Replication of Nonautonomous Retroelements in Soybean Appears to Be Both Recent and Common[1][W][OA]

Adam Wawrzynski, Tom Ashfield, Nicolas W.G. Chen, Jafar Mammadov[2], Ashley Nguyen, Ram Podicheti, Steven B. Cannon, Vincent Thareau, Carine Ameline-Torregrosa, Ethalinda Cannon, Ben Chacko, Arnaud Couloux, Anita Dalwani, Roxanne Denny, Shweta Deshpande, Ashley N. Egan, Natasha Glover, Stacy Howell, Dan Ilut, Hongshing Lai, Sara Martin del Campo, Michelle Metcalf, Majesta O'Bleness, Bernard E. Pfeil, Milind B. Ratnaparkhe, Sylvie Samain, Iryna Sanders, Béatrice Ségurens, Mireille Sévignac, Sue Sherman-Broyles, Dominic M. Tucker, Jing Yi, Jeff J. Doyle, Valérie Geffroy, Bruce A. Roe, M.A. Saghai Maroof, Nevin D. Young, and Roger W. Innes*

Department of Biology, Indiana University, Bloomington, Indiana 47405 (A.W., T.A., R.P., A.D., S.H., S.M.d.C., M.M., R.W.I.); Institut de Biotechnologie des Plantes, UMR CNRS 8618, INRA, Université Paris Sud, 91 405 Orsay, France (N.W.G.C., V.T., M.S., V.G.); Department of Crop and Soil Environmental Sciences, Virginia Tech, Blacksburg, Virginia 24061 (J.M., A.N., N.G., M.B.R., D.M.T., M.A.S.M.); Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota 55108 (S.B.C., C.A.-T., E.C., B.C., R.D., N.D.Y.); U.S. Department of Agriculture-Agricultural Research Service and Department of Agronomy (S.B.C.), and Virtual Reality Application Center (E.C.), Iowa State University, Ames, Iowa 50011; Genoscope/CEA-Centre National de Séquençage, 91 057 Evry, France (A.C., S.S., B.S.); Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019 (S.D., H.L., M.O., I.S., J.Y., B.A.R.); L.H. Bailey Hortorium, Department of Plant Biology, Cornell University, Ithaca, New York 14853 (A.N.E., D.I., B.E.P., S.S.-B., J.J.D.); CSIRO Plant Industry, Canberra, Australian Capital Territory 2601, Australia (B.E.P.); and Division of Plant Sciences, University of Missouri, Columbia, Missouri 65211 (M.B.R.)

Retrotransposons and their remnants often constitute more than 50% of higher plant genomes. Although extensively studied in monocot crops such as maize (*Zea mays*) and rice (*Oryza sativa*), the impact of retrotransposons on dicot crop genomes is not well documented. Here, we present an analysis of retrotransposons in soybean (*Glycine max*). Analysis of approximately 3.7 megabases (Mb) of genomic sequence, including 0.87 Mb of pericentromeric sequence, uncovered 45 intact long terminal repeat (LTR)-retrotransposons. The ratio of intact elements to solo LTRs was 8:1, one of the highest reported to date in plants, suggesting that removal of retrotransposons by homologous recombination between LTRs is occurring more slowly in soybean than in previously characterized plant species. Analysis of paired LTR sequences uncovered a low frequency of deletions relative to base substitutions, indicating that removal of retrotransposon sequences by illegitimate recombination is also operating more slowly. Significantly, we identified three subfamilies of nonautonomous elements that have replicated in the recent past, suggesting that retrotransposition can be catalyzed in trans by autonomous elements elsewhere in the genome. Analysis of 1.6 Mb of sequence from *Glycine tomentella*, a wild perennial relative of soybean, uncovered 23 intact retroelements, two of which had accumulated no mutations in their LTRs, indicating very recent insertion. A similar pattern was found in 0.94 Mb of sequence from *Phaseolus vulgaris* (common bean). Thus, autonomous and nonautonomous retrotransposons appear to be both abundant and active in *Glycine* and *Phaseolus*. The impact of nonautonomous retrotransposon replication on genome size appears to be much greater than previously appreciated.

[1] This work was supported by the National Science Foundation (Plant Genome Research Program grant no. DBI–0321664 to R.W.I., M.A.S.M., N.D.Y., B.A.R., and J.J.D. and Systematics award no. DEB–0516673 to A.N.E.) and by Genoscope/CEA-Centre National de Séquençage (grant to V.G.).

[2] Present address: Trait Genetics and Technology, Dow AgroSciences LLC, Indianapolis, IN 46268.

* Corresponding author; e-mail rinnes@indiana.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: Roger W. Innes (rinnes@indiana.edu).

[W] The online version of this article contains Web-only data.

[OA] Open access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.108.127910

Transposable elements are abundant components of plant genomes. They are typically divided into two groups based on their mechanism of transposition. Class I transposons transpose via an RNA intermediate and must therefore use reverse transcriptase (RT) during the replication process. Class II transposons do not have an RNA intermediate and usually use a cut-and-paste mechanism for transposition (Wicker et al., 2007). Elements of both classes have had major impacts on genome structure, apparently not only promoting mutations of genes and affecting gene regulatory sequences but also playing a substantial role in the creation of new genes by "exon-shuffling" and retrotransposition (Jin and Bennetzen, 1994; Jiang et al., 2004; Bennetzen, 2005;

Morgante et al., 2005; Zabala and Vodkin, 2005; Wang et al., 2006).

Retrotransposons and their remnants often constitute more than 50% of higher plant genomes and can be as high as 90% (Bennetzen et al., 2005; Sabot and Schulman, 2006). Because the majority of such elements appear to have inserted in the last few million years, it was once believed that there had been a relatively recent burst in retrotransposon activity that led to a recent expansion in plant genome sizes. However, it is now clear that genome expansion resulting from retrotransposon activity is counteracted by spontaneous deletions resulting from unequal homologous recombination and illegitimate recombination events (Ma et al., 2004; Bennetzen, 2005; Vitte and Bennetzen, 2006). Plant genomes appear to differ not only in the content of their repetitive fraction but in the dynamics of DNA removal as well. The latter may be estimated by examining the ratio of intact and possibly active elements to their fragmented or recombined counterparts.

The specific families of retrotransposons present in different plant species and their relative abundance varies tremendously, indicating that they are rapidly evolving and may undergo bursts of activity. In addition, most elements are represented by both autonomous (full-length elements encoding all proteins necessary for transposition) and nonautonomous (mutated elements lacking one or more proteins required for transposition) versions in the same genome, with both types varying even among individuals of the same species. These observations, combined with the presence of retrotransposon-derived mRNA, indicate that many elements are still active. Because retroelement sequences decay at a rapid rate, it can be difficult to identify and properly annotate their positions, especially using automated tools. This has led to frequent overestimation of genic sequences in genome annotations (Bennetzen et al., 2004). Because of their impact on genome size and structure, however, proper annotation of retrotransposon-derived sequences in genomes is especially important in terms of studying genome-wide mechanisms of sequence evolution.

As part of the National Science Foundation (NSF)-funded project Comparative Analysis of Legume Genome Evolution, we have generated approximately 4 megabases (Mb) of genomic sequence derived from two varieties of soybean (*Glycine max*), which we are comparing to orthologous regions of a wild perennial relative of soybean (*Glycine tomentella*) and to common bean (*Phaseolus vulgaris*; scientific names will be used for clarity; Innes et al., 2008). These comparisons have allowed us to estimate the impact of retrotransposons on *G. max* genome evolution. In addition, because the two *Glycine* species share a genome duplication event that occurred approximately 10 to 14 million years ago (Shoemaker et al., 2006), we were able to evaluate how duplicated regions differed in their subsequent retrotransposon activity and whether such differences were shared between these two species, which themselves diverged 5 to 7 million years ago (Innes et al., 2008).

RESULTS AND DISCUSSION

Strategy for Identifying Long Terminal Repeat-Retrotransposons

The majority of retrotransposons in plants and animals contain long terminal repeats (LTRs), which are generated during the transposition process. LTRs thus provide a convenient signature when searching genomic sequence for the presence of retrotransposons. We used a combination of publicly available programs that search for repeats, along with manual BLAST searches (Altschul et al., 1997) for homology to known retroelement sequences, to identify LTR-containing retrotransposons (see "Materials and Methods"). Approximately 3.7 Mb of *G. max* genomic sequence were searched, including 1 Mb from the *Rpg1-b* region on molecular linkage group F (homoeologue I [H1]) and 0.87 Mb of homoeologous sequence (H2) on molecular linkage group E. To sample other areas of the soybean genome, we also analyzed 1.85 Mb derived from bacterial artificial chromosome clones (BACs) not assigned to a particular location but available through the National Center for Biotechnology Information (NCBI) high-throughput genomic sequence database.

LTR-retrotransposons were classified as intact when they possessed two full-length LTRs flanked by target-site duplications (TSDs), a recognizable primer binding site, and a polypurine tract. Intact elements were additionally classified as autonomous if they contained intact *Gag* and *Pol* open reading frames (ORFs). *Gag* encodes the structural protein required for nucleocapsid formation, while *Pol* encodes a polyprotein containing an RT domain, an integrase domain, and an aspartic proteinase domain, which is responsible for posttranslational processing of the *Pol* ORF product. Intact elements lacking complete *Gag* and *Pol* ORFs were classified as nonautonomous. LTRs were classified as solo-LTRs when they contained sequence similarity to previously identified LTRs, appeared to be full length, were not associated with a second LTR, and were flanked by TSDs. Solo-LTRs are believed to arise by homologous recombination between LTRs of an individual element, resulting in deletion of the intervening retroelement sequence. All other elements with similarity to retrotransposon sequences, but judged not to be intact, were classified as remnants.

Glycine and *Phaseolus* Contain Many Retrotransposon Families with Recent Insertions

We identified 45 intact LTR-retrotransposons in *G. max*, 23 in *G. tomentella*, and seven in *P. vulgaris* (Table I; Supplemental Table S2). All LTR-transposons with recognizable *Gag-Pol* domains fell into two superfamilies, *Ty1/copia*-like and *Ty3/gypsy*-like, based on the order of the protein domains contained within the Pol polyprotein (Wicker et al., 2007). In *Ty1/copia*-like elements, the integrase domain appears N terminal to the RT domain, whereas in *Ty3/Gypsy*-like elements, the integrase domain appears after the RT domain (Fig. 1A).

Table I. *Summary of LTR-retrotransposon data from this study and from the indicated published data*[a]

nd, Not determined.

	No. of Intact Elements with TSDs	% Intact with TSDs	% Solo LTRs with TSDs	% Fragmented	No. of Intact Elements Dated	Insertion Date, Mya (Average)	Ts:Tv	(Ts+Tv):Indel
G. max[b]	45	55	7	38	45	1.3	2.41	12.81
G. tomentella[b]	23	52	7	41	22	0.8	2.56	13.7
P. vulgaris[b]	7	nd	nd	nd	7	0.8	1.89	nd
Arabidopsis[c,d]	nd	30	35	35	87	1.9	nd	nd
Rice[c,e]	nd	24	35	41	260	1.3	nd	nd
L. japonicus[c]	nd	57	10	33	13	0.3	2.5	7.3
M. truncatula[c]	nd	37	21	42	19	0.2	2.4	3.6
Maize[c]	nd	61	9	30	47	0.7	3.9	8.2
Barley[c]	nd	35	17	48	17	1.3	1.6	12.0
T. monococcum[c]	nd	50	10	40	30	1.0	1.9	8.7

[a]Percentages are relative to total number of elements identified that contain at least a partial LTR sequence with homology to an intact LTR-retrotransposon. Remnants lacking any LTR homology were not included in the calculations used in this table. Ts, Transitions; Tv, transversions; TSD, target site duplication. [b]Summary data taken from Supplemental Tables S2 and S3. [c]Data taken from tables 1 and 4 of Vitte and Bennetzen (2006). [d]Data taken from table 1 of Devos et al. (2002). [e]Data taken from table 2 of Ma et al. (2004).

We further classified the LTR-retrotransposons into 41 families based on their LTR sequences (Supplemental Table S2). We used LTR sequences to classify families rather than more commonly used RT sequences for two reasons. First, many nonautonomous retrotransposon sequences lack an intact RT domain. Second, RT domain sequences diverge at a slower rate than LTR sequences, making it difficult to distinguish more recently diverged families based on the RT domain alone. Following the guidelines for transposable element annotation proposed by Wicker et al. (2007), we grouped elements into the same family when their LTRs shared >80% identity across at least 80% of their length. Using these criteria, we grouped the *G. max* elements into 20 families (Supplemental Table S2). Only three of these families contained previously described *G. max* retrotransposons: *SIRE*-, *Diaspora*-, and *Calypso*-like elements (Wright and Voytas, 2002; Laten et al., 2003; Yano et al., 2005).

At the time of retrotransposon insertion, the two LTR sequences are identical. It is thus possible to estimate the time since insertion by aligning the two LTR sequences of each element and counting the number of nucleotide substitutions (see "Materials and Methods"). Eight of the 20 *G. max* families contained elements that appear to have inserted within the last one million years, and three elements, each from a different family, contained identical LTRs, indicating that the insertion events were very recent (Supplemental Table S2). In addition, we identified insertion events in cv Williams 82 that are absent from line PI 96983 and vice versa (Innes et al., 2008). Furthermore, we identified *G. max* EST sequences with over 90% DNA sequence identity to elements in 10 of the 20 *G. max* families (Supplemental Table S2). Of particular note are families 1 (SIRE-like), 9, and 10, all of which had EST matches with 97% or higher identity and insertions unique to cv Williams 82 or PI 96983. Thus, at least some elements in these families are being actively expressed and are likely generating new insertions.

In *G. tomentella*, we grouped the 23 intact elements into 16 families (Supplemental Table S2). Similar to *G. max*, nine of these 16 families contain elements that had inserted within the last million years, and two elements contained identical LTRs. We grouped the seven *P. vulgaris* elements into five families, including one element with identical LTRs and one that had inserted approximately 600,000 years ago. Thus, all three legume species characterized contain multiple retrotransposon families that have been active in the recent past.

Because of how rapidly LTR sequences diverge, it was not possible to align LTR sequences of elements from different families accurately and hence was not possible to construct phylogenetic trees based on the LTR sequences. We therefore used the RT domains to construct phylogenetic trees using Bayesian analyses (see "Materials and Methods"), splitting the *copia*-like and *gypsy*-like elements into separate trees (Fig. 2, A and B). The LTR-based families, indicated by shaded ovals in Figure 2, grouped together at the terminal branches of the RT trees, indicating that there has been little to no recombination between elements belonging to different RT clades. The *copia*-like elements exhibited

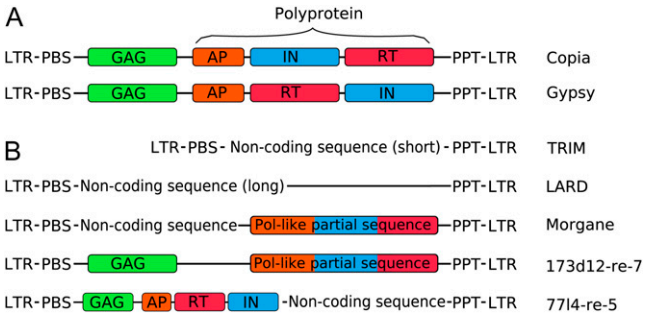

Figure 1. Nonautonomous LTR-retroelements are derived from autonomous elements. A, Structure of autonomous *copia*- and *gypsy*-like LTR-retroelements. B, Structure of nonautonomous derivatives of LTR-retrotransposons, including TRIMs, LARDs, Morganes, and examples from our study. AP, Asp protease; IN, integrase; PBS, primer binding site; PPT, poly-purine track.

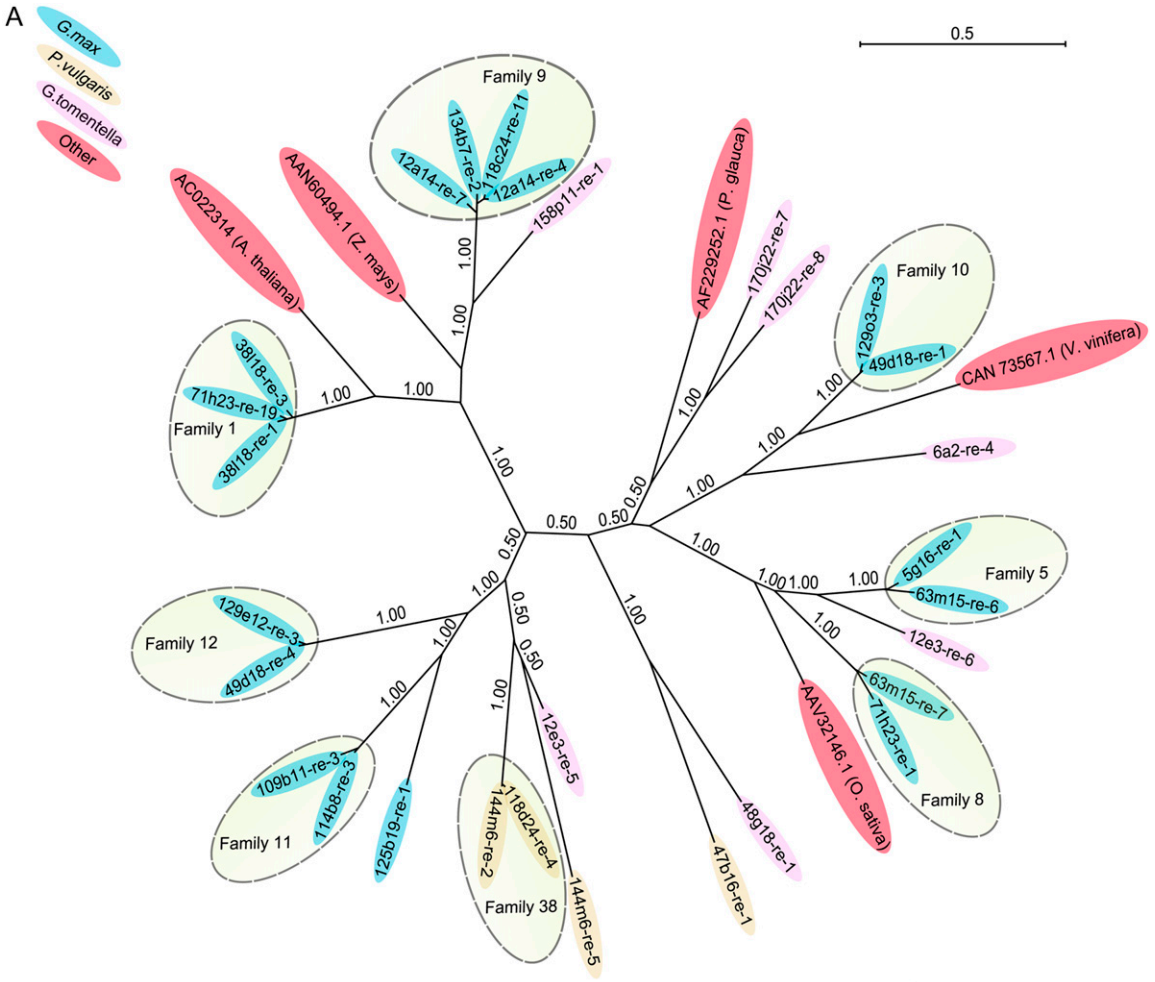

Figure 2. Phylogenetic trees showing the relationship of *G. max* LTR-retroelements to retroelements found in other plant species. A, Bayesian tree derived from the RT domains of *copia*-like LTR-retrotransposons. B, Bayesian tree derived from the RT domains of *gypsy*-like LTR-retrotransposons. Species of origin are indicated by color-coding and elements belonging to the same LTR family are indicated by shaded ovals. Elements not in shaded ovals belong to other families as indicated in Supplemental Table S2. Numbers indicate posterior probabilities, and the scale indicates nucleotide substitutions per site.

a high level of diversity in their RT domains, with no easily recognizable super-clades. In contrast, the *gypsy*-like elements formed three distinct clades. A recent analysis of retrotransposon content in the model legume *Medicago truncatula* uncovered similar patterns of diversity, with the Copia superfamily being significantly more diverse than the Gypsy superfamily (Wang and Liu, 2008). Similarly, an analysis of the retroelement content in garden pea (*Pisum sativum*) revealed a greater diversity of copia elements than gypsy elements (Macas et al., 2007), suggesting that this general pattern arose prior to the split between the *Glycine* lineage and the *Medicago/Pisum* lineage. Despite this similarity in pattern, comparison of abundant repeat families between *M. truncatula* and *G. max* found low levels of sequence similarity, with no major repeat families shared between the two species other than rDNA (Macas et al., 2007; Swaminathan et al., 2007).

To see how these legume retrotransposons were related to previously described retrotransposons from other plant species, we used representative RT sequences from divergent branches of each tree to search the NCBI nonredundant database for related RT sequences. Top hits were then added to the RT alignment and new trees constructed. As shown in Figure 2, A and B, these additional sequences were dispersed throughout the two RT trees, indicating that *Glycine* and *Phaseolus* contain a diversity of retrotransposons that are distributed widely among angiosperms. Assuming that these elements have not been transferred horizontally between species, it would suggest that at least some of these lineages predate the split between monocots and eudicots. This conclusion is supported by a recent phylogenetic analysis of copia elements from wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), rice (*Oryza sativa*), and Arabidopsis (*Arabidopsis thaliana*),

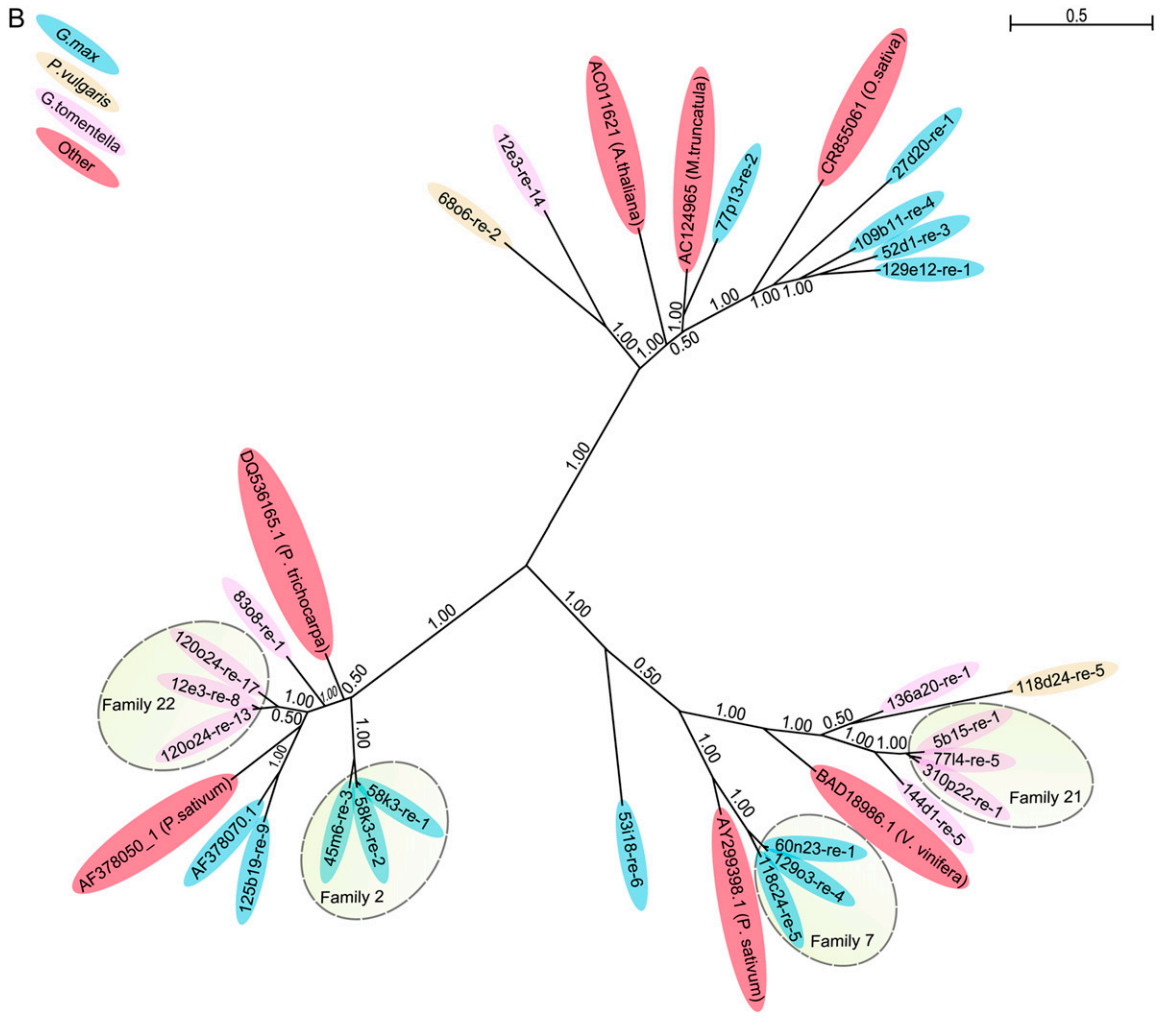

Figure 2. (*Continued.*)

which revealed the presence of six distinct copia lineages that predate the monocot-dicot split (Wicker and Keller, 2007).

Within one of the three clades of *gypsy*-like elements, we identified several that contained a chromatin organization modifier (pfam 00385, CHROMO) domain, which is a hallmark of the CHROMO domain-containing retrotransposons, also known as *Chromoviruses*. The CHROMO domain is part of the integrase domain and is located just upstream of the putative polypurine tract. It is thought to be involved in binding to methylated histone tails and/or to RNA (Nielsen et al., 2002). Four elements from *G. max* (129e12-re-1, 52d1-re-3, 109b11-re-4, and 77p13-re-2) contained a CHROMO domain. CHROMO domain retrotransposons are distributed widely among eukaryotes, and examples can be found in Arabidopsis, *Medicago*, and rice (Fig. 2B) as well as in animals and fungi, suggesting that they form an ancient family of gypsy-like elements (Marin and Llorens, 2000; Kordis, 2005).

Independent Increases in Retrotransposon Content in H2 in Both *G. max* and *G. tomentella*

As stated above, *G. max* and *G. tomentella* diverged approximately 5 to 7 million years ago (Innes et al., 2008) and share a whole-genome duplication event that occurred approximately 10 to 14 million years ago (Schlueter et al., 2004, 2006; Innes et al., 2008). At the time of the whole-genome duplication event, it is assumed that the resulting homoeologous chromosomes were very similar in terms of gene and retrotransposon content, particularly if *G. max* is derived from an autotetraploid event, as some current data suggest (Straub et al., 2006). Comparison of retrotransposon content in H1 to H2 in *G. max* revealed striking differences in content and number (Supplemental Table S2; Innes et al., 2008), with H2 containing many more insertions than H1. Significantly, we also observed this pattern in *G. tomentella*. The preferential accumulation of retrotransposons in H2 appears to have occurred independently in *G. max* and *G. tomen-*

tella, because the majority of the elements that we identified inserted after the speciation event that gave rise to these two species (Supplemental Table S2). These data indicate that H2 is more prone to retrotransposon accumulation than H1 in both species. Fluorescence in situ hybridization analyses in *G. max*, along with preliminary analyses of the *G. max* whole-genome shotgun sequence (Soybean Genome Project, DoE Joint Genome Institute; http://www.phytozome.net/soybean.php), indicate that the H2 region is located near a centromere, while the H1 region is not (Innes et al., 2008). These observations suggest that the H2 region has been translocated to a pericentromeric position, which may be promoting retrotransposon accumulation (Innes et al., 2008). It seems likely, therefore, that this translocation event occurred sometime after the divergence of H1 and H2, but before the divergence of *G. max* and *G. tomentella*, and thus predisposed H2 to retrotransposon accumulation in both species.

Relative Abundance of Intact Elements

We analyzed approximately 3.7 Mb of genomic sequence from *G. max* and identified 45 intact elements, which corresponds to an average density of 12.2 elements per Mb. We do not think this density is an overestimate, as only about 25% of the sequence analyzed came from known pericentromeric BACs (i.e. H2; Supplemental Table S2), while the soybean genome is thought to be made up of 40% to 60% repetitive DNA (Goldberg, 1978; Gurley et al., 1979; Swaminathan et al., 2007). Nevertheless, if we exclude the H2 sequence from the calculation, we identified 22 elements in 2.83 Mb, or an average of 7.8 intact elements per Mb. This average is still much higher than in *M. truncatula*, where only 2.3 elements per Mb were identified (Wang and Liu, 2008), and in rice, where the density across the whole genome was found to be 0.84 elements per Mb (Gao et al., 2004). It should be noted, however, that the *M. truncatula* value may be an underestimate, as it is derived from BAC sequences from the *M. truncatula* genome project, which is focused on gene-rich regions (Young et al., 2005).

The *Glycine* Genome Appears to Be Expanding

The remarkable variation in nuclear genome size of flowering plants is associated mainly with the size of the repetitive element fraction, especially LTR-retrotransposons (Bennetzen, 2005; Ammiraju et al., 2007). Increases in LTR-retrotransposon content are counterbalanced by internal genomic forces driving DNA removal, such as unequal crossing-over between homologous sequences and illegitimate recombination resulting from multiple mechanisms, including repairs of double-strand breaks (nonhomologous end-joining) and slipstrand mispairing. The rates of both genome expansion and genome contraction processes appear to vary between species (Devos et al., 2002; Ma et al., 2004; Bennetzen et al., 2005; Vitte and Bennetzen, 2006), allowing some genomes to shrink while others expand.

The rate of DNA removal caused by homologous recombination between LTR sequences of an individual element can be estimated by calculating the ratio of intact LTR-retroelements to solo-LTRs (Devos et al., 2002; Ma et al., 2004; Bennetzen et al., 2005; Vitte and Bennetzen, 2006). Analysis of our *G. max* and *G. tomentella* data revealed ratios of 8.0 and 7.7, respectively (Supplemental Table S3; *P. vulgaris* was not included in these calculations due to the low number of both intact retroelements and solo-LTRs identified in the sequenced regions). These ratios are much higher than those calculated for Arabidopsis (0.9), rice (0.7), and *Medicago* (1.8) and are similar to maize (*Zea mays*; 6.8) and *Lotus japonicus* (5.7; Vitte and Bennetzen, 2006). The low frequency of solo-LTRs compared to intact LTR-retroelements in our analysis indicates that homologous recombination between LTRs is not a major force driving removal of retroelement sequences in *Glycine*, at least in the regions analyzed.

DNA loss through illegitimate recombination is likely the stronger force driving DNA removal in plants (Devos et al., 2002; Ma et al., 2004; Grover et al., 2008). Such recombination events are typically associated with small deletions and can be detected by aligning LTR sequences from single elements. The relative frequency of these events can be estimated by comparing the ratio of base substitutions to insertion/deletion events, with higher ratios indicating lower rates of DNA removal via illegitimate recombination. *G. max* and *G. tomentella* had ratios of 12.8 and 13.7, respectively (Supplemental Table S2), significantly higher than that reported for previously analyzed plant species such as maize (8.2) and *Medicago* (3.6; Vitte and Bennetzen, 2006; Table I). These high ratios, combined with the low frequency of solo-LTRs, suggest that DNA loss rates from *Glycine* are lower than those reported for other plant species. This apparently low DNA loss rate combined with what appears to be rapid increases in retrotransposon content suggests that the genomes of *G. max* and *G. tomentella* are still expanding and doing so independently since their divergence from a common ancestor. It should be noted, however, that 51% of the LTR pairs analyzed in *G. max* are located in a pericentromeric region (H2) that has undergone a dramatic accumulation of retroelements in the last 10 to 14 million years (Innes et al., 2008); thus, the overall rate of genome expansion is likely to be lower than that indicated by our dataset. Nevertheless, if only the non-H2 elements are considered, we still observe a high ratio of base substitutions to insertion/deletions (10.9) and a low frequency of solo LTRs (two solo LTRs total compared to three in the H2 region; Supplemental Tables S2 and S3), suggesting that the *Glycine* genome overall is still expanding.

This conclusion would seem at odds with the general observation that polyploid genomes tend to be smaller than the sum of the genome sizes of their diploid ancestors (Soltis and Soltis, 1999; Ozkan et al., 2003; Gu

et al., 2006). However, a recent analysis of insertion/deletion events in diploid and polyploid cotton (*Gossypium* spp.) suggests that the genomes of both the diploid ancestors and of the polyploid derivative are adding DNA (via retroelement replication) faster than they are losing it (via homologous and illegitimate recombination; Grover et al., 2008). The diploids appear to be accumulating DNA faster than the polyploid, though, hence giving an overall appearance that the polyploid species is losing DNA relative to its diploid progenitors.

Transition to Transversion Ratios in *Glycine* LTRs Indicate a High Rate of Retrotransposon Methylation

Analysis of LTR mutation patterns can also be used to gain insight into whether these sequences are typically methylated. Methylated LTR sequences are more likely to accumulate transition mutations than transversion mutations due to the high frequency at which 5-methyl cytosine can be replaced by thymine during DNA replication (Vitte and Bennetzen, 2006); thus, the ratio of transition mutations to transversion mutations (Ts:Tv) is often used as an indicator of DNA methylation. The majority of LTRs in all species studied to date show a Ts:Tv ratio higher than that of nontransposon-related coding sequences (SanMiguel et al., 1998; Vitte and Bennetzen, 2006). The Ts:Tv ratios observed for LTRs from both *G. max* and *G. tomentella* were 2.4 (760:316) and 2.6 (266:104), respectively (Supplemental Table S2), which are similar to those reported previously for the LTRs found in the legumes *L. japonicus* and *M. truncatula* (2.4 and 2.5; Vitte and Bennetzen, 2006). Comparison of 15 *G. max* and *G. tomentella* protein coding genes on H1 (genes A through O in Innes et al., 2008) revealed a Ts:Tv ratio of 1.8 (378:212), while comparison of eight genes on H2 (genes A, B, D, G, H, J, K, and O) gave a Ts:Tv ratio of 1.7 (255:147). Similarly, comparison of *G. max* 'Williams 82' H1 and H2 (genes A through O) gave a Ts:Tv ratio of 1.7 (700:409). Based on an expected Ts:Tv ratio of 1.7, the ratio of 2.4 (760:316) for retroelements in *G. max* is significantly different from low copy genes (chi-square = 27.5; $P < 0.0001$). The elevated ratio found in the LTRs relative to the low copy genes leads us to conclude that the majority of *Glycine* retrotransposon LTRs have become methylated. It is also noteworthy that the H2 low copy genes do not display an elevated Ts:Tv ratio, suggesting that the low copy genes have not become methylated despite their pericentromeric location.

Nonautonomous LTR-Retrotransposons Appear to Be Replicating in *G. max* and *G. tomentella*

The terms autonomous and nonautonomous were first applied to DNA-based transposons to distinguish between elements that encoded all necessary proteins for transposition versus those that relied on other elements to provide transposition functions (McClintock, 1950; Fedoroff et al., 1983). Typically, nonautonomous elements are derived from autonomous elements via deletion of transposase genes in the case of DNA-based transposons or deletion of *Gag* and *Pol* genes in the case of retrotransposons. Retrotransposons frequently suffer deletions that render their *Gag-Pol* ORFs nonfunctional. However, it has only recently been established that such elements can still be replicated, presumably by Gag and Pol proteins provided by other elements in the genome (Witte et al., 2001; Kalendar et al., 2004; Sabot et al., 2006). Replication of nonautonomous retrotransposons has been inferred by the presence of genetically uniform families of elements lacking functional *Gag-Pol* genes and displaying recent insertions. Examples include the terminal repeats in miniature (TRIMs), large retrotransposon derivatives (LARDs), and so-called Morganes families (Fig. 1B; Witte et al., 2001; Kalendar et al., 2004; Sabot et al., 2006). In each of these cases, however, the element(s) providing the trans-acting Gag and Pol products have not been identified. There are two reports of putative pairs of autonomous and nonautonomous retroelements in plants, *Dasheng* and *RIRE2* from rice (Jiang et al., 2002) and *BARE-1* and *BARE-2* from barley (Tanskanen et al., 2007), but in both of these cases, there is substantial sequence divergence between the autonomous and nonautonomous families and no direct evidence that the transposition functions are being provided by the autonomous member.

We identified several apparently replicating nonautonomous retrotransposon families in the genomes of *G. max* and *G. tomentella*, including one family of elements containing both autonomous and nonautonomous members. Family 6 from *G. max* is an example of a family for which no autonomous members were found in the sequences we analyzed (Supplemental Table S2). All elements in this family possessed similar LTRs, primer binding sites, and polypurine tracks. This family appears to be one of the most numerous in the *G. max* genome, as we identified a total of five intact copies on three different BAC clones. Support for this conclusion was obtained by searching a database of low-pass whole-genome 454 DNA sequence reads in which high copy number repetitive elements have been assembled into contigs (http://stan.cropsci.uiuc.edu/sequencing.php; Swaminathan et al., 2007). Using the LTR from family 6 element 45m6-re-2 as a query, we identified a contig in this database that was 94% identical over the entire length of the LTR (contig 80367). This contig was 6.3 kb long and was made up of 1,067 reads. Using the formula described in Swaminathan et al. (2007), this level of read redundancy corresponds to a copy number of 265 genome wide. We also identified a second contig, 80354, that was 93% identical to the LTR of 45m6-re-2, which had an estimated copy number of 252. Thus, family 6 likely has a copy number of at least 500 in the soybean genome. Element 45m6-re-2 had nearly identical LTRs (Supplemental Table S2) but a highly degenerated *Gag-Pol* region; thus, we infer that it was recently inserted and that the Gag and Pol func-

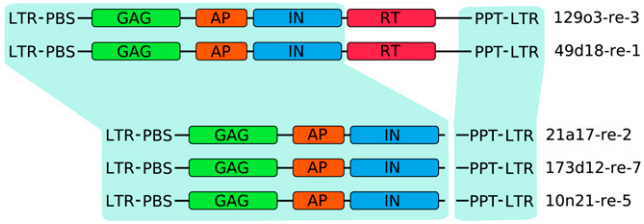

Figure 3. LTR-retrotransposon family 10 is made up of both autonomous and nonautonomous elements. Graphical representations of the five elements that make up family 10 are shown.

tions must have been supplied by another element located elsewhere in the genome.

In contrast to family 6, family 10 appears to have both autonomous and nonautonomous members. We identified five members of this family, three of which were classified as nonautonomous, while two appeared to be fully autonomous (Fig. 3). Both classes appear to have been active recently, as two of the nonautonomous class and one of the autonomous class members contained only a single nucleotide difference in their LTRs. The three nonautonomous members of this family were nearly identical to each other across their whole length but, compared to the two autonomous elements, were missing approximately 2 kb that spanned the RT domain. All five elements of this family were >97% identical to each other across the entirety of their shared sequence. Phylogenetic analysis using just this shared sequence showed that the three nonautonomous elements clustered closely together and were equally related to the two autonomous elements (data not shown), further supporting a model in which the autonomous and nonautonomous elements are both replicating. To our knowledge, this is the first example of a nonautonomous family of LTR-retroelements that appears to be recently derived from an autonomous "parent" element.

Replication of Nonautonomous Retroelements Is Likely Having a Large Impact on Genome Size in *Glycine* Species

As described above, we identified several different families of nonautonomous retrotransposons that appear to be actively replicating in the genomes of *G. max* and *G. tomentella.* When combined with the previously identified nonautonomous elements such as TRIMs, LARDs, and Morganes (Fig. 1B), there appears to be a great diversity in the structures of such elements. This suggests that almost any element with intact LTRs, primer binding site, and polypurine track may be capable of replication when appropriate Gag and Pol proteins are provided in trans. It is tempting to speculate that nonautonomous families of retrotransposons can arise anytime that active autonomous members are present. This resembles the quasispecies concept in the evolution of retroviruses and RNA viruses (Domingo et al., 1985), which has also been applied to the evolution of retrotransposons (Casacuberta et al., 1995; Sabot and Schulman, 2006). The replication of RNA by RT is an error-prone process; thus, replication of retroviruses inevitably leads to generation of many different mutant variants (quasispecies), which depend on their active and autonomous cousins for replication functions. There is no reason why the same should not occur with retrotransposons, and our findings support this hypothesis. If any element with intact LTRs, primer binding site, and polypurine track can be replicated, this provides a mechanism by which retrotransposon-related sequences in plant genomes may be driven to very high copy numbers by autonomous elements.

Apparent "Hitchhiking" of Unrelated DNA Sequences within LTR-Retrotransposons

Families 21 and 22 from *G. tomentella* were unique among the families we characterized in that all elements in these families contained a large insertion of apparently noncoding sequence downstream of the *Pol* ORF. The inserted sequence differed between the two families but was highly conserved within each family. Both families contained elements that had inserted recently, as well as elements that had inserted much earlier; thus, the inserted sequences have been replicated along with these elements for millions of years. This implies that LTR-retrotransposons are capable of replicating other unrelated DNA sequences and could potentially pick up functional genes. Although both families 21 and 22 are *gypsy*-like elements, the insertion and replication of additional DNA sequence downstream of the *Pol* ORF has also been described in the *copia*-like *SIRE* elements, which contain an additional ORF in an equivalent position (Laten et al., 1998; Holligan et al., 2006). Additionally, insertion and replication of a large ORF upstream of the *Gag-Pol* genes has been reported in *gypsy*-like elements named Ogre from the legumes pea and *Vicia pannonica* (Neumann et al., 2006; Macas et al., 2007).

We observed a possible example of such retrotransposon hitchhiking in the *gypsy*-like family 28 from *G. tomentella.* A single element in this family on BAC clone gtt1-129o17 (AC188784.13) contained an insertion of approximately 10.5 kb. The origin of this insertion is unclear, but it contains a mixture of noncoding se-

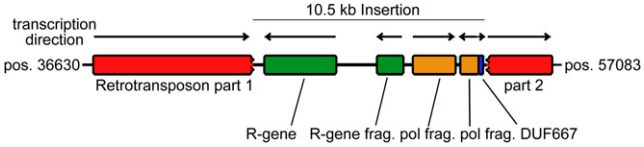

Figure 4. An LTR-retroelement with a disease resistance gene insertion. LTR-retrotransposon 129o17-re-2 located on *G. tomentella* BAC gtt1-129o17 contains an approximately 10.5-kb insertion that includes a full-length plant disease resistance gene belonging to the NB-LRR family. Numbers indicate nucleotide position relative to the complete BAC sequence (accession no. AC188784.13).

quence, several gene fragments, and one full-length nucleotide binding-Leu-rich repeat (NB-LRR) disease resistance-like gene (Fig. 4). We believe that this 10.5-kb region is contained within a single retrotransposon element based on the structure of the LTR sequences, which are 97% identical to each other and are flanked by a target site duplication (TAAGT/TAAGT). These LTRs are 85% identical to other members of family 28 that lack the 10.5-kb insertion. Based on similarity to nearby NB-LRR sequences, the NB-LRR gene within this element may be flanked by appropriate promoter and terminator sequences. What is not clear is whether this retrotransposon can still be replicated, because we did not identify any other copies of this family that carried an NB-LRR gene. However, recent work on the legume *V. pannonica* has shown that Ogre elements larger than 25 kb can be replicated at a high frequency (Neumann et al., 2006). If the gtt1-129o17 element were replicated, this would represent a new mechanism for duplicating and dispersing disease resistance genes throughout a plant genome. Intriguingly, both transcription and transposition of the tobacco retroelement Tnt1 can be induced by fungal elicitors (Melayah et al., 2001), suggesting that pathogen infection could promote retroelement multiplication. If transposition of the gtt1-129o17 element were induced by pathogen infection, it would provide a link between pathogen infection and creation of new disease resistance genes.

Long Interspersed Nuclear Elements of *G. max*, *G. tomentella*, and *P. vulgaris*

Long interspersed nuclear elements (LINEs) represent a non-LTR class of retroelements found throughout eukaryotes (Eickbush, 1992). Xiong and Eickbush's (1990) cladistic studies suggest that the first LTR-retroelements arose through the acquisition of LTRs by LINEs, therefore making them appear to be the oldest class of eukaryotic retroelements. Compared to LTR-retroelements, there have been relatively few analyses performed on LINEs in plants. For those

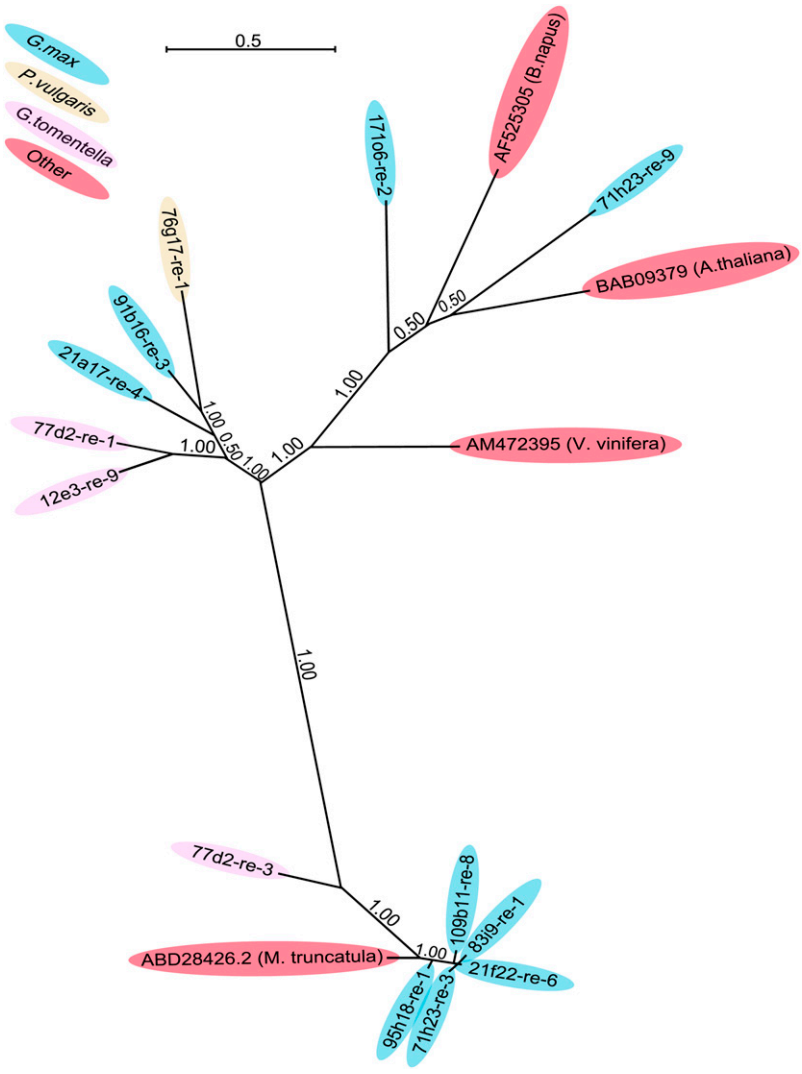

Figure 5. Phylogenetic tree showing the relationship of *G. max* LINEs to elements found in other plant species. The RT domains were used to construct this Bayesian tree. Species of origin are indicated by color-coding. Numbers indicate posterior probabilities, and the scale indicates nucleotide substitutions per site.

species that have been studied (maize, barley, Arabidopsis, lotus [*L. japonicus*], and sugar beet [*Beta vulgaris*]), LINEs appear to be more diverse but less numerous than LTR-retroelements (Schwarz-Sommer et al., 1987; Schmidt et al., 1995; Wright et al., 1996; Vershinin et al., 2002; Holligan et al., 2006). To identify potential LINEs in our BAC sequences, we used BLASTX to search the NCBI nonredundant protein database for similarity to previously characterized LINEs using the BAC DNA sequences as queries. We identified multiple LINE-like elements in *G. max*, *G. tomentella*, and *P. vulgaris.* As observed in other plant species, LINEs were much less common than LTR-retroelements. A total of 21 putative LINEs, including remnants, were identified among 36 *G. max* BACs analyzed. These elements were widely dispersed as only two BACs contained more than one element, and both of these had just two elements. To assess the diversity of these LINEs, we constructed an RT-based phylogenetic tree and included representative LINEs from maize, barley, Arabidopsis, and sugar beet (Fig. 5). This analysis revealed that the *G. max* LINEs are quite diverse. The LINEs from other plant species were distributed throughout the tree, suggesting that the *G. max* LINES are of ancient origin.

CONCLUSION

The analyses presented above show that the *G. max* genome has been heavily impacted by the activity of retroelements and likely continues to be shaped by their replication. Of most significance is our identification of three different nonautonomous families that have undergone recent replication. This observation suggests that rapid expansion of genome size can be driven by both autonomous and nonautonomous elements. A second striking feature of our dataset is the relatively low frequency of insertion/deletion events observed in the LTRs of both *G. max* and *G. tomentella* compared to previously characterized plant species, including the legume *M. truncatula* (Table I). Although the underlying cause for this is not known, it suggests that the *G. max* genome is likely still expanding. Finally, the identification of a retroelement carrying an NB-LRR disease resistance-like gene provides a potential new mechanism for the rapid evolution of new resistance genes.

MATERIALS AND METHODS

BAC and Retroelement Sequences

All BAC sequences were obtained from either the High Throughput Genomic Sequence database or the nonredundant nucleotide database maintained by NCBI. The majority of these sequences were generated as part of the NSF-funded project "Comparative Analysis of Legume Genome Evolution" (grant no. DBI–0321664; Innes et al., 2008). Accession numbers for each BAC are provided in Supplemental Table S1.

Identifying Retroelements

Approximately 3.7 Mb of *Glycine max* genomic sequence were searched, including 1 Mb from H1 from the NSF project (Innes et al., 2008), about 0.85 Mb of H2, and 1.85 Mb derived from BACs not assigned to a particular genomic location that have been sequenced as part of an ongoing project in the R. Shoemaker laboratory. Where BACs covered overlapping regions, only the unique sequences were counted in determining the total area analyzed and total elements identified. We used the program LTR_STRUC as the first step in identifying retrotransposon sequences (McCarthy and McDonald, 2003). LTRs from the elements identified by LTR-STRUC were used as queries in BLAST searches (Altschul et al., 1997). These BACs were also searched for the presence of retrotransposon-related genes using the BAC sequences as a query to search the NCBI nonredundant database using BLASTX. Regions of homology to known retrotransposon-like sequences (e.g. RT, integrase, etc.) were then manually evaluated for the presence of LTRs. In addition, we used the REPuter and RepeatMasker programs to identify repeated sequences (Kurtz et al., 2001; Smit et al., 1996–2008). These additional searches uncovered several intact elements missed by the LTR_STRUC program. Essentially the same approach was used to identify retrotransposons in 1.6 Mb of genomic sequence from *Glycine tomentella* (0.5 Mb from H1, 0.35 Mb from H2, and 0.75 Mb of *G. tomentella* sequence from BACs not yet assigned a genomic location) and 0.94 Mb of sequence from *Phaseolus vulgaris* (Supplemental Table S1). To identify potential LINEs, we used BLASTX to search the NCBI nonredundant protein database for similarity to previously characterized LINEs using the BAC DNA sequences as queries. All elements identified by the above approaches were deposited in a local database, and BAC DNA sequences were then searched for homology to this database using BLASTN (Altschul et al., 1997). To group retrotransposons into families, all LTR sequences were compared to each other in pair-wise BLASTN comparisons. Elements that shared a minimum of 80% sequence identity over at least 80% of the length of the shortest LTR were grouped into the same family per the recommendations of Wicker et al. (2007).

Sequence Alignments and Phylogenetic Tree Construction

Multiple sequence alignments were performed using ClustalX (Jeanmougin et al., 1998) and the MEGA software package version 3.1 followed by manual adjustments to optimize the alignments (Kumar et al., 2004). Transition and transversion mutation rates were also calculated using the MEGA software package. Trees were generated using the MrBayes software package version 3.1.2 (Ronquist and Huelsenbeck, 2003) using the General Time Reversible DNA substitution model with gamma-distributed rate variation across sites and a proportion of invariable sites (General Time Reversible +I+G model). We performed paired runs with four chains each with sampling every 100 generations. The priors for each analysis were the program's defaults. All runs started with a random tree and were run for 5 million generations. After elimination of the first 25% of runs, which included the burn-in phase, the remaining iterations were summarized in a consensus tree with posterior probabilities as nodal support.

Dating LTR-Retrotransposon Insertion Times

The insertion times of LTR-retroelements were dated by aligning their 5′ and 3′ LTR sequences and identifying transition and transversion substitutions using the MEGA software package version 3.1(SanMiguel et al., 1998). The time since element insertion was calculated using the formula $T = K/2r$, where T = time, K = distance calculated using Kimura's two parameter model as implemented within the MEGA software package, and r = substitution rate. Kimura's two parameter model corrects for multiple hits (Kimura, 1980). Two values for the substitution rate were used and are shown in Supplemental Table S2: 5.1×10^{-9} (the average synonymous substitution rate estimated for genic sequences in *G. max*; Pfeil et al., 2005) and 1.3×10^{-8}, which is the value used by Vitte and Bennetzen (2006) in Table I. The latter value takes into account the observation that LTR sequences accumulate mutations at a higher rate than silent sites in standard housekeeping genes, possibly because of the high rate of cytosine methylation observed in LTR sequences (Vitte and Bennetzen, 2006).

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession numbers FJ197979 to FJ198023 (*G. max* LTR-retrotransposons), FJ402900 to FJ402922 (*G. tomentella* LTR-retrotansposons), and FJ402923 to FJ402929 (*P. vulgaris* LTR-retrotransposons) and are also listed in

Supplemental Table S2. Accession numbers for the LINEs analyzed in Figure 5 can be found under accession numbers FJ402887 to FJ402899.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Table S1. BAC clones screened for retroelement-related sequences.

Supplemental Table S2. Analysis of paired LTRs.

Supplemental Table S3. Numbers of intact elements, solo-LTRs, and fragmented elements.

ACKNOWLEDGMENTS

We thank Randy Shoemaker, Barbara Baker, and Chris Pires for serving on the advisory committee for this project. We thank Mounier Elharam and Jennifer Lewis at the University of Oklahoma's Advanced Center for Genome Technology (ACGT) for contributing to the DNA sequencing on the ABI 3730 s and Steve Kenton, Shaoping Lin, and Ying Fu for their helpful discussions on sequencing through difficult regions. Computer support was provided by the Indiana University Information Technology Services Research Database Complex, the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation, and the ACGT.

Received August 10, 2008; accepted October 22, 2008; published October 24, 2008.

LITERATURE CITED

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar DS, et al (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. Plant J **52:** 342–351

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev **15:** 621–627

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol **7:** 732–736

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot (Lond) **95:** 127–132

Casacuberta JM, Vernhettes S, Grandbastien MA (1995) Sequence variability within the tobacco retrotransposon *Tnt1* population. EMBO J **14:** 2670–2678

Devos KM, Brown JK, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res **12:** 1075–1079

Domingo E, Martinez-Salas E, Sobrino F, de la Torre JC, Portela A, Ortin J, Lopez-Galindez C, Perez-Brena P, Villanueva N, Najera R, et al (1985) The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance: a review. Gene **40:** 1–8

Eickbush TH (1992) Transposing without ends: the non-LTR retrotransposable elements. New Biol **4:** 430–440

Fedoroff N, Wessler S, Shure M (1983) Isolation of the transposable maize controlling elements *Ac* and *Ds*. Cell **35:** 235–242

Gao L, McCarthy EM, Ganko EW, McDonald JF (2004) Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequences. BMC Genomics **5:** 18

Goldberg RB (1978) DNA sequence organization in the soybean plant. Biochem Genet **16:** 45–68

Grover CE, Yu Y, Wing RA, Paterson AH, Wendel JF (2008) A phylogenetic analysis of indel dynamics in the cotton genus. Mol Biol Evol **25:** 1415–1428

Gu YQ, Salse J, Coleman-Derr D, Dupin A, Crossman C, Lazo GR, Huo N, Belcram H, Ravel C, Charmet G, et al (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. Genetics **174:** 1493–1504

Gurley WB, Hepburn AG, Key JL (1979) Sequence organization of the soybean genome. Biochim Biophys Acta **561:** 167–183

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The transposable element landscape of the model legume *Lotus japonicus*. Genetics **174:** 2215–2228

Innes RW, Ameline-Torregrosa C, Ashfield T, Cannon E, Cannon SB, Chacko B, Chen NWG, 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

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci **23:** 403–405

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature **431:** 569–573

Jiang N, Jordan IK, Wessler SR (2002) *Dasheng* and *RIRE2*. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. Plant Physiol **130:** 1697–1705

Jin YK, Bennetzen JL (1994) Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize. Plant Cell **6:** 1177–1186

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics **166:** 1437–1450

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16:** 111–120

Kordis D (2005) A genomic perspective on the chromodomain-containing retrotransposons: chromoviruses. Gene **347:** 161–173

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform **5:** 150–163

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res **29:** 4633–4642

Laten HM, Havecker ER, Farmer LM, Voytas DF (2003) *SIRE1*, an endogenous retrovirus family from *Glycine max*, is highly homogeneous and evolutionarily young. Mol Biol Evol **20:** 1222–1230

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

Ma J, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res **14:** 860–869

Macas J, Neumann P, Navratilova A (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. BMC Genomics **8:** 427

Marin I, Llorens C (2000) *Ty3/Gypsy* retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. Mol Biol Evol **17:** 1040–1049

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics **19:** 362–367

McClintock B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA **36:** 344–355

Melayah D, Bonnivard E, Chalhoub B, Audeon C, Grandbastien MA (2001) The mobility of the tobacco *Tnt1* retrotransposon correlates with its transcriptional activation by fungal factors. Plant J **28:** 159–168

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet **37:** 997–1002

Neumann P, Koblizkova A, Navratilova A, Macas J (2006) Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics **173:** 1047–1056

Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, Kouzarides T, Murzin AG, Murzina NV, Laue ED (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. Nature **416:** 103–107

Ozkan H, Tuna M, Arumuganathan K (2003) Nonadditive changes

in genome size during allopolyploidization in the wheat (aegilops-triticum) group. J Hered **94:** 260–264

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

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572–1574

Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity **97:** 381–388

Sabot F, Sourdille P, Chantret N, Bernard M (2006) *Morgane*, a new LTR retrotransposon group, and its subfamilies in wheats. Genetica **128:** 439–447

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet **20:** 43–45

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 BE, Schlueter SD, Shoemaker RC (2006) Sequence conservation of homeologous bacterial artificial chromosomes and transcription of homeologous genes in soybean (*Glycine max* L. Merr.). Genetics **174:** 1017–1028

Schmidt T, Kubis S, Heslop-Harrison JS (1995) Analysis and chromosomal localization of retrotransposons in sugar beet (*Beta vulgaris* L.): LINEs and *Ty1-copia*-like elements as major components of the genome. Chromosome Res **3:** 335–345

Schwarz-Sommer Z, Leclercq L, Gobel E, Saedler H (1987) *Cin4*, an insert altering the structure of the *A1* gene in *Zea mays*, exhibits properties of nonviral retrotransposons. EMBO J **6:** 3873–3880

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

Smit AFA, Hubley R, Green P (1996–2008) RepeatMasker Open-3.0. http://www.repeatmasker.org

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. Trends Ecol Evol **14:** 348–352

Straub SC, 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

Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. BMC Genomics **8:** 132

Tanskanen JA, Sabot F, Vicient C, Schulman AH (2007) Life without GAG: the *BARE-2* retrotransposon as a parasite's parasite. Gene **390:** 166–174

Vershinin AV, Druka A, Alkhimova AG, Kleinhofs A, Heslop-Harrison JS (2002) LINEs and *gypsy*-like retrotransposons in *Hordeum* species. Plant Mol Biol **49:** 1–14

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA **103:** 17638–17643

Wang H, Liu JS (2008) LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. BMC Genomics **9:** 382

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell **18:** 1791–1802

Wicker T, Keller B (2007) Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res **17:** 1072–1081

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet **8:** 973–982

Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (*TRIM*) are involved in restructuring plant genomes. Proc Natl Acad Sci USA **98:** 13778–13783

Wright DA, Ke N, Smalle J, Hauge BM, Goodman HM, Voytas DF (1996) Multiple non-LTR retrotransposons in the genome of *Arabidopsis thaliana*. Genetics **142:** 569–578

Wright DA, Voytas DF (2002) *Athila4* of Arabidopsis and *Calypso* of soybean define a lineage of endogenous plant retroviruses. Genome Res **12:** 122–131

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J **9:** 3353–3362

Yano ST, Panbehi B, Das A, Laten HM (2005) *Diaspora*, a large family of *Ty3-gypsy* retrotransposons in *Glycine max*, is an envelope-less member of an endogenous plant retrovirus lineage. BMC Evol Biol **5:** 30

Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA, Tabata S (2005) Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol **137:** 1174–1181

Zabala G, Vodkin LO (2005) The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the *CACTA* superfamily. Plant Cell **17:** 2619–2632