11-2009

Molecular and Chromosomal Evidence for Allopolyploidy in Soybean

Navdeep Gill

Seth Findley

Jason G. Walling

Christian Hans

Jianxin Ma

See next page for additional authors

Authors

Navdeep Gill, Seth Findley, Jason G. Walling, Christian Hans, Jianxin Ma, Jeff Doyle, Gary Stacey, and Scott A. Jackson

Molecular and Chromosomal Evidence for Allopolyploidy in Soybean[1][OA]

Navdeep Gill, Seth Findley, Jason G. Walling, Christian Hans, Jianxin Ma, Jeff Doyle, Gary Stacey, and Scott A. Jackson*

Department of Agronomy (N.G., J.G.W., C.H., J.M., S.A.J.) and Interdisciplinary Life Science Program (N.G., S.A.J.), Purdue University, West Lafayette, Indiana 47907; Division of Plant Sciences, Bond Life Science Center, University of Missouri, Columbia, Missouri 65211 (S.F., G.S.); and Department of Plant Biology, Cornell University, Ithaca, New York 14853 (J.D.)

Recent studies have documented that the soybean (*Glycine max*) genome has undergone two rounds of large-scale genome and/or segmental duplication. To shed light on the timing and nature of these duplication events, we characterized and analyzed two subfamilies of high-copy centromeric satellite repeats, CentGm-1 and CentGm-2, using a combination of computational and molecular cytogenetic approaches. These two subfamilies of satellite repeats mark distinct subsets of soybean centromeres and, in at least one case, a pair of homologs, suggesting their origins from an allopolyploid event. The satellite monomers of each subfamily are arranged in large tandem arrays, and intermingled monomers of the two subfamilies were not detected by fluorescence in situ hybridization on extended DNA fibers nor at the sequence level. This indicates that there has been little recombination and homogenization of satellite DNA between these two sets of centromeres. These satellite repeats are also present in *Glycine soja*, the proposed wild progenitor of soybean, but could not be detected in any other relatives of soybean examined in this study, suggesting the rapid divergence of the centromeric satellite DNA within the *Glycine* genus. Together, these observations provide direct evidence, at molecular and chromosomal levels, in support of the hypothesis that the soybean genome has experienced a recent allopolyploidization event.

At least 50% to 70% of land plants are estimated to be polyploid (Wendel, 2000), which may be an underestimate because recent genome sequencing has revealed that even putatively diploid genomes contain vestiges of polyploidy, or at least large segmental duplications (Vision et al., 2000; Paterson et al., 2004). Polyploidy is especially prevalent among crop plants, suggesting that polyploidy may provide an advantage in domestication (Udall and Wendel, 2006). Despite the incidence of polyploidy and the logical extension of its importance in evolution and adaptation, relatively little is known about how and when these major events occurred and why they were fixed in many plants, including soybean (*Glycine max*).

There have been several rounds of polyploidization and/or segmental duplication in soybean on the basis of chromosome number (Lackey, 1980), multiple hybridizing RFLP fragments (Shoemaker et al., 1996), and analysis of duplicated ESTs (Blanc and Wolfe, 2004; Schlueter et al., 2004). Whole-chromosome homology, as revealed by fluorescence in situ hybridization (FISH), provides additional evidence for polyploidy in soybean (Walling et al., 2005). Collectively, these data indicate that there have been two large-scale duplication events that occurred within the last 50 million years.

Centromeres of multicellular eukaryotes are generally composed of high-copy, satellite repeats such as the α-satellites of human (Willard, 1985), CentO of rice (*Oryza sativa*; Dong et al., 1998), the pAL1 satellites of Arabidopsis (*Arabidopsis thaliana*; Murata et al., 1994), and CentC of maize (*Zea mays*; Ananiev et al., 1998). Despite the importance of centromeres for chromosomal segregation in cell division, DNA sequences at centromeres are, paradoxically, not well conserved, even within a single genus (Lee et al., 2005). Thus, in an allopolyploid, formed by hybridization between two species, it is not unexpected that the centromeres differ at the sequence level, as seen in allopolyploids of Arabidopsis formed within the past approximately 300,000 years (Kamm et al., 1995; Pontes et al., 2004).

The ancestor of soybean and the remainder of the genus *Glycine* has been hypothesized to have been formed via a polyploid event within the last 15 million years (Shoemaker et al., 2006); however, it remains unclear whether the event was allo- or autopolyploid (Kumar and Hymowitz, 1989; Straub et al., 2006). Here we report that the soybean genome harbors two

[1] This work was supported by the National Science Foundation (grant nos. DBI 0836196 and 0501877 to S.A.J., and grant no. 0516673 to J.D.).

* Corresponding author; e-mail sjackson@purdue.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Scott A. Jackson (sjackson@purdue.edu).

www.plantphysiol.org/cgi/doi/10.1104/pp.109.137935

classes of centromere-specific repeats that mark subsets of chromosomes, suggesting that *Glycine* is allopolyploid.

RESULTS

Identification and Cytological Verification of Centromeric Satellite Repeats in Soybean

To identify centromeric satellite repeats, we constructed a whole-genome shotgun (WGS) library of soybean, and generated 1,454 WGS sequences (GenBank CL867099–CL868434; Lin et al., 2005). Subsequently, a repeat database (http://www.soymap.org) was constructed using a repeat-finding program, RECON (Bao and Eddy, 2002). Because satellite repeats have been found to be a class of DNA that has the highest copy number in completely sequenced genomes of plants, such as Arabidopsis and rice, we hypothesized that the highest-copy-number repeats in the WGS database of soybean might represent centromeric satellite repeats. Therefore, we isolated and analyzed the largest family of repeats revealed by RECON that were composed mostly of 92-bp monomers and were found to be arranged in a head-to-tail pattern in both WGS sequences and RECON-assembled sequences, a feature typical of centromeric satellite repeats (Jiang et al., 2003). A 92-bp sequence from this family was isolated and an oligo-based probe was developed for FISH and found to localize to centromeric regions of 14 out of 20 chromosome pairs (Fig. 1, A–D). In accordance with commonly accepted nomenclature of plant centromeric sequences, this sequence is hereafter referred to as CentGm-1. We used CentGm-1 to search against the nonredundant nucleotide sequence database deposited in GenBank, and found that CentGm-1 is highly similar (>95%) to previously reported SB92 repeats (Vahedian et al., 1995). SB92 repeats were found to hybridize to four to five genomic locations, whereas, in this study it hybridized to 14 centromere pairs, probably a result of more reliable FISH results due to better chromosome preparation techniques.

Since most plant centromeres consist of large arrays of species-specific satellite repeats (Martinez-Zapater et al., 1986; Ananiev et al., 1998; Dong et al., 1998), we asked whether other centromeric satellite repeats existed in the centromeres that were not detected by CentGm-1. Therefore, we searched the repeat database

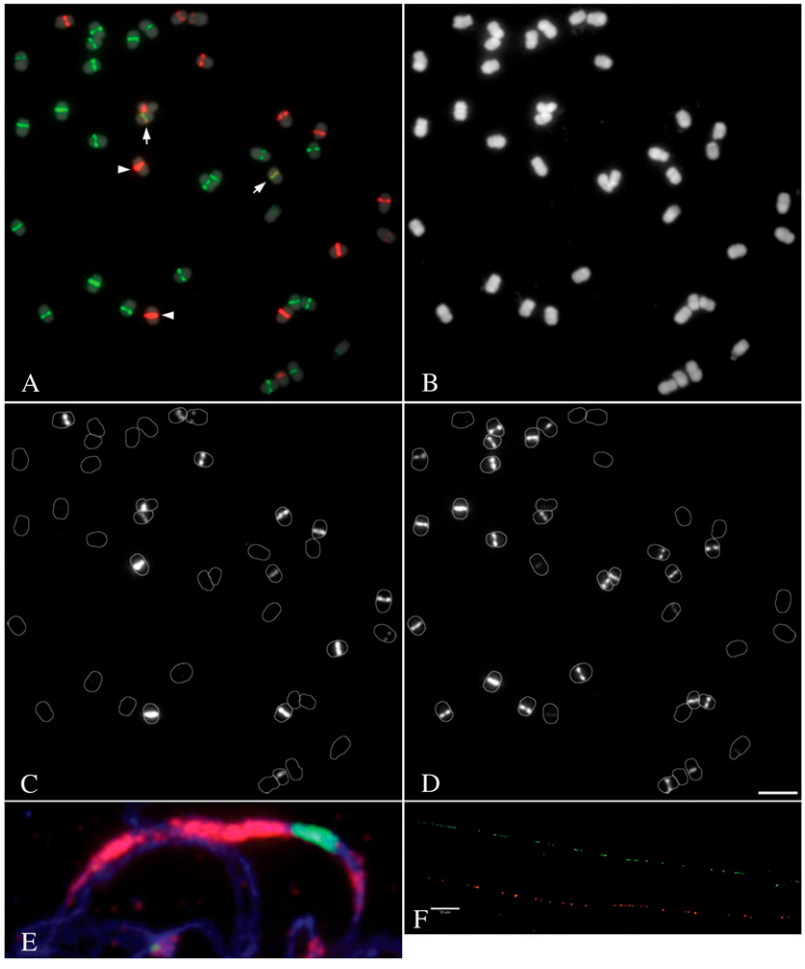

Figure 1. FISH analysis of centromere repeats in soybean. A, Merged image of B to D, CentGm-1 (green) and CentGm-2 (red; arrows indicate both sequences at the same centromere and arrowheads indicate both sequences, but predominantly CentGm-2). B, DAPI-stained chromosomes. C, CentGm-2-labeled red fluorophore marks eight pairs of centromeres (two pairs overlap with CentGm-1). D, CentGm-1 labeled with green fluorophore marks 14 sets of centromeres (bar = 5 μm). E, Centromeric region of a soybean chromosome at the premeiotic pachytene stage. BAC 76J21 is shown in red (labeled digoxigenin detected with rhodamine anti-digoxigenin) marking the pericentromeric repeats whereas CentGm-2 (labeled biotin detected with AF498) marks the primary constriction (centromere). F, FISH of CentGm-2 (labeled biotin detected with two layers of AF498 streptavidin using goat anti-streptavidin as an intermediate) and CentGm-1 (labeled digoxigenin detected with mouse anti-digoxigenin followed by AF568 anti-mouse) on extended DNA fibers from soybean showing little or no colocalization of signal (bar = 10 μm).

that we obtained from the WGS sequences using a tandem repeat finding program (Benson, 1999), and identified another subfamily of high-copy satellite repeats similar in structure to CentGm-1. This subfamily of repeats, named CentGm-2, is composed of 91-bp monomers, sharing approximately 80% sequence identity on average with CentGm-1. In FISH experiments, an oligo probe designed for CentGm-2 localized to eight chromosome pairs (Fig. 1, A–D). A combined FISH experiment with both repeats revealed that the two repeats mark 18 nonoverlapping sets of chromosomes at the centromeric regions (Fig. 1A). Two pairs of centromeres had both sequences; however, one pair was primarily composed of CentGm-1 (Fig. 1D, arrowheads).

To confirm that these sequences were enriched specifically in the centromeric regions and not pericentromeric regions, we utilized a previously isolated bacterial artificial chromosome (BAC), 76J21, that localizes to the pericentromeric regions of all 40 chromosomes (Lin et al., 2005) and contains previously described soybean repeats STR120 (Morgante et al., 1997) and the retrotransposon SIRE1 (Laten et al., 1998). FISH of BAC 76J21 in combination with CentGm-1 and CentGm-2 repeats to premeiotic pachytene chromosomes revealed that both centromeric repeats marked the primary constrictions exclusively (Fig. 1E). Moreover, CentGm-1 localized to the leading edges of chromosomes at mitotic anaphase (Fig. 2A). Due to the squashing of the cell during preparation, some of the chromosomes have been rearranged relative to the mitotic poles/plate, but a majority of the chromosomes still have the CentGm-1 localized on the leading edge that would interact with the spindle apparatus.

CentGm-1 and CentGm-2 Are Found on a Pair of Homologs in Soybean

Previously, we identified potential chromosome homologs in soybean by cross hybridization of BACs to duplicated segments within the soybean genome (Walling et al., 2005). Although CentGm-1 and CentGm-2 are specific to mostly nonoverlapping sets of chromosomes, we asked whether the two centromeric repeats would differentially target the previously identified homologs. Therefore, two BAC clones that cross hybridize to these two homologs were combined in a single FISH experiment with the two centromeric sequences, CentGm-1 and CentGm-2. The results indicated that each pair of homologs is indeed marked by a different centromeric repeat (Fig. 2B). The chromosome corresponding to linkage group 19 is marked by CentGm-1 and the other unidentified chromosome by Cent-Gm-2.

Phylogenetic Analysis of the Centromeric Satellite Repeats in Soybean

The two monomeric sequences used for FISH analysis were used to search against the WGS database by BLASTN and identified 1,001 CentGm-1 and 712 CentGm-2 monomers. Four-hundred sixty-four intact monomer sequences were randomly chosen, aligned, and a distance tree was constructed. Two major clusters were found (Fig. 3A), supporting the hypothesis that there are two major groups of centromeric repeats in the genome with greater than 90% bootstrap support. These results were supported by parsimony analyses (data not shown). The overall mean distance of monomers within clades was 0.134 ± 0.015 and 0.131 ± 0.018 for CentGm-1 and CentGm-2, respectively; whereas the mean distance between the monomers of CentGm-1 and CentGm-2 was 0.275 ± 0.054. Although a cutoff of 60% sequence identity and 80% match length was employed in the BLASTN searches described above, we did not find any other satellite repeats related to CentGm-1 and CentGm-2.

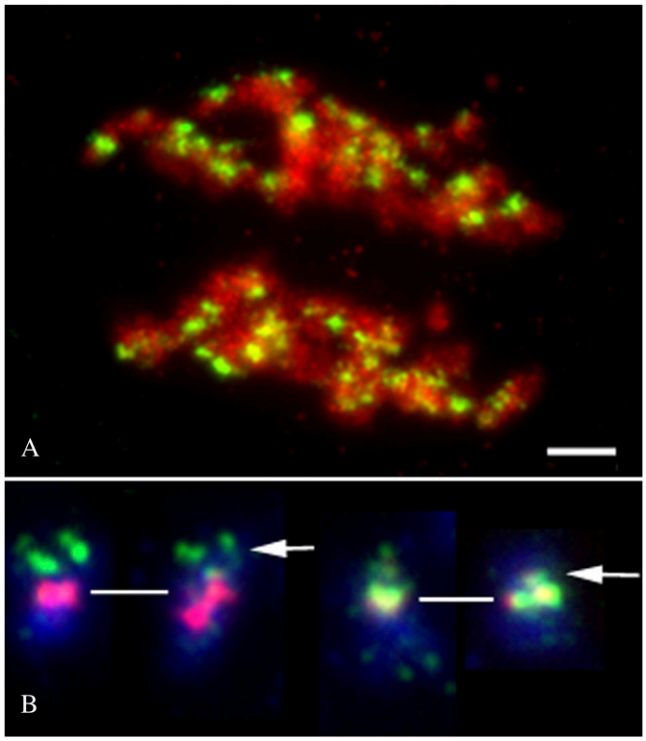

Figure 2. FISH analysis of CentGm-1 repeats during mitosis and CentGm-1 and CentGm-2 repeats on homologous chromosomes. A, CentGm-1 localized to leading edges of chromosomes at mitotic anaphase. CentGm-1 labeled with biotin and detected with AF488 streptavidin (green channel) and chromosomes were counterstained with propidium iodide (red channel); bar = 5 μm. B, Analysis of the distribution of the two centromeric repeats in a pair of homologous chromosomes in soybean. BACs 161O23 and 38L01 labeled with biotin and detected with AF488 streptavidin (green channel, arrows) that cross hybridize to duplicated regions in two homologs (Walling et al., 2005) were used in conjunction with CentGm-1 (labeled with digoxigenin and detected with rhodamine anti-digoxigenin, red channel) and CentGm-2 (labeled and detected with both fluorophores to result in yellow). One pair of homologs is marked by CentGm-1 (linkage group E), while the other pair is marked by CentGm-2. Lines show positions of centromeres between homologous chromosomes.

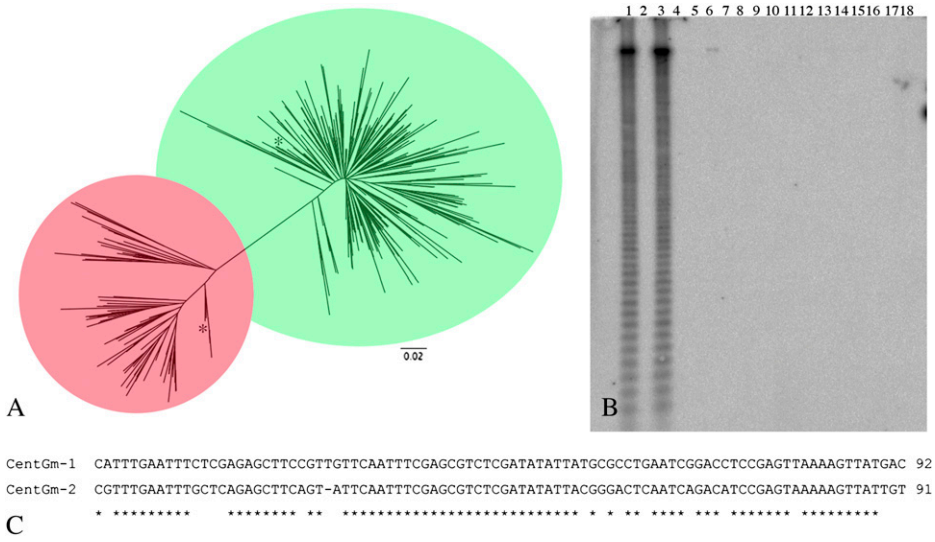

Figure 3. Evolution of CentGm repeats. A, Neighbor-joining tree of 464 CentGm-2/CentGm-1 repeats derived from screening the genome shotgun sequence library with both repeats (60% seqid, 80% length). One-hundred twenty-seven sequences clustered into CentGm-2-like repeats (red shading) and 337 into CentGm-1-like repeats (green shading). Sequences used for probe synthesis are shown with asterisks in either cluster. B, Southern analysis of CentGm-1 (same monomer shown with asterisk in section A) in a set of *Glycine* species: 1, soybean; 2, *Vigna radiata*; 3, *G. soja*; 4, *Glycine arenaria*; 5, *Glycine latifolia*; 6, *G. pescadrensis*; 7, *Glycine rubiginosa*; 8, *Glycine cyrtaloba*; 9, *Glycine tomentella*; 10, *Glycine curvata*; 11, *Glycine stenophita*; 12, *Glycine latrobeana*; 13, *Glycine canescens*; 14, *Glycine clandestina*; 15, *Glycine pindanica*; 16, *Glycine argyrea*; 17, *Glycine falcata*; and 18, *Glycine tabacina*. C, Sequence alignment of the two representative CentGm-2 and CentGm-1 repeats (shown with asterisks in section A).

The Organization of Centromeric Repeat Arrays in Soybean

The observation that two pairs of soybean centromeres contain both CentGm-1 and CentGm-2 was intriguing. To further shed light on the organization of centromeric satellite arrays, we conducted three independent fiber-FISH experiments with two slides each using differentially labeled CentGm-1 and CentGm-2 sequences as probes. As illustrated in Figure 1F, all large fiber segments examined (e.g. more than a megabase in size) are composed of either CentGm-1 or CentGm-2, suggesting an absence of extensive rearrangement and reshuffling of CentGm-1 and CentGm-2 within centromeres. This parallels the observation that no WGS clones screened in this study contained both CentGm-1 and CentGm-2 sequences.

We performed computational analyses on five shotgun sequencing clones that contained CentGm sequences on both ends of the clone (three clones with CentGm-1 and two with CentGm-2 on both ends) to determine the level of monomer variation within defined regions (regions the size of a shotgun clone, approximately 4 kb). The amount of variation between monomers from either end of a clone, using the Kimura model, mirrored the variation seen within monomer clades described previously. For CentGm-1 the overall mean distances ranged from 0.096 to 0.136 and for CentGm-2 the two comparisons were 0.076 and 0.096.

Rapid Divergence of Centromeric Satellite Repeats within the *Glycine* Genus

DNA from a representative set of *Glycine* species was Southern blotted and probed with both centromeric repeats. Both CentGm-1 and CentGm-2 (data not shown) hybridized only to genomic DNA of the cultigen soybean and its intercompatible annual relative and immediate progenitor, *Glycine soja* (Fig. 3B). It is most likely that the centromeric repeats are arranged in the same patterns in *G. soja* as in soybean, as seen in 92/91-bp ladders on Southern blots (CentGm-1 shown in Fig. 3A, lane 1). Ladders arise from tandemly arrayed repeats that either do not completely digest or have sequence mutations such that monomers, dimers, and higher multimers are seen on the Southern. The absence of hybridization to CentGm-1 and CentGm-2 in the other relatives suggests the rapid divergence of centromeric satellite repeats within *Glycine*, as seen in *Oryza* (Lee et al., 2005). The faint band seen in *Glycine pescadrensis* (Fig. 3B, lane 6) is likely due to nonspecific hybridization of high M_r DNA, since the same signal is seen with other noncentromeric DNA probes (data not shown).

DISCUSSION

The chromosome number of soybean and other diploid *Glycine* species is $2n = 40$, which is doubled

relative to its phaseoloid legume relatives (e.g. *Phaseolus*, *Vigna*), most of which are $2n = 20$ or $2n = 22$. Other lines of evidence indicate that the soybean genome has undergone two large-scale genome-wide duplication or polyploidization events (Shoemaker et al., 1996; Blanc and Wolfe, 2004; Pagel et al., 2004; Schlueter et al., 2004, 2006; Walling et al., 2005; Innes et al., 2008). Phylogenetic analyses of homologous gene pairs have shown that the older of the two putative polyploid duplications is shared by *Glycine*, *Medicago*, and *Lotus*, and therefore occurred no later than in their common ancestor (Pfeil et al., 2005; Cannon et al., 2006), which is estimated from chloroplast phylogenies to have existed around 54 million years ago (MYA; Lavin et al., 2005). This date is more consistent with a 46 MYA estimate (Schlueter et al., 2004) than a 16 to 17 MYA estimate (Blanc and Wolfe, 2004), and suggests that the dates of 10 to 15 MYA for the more recent duplication (Schlueter et al., 2004) is probably a closer estimate than the 3 to 5 MYA estimate (Blanc and Wolfe, 2004) as discussed by Shoemaker et al. (2006).

Regardless of the divergence dates of the progenitor genomes, an unresolved question is whether these polyploid events were fundamentally allo- or auto- in nature. Phylogenetic analysis can identify a pattern consistent with allopolyploidy, in which each homolog is most closely related to a different progenitor ortholog. No such pattern has been observed in gene phylogenies of *Glycine* (Straub et al., 2006). This pattern could be due to autopolyploidy, or, as we suggest is more likely, to the extinction of diploid progenitors (Fig. 4); in the latter case, gene phylogenies simply cannot resolve the issue.

In this study, we employed a novel approach to address this question and found two soybean centromere-specific satellite repeat classes that have mostly nonoverlapping distributions. The presence of two different centromeric repeat classes in the soybean genome suggests the existence of two subgenomes, which were already differentiated from one another cytologically, that were brought together by hybridization. In plants, such fixed hybridity defines genetic allopolyploidy, in which homologous (Huskins, 1932) chromosomes generally do not pair, and thus show disomic segregation. In contrast, genetic autopolyploids possess chromosomes similar enough to pair at meiosis, whether as multivalents or as random pairs of bivalents, and thus show polysomic segregation. This permits continued interaction of homologous sequences across all parental chromosomes, and can lead to recombination and segregational loss of parental sequences. However, even in allopolyploids, homologous sequences can continue to interact: many repeat families show homogenization across nonhomologous chromosomes (concerted evolution, e.g. through ectopic gene conversion). Indeed centromeric repeat families are a classic example of this phenomenon (Alexandrov et al., 1988; King et al., 1995; Galian and Vogler, 2003; Hall et al., 2005), and the similarity of repeats within the CentH-1 and CentH-2 classes, two

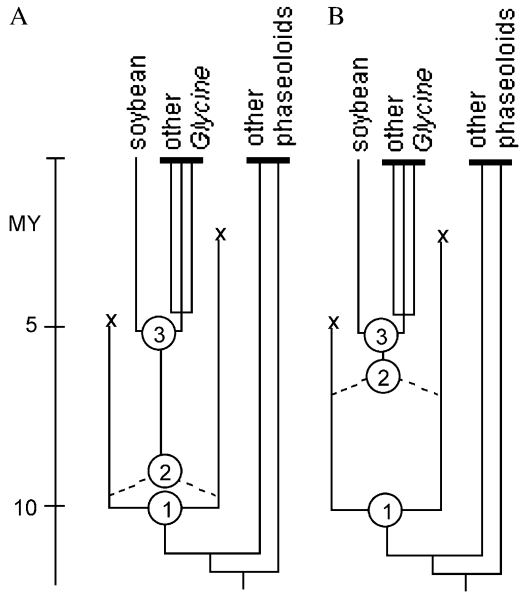

Figure 4. Polyploid evolution in *Glycine*. The diploid progenitor genomes of *Glycine* diverged at point 1, which was followed by polyploidy at point 2, which led to the modern chromosome number of $2n = 40$ in diploid *Glycine* species, including soybean. The divergence of these now-extinct x progenitors occurred about twice as long ago as the divergence between the soybean lineage and the lineage of perennial *Glycine* species. The polyploidy event could have occurred anywhere between points 1 and 3. In A, polyploidy occurred very close to the time of divergence of progenitor genomes, possibly within a single species (taxonomic autopolyploidy); close relationship of progenitor genomes could permit pairing among their chromosomes (genetic autopolyploidy) and could lead to segregational loss of parental sequences or homogenization of repeats across parental genomes. In B, polyploidy occurred considerably later than the divergence of parental genomes; parents likely would be differentiated genetically and taxonomically when hybridization occurred (taxonomic allopolyploidy), also leading to disomic pairing (genetic allopolyploidy). MY, Million years.

centromere satellites of allopolyploid *Arabidopsis suecica*, indicates that these repeats can interact across nonhomologous chromosomes.

The surprising finding in *Glycine* is the low level of recombination or homogenization of the two subfamilies of satellites, which have persisted in the same genome for at least 5 million years (Fig. 4). The homogenization of centromeric satellite repeats has been found to be a relatively rapid process. For instance, extensive rearrangement and reshuffling of CentO satellite repeats in rice centromeres has occurred within the last half-million years (Ma and Bennetzen, 2006; Ma and Jackson, 2006; Ma et al., 2007). In rice and Arabidopsis, both of which are believed to be paleoployploids or extensive paleoaneuploids (Vision et al., 2000; Vandepoele et al., 2003), only single families of centromeric satellite repeats, e.g. CentO in rice (Dong et al., 1998) and pAL1 in Arabidopsis (Kamm et al., 1995), were found in respective genomes. Even in maize, an allotetraploid

formed about 5 MYA from two diploid progenitors that diverged from a common diploid ancestor about 5 to 12 MYA (Gaut and Doebley, 1997; Swigonova et al., 2004), only CentC centromeric satellite repeats were identified and found in all centromeric regions of maize (Ananiev et al., 1998; Jiang et al., 2003). This indicates that the centromeric satellite repeats from the two diploid progenitors of maize, which had evolved independently for about 7 million years before their reunion, have been highly homogenized in the maize genome, probably by conversion, intercentromeric DNA exchanges, and numerous intracentromeric rearrangements (Ma and Bennetzen, 2006; Ma and Jackson, 2006).

The fact that the CentGm-1 probe targets more centromeres than CentGm-2 is intriguing. Assuming that the CentGm-1 and CentGm-2 donor genomes contributed equal numbers of chromosomes upon the formation of the polyploid genome and subsequent normal cell division, we would not expect unequal numbers of chromosomes carrying the two satellite repeats. It is possible that the CentGm-1 progenitor had more chromosomes than the CentGm-2 progenitor, although the real scenario cannot be revealed based solely on our current data. This is not a unique observation, since it was observed in several Arabidopsis species that there were multiple centromeric repeats that marked unequal subsets of centromeres (Kawabe and Nasuda, 2006).

The most likely explanation for the unequal distribution of centromere types in soybean is that paleopolyploid *Glycine* originated as a cross between two now presumed to be extinct $2n = 20$ plants (Fig. 4), followed by partial homogenization of one centromeric repeat class by the other. The presence of both repeat classes in two pairs of centromeres suggests that the formation of chimeric tandem repeats, rather than homogenization, has also been the outcome of interactions between chromosomes bearing different repeat classes.

Given the rapid evolution of species-specific centromeric heterochromatin repeat sequences (Lee et al., 2005), it is not initially surprising that there was no hybridization of soybean repeats to genomic DNA of perennial soybean species. However, this observation must be reconciled with the even longer-term maintenance of two repeat classes in the soybean genome. The perennial *Glycine* species, like soybean, are $2n = 40$ (one species is $2n = 38$), and share a common ancestor with the soybean lineage roughly 5 MYA, based on silent site divergences of numerous nuclear genes (Innes et al., 2008). Thus, these species should possess centromeric repeats homologous with CentGm-1 and CentGm-2. Assuming clock-like rates of evolution, their CentGm-1-like repeats should be approximately only half as diverged from CentGm-1 as CentGm-1 and CentGm-2 are from each other. The same should be true of their CentGm-2-like repeats. Sequences roughly 90% similar should have been detectable by genomic Southern hybridization, but no hybridization was observed. This could be explained by homogenization between CentGm-1 and CentGm-2 throughout the history of soybean, though at a lower level than homogenization within either repeat family. This could cause both sequence families to diverge concertedly from homologous CentGm-like sequences in perennial relatives. This hypothesis predicts that perennial *Glycine* species may also possess two clusters of centromeric repeats that would be more similar to one another than to CentGm-1 and CentGm-2.

Because it is unknown how much interaction has occurred between CentGm-1 and CentGm-2 since they were brought together in the same genome, we cannot determine how divergent these repeats were at the time of hybridization. If they were already well differentiated (as suggested by their persistence as separate groups), this would be most consistent with the progenitors belonging to different species, suggesting taxonomic allopolyploidy. The unequal number of centromeres bearing the different repeat types would then be part of the rearrangement process that has led to the scrambling of the soybean genome such that homologous regions are scattered among different chromosomes, consistent with other mapping in soybean (Schlueter et al., 2006; Shoemaker et al., 2006). However, the unequal number of homologous centromere classes could also be attributable to random segregational loss of parental chromosomes during a period of tetrasomic chromosome association in a genetic autopolyploid, or the newly formed polyploid *Glycine* could have been a segmental allopolyploid, with attributes of both auto- and allopolyploidy at different loci. One explanation for the observed centromeric structure in soybean would be some type of allopolyploidy followed by diploidization leading to a present day pseudodiploid, similar to what was recently hypothesized for *Boechera holboellii* (Kantama et al., 2007). Ultimately, the exact mode of origin and thus the formal distinction is less important than the observation that homologous variation currently exists in the *Glycine* genome, and that the soybean is a fixed polyploid hybrid for centromeres, as it is for many other loci.

MATERIALS AND METHODS

Identification of Centromeric Repeats from Genome Shotgun Sequence Data

A total of 25,082 soybean (*Glycine max*) shotgun sequences comprising approximately 11.4 MB were used for the de novo identification of repeats using RECON (Bao and Eddy, 2002). The most frequently occurring families from the RECON output were selected and annotated using BLASTN ($e = 10^{-4}$) with the nonredundant database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). Families with satellite repeats were selected, and all the satellite repeats were combined into a single family. These satellite sequences were then assembled into contigs using PHRAP (http://www.phrap.org/phredphrapconsed.html). The contigs were searched for tandem repeats using the program Tandem Repeats Finder (Benson, 1999). High-copy number repeats were selected from the Tandem Repeats Finder output.

FISH of Centromeric Repeats

Chromosome preparations and FISH followed published procedures (Kato et al., 2004), with the following modifications. Soybean seeds (cv Williams 82) were surfaced sterilized overnight and then germinated for 3 d at 24°C in petri dishes on germination medium. Seedling roots were excised and subjected to pressurized nitrous oxide treatment (Kato, 1999) at room temperature for 55 min to induce mitotic arrest. Roots were then quickly fixed in ice-cold 90% acetic acid for 10 min and washed six times with distilled water. The terminal 2 to 3 mm of each root tip was then excised and individually transferred to microfuge tubes containing 20 μL of enzyme solution containing 1% Pectolyase Y-23 (MP Biomedicals) and 2% Onozuka R-10 cellulase (Research Products International) in citric buffer (10 mM sodium citrate, 10 mM sodium EDTA, pH 5.5) and digested for 55 min in a 37°C water bath. Tubes were transferred to wet ice and washed twice with 70% ethyl alcohol. Root tips were then macerated in the residual approximately 100 μL of liquid using a dissecting probe and then centrifuged at 2,000 relative centrifugal force for 4 s and the supernatant poured off onto a paper towel. Each cell pellet was then resuspended in 35 μL of room temperature glacial acetic acid. To prepare individual slides for hybridization, 5 μL of cell suspension was applied to each glass slide (Gold Seal) in a moistened paper towel-covered humid chamber. Once the solvent had evaporated (5–10 min), slides were to UV crosslinked under optimal crosslink setting in a model XL-1000 UV crosslinker (Spectronics Corporation). Next, 10 μL of blocking solution (2× SSC, 1× Tris-EDTA, 10 μg/μL sonicated salmon sperm DNA) and a flexible plastic coverslip (Fisher Scientific) were applied. Groups of slides were then denatured in a covered tray in a covered boiling water bath for 5 min and then transferred to a chilled metal plate on wet ice. Ten microliters of the following hybridization solution was then applied per slide: 2× SSC, 1× Tris-EDTA, and 0.2 ng/μL (2 ng total) of each fluorescently labeled DNA oligonucleotide (Integrated DNA Technologies). The following oligonucleotide pairs were used in combination: Pr91-C:AGTAAAAAGTTATTGTCGTTTGAATTT (CentGm-2) and Pr92-C:AGTTAAAAGTTATGACCATTTGAATTT (CentGm-1). The Pr91 oligonucleotide was labeled with red fluor (5′ TEX 615) and the Pr92 oligonucleotide was labeled with green fluor (5′ 6-FAM). Slides were hybridized overnight at 55°C in a humidified hybridization chamber. For posthybridization washes, slides were first dipped in room temperature 2× SSC to remove cover glasses and then washed for 5 min in 2× SSC at 55°C for 5 min in a 55°C air incubator. Wash solution was removed and chromosome spreads were covered with 10 μL of 4′,6-diamino-2-phenilinodole (DAPI) in Vectashield mounting medium (Vector Laboratories) and a glass coverslip.

Images were collected on an Olympus BX61 microscope using Applied Spectral Imaging (Vistas) software and a COOL-1300QS CCD camera (VDS Vosskühler). Raw TIF format image files were imported into Adobe Photoshop CS2; the resolution was then increased from 72 to 200 dpi. Images were converted from 8- to 16-bit mode. Next, using the levels menu, cytoplasmic background was subtracted using the Set Black Point tool, and images were converted back to 8-bit mode. To construct the final images, the blue (DAPI) channel was removed and replaced with 100% black fill and then reintroduced as a separate grayscale layer that was then set to 25% opacity, thereby converting chromosomes from blue to gray.

For meiotic chromosome and DNA fiber FISH, plants (cv Williams 82) were grown under standard greenhouse conditions (16-h daylength and 27°C daytime temperature). Florets were collected for meiotic chromosome preparations according to Walling et al. (2005) and stored in Carnoy's solution at 4°C until used to prepare chromosome spreads. Nuclei extraction and fiber FISH were performed as described previously (Jackson et al., 1998).

Plasmid clones of CentGm-1 and CentGm-2 were purified using Qiagen miniprep kit according to manufacturer's instructions. Approximately one microgram of purified plasmid DNA was labeled with either digoxigenin or biotin using Nick translation kits (Roche). The DNA labeling reaction was kept at 15°C for 2 h and then cleaned using Qiagen PCR columns.

FISH of plasmid clones on DNA fibers (fiber-FISH) was performed as previously described (Jackson et al., 1998). AlexaFluor (AF) 488 streptavidin (Invitrogen) detected the biotin label. This signal was amplified by layering goat anti-streptavidin conjugated with biotin (Vector Laboratories) followed by a second application of AF488 streptavidin. Digoxigenin labels were detected using mouse anti-digoxigenin (Roche) followed by AF568 anti-mouse (Invitrogen). Mitotic and meiotic chromosome FISH was performed as previously described (Jiang et al., 1995). Biotin-labeled probes were detected using a single layer of AF488 streptavidin, and the digoxigenin-labeled probes were detected using a single layer of sheep anti-digoxigenin conjugated with rhodamine (Roche). Posthybridization formamide treatments and stringency washes were the same. FISH images were captured using a Photometrics Cool Snap HG camera attached to a Nikon Eclipse 80i fluorescence microscope. Images were adjusted and analyzed using Metamorph (Universal Imaging). Further cropping and labeling of images was performed using Adobe Photoshop CS v. 8.0 for Macintosh.

Computational Analysis of Centromeric Repeats

CentGm-1 and CentGm-2 were used to query the soybean genome shotgun sequences using BLASTN at default parameters to get 1,337 and 1,260 hits, respectively. These hits were parsed using cutoffs of 80% length and 60% sequence identity, which reduced the hit number to 1,001 and 712, respectively. These criteria were chosen so as to retain potentially diverged sequences. Since CentGm-1 and CentGm-2 share approximately 80% similarity with each other, we expected both sets of sequences to be represented at this stringency. Sequences corresponding to these hits were extracted using custom PERL scripts. Since the minimum length was 80%, some of the hits were truncated, and we therefore chose a random set of 464 nontruncated sequences. The 464 sequences were aligned using ClustalX (Thompson et al., 1994) using default options. The alignment was manually edited in Jalview (Clamp et al., 2004) and a neighbor-joining (Saitou and Nei, 1987) tree was constructed from this alignment using a 1,000 replicate bootstrap analysis (Kumar et al., 2004). FigTree version1.2.2 (http://tree.bio.ed.ac.uk/software/figtree/) was used to draw and view the unrooted neighbor-joining tree.

To determine between the two clusters, a single consensus sequence for each cluster was created using multiple sequence alignments from the program Emma (http://www.hku.hk/bruhk/emboss/emma.html), an interface to the ClustalW program. The consensus sequences were then manually curated and aligned to each other using MEGA (Kumar et al. 2004). Distance between the two clusters was calculated using the Kimura-2 parameter model and complete deletion option.

Southern Analysis

Plant genomic DNA was extracted from young leaf tissue using a standard cetyl trimethyl ammonium bromide extraction protocol. For each species, 1 μg of plant genomic DNA was restriction digested with 6 units of *Hin*dIII (New England Biolabs) in a 37°C water bath overnight and separated on a 0.8% agarose gel. DNA from the gels was blotted onto Zeta-Probe GT genomic tested blotting membrane (Bio-Rad). The membrane was prehybridized for at least 30 min in Church hybridization buffer (1% bovine serum albumin/1 mM EDTA/7% SDS/0.5 M sodium phosphate) at 58°C. Probes were prepared using the Rediprime II random prime labeling system (Amersham Biosciences). Before the probes were used for hybridization, they were purified using the QIAquick nucleotide removal kit (Qiagen). The probe was hybridized to the membrane at 58°C overnight. After hybridization, the membrane was washed in 1.5× SSC/0.1% SDS for 30 min at 58°C, then in 1× SSC/0.1% SDS for 30 min. The membrane was exposed overnight to a Fujifilm BAS-MS imaging plate and digitally scanned using a Fuji FLA-5000 bio imaging analyzer.

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession numbers CL866971 to CL866979, CL866987 to CL868441, CL876820 to CL876828, CL877016 to CL884613, and CL884615 to CL900625.

ACKNOWLEDGMENTS

We thank Zhanyuan Zhang (University of Missouri Plant Transformation Core Facility) for use of facilities, James H. Birchler (Division of Biological Sciences, University of Missouri), and two anonymous reviewers.

Received March 2, 2009; accepted July 9, 2009; published July 15, 2009.

LITERATURE CITED

Alexandrov IA, Mitkevich SP, Yurov YB (1988) The phylogeny of human chromosome specific alpha satellites. Chromosoma **96:** 443–453

Ananiev EV, Phillips RL, Rines HW (1998) Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions. Proc Natl Acad Sci USA **95:** 13073–13078

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res **12:** 1269–1276

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27:** 573–580

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667–1678

Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T, et al (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. Proc Natl Acad Sci USA **103:** 14959–14964

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics **20:** 426–427

Dong F, Miller JT, Jackson SA, Wang GL, Ronald PC, Jiang J (1998) Rice (Oryza sativa) centromeric regions consist of complex DNA. Proc Natl Acad Sci USA **95:** 8135–8140

Galian J, Vogler AP (2003) Evolutionary dynamics of a satellite DNA in the tiger beetle species pair Cicindela campestris and C. maroccana. Genome **46:** 213–223

Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci USA **94:** 6809–6814

Hall SE, Luo S, Hall AE, Preuss D (2005) Differential rates of local and global homogenization in centromere satellites from Arabidopsis relatives. Genetics **170:** 1913–1927

Huskins CL (1932) A cytological study of Vilmoren's unfixable dwarf wheat. Genetics **25:** 113–124

Innes RW, Ameline-Torregrosa C, Ashfield T, Cannon E, Cannon SB, Chacko B, Chen NW, Couloux A, Dalwani A, Denny R, et al (2008) Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. Plant Physiol **148:** 1740–1759

Jackson SA, Wang ML, Goodman HM, Jiang J (1998) Fiber-FISH analysis of repetitive DNA elements in Arabidopsis thaliana. Genome **41:** 566–572

Jiang J, Birchler JA, Parrott WA, Dawe RK (2003) A molecular view of plant centromeres. Trends Plant Sci **8:** 570–575

Jiang J, Gill BS, Wang GL, Ronald PC, Ward DC (1995) Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes. Proc Natl Acad Sci USA **92:** 4487–4491

Kamm A, Galasso I, Schmidt T, Heslop-Harrison JS (1995) Analysis of a repetitive DNA family from Arabidopsis arenosa and relationships between Arabidopsis species. Plant Mol Biol **27:** 853–862

Kantama L, Sharbel TF, Schranz ME, Mitchell-Olds T, de Vries S, de Jong H (2007) Diploid apomicts of the Boechera holboellii complex display large-scale chromosome substitutions and aberrant chromosomes. Proc Natl Acad Sci USA **104:** 14026–14031

Kato A (1999) Air drying method using nitrous oxide for chromosome counting in maize. Biotech Histochem **74:** 160–166

Kato A, Lamb JC, Birchler JA (2004) Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci USA **101:** 13554–13559

Kawabe A, Nasuda S (2006) Polymorphic chromosomal specificity of centromere satellite families in Arabidopsis halleri ssp. gemmifera. Genetica **126:** 335–342

King K, Jobst J, Hemleben V (1995) Differential homogenization and amplification of two satellite DNAs in the genus Cucurbita (Cucurbitaceae). J Mol Evol **41:** 996–1005

Kumar PS, Hymowitz T (1989) Where are the diploid (2n=2x=20) genome donors of Glycine Willd. (Leguminosae, Papilionoideae)? Euphytica **40:** 221–226

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform **5:** 150–163

Lackey JA (1980) Chromosome numbers in the Phaseoleae (Fabaceae: Faboideae) and their relationship to taxonomy. Am J Bot **67:** 595–602

Laten HM, Majumdar A, Gaucher EA (1998) SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci USA **95:** 6897–6902

Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. Syst Biol **54:** 575–594

Lee HR, Zhang W, Langdon T, Jin W, Yan H, Cheng Z, Jiang J (2005) Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. Proc Natl Acad Sci USA **102:** 11793–11798

Lin JY, Jacobus BH, SanMiguel P, Walling JG, Yuan Y, Shoemaker RC, Young ND, Jackson SA (2005) Pericentromeric regions of soybean (Glycine max L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics **170:** 1221–1230

Ma J, Bennetzen JL (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc Natl Acad Sci USA **103:** 383–388

Ma J, Jackson SA (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. Genome Res **16:** 251–259

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends Genet **123:** 134–139

Martinez-Zapater JM, Estelle MA, Somerville CR (1986) A high repeated DNA sequence in Arabidopsis thaliana. Mol Gen Genet **204:** 417–423

Morgante M, Jurman I, Shi L, Zhu T, Keim P, Rafalski JA (1997) The STR120 satellite DNA of soybean: organization, evolution and chromosomal specificity. Chromosome Res **5:** 363–373

Murata M, Ogura Y, Motoyoshi F (1994) Centromeric repetitive sequences in Arabidopsis thaliana. Jpn J Genet **69:** 361–370

Pagel J, Walling JG, Young ND, Shoemaker RC, Jackson SA (2004) Segmental duplications within the Glycine max genome revealed by fluorescence in situ hybridization of bacterial artificial chromosomes. Genome **47:** 764–768

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA **101:** 9903–9908

Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst Biol **54:** 441–454

Pontes O, Neves N, Silva M, Lewis MS, Madlung A, Comai L, Viegas W, Pikaard CS (2004) Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid Arabidopsis suecica genome. Proc Natl Acad Sci USA **101:** 18240–18245

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:** 406–425

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome **47:** 868–876

Schlueter JA, Scheffler B, Schlueter SD, Shoemaker RC (2006) Sequence conservation of homeologous BACs and expression of homeologous genes in soybean (Glycine max L Merr). Genetics **174:** 1017–1028

Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, et al (1996) Genome duplication in soybean (Glycine subgenus soja). Genetics **144:** 329–338

Shoemaker RC, Schlueter J, Doyle JJ (2006) Paleopolyploidy and gene duplication in soybean and other legumes. Curr Opin Plant Biol **9:** 104–109

Straub SCK, Pfeil BE, Doyle JJ (2006) Testing the polyploid past of soybean using a low-copy nuclear gene—is glycine (Fabaceae: Papilionoideae) an auto- or allopolyploid? Mol Phylogenet Evol **39:** 580–584

Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen J, Messing J (2004) On the tetraploid origin of the maize genome. Comp Funct Genomics **5:** 281–284

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673–4680

Udall JA, Wendel JF (2006) Polyploidy and crop improvement. Crop Sci **46:** S-3–S-14

Vahedian M, Shi L, Zhu T, Okimotot R, Danna K, Keim P (1995) Genomic organization and evolution of the soybean SB92 satellite sequence. Plant Mol Biol **29:** 857–862

Vandepoele K, Simillion C, Van de Peer Y (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell **15:** 2192–2202

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

Walling JG, Shoemaker RC, Young ND, Mudge J, Jackson SA (2005) Chromosome level homeology in paleopolyploid soybean (Glycine max) revealed through integration of genetic and chromosome maps. Genetics **172:** 1893–1900

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol **42:** 225–249

Willard HF (1985) Chromosome-specific organization of human alpha satellite DNA. Am J Hum Genet **37:** 524–532