Differential Accumulation of Retroelements and Diversification of NB-LRR Disease Resistance Genes in Duplicated Regions following Polyploidy in the Ancestor of Soybean[1][W][OA]

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

Department of Biology, Indiana University, Bloomington, Indiana 47405 (R.W.I., T.A., A.D., S.H., S.M.d.C., M.M., R.P., A.W.); Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota 55108 (C.A.-T., E.C., S.B.C., B.C., R.D., N.D.Y.); Virtual Reality Application Center, Iowa State University, Ames, Iowa 50011 (E.C.); United States Department of Agriculture-Agricultural Research Service and Department of Agronomy, Iowa State University, Ames, Iowa 50011 (S.B.C.); Institut de Biotechnologie des Plantes, UMR CNRS 8618, INRA, Université Paris Sud, 91 405 Orsay, France (N.W.G.C., M.S., V.T., V.G.); Genoscope/Commissariat à l'Energie Atomique-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.); Department of Crop and Soil Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061 (N.G., J.M., A.N., M.B.R., D.M.T., M.A.S.M.); Department of Agronomy, Purdue University, West Lafayette, Indiana 47907 (C.S.H., S.J., J.W.); Commonwealth Scientific and Industrial Research Organization 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.)

The genomes of most, if not all, flowering plants have undergone whole genome duplication events during their evolution. The impact of such polyploidy events is poorly understood, as is the fate of most duplicated genes. We sequenced an approximately 1 million-bp region in soybean (*Glycine max*) centered on the *Rpg1-b* disease resistance gene and compared this region with a region duplicated 10 to 14 million years ago. These two regions were also compared with homologous regions in several related legume species (a second soybean genotype, *Glycine tomentella*, *Phaseolus vulgaris*, and *Medicago truncatula*), which enabled us to determine how each of the duplicated regions (homoeologues) in soybean has changed following polyploidy. The biggest change was in retroelement content, with homoeologue 2 having expanded to 3-fold the size of homoeologue 1. Despite this accumulation of retroelements, over 77% of the duplicated low-copy genes have been retained in the same order and appear to be functional. This finding contrasts with recent analyses of the maize (*Zea mays*) genome, in which only about one-third of duplicated genes appear to have been retained over a similar time period. Fluorescent in situ hybridization revealed that the homoeologue 2 region is located very near a centromere. Thus, pericentromeric localization, per se, does not result in a high rate of gene inactivation, despite greatly accelerated retrotransposon accumulation. In contrast to low-copy genes, nucleotide-binding-leucine-rich repeat disease resistance gene clusters have undergone dramatic species/homoeologue-specific duplications and losses, with some evidence for partitioning of subfamilies between homoeologues.

[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 by a grant from Genoscope/Commissariat à l'Energie Atomique-Centre National de Séquençage (to V.G.). A.N.E. was supported by a National Science Foundation Systematics Award (grant no. DEB–0516673).

[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.108.127902

The comparative approach to studying genes and genomes is a powerful method for addressing both

fundamental and applied questions in genome evolution (Paterson, 2006; Schranz et al., 2007; Margulies and Birney, 2008). To realize the full potential of this approach, however, requires consideration of the variable rates at which different aspects of genome evolution occur, which in turn requires comparison of multiple species positioned at varying phylogenetic distances (Schranz et al., 2007). Important questions in genome evolution, particularly about the evolution of gene families and genome structure, can be addressed most effectively by analysis of large contiguous blocks of DNA sequence from multiple species (Moreno et al., 2008). Few such analyses have been performed on plant genomes (Lai et al., 2004; Swigonova et al., 2004; Ma et al., 2005; Mudge et al., 2005), however, because of the wide phylogenetic distances between the best-sampled genomes: Arabidopsis (*Arabidopsis thaliana*), maize (*Zea mays*), and rice (*Oryza sativa*), and, more recently, *Medicago truncatula*, poplar (*Populus trichocarpa*), grape (*Vitis vinifera*), and *Carica papaya*. Here, we describe comparison of a 1-Mb region centered on the *Rpg1-b* disease resistance gene of soybean (*Glycine max*; Ashfield et al., 2004) with homologous regions in three other legume species.

Soybean is an attractive choice for genome evolution studies because it is a major food crop, it is a legume (a large and diverse plant family that is both ecologically and economically important; Doyle and Luckow, 2003; Singh et al., 2007), and, of particular relevance to this study, it is an ancient polyploid (Shoemaker et al., 2006). Polyploidy (genome duplication) is thought to be a driving force behind the rapid diversification of angiosperms and their ability to successfully colonize new niches (Hegarty and Hiscock, 2008; Leitch and Leitch, 2008). Many of our major crop plants (e.g. wheat [*Triticum aestivum*], maize, cotton [*Gossypium hirsutum*], and soybean) are of polyploid origin; thus, the role of polylploidy in shaping plant genome structure and function is of fundamental and practical interest (Bowers et al., 2003; Paterson, 2005). The molecular mechanisms underlying the success of polyploid plants are currently under intensive investigation (Hegarty and Hiscock, 2008; Leitch and Leitch, 2008), but recent data indicate that, following polyploidy, large-scale genetic (transposable element activity, chromosome translocations, insertions, and deletions) and epigenetic (histone modifications and DNA methylation) alterations occur (Adams et al., 2003, 2004; Gaeta et al., 2007; Doyle et al., 2008), some of which differentially affect duplicated gene copies and are potentially associated with the acquisition of adaptive traits (Dubcovsky and Dvorak, 2007).

The soybean genome has undergone at least two rounds of whole genome duplication, one estimated to have occurred 10 to 14 million years ago (mya) and a second more ancient event estimated to have occurred 50 to 60 mya (Shoemaker et al., 2006). By examining a 1-Mb region, we hoped to compare and contrast the evolutionary fates of low-copy and high-copy gene families following a polyploidy event. Our initial goal was to compare the *Rpg1-b* region of soybean with the region duplicated during the 10-mya polyploidy event. To assign polarity to any differences found between these two homoeologous regions (regions diverged due to polyploidy), we compared both regions with the single orthologous region in *Phaseolus vulgaris* (common bean), which diverged from soybean approximately 20 mya (Lavin et al., 2005) and has not undergone a subsequent polyploidy event (Fig. 1). We also compared these regions with the orthologous regions in *Glycine tomentella*, a wild perennial relative of soybean that diverged from soybean approximately 5 mya (see below). This comparison allowed us to place a time frame on when differences between homoeologues arose. Finally, we also compared these regions with orthologous regions in *Medicago* and in a second accession of soybean (PI96983), allowing us to detect more ancient (approximately 54 mya) and very recent events.

One of our specific goals was to assess the impact of polyploidy on the evolution of disease resistance genes, which are among the most rapidly evolving and polymorphic genes known in plants (Meyers et al., 1999, 2003). We focused our analyses on the *Rpg1-b* region, located on molecular linkage group F (chromosome 13), because it is highly enriched in resistance genes belonging to the nucleotide-binding-leucine-rich repeat

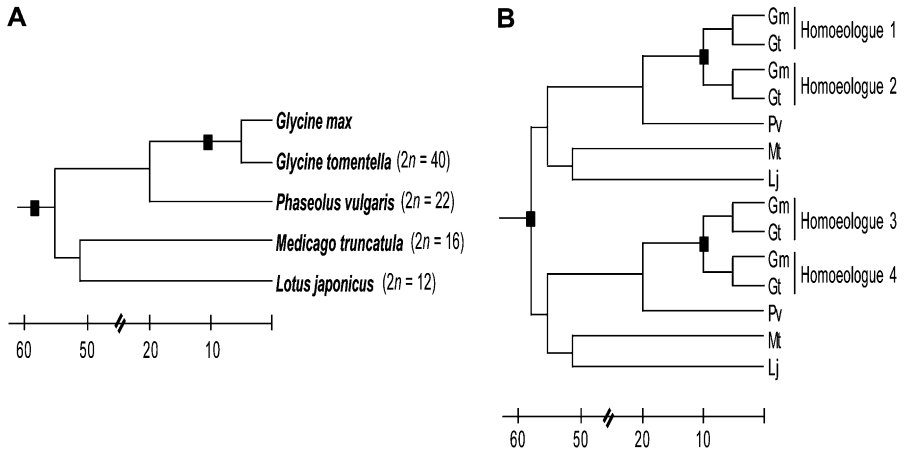

Figure 1. Phylogenetic relationships of taxa and their genomes. Divergence times (million years) are from Lavin et al. (2005). A, Species relationships. Black squares mark polyploidy events. B, Expected relationships among genomes and their genes. Black squares mark duplications produced by polyploidy events. Two successive whole genome duplications in the ancestor of *Glycine* lead to the expectation of a maximum of four homoeologues in modern *Glycine* species.

(NB-LRR) family (Ashfield et al., 2003, 2004), and we selected PI96983 for analysis because it had previously been shown to contain a suite of NB-LRRs in the *Rpg1-b* region quite distinct from Williams 82, based on Southern-blot hybridizations (Jeong et al., 2001). In addition, testing of PI96983 revealed that it did not contain a functional *Rpg1-b* gene, which confers resistance to *Pseudomonas syringae* pv *glycinea* strains that express the type III effector gene *avrB* (Ashfield et al., 1995, 2004). Conversely, the *Rsv1* gene, which confers resistance to *Soybean mosaic virus* and has been mapped to the *Rpg1-b* region, is found in PI96983 but not in Williams 82 (Gore et al., 2002; Hayes et al., 2004). We thus expected to find structural polymorphisms involving NB-LRRs and wondered if these extended to other classes of genes in this region.

These comparisons allowed us to address several fundamental questions relating to NB-LRRs, polyploidy, and genome evolution in legumes. Do rates of gene evolution vary between homoeologous regions? To what extent have genome rearrangements occurred in these homoeologous segments? Does one member of a homoeologous chromosome pair rearrange preferentially, or are the rearrangements evenly distributed? Are some duplicated copies of particular gene classes (e.g. NB-LRRs) preferentially lost following polyploidy? Do NB-LRRs behave differently from other clustered gene families? What impact have retrotransposons had on the evolution of this region, and do homoeologues differ in this regard?

RESULTS

Assembly of Bacterial Artificial Chromosome Contigs for Sequencing

We assembled and sequenced an approximately 1-Mb bacterial artificial chromosome (BAC) contig from soybean cv Williams 82 centered on the *Rpg1-b* gene (Supplemental Fig. S1; see "Materials and Methods"; Ashfield et al., 1998, 2003). To minimize gaps, this contig was assembled from two different BAC libraries (referred to here as gmw1 and gmw2) that employed different restriction enzymes in their creation. Combined, these libraries represented 17.4 genome equivalents (Marek and Shoemaker, 1997; Marek et al., 2001). Despite these precautions, we were unable to span one genomic interval that is highly unstable in BAC vectors in *Escherichia coli*, resulting in two BAC contigs that were tightly linked genetically (Supplemental Fig. S1). Using fiber fluorescent in situ hybridization (FISH) analysis, we estimated the gap between these two contigs to be between 50 and 60 kb (data not shown). While completing our analyses, the first draft of the soybean genome sequence was released (7-fold coverage; Soybean Genome Project, Department of Energy Joint Genome Institute; http://www.phytozome.net/soybean.php), which enabled us to compare our BAC-based sequence data with the whole genome shotgun (WGS) sequencing data for this region. To distinguish this WGS draft from future assemblies, we will refer to the scaffolds from this draft as "7x" scaffolds. Both of the BAC contigs were located within 7x scaffold 172 of the WGS sequence, and the gap was found to be 57,562 nucleotides.

Supplemental Figure S1 shows the specific BAC clones that were selected for sequencing. Assembly of these BAC sequences into supercontigs resulted in 553,148 bp of unique sequence in the left contig and 453,942 bp in the right contig. Combined with the gap sequence from the WGS data, this represented 1,064,642 bp in total. This megabase region served as our reference sequence for identifying the homoeologous region(s) in cv Williams 82, both homoeologues in soybean line PI96983 and *G. tomentella* accession G1403, and the single orthologous region in *P. vulgaris* accession G19833 and *M. truncatula* var Jemalong.

To identify these homologous regions, we screened BAC libraries using DNA hybridization probes derived from low-copy protein-coding genes identified in the Williams 82 sequence (Supplemental Table S1; Supplemental Fig. S1). BAC clones that hybridized to two or more probes were then fingerprinted and end sequenced. A combination of fingerprint information, probe hybridization patterns, and end sequence information was used to assemble contigs and identify a minimum tiling path for sequencing. Supplemental Figure S1 shows a physical map of all of the BACs selected for sequencing and the probes that hybridized to each. Note that we identified BAC clones in soybean cv Williams 82 that appear to represent homoeologous regions derived from two different whole genome duplication events (Fig. 1; see below; Shoemaker et al., 2006). For clarity, we refer to the BAC contigs derived from the most recent duplication as homoeologue 2 (H2) and the BAC contig derived from the more ancient duplication as homoeologue 3 (H3), with the reference sequence being homoeologue 1 (H1; Fig. 1). We did not identify BAC clones corresponding to homoeologue 4 (Fig. 1) in our library screen, presumably because this region has become fragmented to the extent that the genes homoeologous to H1 are separated by more than one BAC length.

Based on probe hybridization patterns, we obtained good coverage of H2 in Williams 82 and the orthologous and homoeologous regions in the other taxa; however, there remained a number of gaps in most BAC contigs (Supplemental Fig. S1). For soybean H2, this was likely due to expansion by retroelement insertions (see below), making it difficult to identify individual BAC clones containing genes homologous to two or more H1 probes. For the other taxa, gaps may be due to the lower depth of the BAC libraries screened (average of eight genome equivalents), the presence of regions that are unstable in *E. coli*, and/or genomic rearrangements relative to the reference Williams 82 sequence.

To address whether genomic rearrangements were a possible cause of the gaps in the soybean H2 contig, we genetically mapped several BAC clones that spanned H2 from Williams 82 using microsatellite markers (Akkaya et al., 1995). We also mapped two BACs

from H3 of Williams 82. Significantly, all BACs identified from H2 mapped to the same location on soybean molecular linkage group E (chromosome 15), approximately 4 cM from microsatellite marker Sat_136 (data not shown), indicating that this region has not been fragmented by chromosomal translocations or inversions. BACs gmw2-129e12 and gmw2-91b16 from H3 mapped to linkage group C2 (chromosome 6), approximately 5 cM from marker Satt357.

The soybean Williams 82 H2 supercontig (the combination of contigs, individual BACs, and gaps corresponding to this region) contained two gaps (Supplemental Fig. S1). To determine the size of these gaps, we again compared our sequence with the 7x WGS sequence. All three BAC contigs from the Williams 82 H2 region were contained within 7x scaffold 55 of the WGS sequence, providing confirmation for our genetic mapping data. To our surprise, both gaps were very large (gap 1 = 819,550 bp and gap 2 = 835,706 bp). This analysis also revealed that the homology between the H1 contig and the WGS 7x scaffold 55 sequence extended 177,494 bp on the left flank of the H2 supercontig and 730,000 bp on the right flank. Thus, the 1,064,642-bp H1 region corresponds to a region of 3,434,337 in H2, raising the question of whether H1 had lost DNA or H2 had gained DNA.

Expansion of H2 Is Caused by Retroelement Insertions and Is Associated with Proximity to a Centromere

To determine the origin of the differences in sequence content between soybean homoeologues H1 and H2, we annotated both sequences and compared their gene contents with the homologous region from *Phaseolus*. BAC sequences were annotated using a semiautomated approach to identify both protein-coding genes and repetitive elements (see "Materials and Methods"). We then aligned BAC contigs based on positions of conserved low-copy genes. Figure 2A and Supplemental Figure S2A show an alignment between homoeologues 1 and 2 of soybean cv Williams 82 and *Phaseolus*. The comparison with *Phaseolus* enabled us to infer whether differences between H2 and H1 represented losses or insertions, as any genes shared between *Phaseolus* and one or both *Glycine* homoeologues presumably were present in their most recent common ancestor. This alignment revealed that H2 contains many more retroelement insertions than either H1 or *Phaseolus*, and as a consequence, the low-copy genes have been spread apart in H2. The degree of expansion is not constant along H2, with some regions affected more than others (Fig. 2A). This expansion explains our failure to identify BAC clones that spanned gap 1 and gap 2 in our H2 contig assembly, as our initial criteria for identifying homoeologous BACs required that they contain at least two low-copy genes found on H1.

Neither *Phaseolus* nor soybean H1 contained the high retrotransposon content observed in soybean H2, suggesting that retrotransposons have accumulated in H2 in the 10 to 14 million years since the divergence of H1 and H2 from their common ancestor. Consistent with this hypothesis, a similar retroelement-mediated expansion of H2 was also observed in *G. tomentella* (Fig. 2B; Supplemental Fig. S2B); thus, the propensity for H2 to accumulate retroelement insertions was conditioned prior to the separation of soybean and *G. tomentella*, which occurred 5 to 7 mya (Fig. 1). A diverse collection of retroelements was found in H2 of both soybean and *G. tomentella*, including *copia*-like and *gypsy*-like long terminal repeat (LTR) retrotransposons, as well as LINE elements (Wawrzynski et al., 2008). The majority of the LTR retroelements identified in both soybean and *G. tomentella*, however, inserted within the last 4 million years, and many within the last 1 million years (Wawrzynski et al., 2008). Thus, the H2 region in both soybean and *G. tomentella* continues to be receptive to a diverse set of retroelement insertions.

In both plants and animals, centromeric and pericentromeric regions of the genome are enriched in repetitive elements, including retroelements (Lin et al., 2005; Ma et al., 2007), which suggested that the H2 region might be adjacent to a centromere. To test this hypothesis directly, we performed standard FISH analysis on soybean mitotic prometaphase chromosome spreads, using H2 BAC gmw2-12n11 as a hybridization probe in conjunction with a centromere-specific probe (SB91; see "Materials and Methods"). These hybridizations revealed that the H2 region in soybean is indeed closely associated with a centromere (Supplemental Fig. S3). To confirm this finding, we queried the preliminary 8x chromosome assemblies produced by the Soybean Genome Project, Department of Energy Joint Genome Institute (S.B. Cannon, unpublished data). The soybean H2 region was found to be within 2.8 Mb of the centromere (defined as the region containing tandem 91- to 92-nucleotide satellite repeats) on chromosome 15 (linkage group E), while the soybean H1 region is approximately 13.9 Mb from the centromere of chromosome 13 (linkage group F). Both chromosomes are approximately 50 Mb in length. We thus hypothesize that sometime after the divergence of the H1 and H2 parent species, and prior to the split between soybean and *G. tomentella*, a chromosomal rearrangement occurred that placed the H2 region adjacent to a centromere. As a consequence, this region began accumulating repetitive elements, including retroelements, at a higher rate than H1 and continued to do so thereafter. These data suggest that centromeric sequences can promote repetitive element accumulation in adjacent sequences when brought into proximity. As discussed further below, this observation is consistent with the "centromere drive" model, whereby expansion of centromeric repeats facilitates the segregation of chromosomes into the egg cell during meiosis (Henikoff and Malik, 2002).

Homoeologous Gene Loss in Soybean Is Biased Toward H2

Given the rapid expansion of retrotransposon content in H2 of soybean, we asked whether this would

Figure 2. (*Figure continues on following page.*)

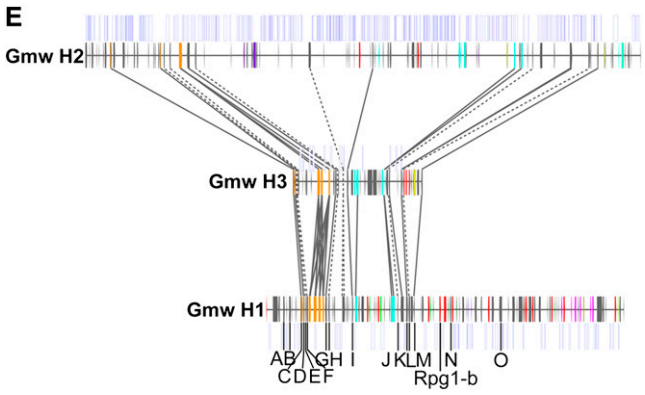

Figure 2. Alignment of homoeologous and orthologous BAC contigs. Vertical filled rectangles represent protein-coding genes other than retrotransposon-associated proteins, and open blue rectangles, shown in a separate track, represent retrotransposons. Solid lines between contigs indicate allelic, homoeologous, or orthologous genes. Supplemental Figure S2 shows a larger version of this figure that includes gene names. A, Comparison of soybean H1 and H2 with *Phaseolus*. Dotted lines between the H1 and H2 contigs indicate homoeologous genes that either are not present in the orthologous region of *Phaseolus* or are positioned outside of the region sequenced in *Phaseolus*. B, Comparison of soybean H1 and H2 with their orthologues in *G. tomentella*. C, Comparison of soybean cv Williams 82 and line PI96983. Dashed rectangles indicate NB-LRR clusters that differ in gene content between these two genotypes. D, Comparison of soybean H1 and H2 and *Medicago*. E, Comparison of soybean H1, H2, and H3.

correlate with a more rapid loss of genes duplicated by polyploidy. To estimate gene loss, we first identified all non-NB-LRR genes present in either H1 or H2 that were also present in a syntenic position in *Phaseolus* (connected by lines in Fig. 2A and Supplemental Fig. S2A), which defined a minimal set of genes present in this region in the ancestral chromosome of H1 and H2. Of these 35 genes, 27 were present in H2, indicating that H2 has lost 23% of its non-NB-LRR genes following polyploidy. Conversely, all 35 genes were retained on H1, indicating that there has been little gene loss from H1 following polyploidy. Thus, gene loss has been strongly biased toward H2, which may be related to the increased retrotransposon activity in this region.

Note that there are 13 predicted low-copy genes present only on H1 and over 70 predicted low-copy genes present only on H2 and not in *Phaseolus* (Supplemental Fig. S2A). We speculate that these represent genes or gene fragments that were transposed into these positions by retroelements or other mobile DNAs. In support of this speculation, the majority of these predicted genes contain open reading frames shorter than 500 bp and are not present in the soybean EST collection; thus, they are unlikely to be functional. We also identified three genes that are conserved in H1 and H2 but are missing from the syntenic position in *Phaseolus* (indicated by dotted lines in Fig. 2A and Supplemental Fig. S2A). These genes presumably were lost from *Phaseolus* after *Phaseolus* and *Glycine* diverged or, alternatively, they were inserted after this time point and before the divergence of H1 and H2.

Homoeologous Gene Pairs Are Diverging at the Same Rate

To confirm that H1 and H2 were indeed derived from the most recent whole genome duplication event, we analyzed nucleotide substitution rates at silent sites (Ks) for 15 low-copy genes spread over the entire aligned region (indicated by letters A–O in Fig. 2A; described in Supplemental Table S3). When averaged over multiple homoeologous gene pairs, such Ks analyses provide an approximate age of genome duplication events (Lynch and Conery, 2000; Shoemaker et al., 2006). The soybean H1:H2 comparison revealed substantial variation in Ks values between genes (range of 0.054–0.184; Supplemental Table S2). This 3.4-fold variation is not surprising; Zhang and colleagues (2002) found that Ks values varied by as much as 13.8-fold in 242 homoeologous gene pairs in Arabidopsis, although 90% were within 2.6-fold. The mean Ks value for the 15 H1:H2 gene pairs examined in soybean cv Williams 82 was 0.122 $\pm$ 0.035 (Table I; Supplemental Table S2). This value compares well with previous analyses of gene duplicates derived from the most recent whole genome duplication event in soybean, where five homoeologous gene pairs showed an average Ks of 0.149 (Schlueter et al., 2006) and 23 homoeologous gene pairs showed an average Ks of 0.147 (Van et al., 2008).

The variation in Ks values between individual gene pairs implies that substitution rates for individual genes vary considerably, even for genes in the same genomic region. To test whether gene conversion might account for this variation, and to determine whether the two homoeologues were undergoing substitutions at the same rate, we compared each homoeologue pair with its common orthologue in *Phaseolus*. Ks analysis revealed nearly identical substitution rates for H1 and H2, with mean Ks values of 0.260 $\pm$ 0.082 for the H1:*Phaseolus* comparison and 0.256 $\pm$ 0.067 for the H2:*Phaseolus* comparison (n = 13 genes for both; Table I; Supplemental Table S2). To rule out gene conversion events between homoeologues, which would reduce the apparent differences between H1 and H2, we calculated the ratio of the H1:H2 Ks value to the H1:*Phaseolus* Ks value for each low-copy gene. In the absence of gene conversion, this ratio should be roughly the same for all genes, assuming relatively constant substitution rates for individual genes subsequent to the divergence between soybean and *Phaseolus*. A gene conversion event between H1 and H2 would artificially lower the Ks value for the H1:H2 comparison relative to the H1:*Phaseolus* comparison. As shown in Supplemental Table S2, 11 of 13 genes showed similar ratios (range, 0.36–0.62), indicating that gene conversion is not a cause of the gene-to-gene variation in substitution rates. Only gene H showed an

unusually low ratio (0.19), while gene D showed a high ratio (0.90). The mean ratio for all 13 genes was 0.49 ± 0.16, consistent with the estimated divergence dates of H1 and H2 (approximately 10 mya) compared with *Phaseolus* and *Glycine* (approximately 20 mya). Because homoeologues have retained nearly identical substitution rates (see phylogenetic analysis below), it suggests that the gene-to-gene variation in substitution rate is a property of the gene that has been maintained after polyploidy, and is not strongly influenced by proximity to the centromere or other external factors.

The conservation of the majority of homoeologous gene duplicates implies that both copies of each homoeologous gene pair continue to function and are under selection (Fig. 2A). This inference was confirmed by analysis of nonsynonymous to synonymous nucleotide substitution ratios (Ka/Ks), which had a mean of 0.31 ± 0.19 for the 15 gene pairs compared (Supplemental Table S2), indicating that these gene duplicates remain under purifying selection. This conclusion was further supported by the identification of ESTs derived from both members of 23 of 34 homoeologous gene pairs analyzed (Supplemental Table S4; six pairs had no identifiable ESTs for either gene), demonstrating that both the H1 and H2 gene copies are transcribed. Thus, while more gene loss has occurred from H2 than from H1, proximity to a centromere and accumulation of repetitive elements on H2 have not caused the inactivation and/or loss of most low-copy genes on H2, nor has it led to pronounced diversifying selection acting on homoeologous genes.

Phylogenetic Analysis of Low-Copy Genes

To evaluate the evolutionary history of low-copy genes in the sequenced regions more carefully, we estimated phylogenies for 15 low-copy genes conserved between soybean H1 and H2, indicated by letters A through O in Figure 2A and Supplemental Figure S2A (see Supplemental Table S3 for individual gene names and descriptions). Maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference (BI) phylogenetic analyses were conducted on all homologues of these 15 low-copy genes (Supplemental Fig. S4). Data sets included all taxa and homoeologous segments, where possible, ranging from 6 to 13 gene sequences per analysis. Phylogenies for each gene were mostly congruent across phylogenetic method, with the exception of genes C and D, for which MP weakly disagreed with ML and BI. All 15 gene phylogenies were consistent with the expected relationships among genes from *Glycine* homoeologues, *Phaseolus*, and *Medicago* (Fig. 1; Supplemental Fig. S4), thus corroborating the assignment of BACs to homoeologues in *Glycine*. These trees also confirmed that low-copy genes on H2 are evolving at roughly the same rate as their homoeologues on H1. Tajima relative rate tests using all sites showed that the branch lengths subtending H1 and H2 did not differ significantly in length for 13 of 15 low-copy genes (Kumar et al., 2004). Gene E showed a modest acceleration in H1 relative to H2 ($P = 0.02$), whereas gene M showed acceleration in H2 relative to H1 ($P = 0.04$).

These phylogenetic analyses also revealed lineage-specific expansion and loss of low-copy genes between both species and homoeologues. For example, gene M, which belongs to the root nodulin MtN21 protein family, has been tandemly duplicated in *Phaseolus*. This event took place after the split between *Phaseolus* and *Glycine*, as the two *Phaseolus* copies are positioned sister to each other in the phylogenetic tree (Supplemental Fig. S4). Similarly, gene E (encoding a member of the 12-oxophytodienoate reductase protein family) has been duplicated independently in *Phaseolus* and in H1 of *G. tomentella* (Gtd H1). The Gtd H2 homoeologue of this gene has been mostly lost, with only a small part (16%) of the gene remaining. The presence of multiple transposable elements surrounding this gene is suggestive of transposable element-mediated gene loss through recombination.

In addition to gene loss possibly mediated by transposable element insertions, pseudogenization is evident in H2. The Gtd H2 homoeologue of gene J exhibits premature stop codons, suggesting a loss of gene

Table I. *Mean Ka and Ks values for low-copy gene comparisons*

Values above the diagonal are Ka and values below are Ks. Numbers in parentheses indicate total number of genes compared. Ka and Ks values for individual gene comparisons can be found in Supplemental Table S1. Gmw, *G. max* cv Williams 82; Gmp, *G. max* line PI96983; Gtd, *G. tomentella* diploid accession G1403; Pva, *P. vulgaris* Andean accession G19833; Mth, *M. truncatula* var Jemalong.

Gene	Gmw H1	Gmw H2	Gmw H3	Gmp H1	Gmp H2	Gtd H1	Gtd H2	Pva	Mth
Gmw H1	–	0.031 (15)	0.136 (7)	0.0024 (15)	0.028 (10)	0.022 (15)	0.033 (8)	0.066 (13)	0.132 (7)
Gmw H2	0.123 (15)	–	0.138 (7)	0.033 (15)	0.0003 (10)	0.036 (15)	0.021 (8)	0.066 (13)	0.137 (7)
Gmw H3	0.488 (7)	0.459 (7)	–	0.137 (7)	0.132 (5)	0.138 (7)	0.144 (4)	0.138 (6)	0.170 (2)
Gmp H1	0.0006 (15)	0.122 (15)	0.488 (7)	–	0.032 (10)	0.025 (15)	0.033 (8)	0.069 (13)	0.141 (6)
Gmp H2	0.126 (10)	0.0029 (10)	0.453 (5)	0.125 (10)	–	0.036 (10)	0.0207 (6)	0.063 (8)	0.132 (4)
Gtd H1	0.064 (15)	0.140 (15)	0.494 (7)	0.065 (15)	0.150 (10)	–	0.034 (8)	0.073 (13)	0.124 (6)
Gtd H2	0.121 (8)	0.064 (8)	0.447 (4)	0.123 (8)	0.060 (6)	0.126 (8)	–	0.064 (7)	0.150 (5)
Pva	0.260 (13)	0.256 (13)	0.500 (6)	0.261 (13)	0.256 (8)	0.260 (13)	0.240 (7)	–	0.119 (5)
Mth	0.487 (7)	0.493 (7)	0.798 (2)	0.488 (7)	0.455 (4)	0.494 (7)	0.521 (5)	0.543 (6)	–

function and transition to a pseudogene. The release of this gene from selective constraints is apparent from its inflated branch length relative to its soybean homologues (Supplemental Fig. S4), implying a higher rate of molecular evolution after speciation. This is also reflected in Ka/Ks ratios. The average Ka/Ks ratio of all pairwise gene J comparisons not involving the Gtd H2 gene J (0.387 ± 0.0075) is statistically different (*t* test, $P < 0.0001$) and smaller than the average of comparisons involving Gtd H2 gene J (0.624 ± 0.0307), implying less selective constraint on Gtd H2 gene J.

Collinearity between Homoeologues Breaks Down around NB-LRR Genes

Although collinearity between H1 and H2 of soybean is quite good for low-copy genes (Fig. 2B; Supplemental Fig. S2B), this is not the case for most NB-LRR genes. NB-LRR genes are commonly grouped into two major phylogenetic subclasses, those containing an N-terminal domain homologous to the Toll and Interleukin-1 receptor (TIR) and those lacking a TIR domain (non-TIR; Meyers et al., 1999). The non-TIR subclass appears to be the more ancient one based on phylogenetic diversity and is found throughout eudicots and monocots and even in some gymnosperms (Meyers et al., 1999). We found both TIR and non-TIR-NB-LRR genes in both the soybean H1 and H2 regions (pink and red boxes, respectively, in Fig. 2A and Supplemental Fig. S2A). TIR genes were confined to a single cluster on the right side of both H1 and H2, in a shared syntenic position relative to low-copy genes. H1 contains seven genes in this cluster, while H2 contains four (Fig. 2A; Supplemental Fig. S2A). In contrast to the TIR genes, the non-TIR-NB-LRR genes are broadly distributed across the H1 and H2 regions and show very poor collinearity. In particular, H2 has reduced numbers of non-TIR-NB-LRRs in the left half of the aligned region relative to H1 (two copies on H2 versus seven on H1). Notably, the cluster on H1 that contains *Rpg1-b* appears to be completely absent from H2 (Fig. 2A; Supplemental Fig. S2A).

In contrast to the left half of the alignment between soybean H1 and H2, in the right half of the alignment H2 contains more non-TIR-NB-LRR genes than H1 (16 copies in H2 versus seven in H1). Although there are a few conserved low-copy genes scattered through this region, the relative positions of the non-TIR-NB-LRR genes do not appear to be conserved in relation to these low-copy genes. The lack of collinearity of the non-TIR-NB-LRRs and their differences in copy number argue for relatively frequent gene gains and losses, which is investigated further below.

Low-Copy Genes Are Well Conserved between Soybean and *G. tomentella*, whereas NB-LRRs and Retroelements Show Major Differences

Alignment of the soybean H1 and H2 regions with the orthologous regions of *G. tomentella* revealed a high level of conservation of low-copy genes but major differences in NB-LRR content and retroelement content as well as differences in copy number of a family of protein kinases on H1 (Fig. 2B). Low-copy gene order on *G. tomentella* H1 is nearly identical to that of soybean. Analysis of 15 conserved gene pairs gave a mean Ks value of 0.064 ± 0.034, consistent with a divergence time of 5 to 7 mya for these two species (Supplemental Table S2). Interestingly, there are examples of low-copy gene loss unique to *G. tomentella* H2 not observed in soybean H2, indicating that homoeologue-specific gene loss has continued subsequent to the divergence of these two species, albeit at a slow rate.

As described above, *G. tomentella* H2 has accumulated a large number of retroelements compared with H1, similar to what was observed in soybean. However, the precise locations of retroelement insertions in *G. tomentella* differ from those in soybean, with the distances between any two low-copy genes varying substantially between these two species. Consistent with this, many of the retroelements in *G. tomentella* have intact LTRs and represent insertions that occurred within the last 4 million years (Wawrzynski et al., 2008), after the split between *G. tomentella* and soybean. Although older insertions are almost certainly present, they are difficult to detect due to the degradation of LTR sequences caused by insertions and deletions.

Differences between Soybean Accessions PI96983 and Williams 82 Are Mostly Confined to NB-LRRs

We also compared soybean cv Williams 82 with a second accession of soybean, PI96983, which was selected because it differs from Williams 82 functionally at several disease resistance loci mapped to the *Rpg1* region (Ashfield et al., 1998; Hayes et al., 2000; Jeong et al., 2001; Gore et al., 2002). A study of single-nucleotide polymorphisms in soybean, however, has shown that domesticated soybean has undergone a significant genetic bottleneck, with total variation at the single-nucleotide polymorphism level being much lower than that observed in maize (Hyten et al., 2006). To see if this lack of variation also applied to a region rich in disease resistance genes, we compared the Williams 82 sequence with its allelic regions in PI96983. As can be seen in Figure 2C, the order of low-copy genes is nearly perfectly conserved between these two accessions. There are differences, however, in the number and position of NB-LRRs and in a small gene family that flanks some NB-LRRs. These differences are likely due to unequal crossover events in tandemly repeated gene clusters, as the differences are confined to clusters of two or more NB-LRRs. There are also a few differences in retrotransposon insertions (Fig. 2C; Supplemental Fig. S2C), which, combined with analysis of LTR sequences (Wawrzynski et al., 2008), indicates that these are likely very recent insertions.

Comparison of Soybean and *M. truncatula*

We also compared the soybean cv Williams 82 H1 and H2 sequences with the recently released *M. truncatula* genome sequence. *Medicago* diverged from *Glycine* about 50 mya, or about 30 mya before the *Phaseolus/Glycine* split (Fig. 1). A region on *Medicago* chromosome 8 was found to share synteny with the H1 and H2 sequences, with the similarity to H1 extending over approximately 700 kb in soybean and 900 kb in *Medicago* (Fig. 2D). This alignment revealed islands of well-conserved low-copy gene order but very poor conservation of NB-LRRs. Also notable is the expansion of three different gene families in *Medicago*: protein kinases, sulfotransferases, and signal peptidases (Fig. 2D). The complete absence of NB-LRRs from the left three-quarters of the aligned region in *Medicago* is particularly striking, given the presence of shared low-copy genes flanking this region. It is not clear whether the *Medicago* lineage lost these NB-LRRs or whether NB-LRRs were inserted in the lineage that gave rise to *Phaseolus* and *Glycine*. Significantly, of all the NB-LRRs found in the *Medicago* genome, the two most similar to soybean *Rpg1-b* are located within 400 kb of the aligned region in *Medicago* (Fig. 3). As discussed in more detail below, the subfamily of NB-LRRs that includes *Rpg1-b* is distributed over nearly 500 kb in soybean, which may account for the seeming lack of collinearity between the Medicago *Rpg1-b* homologues and soybean *Rpg1-b*. On the right side of the alignment shown in Figure 2D, there is a cluster of TIR-NB-LRRs that is located in a syntenic position in *Medicago* and soybean. Consistent with this syntenic position, these *Medicago* genes are the most closely related to these soybean TIR-NB-LRRs of any NB-LRRs in the *Medicago* genome (data not shown). Phylogenetic analysis indicates that this cluster in both *Medicago* and *Glycine* has undergone duplications subsequent to the split between *Medicago* and *Glycine* but that the common ancestor of *Medicago* and *Glycine* contained at least two TIR-NB-LRRs at this position, which gave rise to the members present in this cluster in both species today (Fig. 3, B and C).

Comparison of Soybean H3 with H1 and H2

As mentioned above, we identified a set of BACs in the Williams 82 library that contained homologues of several H1 low-copy genes but appeared to represent a much older duplication event than the divergence of homoeologues 1 and 2. To estimate the time of this duplication, we determined Ks values for all gene duplicates (seven pairs). The average Ks value for H1:H3 comparisons was 0.488 ± 0.098 (Supplemental Table S2), which compares with an average Ks value of 0.122 ± 0.035 for H1:H2 comparisons. Thus, the older duplication (H3) is roughly four times as old as the H2 duplication, which would place the duplication event approximately 40 to 50 mya. *Medicago* and *Glycine* share a presumed genome-wide duplication event estimated to have occurred 50 to 60 mya (Fig. 1; Mudge et al., 2005; Shoemaker et al., 2006); thus, the H3 BACs identified by us are most likely derived from that event.

Alignment of H3 with H1 and H2 also revealed conserved synteny (Fig. 2E). Of the 35 non-NB-LRR genes found on H3 (counting tandemly duplicated carbohydrate transporter and kinases as single genes), 11 are found on H1, H2, or both. Notably, H3 contains a small cluster of non-TIR-NB-LRRs in a position roughly equivalent to the *Rpg1-b* cluster in H1. Because the H3 duplication predates the divergence of *Medicago* and *Glycine* (Pfeil et al., 2005), this finding further supports our conclusion that NB-LRRs were present in this region in the common ancestor of *Medicago* and *Glycine*.

Phylogenetic Analysis of NB-LRRs Reveals that NB-LRR Clusters Have Been Conserved since before the Divergence of *Phaseolus* and *Glycine*, Yet Are Undergoing a Rapid Birth and Death Process

To understand the evolution of the NB-LRRs in the sequenced region better, we performed phylogenetic analyses using the NB region of each NB-LRR, which is the most highly conserved domain. We first divided the NB-LRRs into TIR and non-TIR subclasses. The NB regions from all members of each subclass from all of the taxa sequenced were aligned and then analyzed for potential recombination events (see "Materials and Methods"). Any genes with apparent recombination events in the NB region were eliminated from the phylogenetic analysis because recombination mixes different histories and leads to misleading phylogenetic inferences. Based on these analyses, we eliminated one non-TIR gene and one TIR gene that showed evidence of recombination within the NB region. The remaining 89 genes were then subjected to Bayesian analysis, and a tree was constructed for each subclass (Fig. 3).

Several important conclusions can be drawn from analysis of Figure 3. First, there clearly has been recent expansion of non-TIR-NB-LRR clusters in *Glycine* H1, *Glycine* H2, and *Phaseolus*, as evidenced by terminal branches with multiple closely related genes from individual taxa. Although most of this recent expansion occurs as tandemly repeated genes (note clusters of similarly colored boxes in Fig. 3), there are two clear cases where recent duplications are spread over several hundred kilobases. This is best illustrated by the soybean *Rpg1-b* subclade on H1 (shown in purple in Fig. 3). *Rpg1-b* is located approximately 200 kb away from the highly similar gene W21F22_29. The latter gene is located approximately 50 kb away and in the opposite orientation from another highly similar gene, W221b6_21, which is >100 kb away from a fourth highly similar gene, W10n21_4. In addition, genes belonging to the 42i18_2 subclade (shown in blue in Fig. 3) are spread over several hundred kilobases of soybean H1. In both examples, there are NB-LRRs from other clades intermixed, as indicated by the

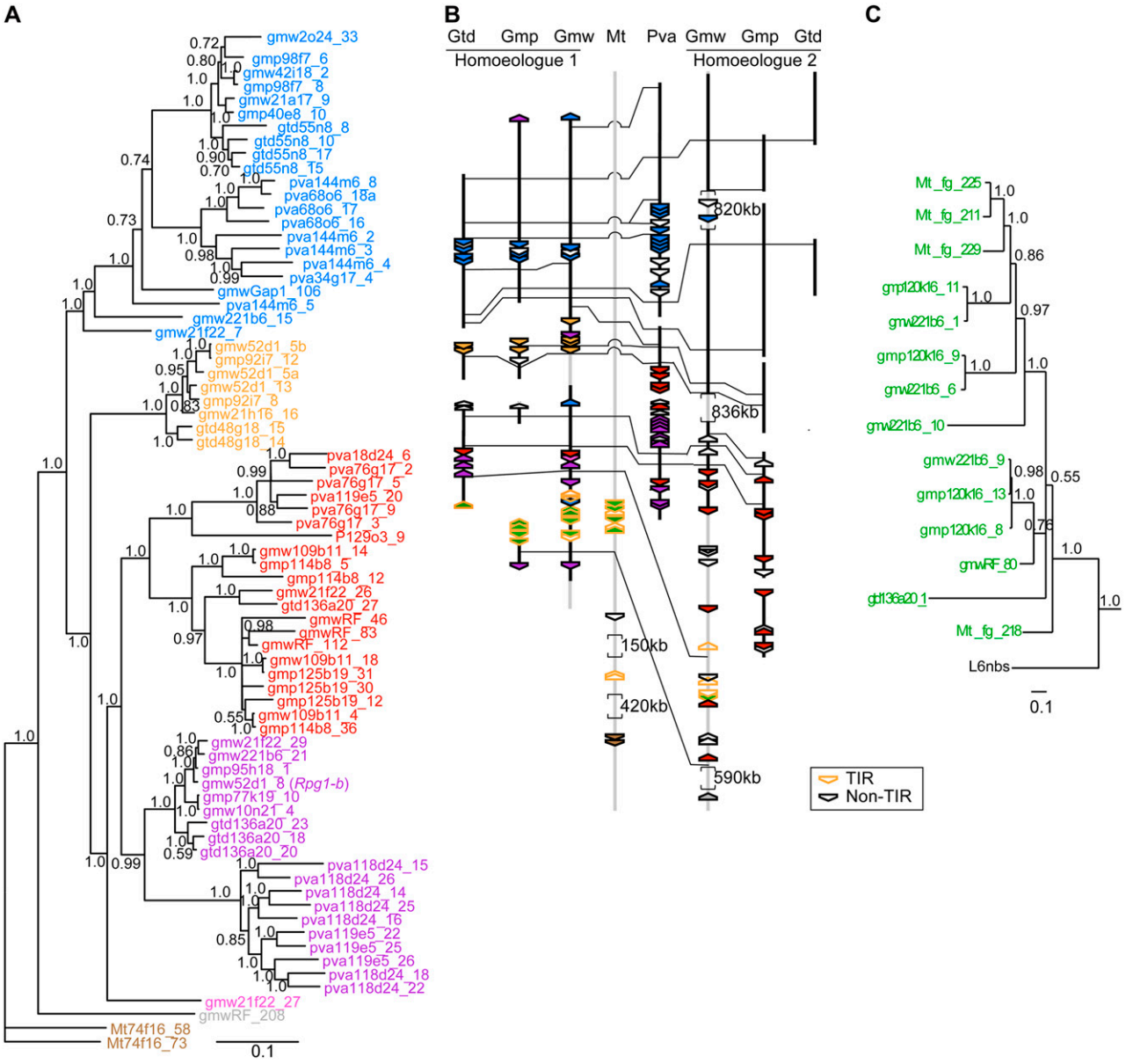

Figure 3. Phylogenetic analysis and physical map of NB-LRR genes. A, Bayesian tree of non-TIR-NB-LRRs of the *Rpg1-b* syntenic region. This tree was constructed using just the NB domains (from the P-loop to the MHD motif). NB-LRR names are colored according to major clades. Gene names are abbreviated as follows: gmw, *G. max* cv Williams 82; gmp, *G. max* line PI96983; gtd, *G. tomentella* diploid accession G1403; pva, *P. vulgaris* Andean accession G19833; Mt, *M. truncatula* var Jemalong. B, Physical map of NB-LRRs of the *Rpg1-b* syntenic region. Vertical black lines represent sequenced BAC contigs from the different genotypes and homoeologues. Vertical gray lines represent chromosome 8 of *Medicago* and WGS 7x scaffolds 172 and 55 of homoeologues 1 and 2 from soybean cv Williams 82. Arrows represent predicted TIR (orange borders) or non-TIR (black borders) NB-LRR genes and their orientation. Fill colors of non-TIR-NB-LRR genes correspond to the colors of the clades defined by the non-TIR tree (A). White NB-LRRs are pseudogenes with partial or total deletion of the NB region, or genes that showed evidence of recombination within the NB region, and thus were not included in the phylogenetic analysis. Horizontal lines represent low-copy genes conserved between species, genotypes, or homoeologues, which were used to align these regions. C, Bayesian tree of TIR NB-LRRs of the *Rpg1-b* syntenic region.

differently colored boxes. How this pattern of gene duplication arose is difficult to explain, but it is unlikely to be a result of unequal crossing over between NB-LRRs, as such events should eliminate the collinearity of intervening low-copy genes when comparing soybean with *G. tomentella* or *Phaseolus*, but collinearity has been maintained.

The phylogenetic and physical analyses revealed that multiple NB-LRR lineages in this region predate the split between *Phaseolus* and *Glycine*. For example,

the blue, purple, and red clades in Figure 3 are all shared between *Phaseolus* and *Glycine*. Interestingly, all three clades have undergone relatively recent expansion in *Phaseolus* as a result of tandem duplication events. In contrast, *Glycine* H2 appears to have lost the purple clade completely and reduced the number of blue clade members. However, there are more members of the red clade on H2 than on H1 in soybean, suggesting that there may have been some partitioning of NB-LRR subfamilies between homoeologues subsequent to the most recent polyploidy event.

The phylogenetic analysis also revealed a clade (represented in orange in Fig. 3) that appears to have recently expanded on *Glycine* H1 but that is absent from both *Phaseolus* and *Glycine* H2. Its absence from both *Phaseolus* and H2 suggests that this clade arose subsequent to the H1:H2 split or, alternatively, that it has been lost from both the *Phaseolus* and H2 lineages. Because this orange clade is positioned between the blue clade and the red/purple clades on the phylogenetic tree (Fig. 3), and all three of the latter clades predate the *Phaseolus*/*Glycine* split, the most likely scenario is that this lineage has been lost independently from this region in *Phaseolus* and *Glycine* H2.

There also appears to have been gene loss in *Medicago*, particularly of the non-TIR-NB-LRR class, or at least a failure to expand. The *Medicago* NB-LRR genes most similar to *Rpg1-b* (Mt74f16_58 and Mt74f16_73) are located approximately 500 kb away from the aligned regions (brown arrows at the bottom of the *Medicago* line in Fig. 3) and group outside all of the non-TIR-NB-LRR genes in the phylogenetic tree shown in Figure 3. A BLAST search of the complete 7x WGS soybean genome sequence using these two *Medicago* NB-LRRs as queries failed to find any additional soybean genes that were more similar than those on H1 and H2; thus, it is unclear whether *Medicago* underwent any significant losses of non-TIR-NB-LRRs or whether these have just expanded in *Glycine* and *Phaseolus*.

Other Multicopy Gene Families Are Evolving More Slowly Than the NB-LRR Family

To determine whether the rapid evolution of the NB-LRR families is a general property of clustered homologous genes or is confined to NB-LRRs in this region, we performed a phylogenetic analysis on two non-NB-LRR gene families that are also located within the sequenced region. The first gene set encodes a family of putative carbohydrate transporters and is represented by four tandemly repeated genes on H1 of both soybean cultivars sequenced and on H1 of *G. tomentella* (orange boxes in Fig. 2 and Supplemental Fig. S2). Note that in Figure 2B and Supplemental Figure S2B, *G. tomentella* paralogues 1 and 2 are not shown because they fall into a gap in the *G. tomentella* BAC contig. The presence of paralogues 1 and 2 in this position was confirmed by screening a second *G. tomentella* BAC library derived from a tetraploid accession, G1134, and then sequencing a homologous BAC. Based on phylogenies, the four carbohydrate transporter genes arose by tandem duplication sometime prior to the split of soybean and *G. tomentella* (Fig. 4A). This observation indicates that, unlike NB-LRRs, these repeated genes are quite stable and are not undergoing rapid birth and death (none since the split of soybean and *G. tomentella* 5–7 mya). It also indicates that there has not been much recombination and gene conversion occurring between these four copies, which should result in concerted evolution and a loss of the orthologous relationships within each of the four gene pairs in *G. tomentella* and soybean.

H2 in soybean contains just one carbohydrate transporter gene. This gene (gmw2-12n11_11) is sister to copy 1 of the Gmw H1 cluster (gmw1-173d12_7), rather than falling outside the four Gmw H1 genes. This arrangement suggests that there may have been four copies of this gene present in the common ancestor of H1 and H2 and that H2 has lost three of the four copies during the last 10 million years. Consistent with this hypothesis, Ks analysis of the four Gmw H1 copies indicates that these duplications arose between 17 and 27 mya (Ks = 0.199–0.297), well before the polyploidy event and spanning the time of divergence between *Phaseolus* and *Glycine*, while the Ks value for the H1 (gmw1-173d12_7):H2(gmw2-12n11_11) comparison is 0.144, as expected for the most recent whole genome duplication.

H3 in soybean also contains a tandem cluster of three carbohydrate transporter genes in the syntenic position (Fig. 2E), suggesting that this cluster dates back at least 50 mya. However, in the phylogenetic tree, these three H3 copies group sister to each other, which indicates that the duplications on H3 occurred after the 50-mya genome duplication event, or alternatively, that there has been some level of recombination/gene conversion occurring among the H3 copies such that the sequences have become homogenized. Ks analysis indicates that gene gmw2-129e12_19 diverged from the other two genes approximately 50 mya (Ks = 0.575–0.611), which is near the same time that the ancient polyploidy event is thought to have occurred. Meanwhile, comparison of the two most closely related genes (gmw2-129e12_14 and gmw2-129e12_16) gives a Ks of 0.205, indicating that this duplication occurred well after the ancient polyploidy event. The large differences in Ks values suggest that recombination among the three genes has been minimal and that there has been independent birth and death occurring in this cluster on H1, H2, and H3, albeit over much larger time frame than was observed for the NB-LRR family.

Phaseolus contains at least two tandem copies (pva1-47b16_2 and _3) of the carbohydrate transporter gene in the syntenic position (Fig. 2A; additional copies could be present in the adjacent gap). Based on branch length (Fig. 4), both of these copies appear to have been evolving at a faster rate than the most closely related *Glycine* homologues (paralogue 2, which in-

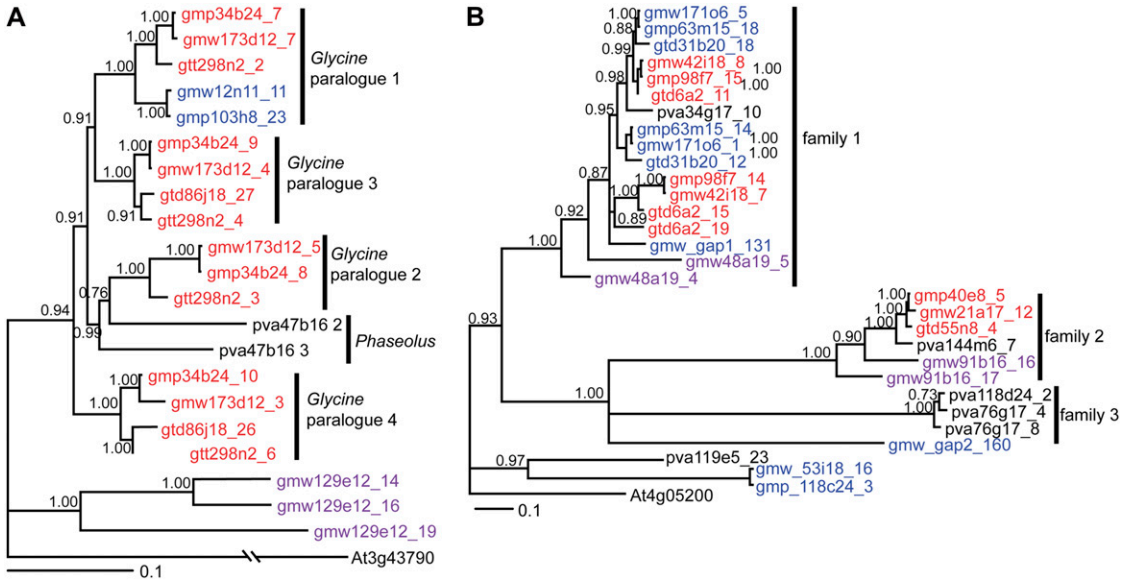

Figure 4. Phylogenetic analysis of carbohydrate transporter and kinase families. A, Carbohydrate transporter family. B, Kinase family. Red text indicates genes from *Glycine* H1, blue text indicates *Glycine* H2, and violet text indicates *Glycine* H3. Gene names are abbreviated as in Figure 3A, with the addition of gtt indicating *G. tomentella* tetraploid species G1134 and At indicating Arabidopsis. Both trees were constructed using Bayesian analysis. Numbers at nodes indicate posterior probabilities.

cludes gmw2-173d12_5). A Tajima relative rates test of each member of the *Glycine* clade confirmed this conclusion ($P < 0.01$ using gmp1-34b24_7 or _9 as the outgroup). The topology of Figure 4A suggests that one of the two *Phaseolus* copies is orthologous to the three *Glycine* paralogue 2 genes, with the other gene representing a duplication lost from *Glycine*. However, the low posterior probability of this node (0.76; Fig. 4A) makes it possible that both genes are coorthologous to the three genes in the *Glycine* clade.

We also analyzed a collection of kinase genes, which are more broadly distributed across soybean H1 (aqua boxes in Fig. 2 and Supplemental Fig. S2). We aligned the kinase domains of all kinases identified in all of the sequenced BAC clones and constructed a phylogenetic tree using Bayesian analysis. This tree revealed three distinct kinase families, one of which was specific to *Phaseolus* (Fig. 4B). Each member of the unique *Phaseolus* kinase family (family 3 in Fig. 4B) is located adjacent to an NB-LRR gene (pva1-76g17_5, pva1-76g17_9, and pva1-118d24_1 in Supplemental Fig. S2A), suggesting that the recent expansion of this kinase family may be related to its location amid an R gene cluster.

The largest kinase family (family 1) included members from *Phaseolus* and all three soybean homoeologues as well as from H1 and H2 of *G. tomentella*. Two of the genes from this family are located adjacent to each other in soybean but oriented in opposite directions (Fig. 2A; Supplemental Fig. S2A; genes gmw1-42i18_7 and gmw1-42i18_8). Phylogenetic analysis (Fig. 4B) revealed that gene gmw1-42i18_8 has a clear orthologue in *G. tomentella* (gtd1-6a2_11), a clear homoeologue on soybean H2 (gmw2-171o6_5), and a clear orthologue in *Phaseolus* (pva1-34g17_10; Fig. 4B), all of which are located in similar syntenic positions, indicating that this gene has been conserved for at least 20 million years. The adjacent gene in soybean, gmw1-42i18_7, appears to be conserved in *G. tomentella* as well (orthologue is gtd1-6a1_15) and is oriented in the opposite direction to gtd1-6a2_11, just as in soybean (Fig. 2B). Thus, this gene pair has been quite stable, similar to the carbohydrate transporter family and unlike NB-LRR clusters. However, *G. tomentella* H1 contains an additional member of this kinase family (gtd1-6a11_19) adjacent to the inverted gene pair. Phylogenetically, this additional kinase is sister to the inverted gene pair; thus, it is likely from an earlier duplication event that may have been lost from soybean. H3 also contains two members of this family in the syntenic position (gmw2-48a19_4 and gmw2-48a19_5). Based on the tree topology, one copy (W48a19_5) is the homoeologue of all of the H1 and H2 genes in this family, consistent with H3 being derived from a much older duplication event. The second copy (W48a19_4) is derived from a still older duplication event.

The last kinase family (family 2) displays a phylogenetic pattern similar to that of the low-copy gene families described above, with clearly identifiable orthologues and homoeologues (Fig. 4B). Specifically, gene gmw1-21a17_12 from soybean cv Williams 82 has a clear orthologue in *Phaseolus* (pva1-144m6_7) and a homoeologue on soybean H3 (gmw2-91b16_16), all of

which occupy similar syntenic positions, indicating that this kinase has occupied this genomic position for over 50 million years. However, there is no H2 homoeologue present in this family, indicating that it has been lost from H2. The phylogenetic analysis also revealed a kinase subclade within family 1 that was unique to *Glycine* H2 (gmw2-171o6_1, gmp1-63m15_14, and gtd1-31b20_12) that did not appear to have homoeologues in H1 or an orthologue in *Phaseolus*. This is suggestive of a gene insertion event that occurred after the 10-mya *Glycine* polyploidy event and prior to the soybean-*G. tomentella* divergence.

In summary, although there have been some duplications and losses across taxa and homoeologues in both the carbohydrate transporter family and kinase families, overall they are not experiencing the high rates of birth, death, and translocation observed in the NB-LRR family. This observation suggests that the rapid evolution of NB-LRRs is not simply a by-product of their clustering. Intriguingly, the one exception to this general rule is a small kinase family in *Phaseolus* that appears to be undergoing duplications alongside an immediately adjacent NB-LRR.

DISCUSSION

The above analyses were designed to address multiple questions regarding the evolution of the soybean genome, particularly in regard to the effects of polyploidy. By comparing the sequences of two homoeologous regions within soybean with the single orthologous region in *Phaseolus*, we were able to hypothesize the polarity of most changes. In addition, by comparing the soybean cv Williams 82 sequences with allelic sequences in soybean line PI96983 and with orthologous sequences in *G. tomentella*, we were able to gain insights into the tempo and mode of changes that are driving homoeologous sequences apart in soybean.

The most significant insights arising from our analyses relate to chromosomal rearrangements within *Glycine* that placed the H2 homoeologue adjacent to a centromere. Whether this occurred as a result of polyploidy is unknown. Homoeologous recombination is known to occur in polyploids (Gaeta et al., 2007) and may be particularly prevalent during the early stages after formation, when the genome is adjusting to the "genomic shock" of doubling and (in the case of allopolyploids) hybridization (Comai, 2000). However, if *Glycine* is an allopolyploid, chromosomal rearrangement could have taken place in the species that gave rise to H2, prior to the polyploid event.

The relocation of the H2 region adjacent to a centromere is correlated with a dramatic increase in retrotransposon content in H2. Although it is well known that plant and animal centromeres are enriched in repetitive sequences, including retrotransposons (Lin et al., 2005; Ma et al., 2007), there is little information on how centromeres become this way or on the rate at which repetitive elements accumulate in pericentromeric regions. Our data suggest that some property of centromeric location promotes rapid retrotransposon accumulation, since the ancestral state of this region appears to have been relatively low in retrotransposon content prior to its translocation to a centromeric region. Because the split between H1 and H2 occurred only about 10 mya, this accumulation of retrotransposons can apparently occur quite rapidly. Possible mechanisms include a chromatin structure that causes a decrease in retrotransposon deletion events (e.g. via recombination among LTRs) or an increase in retrotransposon insertion events or both.

Expansion of retroelement content in the centromeric region may be selected for during meiosis, as these repeats are thought to promote microtubule binding, which then increases the frequency that a given chromosome ends up in the egg nucleus (Henikoff et al., 2001; Henikoff and Malik, 2002; Malik and Henikoff, 2002). Competition between homologous chromosomes for segregation into the egg nucleus would thus favor chromosomes with a greater number of repeats in the centromere and lead to relatively rapid expansion. In the *Glycine* H2 region, this expansion began during the last 10 million years and appears to be continuing based on the dates of retroelement insertions and the relatively low frequency of solo LTR sequences, which are indicative of retrotransposon deletion by homolologous recombination between LTRs (Wawrzynski et al., 2008).

In addition to containing high levels of repetitive DNA, pericentromeric regions typically are heterochromatic in structure (i.e. highly condensed). As a consequence, pericentromeric regions are often regarded as being low in both gene content and gene expression. Therefore, it is striking that the majority of the H2 low-copy genes have been conserved and continue to be expressed. Thus, pericentromeric location, per se, does not cause gene silencing or rapid loss of genes. Moreover, the synonymous mutation rates (Ks) observed in H2 low-copy genes were nearly identical to the rates observed in their homoeologues on H1 (Supplemental Table S2); thus, pericentromeric location does not alter substitution frequency, either.

Instead, our data indicate that synonymous substitution frequencies are determined by intrinsic properties of individual genes rather than extrinsic forces such as genomic context. We observed a wide variation in Ks values for individual low-copy genes on soybean H1 when comparing these genes with *Phaseolus* (2.75-fold; Supplemental Table S2), but there was little difference between homoeologous pairs (maximum fold difference of 1.29). Thus, the properties of individual genes that determine synonymous substitution rates must be conserved following polyploidization. Although the cause of gene-to-gene synonymous variation is unknown, Zhang and colleagues (2002) ruled out base composition, codon usage bias, and other selectively driven explanations, leaving expres-

sion level as a leading explanation for synonymous rate variation among Arabidopsis homoeologues.

In the soybean regions that we analyzed, the majority (approximately 77%) of low-copy gene duplicates derived from the most recent polyploidy event have been maintained over the course of 10 million years. A similarly high level of duplicate retention was also observed in two prior comparisons of homoeologous soybean BAC sequences (Schlueter et al., 2006; Van et al., 2008), indicating that this level of duplicate retention is not unique to the *Rpg1* region. Homoeologous genes in the regions we analyzed appear to be under purifying selection, which is also true for other homoeologous regions in soybean (Schlueter et al., 2006; Van et al., 2008). The level of retention observed in our region is higher than would be predicted based on the Arabidopsis genome (Maere et al., 2005), where homoeologous regions have experienced considerable and biased gene loss (Thomas et al., 2006). For a Ks of 0.12, the models of Maere and colleagues (2005) predict that as many as 40% of homoeologous genes should have been lost. There are many hypotheses for why retention of homoeologous copies in polyploids is higher than would be predicted given the expected loss of redundant genes. The evolution of new functions (neofunctionalization) or the partitioning of existing functions between gene copies (subfunctionalization) are potential outcomes for any duplicate gene pair and lead to the selection-driven preservation of both copies (Lynch and Force, 2000). The key feature of polyploids is the duplication of all genes, leading to theories concerning concerted divergence of expression networks (Blanc and Wolfe, 2004) and dosage balance hypotheses (Thomas et al., 2006). Testing these hypotheses for the soybean genome will require extensive data on the expression of homoeologous copies.

A 77% retention rate for homoeologous gene duplicates in soybean appears to be much higher than that reported for maize (Bruggmann et al., 2006), where only 20% to 35% of homoeologous duplicates have been retained, even though the divergence time of the maize homoeologues (approximately 11.9 mya; Swigonova et al., 2004) is roughly the same as that of the soybean homoeologues (approximately 10 mya). The maize retention rate estimate comes from a study similar to ours in which a 6.6-Mb region of maize chromosome 9 was aligned with the homoeologous region on chromosome 1 and with the single orthologous region in rice (chromosome 3). Of the 475 predicted nonrepetitive maize genes in this study, 133 (28%) were not present in either the rice orthologous region or the maize homoeologous region. Significantly, of these 133 genes, 90 (68%) were found to be located elsewhere in the rice genome, suggesting that these genes had undergone a transposition event in either maize or rice. It is thus possible that the reduced rate of apparent duplicate retention in maize relative to soybean is partly an artifact of increased gene mobility in maize whereby collinearity is lost, but duplicates in fact are being retained. Regardless, the soybean genome appears to have retained gene collinearity following polyploidy to a much greater extent than has maize over a similar time scale.

The maize study also analyzed differences in retrotransposon content between homoeologues (Bruggmann et al., 2006). As we observed in soybean, the two maize homoeologues differed significantly in location of retroelements and in local expansion of retroelement copies. Overall, maize chromosome 1 in this region has expanded significantly more than maize chromosome 9, with approximately 2.5 Mb of chromosome 9 aligning with 7.6 Mb of chromosome 1. The majority of the extra sequence in chromosome 1 is due to retrotransposon content (Bruggmann et al., 2006). Thus, in both soybean and maize, homoeologous chromosomes appear to follow independent paths relative to retrotransposon accumulation. As in soybean, the vast majority of intact maize LTR retrotransposons were inserted subsequent to the polyploidy event, but this likely reflects only the high rate of decay that retrotransposons undergo after insertion (Ma et al., 2004) rather than increased activity following polyploidy.

In contrast to the low-copy genes, NB-LRR-encoding genes are not well conserved between soybean H1 and H2. Most notably, there appears to have been reciprocal loss of these genes between H1 and H2. Loss of NB-LRR-encoding genes following polyploidy appears to be a general phenomenon, as NB-LRRs are highly underrepresented in duplicated regions of the Arabidopsis genome (Cannon et al., 2004; Nobuta et al., 2005). This observation suggests that there may be a fitness cost associated with carrying excess NB-LRRs. In the case of allopolyploidy, this fitness cost may in part be caused by an autoimmune-like phenomenon. Most NB-LRR proteins are believed to detect pathogens by detecting pathogen-induced modification of other host proteins (Innes, 2004). NB-LRRs likely coevolve with these other host proteins. When genomes are combined by allopolyploidy, proteins from one genome may be recognized by NB-LRRs from the other genome as being modified by a pathogen. Recently, Bomblies et al. (2007) showed that 2% of intraspecific crosses between Arabidopsis ecotypes yielded F1 plants that displayed necrotic phenotypes similar to those induced by pathogen recognition. Analysis of one of these crosses established that necrosis was dependent on an NB-LRR-encoding gene from one of the parents and a second gene from the other parent. Thus, there is a precedent for the involvement of NB-LRRs in incompatibilities between genotypes or genomes. Such self-recognizing NB-LRRs should be quickly selected out of the population in subsequent generations. Further support for this autoimmunity model comes from work on the *RPM1* gene of Arabidopsis, which was found to exert a fitness cost when transgenically moved from one Arabidopsis genotype into a second Arabidopsis genotype (Tian et al., 2003).

Although the loss of NB-LRRs primarily occurred from H2, one NB-LRR subfamily was preferentially lost from H1 compared with H2 (red genes in Fig. 3). Such partitioning of NB-LRR subfamilies between homoeologues should facilitate sequence divergence, as it would be expected to reduce unequal crossover and gene conversion events between NB-LRR copies (Mondragon-Palomino and Gaut, 2005), forces that would tend to homogenize tandemly arrayed genes. Thus, partitioning of NB-LRRs between homoeologues could be a mechanism for preserving diversity.

Independent of polyploidy, our data show that NB-LRR gene clusters in both *Glycine* and *Phaseolus* are rapidly evolving, as evidenced by the phylogenetic trees shown in Figure 3. These trees display clusters of genes from the same taxa at terminal nodes, indicating recent duplication events. In addition, alignment of the two soybean genotypes shows significant changes in NB-LLR gene number and arrangement (dashed red boxes in Fig. 2C and Supplemental Fig. S2C). Rapid birth and death of NB-LRR genes have been observed in many plant species and are usually attributed to unequal crossover events, both within and between NB-LRR genes in a cluster (Michelmore and Meyers, 1998; Noel et al., 1999; Chin et al., 2001; Nagy and Bennetzen, 2008). We are currently analyzing the NB-LRR genes in our data set and their flanking sequences to assess the impact of various recombination events on the evolution of this gene family. These analyses will be reported in a subsequent paper.

Although the tandem arrangement of NB-LRR genes is thought to be necessary for the rapid birth and death events observed in NB-LRR clusters (e.g. to facilitate unequal crossover events), it must not be sufficient, because we observed a cluster of carbohydrate transporter-like genes in the *Glycine* H1 contigs that are surprisingly stable, with four copies being maintained since at least the split between soybean and *G. tomentella* (Fig. 4). In addition, phylogenetic analysis indicates that there has been little, if any, concerted evolution occurring among these four genes (Fig. 4), suggesting that gene conversion events are rare. It is unclear at present why some tandem gene clusters appear to recombine frequently and others do not. Phylogenetic analyses of 50 large gene families in the Arabidopsis genome revealed a large variation in apparent tandem duplication rates among families (Cannon et al., 2004). Interestingly, the majority of the most rapidly evolving families were associated with pathogen defense and included NB-LRR genes, the Major Latex Protein family (related to the pathogen-inducible PR10 proteins; Osmark et al., 1998), the Germin family (Membre et al., 2000), subtilisin-like Ser proteases (Jorda and Vera, 2000), and the pathogenesis-related PR1 family (Mitsuhara et al., 2008). These findings suggest that the apparent differences in duplication rates (and hence unequal crossover rates) may be partly explained by pathogen-mediated selection.

MATERIALS AND METHODS

BAC Libraries

All BAC libraries used in this project are available through the Clemson University Genomics Institute (https://www.genome.clemson.edu/cgi-bin/orders/). Two libraries of soybean (*Glycine max* 'Williams 82') were used in this project. The gmw1 library (CUGI GM_WBa) was constructed in R. Shoemaker's laboratory (Iowa State University) and contains 5.4 genome equivalents. The gmw2 library (CUGI GM_WBb) was constructed at the Clemson University Genomics Institute and contains 12 genome equivalents. The soybean PI96983 library (gmp1; CUGI GM_PBb) was constructed by BIO S&T and contains 6.8 genome equivalents. The *Glycine tomentella* diploid accession G1403 library (gtd1; CUGI GT_GBa) and tetraploid accession G1134 library (gtt1; CUGI GT_GBb) were also made by BIO S&T and contain 9.7 genome equivalents and 8 genome equivalents, respectively. The *Phaseolus vulgaris* accession G19833 library (pva1; CUGI PV_GBa) was made by Matthew Blair at the International Center for Tropical Agriculture and contains 12 genome equivalents. Additional library details, such as average insert sizes and restriction enzymes used, can be obtained from the Clemson University Genomics Institute Web site.

BAC Contig Assembly

Assembly of the H1 BAC contig from soybean cv Williams 82 was initiated during the cloning of the *Rpg1-b* disease resistance gene (Ashfield et al., 2003). Extending this BAC contig to span a full megabase region was accomplished using low-copy genes near BAC ends as PCR-based probes to screen the gmw1 and gmw2 BAC libraries. Positive BACs were end sequenced using BAC DNA as template and fingerprinted using a high-information-content fingerprinting protocol (Luo et al., 2003) and an ABI3730 automated DNA sequencing instrument (Applied Biosytems). Fingerprints were used to confirm overlapping BACs and to estimate BAC sizes. BACs that maximally extended the existing contig in both directions were selected for sequencing, then the entire process was repeated until the full megabase region was spanned. Despite screening two gmw libraries representing over 17 genome equivalents, we were left with a gap in our BAC contig located roughly in the middle. A single additional BAC was identified in the gmw1 library that appeared to extend into this gap (gmw1-21h16). However, the *Escherichia coli* strain carrying this BAC clone grew very slowly on Luria-Bertani agar plates and in Luria-Bertani liquid. BAC DNA preparations derived from different single-colony isolates of this clone showed different restriction digest patterns, indicating that this clone was undergoing deletion events during culture. The slow growth of this clone and its instability suggest that this region of the soybean genome contains a DNA sequence that is toxic to *E. coli*, which would account for its poor representation in BAC libraries.

After sequencing the Williams 82 BACs, we identified low-copy protein-coding genes conserved in Arabidopsis (see "Annotation Protocols" below). A low-copy number in soybean was verified by searching The Institute for Genomic Research (TIGR) soybean Transcript Assembly database (http://tigrblast.tigr.org/euk-blast/plantta_blast.cgi). A subset of these low-copy gene sequences were then used as DNA hybridization probes (Supplemental Table S1; Supplemental Fig. S1) to screen BAC libraries of soybean cv Williams 82, soybean line PI96983, *G. tomentella* diploid accession G1403, *G. tomentella* tetraploid accession G1134, and *P. vulgaris* accession G19833. BAC clones that hybridized to two or more probes were then fingerprinted and end sequenced. A combination of fingerprint information, probe hybridization patterns, and end sequence information was used to assemble contigs and identify a minimum tiling path for sequencing. For *G. tomentella* tetraploid accession G1134, only a single BAC containing the carbohydrate transporter gene family was analyzed.

BAC Sequencing

The detailed procedures for large-insert genomic DNA isolation, random shotgun cloning, fluorescence-based DNA sequencing, and subsequent analysis have been described previously (Bodenteich et al., 1993; Chissoe et al., 1995; Roe, 2004). Briefly, BAC DNA was isolated free from host genomic DNA via a cleared lysate-acetate precipitation-based protocol (Roe, 2004). Subsequently, 50-μg portions of purified BAC DNA were randomly sheared and made blunt ended (Sambrook et al., 1989; Bodenteich et al., 1993; Roe, 2004). After kinase treatment and gel purification, fragments in the 1- to 3-kb range

were ligated into *Sma*I-cut, bacterial alkaline phosphatase-treated pUC18 (Pharmacia), and *E. coli* strain XL1BlueMRF′ (Stratagene) was transformed by electroporation. A random library of approximately 1,200 colonies was picked from each transformation, grown in Terrific Broth medium (Sambrook et al., 1989) supplemented with 100 μg mL^{-1} ampicillin for 14 h at 37°C with shaking at 250 rpm, and plasmid DNA was isolated by a cleared lysate-based protocol using a Zymark SciClone robot (Bodenteich et al., 1993; Roe, 2004).

Sequencing reactions were performed as described previously (Chissoe et al., 1995) using *Thermus aquaticus* (*Taq*) DNA polymerase and either the Amersham ET fluorescence-labeled terminator or the Perkin-Elmer Cetus fluorescence-labeled Big Dye Taq terminator sequencing kit at a 1:16 dilution. The reactions were incubated for 60 cycles in a Perkin-Elmer Cetus DNA Thermocycler 9600, and after removal of unincorporated dye terminators by ethanol precipitation followed by a 70% ethanol wash, the fluorescence-labeled nested fragment sets were resolved by electrophoresis on ABI 3700 Capillary DNA Sequencers. After base calling with Phred (Ewing et al., 1998), the analyzed data were transferred to a Sun Workstation Cluster and assembled using Phrap (Ewing and Green, 1998). Overlapping sequences (contigs) were analyzed using Consed (Gordon et al., 1998). Gap closure and proofreading were performed using either custom primer walking or PCR amplification of the region corresponding to the gap in the sequence followed by subcloning into pUC18 and cycle sequencing with the universal pUC primers via *Taq* terminator chemistry. In some instances, additional synthetic custom primers were synthesized and used, when necessary to obtain at least 3-fold coverage for each base.

All sequenced BACs have been deposited in GenBank and assigned accession numbers (Supplemental Table S5).

Annotation Protocols

Genes were predicted using the dicot (Arabidopsis) matrix of FGENESH (Salamov and Solovyev, 2000; http://www.softberry.com). Predicted exons were then used as queries to perform BLAST searches of the TIGR Plant Transcript Assemblies database (http://plantta.tigr.org/), the Universal Protein Resource UNIPROT UniRef90 nonredundant protein database (http://www.pir.uniprot.org/), and the complete set of predicted Arabidopsis proteins (TAIR6 [for The Arabidopsis Information Resource] genome release; TAIR6_pep_20060906.fasta; ftp://ftp.arabidopsis.org/home/tair/Genes/). Genes with one or more exons that had hits with maximum expect (E) values of 1e-5 in the UNIPROT and/or TAIR databases and/or 1e-10 in the TIGR transcript assembly database were marked as supported gene predictions.

Repetitive sequences, including retrotransposons, were identified through a multistep iterative process. We used the program LTR_STRUC as the first step in identifying retrotransposons in sequenced BACs (McCarthy and McDonald, 2003). LTRs from the elements identified by LTR-STRUC were then used as queries in BLAST searches of all BACs. These BACs were also searched for the presence of retrotransposon-related genes using the BAC sequences as a query to search the NCBI nonredundant protein database using BLASTX. Regions of homology to known retrotransposon-like sequences (e.g. reverse transcriptase, 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; A. Smit, R. Hubley, and P. Green, unpublished data). These additional searches uncovered several intact elements missed by the LTR_STRUC program. All identified repetitive sequences were then loaded into a local database. The exons predicted by FGENESH were then searched against this database using BLASTN, and genes with hits with maximum E-values of 1e-10 were marked as repetitive.

Chromosomal Alignment Methods

Genomic regional alignments were generated using similarity comparisons of predicted proteins. Synteny images were generated using custom Perl scripts (available on request) and the GD-SVG image library. Gene correspondences were calculated using BLASTALL (Altschul et al., 1997) on peptide sequences from repeat-filtered supported FGENESH gene calls (Salamov and Solovyev, 2000), with a maximum E-value of 1e-10 (Supplemental Fig. S2, A, D, and E) or 1e-20 (Supplemental Fig. S2, B and C). For visual clarity, correspondence lines were manually removed for the gene families indicated by colored boxes in Figure 2 and Supplemental Figure S2. A phylogenetic analysis was then performed for each family, and where clear orthology could be established, lines were added back to indicate orthologous or coortholo-gous relationships. When alignments indicated that a low-copy gene was missing from a given BAC sequence, the sequence was rechecked using the low-copy gene as a query to perform a TBLASTX search of the entire BAC sequence. This additional step occasionally identified genes that had been missed by FGENESH or that contained an exon from a repetitive sequence and had thus been filtered. In such cases, the genes were added back to maximize the alignment. The alignments occasionally revealed that FGENESH had split a single gene into two genes on one chromosome but not its allele/orthologue/homoeologue on the other chromosome. In these cases, we fused the two gene calls into a single gene but left both FGENESH names in Supplemental Figure S2. To identify retroelements unique to Gmw or Gmp in Figure 2C and Supplemental Figure S2C, differences in DNA sequence between Gmw and Gmp were identified using mVISTA (Mayor et al., 2000). These differences were then checked for the presence of an intact LTR retroelement or an apparently full-length LINE element by comparison with our retroelement database.

Phylogenetic Analysis of Low-Copy Genes

Exons from 15 conserved low-copy genes spanning the region were aligned, along with available *Medicago* orthologues, using MUSCLE (Edgar, 2004) with default settings and then corrected by eye using Se-Al (A. Rambaut, unpublished data). Outgroup sequences from various taxa outside Fabaceae were included when *Glycine* H3 was incorporated. Gene phylogenies were estimated using MP, ML, and BI. MP used PAUP* 4.0b (Swofford, 2002) with 1,000 random addition searches and TBR branch swapping. Equally parsimonious trees were summarized in a strict consensus tree. Nodal support was estimated using 500 bootstrap replicates, each with 10 random addition sequences and TBR branch swapping. ML was conducted using PHYML (Guindon and Gascuel, 2003) accessed through the PHYML online Web server (Guindon et al., 2005) with 500 bootstrap replicates. Nucleotide substitution models for ML and BI were determined using modeltest 3.7 (Posada and Crandall, 1998) according to the Akaike information criterion. Bayesian analyses used MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001), with runs consisting of four chains run for 1 to 5 million generations sampled every 1,000 generations. The prior for each analysis was of equal probability. All runs started with a random tree. After elimination of the burn-in phase, the remaining iterations were summarized in a consensus tree with posterior probabilities as nodal support.

Ka/Ks Calculations

Ks calculations were estimated on the low-copy gene alignments used for phylogeny estimation. Before analysis, each alignment was checked for the correct reading frame. Ks was then determined using the yn00 algorithm implemented in PAML 3.15 (Yang, 1997).

Tajima Relative Rate Tests

In-frame alignments of low-copy genes including soybean H1 and H2 and *Phaseolus* copies were submitted to MEGA (Kumar et al., 2004). Relative rate tests were performed using all sites and changes.

Phylogenetic Analysis of NB-LRRs

NB-LRRs were subdivided into TIR and non-TIR classes, and a separate phylogenetic analysis was performed on each class. An approximately 900-bp region spanning from the P-loop (VGMGG in Rpg1-b) to the MHD motif (MHDLL in Rpg1-b) was used to construct phylogenetic trees. Amino acid sequences were initially aligned using ClustalW (Thompson et al., 1994) implemented in BioEdit version 7.0.5.3 (Hall, 1999). Alignments were optimized extensively by manual gap insertion and then reconversion to nucleotide sequence. We then checked for evidence of recombination among NB-LRRs using a suite of programs implemented in RDP version 3.15 (Martin et al., 2005b), RDP (Martin and Rybicki, 2000), Geneconv (Padidam et al., 1999), Chimaera (Posada and Crandall, 2001), and Bootscan (Martin et al., 2005a). Default parameter settings were used for each method except as follows: RDP (internal reference sequence), Bootscan (window = 150, step = 15, neighbor-joining trees, 200 replicates, 90% cutoff, J&N model with Ti:Tv = 2, coefficient of variation = 2). Network displays of phylogenetic signal were also used to

visualize the reticulation due to recombination and other causes. Splits Tree version 4.5 (Huson and Bryant, 2006) using the neighbor net algorithm (and p distances among sequences) was used to display the reticulation. Sequences that showed significant evidence of recombination were eliminated from further analysis. Phylogenetic analyses were performed using MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003). We explored the effect of model choice on trees produced using four different models: (1) HKY, (2) HKY+G, (3) GTR+I+G, and (4) GTR+I+G with codons. Each analysis, except for the codon model, was run in paired runs, with 10 chains each, three separate times to 5 million generations, and trees and posterior probabilities of clades were checked for consistency within and among analyses. The codon model analysis was run once only to 5 million generations because of computing limits. As the analysis including a codon model failed to improve on the likelihood score of the next most complex model, we discarded these results. The GTR+I+G analysis was favored by Bayes factors (Kass and Raftery, 1995) and used to produce Figure 3.

Phylogenetic Analysis of Carbohydrate Transporter and Protein Kinase Gene Families

For the carbohydrate transporter tree, we included genes from the neotetraploid *G. tomentella* accession G1134 on BAC clone gtt1-298n2, as we had a gap in our BAC contig from *G. tomentella* accession G1403 covering two of the four transporter genes on H1. Amino acid sequences were initially aligned using ClustalW (Thompson et al., 1994) as implemented in BioEdit version 7.0.5.3 (Hall, 1999). Alignments were optimized extensively by manual gap insertion and then reconversion to nucleotide sequence. Transporter sequences were trimmed at the 5′ and 3′ ends to eliminate regions of poor alignment. Kinase sequences were trimmed to include only the kinase domain. Aligned nucleotide sequences were then subjected to Bayesan analysis using MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003) and a GTR+I+G model. We performed paired runs with four chains each and ran the analysis for 4 million generations with sampling every 100 generations. The prior for each analysis was of equal probability. All runs started with a random tree. 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.

FISH

Soybean plants (cv Williams 82) were grown under standard greenhouse conditions (16-h daylength and 27°C daytime temperature). Root tips for somatic chromosome preparations were sampled and treated with 8-hydroxyquinoline according to previously published methods (Walling et al., 2005) and were stored in Carnoy's solution at 4°C until used for preparation of chromosome spreads. Nuclei extraction and fiber FISH were performed as described previously (Jackson et al., 1998).

Plasmid/BAC clones were purified using Qiagen maxiprep kits according to the manufacturer's instructions. Approximately 1 μg 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, after which unincorporated nucleotides were removed using Qiagen PCR columns.

FISH of BAC clones onto DNA fibers (fiber-FISH) was performed as described (Jackson et al., 1998). AlexaFluor 488 streptavidin (Invitrogen) was used to detect the biotin label. This signal was amplified by layering goat anti-streptavidin conjugated with biotin (Vector Laboratories) followed by another application of AlexaFluor 488 streptavidin. Digoxigenin labels were detected using mouse anti-digoxigenin (Roche) followed by AF568 anti-mouse (Invitrogen).

Mitotic chromosome FISH was performed as described previously (Jiang et al., 1995). Centromeres were detected using the centromeric repeat probe SB91 (5′-CGTTTGAATTTGCTCAGAGCTTCAGTATTCAATTTCGAGCGTC-TCGATATATTACGGGACTCAATCAGACATCCGAGTAAAAAGTTATTGT-3′), which is a homologue of SB92 (Vahedian 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). 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). Cropping and labeling of images were performed using Adobe Photoshop CS version 8.0 for Macintosh.

Genetic Mapping

BAC and BAC end sequences were searched for two, three, or four nucleotide repeat motifs that had a minimum repeat length of 15 using the SSRIT script (Temnykh et al., 2001). Primer pairs flanking these microsatellites were identified using Primer3 version 4.0 software (Rozen and Skaletsky, 2000) using the default settings and amplifying a product ranging in size from 200 to 250 bp. Marker amplification and PAGE were performed as described (Saghai Maroof et al., 1994). Primers that yielded polymorphic products were mapped in one or both of the following mapping populations for which densely saturated molecular maps were available: a F2:3 *G. max* × *G. max* population (Gore et al., 2002) or a *G. max* × *Glycine soja* recombinant inbred line population (Maughan et al., 2000). Mapmaker 3.0b (Lander et al., 1987) was used to construct linkage groups. The initial grouping was performed using the "group" command at a log of the odds score of 3.0 with a maximum Haldane distance of 50 cM. The "order" and "compare" commands were used to determine the most probable marker order in both populations. Markers not meeting the threshold criteria were placed in intervals using the "try" command.

Sequence data from this article can be found in the GenBank/EMBL data libraries under the accession numbers listed in Supplemental Tables S1 and S5.

Supplemental Data

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

Supplemental Figure S1. Physical map of BAC contigs assembled and sequenced.

Supplemental Figure S2. Enlarged alignments of homoeologous and orthologous BAC contigs.

Supplemental Figure S3. FISH localization of soybean H2.

Supplemental Figure S4. Phylogenetic analyses of low-copy genes.

Supplemental Table S1. Low-copy probes used to screen BAC libraries.

Supplemental Table S2. Ka and Ks values for low-copy gene comparisons.

Supplemental Table S3. Low-copy gene names and best BLAST hits.

Supplemental Table S4. ESTs corresponding to each member of homoeologous gene pairs.

Supplemental Table S5. BAC clones sequenced.

ACKNOWLEDGMENTS

We thank Randy Shoemaker, Barbara Baker, and Chris Pires for serving on the advisory committee for this project. We also thank Randy Shoemaker for help with screening of BAC libraries. We thank Mounier Elharam and Jennifer Lewis at the University of Oklahoma's Advanced Center for Genome Technology for contributing to the DNA sequencing on the ABI3730 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 Advanced Center for Genome Technology.

Received August 10, 2008; accepted October 6, 2008; published October 8, 2008.

LITERATURE CITED

Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci USA **100:** 4649–4654

Adams KL, Percifield R, Wendel JF (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics **168:** 2217–2226

Akkaya MS, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci **35:** 1439–1445

Altschul SF, Madden TL, Schaffer 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

Ashfield T, Bocian A, Held D, Henk AD, Marek LF, Danesh D, Penuela S, Meksem K, Lightfoot DA, Young ND, et al (2003) Genetic and physical localization of the soybean *Rpg1-b* disease resistance gene reveals a complex locus containing several tightly linked families of NBS-LRR genes. Mol Plant Microbe Interact **16:** 817–826

Ashfield T, Danzer JR, Held D, Clayton K, Keim P, Saghai Maroof MA, Webb PM, Innes RW (1998) *Rpg1,* a soybean gene effective against races of bacterial blight, maps to a cluster of previously identified disease resistance genes. Theor Appl Genet **96:** 1013–1021

Ashfield T, Keen NT, Buzzell RI, Innes RW (1995) Soybean resistance genes specific for different *Pseudomonas syringae* avirulence genes are allelic, or closely linked, at the *RPG1* locus. Genetics **141:** 1597–1604

Ashfield T, Ong LE, Nobuta K, Schneider CM, Innes RW (2004) Convergent evolution of disease resistance gene specificity in two flowering plant families. Plant Cell **16:** 309–318

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell **16:** 1679–1691

Bodenteich A, Chissoe S, Wang YF, Roe BA (1993) Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. *In* JC Venter, ed, Automated DNA Sequencing and Analysis Techniques. Academic Press, London, pp 42–50

Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, Weigel D (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. PLoS Biol **5:** e236

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

Bruggmann R, Bharti AK, Gundlach H, Lai J, Young S, Pontaroli AC, Wei F, Haberer G, Fuks G, Du C, et al (2006) Uneven chromosome contraction and expansion in the maize genome. Genome Res **16:** 1241–1251

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol **4:** 10

Chin DB, Arroyo-Garcia R, Ochoa OE, Kesseli RV, Lavelle DO, Michelmore RW (2001) Recombination and spontaneous mutation at the major cluster of resistance genes in lettuce (*Lactuca sativa*). Genetics **157:** 831–849

Chissoe SL, Bodenteich A, Wang YF, Wang YP, Burian D, Clifton SW, Crabtree J, Freeman A, Iyer K, Jian L, et al (1995) Sequence and analysis of the human ABL gene, the *BCR* gene, and regions involved in the Philadelphia chromosomal translocation. Genomics **27:** 67–82

Comai L (2000) Genetic and epigenetic interactions in allopolyploid plants. Plant Mol Biol **43:** 387–399

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet **42:** (in press)

Doyle JJ, Luckow MA (2003) The rest of the iceberg: legume diversity and evolution in a phylogenetic context. Plant Physiol **131:** 900–910

Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. Science **316:** 1862–1866

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:** 1792–1797

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8:** 186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8:** 175–185

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC (2007) Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. Plant Cell **19:** 3403–3417

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res **8:** 195–202

Gore MA, Hayes AJ, Jeong SC, Yue YG, Buss GR, Maroof S (2002) Mapping tightly linked genes controlling potyvirus infection at the *Rsv1* and *Rpv1* region in soybean. Genome **45:** 592–599

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:** 696–704

Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online: a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res **33:** W557–559

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser **41:** 95–98

Hayes AJ, Jeong SC, Gore MA, Yu YG, Buss GR, Tolin SA, Maroof MA (2004) Recombination within a nucleotide-binding-site/leucine-rich-repeat gene cluster produces new variants conditioning resistance to soybean mosaic virus in soybeans. Genetics **166:** 493–503

Hayes AJ, Yue YG, Saghai Maroof MA (2000) Expression of two soybean resistance gene candidates shows divergence of paralogous single-copy genes. Theor Appl Genet **101:** 789–795

Hegarty MJ, Hiscock SJ (2008) Genomic clues to the evolutionary success of polyploid plants. Curr Biol **18:** R435–R444

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. Science **293:** 1098–1102

Henikoff S, Malik HS (2002) Centromeres: selfish drivers. Nature **417:** 227

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:** 754–755

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol **23:** 254–267

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA **103:** 16666–16671

Innes RW (2004) Guarding the goods: new insights into the central alarm system of plants. Plant Physiol **135:** 695–701

Jackson SA, Wang ML, Goodman HM, Jiang J (1998) Application of fiber-FISH in physical mapping of *Arabidopsis thaliana*. Genome **41:** 566–572

Jeong SC, Hayes AJ, Biyashev RM, Saghai Maroof MA (2001) Diversity and evolution of a non-TIR-NBS sequence family that clusters to a chromosomal "hotspot" for disease resistance genes in soybean. Theor Appl Genet **103:** 406–414

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

Jorda L, Vera P (2000) Local and systemic induction of two defense-related subtilisin-like protease promoters in transgenic Arabidopsis plants: luciferin induction of *PR* gene expression. Plant Physiol **124:** 1049–1058

Kass R, Raftery A (1995) Bayes factors. J Am Stat Assoc **90:** 773–795

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

Lai J, Ma J, Swigonova Z, Ramakrishna W, Linton E, Llaca V, Tanyolac B, Park YJ, Jeong OY, Bennetzen JL, et al (2004) Gene loss and movement in the maize genome. Genome Res **14:** 1924–1931

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174–181

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

Leitch AR, Leitch IJ (2008) Genomic plasticity and the diversity of polyploid plants. Science **320:** 481–483

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

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics **82:** 378–389

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

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473

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

Ma J, SanMiguel P, Lai J, Messing J, Bennetzen JL (2005) DNA rearrangement in orthologous *orp* regions of the maize, rice and sorghum genomes. Genetics **170:** 1209–1220

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends Genet **23:** 134–139

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA **102:** 5454–5459

Malik HS, Henikoff S (2002) Conflict begets complexity: the evolution of centromeres. Curr Opin Genet Dev **12:** 711–718

Marek LF, Mudge J, Darnielle L, Grant D, Hanson N, Paz M, Huihuang Y, Denny R, Larson K, Foster-Hartnett D, et al (2001) Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome **44:** 572–581

Marek LF, Shoemaker RC (1997) BAC contig development by fingerprint analysis in soybean. Genome **40:** 420–427

Margulies EH, Birney E (2008) Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. Nat Rev Genet **9:** 303–313

Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics **16:** 562–563

Martin DP, Posada D, Crandall KA, Williamson C (2005a) A modified Bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses **21:** 98–102

Martin DP, Williamson C, Posada D (2005b) RDP2: recombination detection and analysis from sequence alignments. Bioinformatics **21:** 260–262

Maughan PJ, Saghai Maroof MA, Buss GR (2000) Identification of quantitative trait loci controlling sucrose content in soybean (*Glycine max*). Mol Breed **6:** 105–111

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics **16:** 1046–1047

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

Membre N, Bernier F, Staiger D, Berna A (2000) *Arabidopsis thaliana* germin-like proteins: common and specific features point to a variety of functions. Planta **211:** 345–354

Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J **20:** 317–332

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell **15:** 809–834

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res **8:** 1113–1130

Mitsuhara I, Iwai T, Seo S, Yanagawa Y, Kawahigasi H, Hirose S, Ohkawa Y, Ohashi Y (2008) Characteristic expression of twelve rice *PR1* family genes in response to pathogen infection, wounding, and defense-related signal compounds (121/180). Mol Genet Genomics **279:** 415–427

Mondragon-Palomino M, Gaut BS (2005) Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. Mol Biol Evol **22:** 2444–2456

Moreno C, Lazar J, Jacob HJ, Kwitek AE (2008) Comparative genomics for detecting human disease genes. Adv Genet **60:** 655–697

Mudge J, Cannon SB, Kalo P, Oldroyd GE, Roe BA, Town CD, Young ND (2005) Highly syntenic regions in the genomes of soybean, *Medicago truncatula*, and *Arabidopsis thaliana*. BMC Plant Biol **5:** 15

Nagy ED, Bennetzen JL (2008) Pathogen corruption and site-directed recombination at a plant disease resistance gene cluster. Genome Res (in press)

Nobuta K, Ashfield T, Kim S, Innes RW (2005) Diversification of non-TIR class NB-LRR genes in relation to whole-genome duplication events in Arabidopsis. Mol Plant Microbe Interact **18:** 103–109

Noel L, Moores TL, van der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JD (1999) Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. Plant Cell **11:** 2099–2112

Osmark P, Boyle B, Brisson N (1998) Sequential and structural homology between intracellular pathogenesis-related proteins and a group of latex proteins. Plant Mol Biol **38:** 1243–1246

Padidam M, Sawyer S, Fauquet CM (1999) Possible emergence of new geminiviruses by frequent recombination. Virology **265:** 218–225

Paterson AH (2005) Polyploidy, evolutionary opportunity, and crop adaptation. Genetica **123:** 191–196

Paterson AH (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. Nat Rev Genet **7:** 174–184

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

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics **14:** 817–818

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA **98:** 13757–13762

Roe B (2004) Shotgun library construction for DNA sequencing. *In* S Zhao, M Stodolsky, eds, Bacterial Artificial Chromosomes, Vol 1: Library Construction, Physical Mapping, and Sequencing. Humana Press, Totowa, NJ, pp 171–187

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

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *In* S Krawetz, S Misener, eds, Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365–386

Saghai Maroof MA, Biyashev RM, Yang GP, Zhang Q, Allard RW (1994) Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics. Proc Natl Acad Sci USA **91:** 5466–5470

Salamov AA, Solovyev VV (2000) *Ab initio* gene finding in Drosophila genomic DNA. Genome Res **10:** 516–522

Sambrook J, Fritsch EF, Maniatis T, editors (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

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

Schranz ME, Song BH, Windsor AJ, Mitchell-Olds T (2007) Comparative genomics in the Brassicaceae: a family-wide perspective. Curr Opin Plant Biol **10:** 168–175

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

Singh RJ, Chung GH, Nelson RL (2007) Landmark research in legumes. Genome **50:** 525–537

Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) Close split of sorghum and maize genome progenitors. Genome Res **14:** 1916–1923

Swofford DL (2002) PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods), Ed 4.0b. Sinauer Associates, Sunderland, MA

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res **11:** 1441–1452

Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res **16:** 934–946

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

Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J (2003) Fitness costs of *R*-gene-mediated resistance in *Arabidopsis thaliana*. Nature **423:** 74–77

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

Van K, Kim DH, Cai CM, Kim MY, Shin JH, Graham MA, Shoemaker RC,

Choi BS, Yang TJ, Lee SH (2008) Sequence level analysis of recently duplicated regions in soybean [*Glycine max* (L.) Merr.] genome. DNA Res **15:** 93–102

Walling JG, Pires JC, Jackson SA (2005) Preparation of samples for comparative studies of plant chromosomes using *in situ* hybridization methods. Methods Enzymol **395:** 443–460

Wawrzynski A, Ashfield T, Chen NWG, Mammadov J, Nguyen A, Podicheti R, Cannon SB, Thareau V, Ameline-Torregrosa C, Cannon E, et al (2008) Replication of nonautonomous retroelements in soybean appears to be both recent and common. Plant Physiol **148:** 1760–1771

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13:** 555–556

Zhang L, Vision TJ, Gaut BS (2002) Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol Biol Evol **19:** 1464–1473