous (3). Our analysis showed that the upper 15–20 cM of both soybean linkage groups corresponded to approximately 6 cM on both arabII and arabIV (Fig. 4). Additionally, two linked markers in soybean (Sct_046 and A060) mapped to single BACs in each duplicated region. Other soybean sequences from this region also mapped to similar regions in arabII and arabIV, although clearly significant rearrangements have occurred in both genomes since their last common ancestor. Six of seven soybean sequences had matches to genes or cDNAs; B101 had a match to a BAC with an expect value of 9e-28. Simulations using the three loci in common between the homologous segments, but not considering that a single BAC in each region contained the same two soybean sequence homologies, show that the probability of all three occurring in two subgenomic regions is approximately 0.002.

Multiple segmental duplications involving the acid phosphatase genes in Arabidopsis. Associations of soybean sequences with acid phosphatase-like sequences in *Arabidopsis* suggest that segmental duplication may be a common event in *Arabidopsis* and may underlie some of the extensive gene duplication that has been reported for both species.

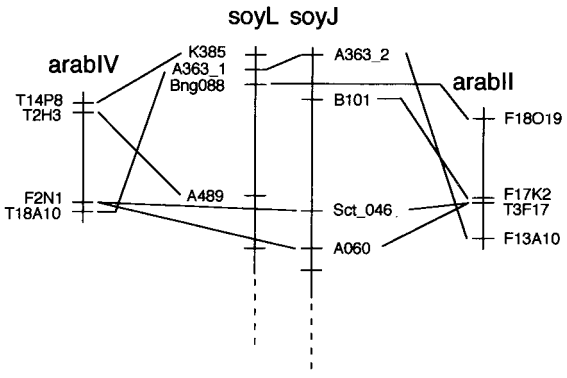

Fig. 4. Synteny between parts of homeologous soybean linkage groups J and L and duplicated segments of *Arabidopsis* chromosomes II and IV located at approximately 80 and 10 cM, respectively. Homologous sequences are connected by solid lines. Homologs in *Arabidopsis* to soybean sequences Sct_046 and A060 are located on a single BACs on both arabII and arabIV.

DNA sequence from soybean *Vsp*27 (18) shows homology to *Arabidopsis* nonpurple AP but not to the purple AP (PAP). In two instances, *Vsp*27 and Bng181 homologs appear on the same *Arabidopsis* BAC on arabI (F21M11) and arabII (T28M21) (Fig. 5). In addition, the Cold Spring Harbor table shows these homologs to be about nine BACs apart on arabV (MBD2 and MRH10). In three cases, *Arabidopsis* PAP-related sequences and

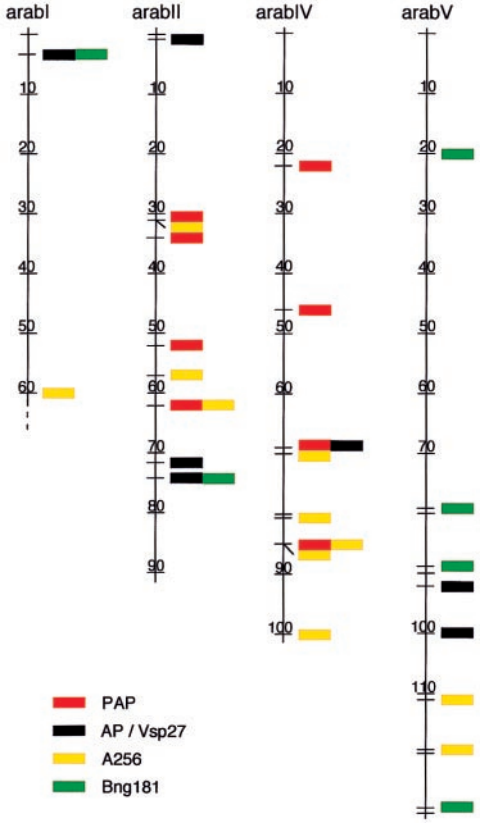

Fig. 5. Locations of AP- and PAP-like sequences and homologs to soybean *Vsp*27, A256, and Bng181 on *Arabidopsis* chromosomes. Numbers on the maps show an approximate scale in centimorgans. *Arabidopsis* homologs to soybean AP/Vsp27 and A256 are tightly linked in three locations. *Arabidopsis* homologs to PAP and soybean Bng181 are tightly linked at five locations.

homologs of soybean A256 map to the same BAC on arabII (F24L7) and arabIV (ESSA AP2 fragment 2) or separated by one BAC (about 140 kb) on arabII (F16F14 and T24I21). In two other cases they are separated by five or seven BACs on arabIV (F13M23 and F20B18) or arabII (T22O13 and T17D12), respectively. In all cases the soybean RFLP homologs and the *Vsp*27 sequence or AP similarity are in different parts of the BAC. Each of these sequences or genes also appears several times in the *Arabidopsis* genome not in close association. Because only about 40% of the *Arabidopsis* genome has been annotated, the BACs we identified by using ENTREZ probably represent only a subset of the AP-like sequences in *Arabidopsis*. Simulations showed that the probability of such tight linkage between A256 and PAP is approximately 0.0003 whereas that for Bng181 and AP is approximately 0.008.

There were no obvious patterns in the expect values for paralogous sequences in the duplicated regions in *Arabidopsis* although both Bng181 and A256 are homologous to cDNAs in *Arabidopsis*. For example, BAC F21M11 on arabI contained homologs to *Vsp*27 (E = 3.1e-24) and Bng181 (E = 2.4e-3) whereas the paralogous sequences in BAC T28M21 on arabII had expect values of 1.5e-2 and 2.2e-11, respectively. This suggests that the genes in the duplicated regions have evolved independently since the duplication event.

The proposed soybean chromosomal rearrangements shown in Fig. 1 explain many of the differences in map order now seen between arabI and soyA2. The proposed telomeric translocation places Bng181 and A256 very close to the AP gene in the ancestral chromosome. It is tempting to hypothesize that this association represents the starting point for several segmental duplications in *Arabidopsis*.

Discussion

We have conducted a preliminary comparative analysis of the soybean genome with that of *Arabidopsis*. Using data from only three linkage groups of soybean and the information currently available from the *Arabidopsis* Genome Initiative, we were able to demonstrate synteny between the two genera and show that duplicated segments spanning 10–20 cM are common in *Arabidopsis*. To make our analysis more powerful, we compared genome organization of soybean and *Arabidopsis* through TBLASTX conceptual translations of the DNA sequences. Such comparisons may have a higher probability of recognizing true evolutionary relationships because (*i*) many changes at the DNA level are not reflected at the amino acid level due to the degeneracy of the genetic code and the functional similarity of some amino acids and (*ii*) subgenic regions or motifs often are widely conserved while the sequences between them are not. Most of the soybean RFLP sequences we analyzed showed homology to BACs from multiple locations in *Arabidopsis*. Because the sequences we used were from random, *Pst*I-generated genomic fragments, and thus were not full-length genes, the homologies we observed were necessarily to subgenic regions. This makes it difficult to know whether we are detecting members of gene families. However, the high incidence of conserved amino acid sequences allowed us to infer the evolutionary relatedness of sequences in soybean and *Arabidopsis* without any direct knowledge of gene function in either organism. This approach revealed extensive genome duplication within the *Arabidopsis* genome. Our results suggest that extensive segmental duplication has occurred during the evolution of this genome or even that, similar to soybean, *Arabidopsis* may be an ancient paleopolyploid.

Despite the substantial colinearity observed between soyA2 and arabI, it is clear that these two chromosomes do not simply represent modern chromosomes that evolved from a single, common ancestor. Many loci on soyA2 have no detectable homolog on arabI, and, although we analyzed sequences from

only three soybean linkage groups, loci from all three had homologs on arabI. This along with the numerous duplications we observed in *Arabidopsis* suggests that considerable chromosomal rearrangement involving small regions likely has occurred in both species. Elucidating the evolutionary history of these chromosomes is complicated further by the one or two genome duplications and subsequent diploidization in the soybean lineage (3). In most cases only one of the 2–4 loci detected by a RFLP probe in soybean has been mapped. Thus, the DNA sequence we used in these comparison is not necessarily that of the mapped locus and may explain why the matches we report between soyA2 and arabI were often but not always the most significant ones of those we observed.

The evidence we present for duplicated chromosomal regions in *Arabidopsis* is based on homologies between the predicted amino acid sequences derived from soybean genomic clones and *Arabidopsis* BACs. One of the decisions we had to make in making these comparisons was whether a given sequence similarity was evolutionarily significant or was simply due to chance. In this study we did not want to miss any weak homologies resulting from ancient duplications or speciation with subsequent divergence of the sequences. Thus, we included any low-scoring *Arabidopsis* sequences whose region of homology to the soybean sequence was both a subset of, and in the same translation frame as, the most significant match and that had expect values of less than 0.025, although such E values normally would not be considered significant. Although this means that potentially some matches might be accepted when they were not actually significant, we found only one soybean/*Arabidopsis* homology in a duplicated region where the probability of the match occurring by chance was less than 1 in 50 (soyJ/A060 on arabII). In total, only eight of the homologies we report (5.5%) had probabilities of being due to chance of less than 1:1,000.

The level of duplicated loci in *Arabidopsis* has been proposed to approximate a basal level of duplicated genes among crucifers (5). Our observation of extensive segmental genomic duplication covering regions in all of the *Arabidopsis* chromosomes for which sequence is available suggests that even a basal genomic level of redundancy in a higher eukaryote may include a high level of ancient genome duplication beyond the single-gene level.

Circumstantial evidence would suggest that all organisms have experienced at least one round of genome duplication in their phylogenetic past. Thus, all eukaryotes probably are ancient polyploids (26). There is a tendency for a polyploid genome to evolve into a diploid state through sequence diversification and chromosomal rearrangement (3, 26–28). This process of "diploidization" may result in changes in the amount of nuclear DNA because of additions and deletions, major genome restructuring because of rearrangements (29), as well as an accumulation of sequence and functional differences (30). Not surprisingly, then, our results indicate that at least one member of each pair of the large duplicated regions we identified in *Arabidopsis* has been rearranged since the duplication event.

Cretaceous fossil records have placed rosids of various types (a lineage that includes legumes) and a capparalean taxon (Capparales include Cruciferae) to about 92 million years ago (17), indicating that divergence of the lineages that gave rise to Cruciferae and legumes probably occurred at about that time. Despite this long period of separation, we were able to detect numerous instances of sequence homology and several regions of synteny between *Arabidopsis* and soybean. Our results, along with those reported previously between various *Brassicas* and *Arabidopsis* (20, 31, 32), *Arabidopsis* and cotton (10), among some legumes (33, 34), and between members of the Solanacea (21), suggest that it should be possible to use the maps and molecular information developed for *Arabidopsis* widely throughout the dicots.

We thank Ms. Brianne Veach for technical assistance, Dr. T. Vision for critical reading of the manuscript, and Dr. D. Ashlock for statistics assistance. Contributions of the Corn Insect and Crop Genetics Research Unit, U.S. Department of Agriculture–Agricultural Research Service, Midwest Area, and Project 3236 of the Iowa Agriculture and Home Economics Experiment Station (Ames, IA) are acknowledged. This is journal paper no. 18657.

1. Arumuganathan, K. & Earle, E. D. (1991) *Plant Mol. Biol. Rep.* **9,** 208–218.
2. Pruitt, R. E. & Meyerowitz, E. M. (1986) *J. Mol. Biol.* **187,** 169–184.
3. Shoemaker, R. C., Polzin, K., Labate, J., Specht, J., Brummer, E. C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J. P., *et al.* (1996) *Genetics* **144,** 329–338.
4. Gurley, W. B., Hepburn, A. G. & Key, J. L. (1979) *Biochim. Biophys. Acta* **561,** 167–183.
5. McGrath, J. M., Jancso, M. M. & Pichersky, E. (1993) *Theor. Appl. Genet.* **86,** 880–888.
6. Kowalski, S. P., Lan, T.-H., Feldmann, K. A. & Paterson, A. H. (1994) *Genetics* **138,** 499–510.
7. Chang, C., Bowman, J. L., DeJohn, A. W., Lander, E. S. & Meyerowitz, E. M. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6856–6860.
8. Rounsley, S. D., Glodek, A. & Sutton, G. (1996) *Plant Physiol.* **112,** 1177–1193.
9. Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villaroel, R., Jesse, T., *et al.* (1999) *FEBS Lett.* **445,** 237–245.
10. Paterson, A. H., Lan, T.-H., Reischmann, K. P., Chang, C., Lin, Y.-R., Liu, S.-C., Burow, M. D., Kowalski, S. P., Katsur, C. S., DelMonet, T. A., *et al.* (1996) *Nat. Genet.* **14,** 380–382.
11. Cregan, P. B., Jarvik, T., Bush, A. L., Shoemaker, R. C., Lark, K. G., Kahler, A. L., Kaya, N., VanToai, T. T., Lohnes, D. G., Chung, J., *et al.* (1999) *Crop Sci.* **39,** 1464–1490.
12. Keim, P. & Shoemaker, R. C. (1988) *Soybean Genet. Newsletter* **15,** 147–148.
13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
14. Gish, W. & States, D. J. (1993) *Nat. Genet.* **3,** 266–272.
15. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
16. Burr, B., Burr, F., Thompson, K., Albertson, M. & Stuber, C. (1988) *Genetics* **118,** 519–526.
17. Gandolfo, M., Nixon, K. & Crepet, W. (1998) *Am. J. Bot.* **85,** 964–974.
18. Staswick, P. E. (1988) *Plant Physiol.* **87,** 250–254.
19. DeWald D. B., Mason, H. S. & Mullet, J. E. (1992) *J. Biol. Chem.* **267,** 15958–15964.
20. Lagercrantz, U. (1998) *Genetics* **150,** 1217–1228.
21. Livingstone, K., Lackney, V. K., Blauth, J. R., van Wijk, R. & Jahn, M. K. (1999) *Genetics* **152,** 1183–1202.
22. Parker, J. E., Coleman, M. J., Szabo, V., Frost, L. N., Schmidt, R., van der Biezen, E. A., Moores, T., Dean, C., Daniels, M. J. & Jones, M. D. (1997) *Plant Cell* **9,** 879–894.
23. McDowell, J. M., Dhandaydham, M., Long, T. A., Aarts, M. G., Holub, E. B. & Dangl, J. L. (1998) *Plant Cell* **10,** 1861–1874.
24. Warren, R. F., Henk, A., Mowery, P., Holub, E. & Innes, R. W. (1998) *Plant Cell* **10,** 1439–1452.
25. Mathews, S. & Sharrock, R. A. (1997) *Plant Cell Environ.* **20,** 666–671.
26. Leipoldt, M. & Schmidtke, J. (1982) in *Genome Evolution,* eds. Dover, G. & Flavell, R. (Academic, New York), pp. 219–236.
27. Song, K., Lu, P., Tang, K. & Osborn, T. (1995) *Proc. Nat. Acad. Sci. USA* **92,** 7719–7723.
28. Lagercrantz, U. & Lydiate, D. (1996) *Genetics* **144,** 1903–1910.
29. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, New York).
30. Pickett, F. B. & Meeks-Wagner, D. R. (1995) *Plant Cell* **7,** 1347–1356.
31. Cavell, A. C., Lydiate, D. J., Parkin, I. A. P., Dean, C. & Trick, M. (1998) *Genome* **41,** 62–69.
32. Conner, J. A., Conner, P., Nasrallah, M. E. & Nasrallah, J. B. (1998) *Plant Cell* **10,** 801–812.
33. Weeden, N. F., Muehlbauer, F. J. & Ladizinsky, G. (1992) *J. Hered.* **83,** 123–129.
34. Boutin, S. R., Young, N. D., Olson, T. C., Yu, Z. H., Shoemaker, R. C. & Vallejos, C. E. (1995) *Genome* **38,** 928–937.