Distinct Copy Number, Coding Sequence, and Locus Methylation Patterns Underlie Rhg1-Mediated Soybean Resistance to Soybean Cyst Nematode^1\[W]\[OPEN\]

David E. Cook^2, Adam M. Bayless, Kai Wang, Xiaoli Guo^3, Qijian Song, Jiming Jiang, and Andrew F. Bent*

Department of Plant Pathology (D.E.C., A.M.B., X.G., A.F.B.) and Department of Horticulture (K.W., J.J.), University of Wisconsin, Madison, Wisconsin 53706; and Soybean Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, Maryland 20705 (Q.S.)

Copy number variation of kilobase-scale genomic DNA segments, beyond presence/absence polymorphisms, can be an important driver of adaptive traits. Resistance to Heterodera glycines (Rhg1) is a widely utilized quantitative trait locus that makes the strongest known contribution to resistance against soybean cyst nematode (SCN), Heterodera glycines, the most damaging pathogen of soybean (Glycine max). Rhg1 was recently discovered to be a complex locus at which resistance-conferring haplotypes carry up to 10 tandem repeat copies of a 31-kb DNA segment, and three disparate genes present on each repeat contribute to SCN resistance. Here, we use whole-genome sequencing, fiber-FISH (fluorescence in situ hybridization), and other methods to discover the genetic variation at Rhg1 across 41 diverse soybean accessions. Based on copy number variation, transcript abundance, nucleic acid polymorphisms, and differentially methylated DNA regions, we find that SCN resistance is associated with multicopy Rhg1 haplotypes that form two distinct groups. The tested high-copy-number Rhg1 accessions, including plant introduction (PI) 88788, contain a flexible number of copies (seven to 10) of the 31-kb Rhg1 repeat. The identified low-copy-number Rhg1 group, including PI 548402 (Peking) and PI 437654, contains three copies of the Rhg1 repeat and a newly identified allele of Glyma18g02590 (a predicted α-SNAP [α-soluble N-ethylmaleimide-sensitive factor attachment protein]). There is strong evidence for a shared origin of the two resistance-conferring multicopy Rhg1 groups and subsequent independent evolution. Differentially methylated DNA regions also were identified within Rhg1 that correlate with SCN resistance. These data provide insights into copy number variation of multigene segments, using as the example a disease resistance trait of high economic importance.

Vascular plants experienced a rapid diversification following land colonization, overcoming biotic and abiotic stresses to occupy diverse niches in a process that continues to the present and includes human-guided plant breeding (Kenrick and Crane, 1997; Steemans et al., 2009; Oh et al., 2012). One mechanism of genetic variation is diversification of the physical genome, at scales broader than isolated DNA base pair changes. This genome structural variation (Feuk et al., 2006) is increasingly recognized for having significant impacts on phenotypes and evolution (Aitman et al., 2006; Perry et al., 2008; Maron et al., 2013). Recent advances in plant genomics have highlighted the role of structural variation in plant adaptation to environmental stress (DeBolt, 2010; Dassanayake et al., 2011; Wu et al., 2012; Olsen and Wendel, 2013).

Copy number variation is an important type of structural variation because of its varied evolutionary impacts, facilitating neofunctionalization, subfunctionalization, and gene dosage effects (Ohno, 1970; Moore and Purugganan, 2005; Flagel and Wendel, 2009; Marques-Bonet et al., 2009). While the majority of duplicated genes are not retained, undergo pseudogenation, or exhibit distinct negative effects (Lynch and Conery, 2000; Demuth and Hahn, 2009; Tang and Amon, 2013), gene duplication has facilitated evolution in diverse organisms (Kondrashov et al., 2002; Conant and Wolfe, 2008). For one of the simplest types of copy number variation, gene duplication, a wide range of resulting adaptations to changing local environmental conditions has been characterized (Triglia et al., 1991; Labbé et al., 2007; Schmidt et al., 2010; Dassanayake et al., 2011; Heimb erg et al., 2013; for review, see Kondrashov, 2012). Single gene copy number amplification has also been observed as an adaptive response to selective pressures (Bass and Field, 2011).

Epigenetic modifications, prominently including differential cytosine methylation, can also significantly impact organismal phenotypes (Chen, 2007; Gohlke et al., 2013; Hernandez-Herraez et al., 2013). While the term...
epigenetic indicates heritable changes in gene activity not caused by changes in DNA sequence, there is increasing appreciation not only of the extent of methylation and other epigenetic marks throughout genomes, but also of the plasticity of these marks (Schmitz et al., 2013b; Ziller et al., 2013).

Domesticated soybean (Glycine max) is an important world commodity, accounting for a majority of the world’s protein-meal and oilseed production (soystats.com). The most economically damaging pathogen of soybean is the soybean cyst nematode (SCN), Heterodera glycines (Niblack et al., 2006). SCNs are obligate endoparasites that cause disease by reprogramming host root cells to form specialized feeding cells termed syncytia, robbing the plant of carbon and adversely affecting yield (Lauritis et al., 1983; Endo, 1984; Young, 1996; Sharma, 1998). SCN is found in all major soybean-growing states in the United States and cannot feasibly be removed (Niblack, 2005). Because the primary control strategies for SCN are crop rotation and planting resistant varieties, significant attention has been focused on the identification, development, and use of soybean germplasm that exhibits resistance to SCN (Diers et al., 1997; Concibido et al., 2004; Brucker et al., 2005b; Wrather and Koenning, 2009; Kim et al., 2010a, 2011). The Rhg1 (for Resistance to Heterodera glycines) locus, sometimes in combination with Rhg4, makes the greatest contribution to resistance in the vast majority of the commercially utilized soybean cultivars that exhibit SCN resistance (Caldwell et al., 1960; Matson and Williams, 1965; Webb et al., 1995; Li et al., 2004; Brucker et al., 2005b; Tylka et al., 2012).

We recently discovered that the SCN resistance conferred by Rhg1 is mediated by a 31-kb segment of DNA that contains four open reading frames and exhibits substantial copy number variation (Cook et al., 2012). A commercial soybean line containing the most widely utilized version of the Rhg1 locus, derived from plant introduction (PI) 88788, contains 10 tandem repeat copies of the 31-kb segment. Only a single copy of this 31-kb block was detected in the SCN-susceptible line Williams 82 and three other SCN-susceptible lines. It is particularly intriguing that three distinct genes within the 31-kb repeat were shown to contribute to SCN resistance (Cook et al., 2012). These genes are Glyma18g02580 (encoding a predicted α-SNAP [α-soluble N-ethylmaleimide-sensitive factor attachment protein] vesicle-trafficking protein), Glyma18g02590 (encoding a predicted α-SNAP [α-soluble N-ethylmaleimide-sensitive factor attachment protein] vesicle-trafficking protein), and Glyma18g02610 (encoding a protein lacking a predicted function). The predicted protein sequences of Glyma18g02580 and Glyma18g02610 were invariant between the examined SCN-resistant and SCN-susceptible alleles, and experimental evidence suggests that these two genes contribute to resistance via enhanced expression arising through copy number variation. The SCN-resistant line derived from PI 88788 did contain an alternative allele of Glyma18g02590, which was also more highly expressed in SCN-resistant lines relative to susceptible lines. In addition to PI 88788, the other primary source of Rhg1-mediated SCN resistance in commercially cultivated soybean varieties is PI 548402 (commonly and throughout this article referred to as Peking). We found that the Peking Rhg1 contains three copies of the 31-kb region, but nucleotide sequences of the genes in Peking Rhg1 were not determined (Cook et al., 2012).

A well-documented epistasis occurs in Peking-derived SCN resistance, in which Peking Rhg1 has low efficacy relative to the Rhg1 from PI 88788, but only if Peking Rhg4 is not simultaneously present (Brucker et al., 2005a; Liu et al., 2012). The responsible gene at Rhg4 was recently discovered to encode a Ser hydroxymethyltransferase (Liu et al., 2012). Peking and PI 437654 (the source of the less used but commercially relevant Hartwig or CystX resistance) contain an Rhg4 allele whose product exhibits altered enzyme kinetics. Impacts of Rhg4 on SCN resistance are difficult to detect when deployed together with the high-copy-number rhg1-b from PI 88788 (Brucker et al., 2005a). It is intriguing and of high economic relevance that SCN populations arise that partially overcome the resistance mediated by certain sources of Rhg1 while remaining sensitive to the resistance conferred by other Rhg1 sources (Niblack et al., 2002; Colgrove and Niblack, 2008). In addition to understanding the biology of trait variation caused by copy number variation, and of traits in multicellular eukaryotes that are conferred by tightly linked blocks of distinct genes, there is substantial practical interest in understanding the variation in SCN resistance caused by different sources of Rhg1, and in the potential to predict, discover, and/or develop more effective versions of Rhg1.

Here, we use quantitative PCR (qPCR), fluorescence in situ hybridization (fiber-FISH), whole-genome sequencing, and DNA methylation analyses to investigate the major SCN resistance locus Rhg1 from a diverse population of soybean lines. We sequenced and analyzed the genomes of six Hg Type Test soybean lines that are widely used to characterize SCN field populations for their capacity to overcome different sources of SCN resistance (Niblack et al., 2002) and also analyzed whole-genome sequence data from 35 diverse soybean lines that are in use as parents in a separate soybean nested association mapping (SoyNAM) project. We discovered three classes of the Rhg1 locus that can be differentiated by gene dosage, copy number, and coding sequence. We also observed differential DNA methylation between resistant and susceptible Rhg1 haplotypes at genes impacting SCN resistance. The collective data allow clearer inferences to be drawn regarding the evolutionary history of the locus and provide a detailed analysis of one of the few confirmed examples in plant or animal biology in which copy number variation of a small multigene segment contributes to a defined adaptive trait.

RESULTS

Commonly Used Sources for Rhg1 Resistance Possess Either a Low Copy Number or a High Copy Number of Rhg1 Repeats as Compared with the Wild-Type Single Copy

To assess the natural variation present at Rhg1, beyond the previous determination that there are 10 and
three copies of the 31-kb Rhg1 repeat in two previously studied lines (Cook et al., 2012), we analyzed five other SCN-resistant lines. Together with PI 88788 and Peking, these seven soybean lines comprise the diagnostic test set in the established Hg Type Test that describes the capacity of SCN populations to overcome different sources of SCN resistance (Niblack et al., 2002). Initial characterization of Rhg1 copy number, using qPCR on genomic DNA, revealed three copy number classes: single copy, low copy (two to four copies), and high copy (more than six copies; Fig. 1A). For lines estimated to contain more than six copies, qPCR produced variable results and unreliable absolute copy number estimates, possibly because it is difficult to reduce qPCR variation below approximately 50% (half of one PCR cycle) between replicate tissue samples. Copy number estimates based on qPCR, however, did consistently identify two different classes for Rhg1 repeats.

To determine the impact that varying Rhg1 copy number has on constitutive transcription, we quantified root transcript abundance using qPCR in the Hg Type Test lines (Niblack et al., 2002). The four genes encoded within the previously identified Rhg1 repeat, Glyma18g02580, Glyma18g02590, Glyma18g02600, and Glyma18g02610, are more highly expressed in each of the seven tested Hg Type Test SCN resistance lines relative to SCN-susceptible Williams 82 (Fig. 1B). The transcript abundance of an adjacent gene that is outside of the 31-kb repeat, Glyma18g02570, had similar transcript abundance across all tested SCN-resistant and SCN-susceptible genotypes. Four of the SCN-resistant genotypes, Peking, PI 90763, PI 89772, and PI 437654, showed similar levels of elevated expression of the repeated genes, while expression was even more elevated in Cloud (PI 548316), PI 88788, and PI 209332 (Fig. 1B). These groupings were the same as those identified for qPCR estimates of DNA copy number and indicate that transcript abundance for these genes scales with gene copy number. One gene in the repeat, Glyma18g02600, was more highly expressed in SCN-resistant lines, but the expression level was similar between genotypes in different copy number classes. However, transcript abundance for this gene was close to the limit of detection for qPCR, was also detected only at very low levels in published RNA sequencing experiments (Severin et al., 2010), and no contribution of this gene to SCN resistance has yet been demonstrated (Cook et al., 2012). The soybean line Cloud, which was placed in the high-copy-number class but estimated to have fewer Rhg1 copies than PI 88788 and PI 209332, also showed lower transcript abundance of Glyma18g02580 and Glyma18g02590 than the other two lines in the high-copy class (Fig. 1B).

**Copy Number at the Rhg1 Locus in the High-Copy Lines Is Dynamic**

To definitively determine Rhg1 copy number in the Hg Type Test lines, we performed fiber-FISH using a diagnostic pair of DNA probes that span the repeat junction and partially overlap (Cook et al., 2012; Walling and

![Figure 1](https://example.com/figure1.png)
Representative fiber-FISH images for soybean lines PI 90763, PI 89772, and PI 437654 shown in Figure 2B (top row) summarize the finding that all three lines contain three copies of the 31-kb Rhg1 locus per haplotype, arranged as head-to-tail direct repeats. These results confirm the copy number estimates from qPCR. More importantly, for soybean lines in the high-copy Rhg1 class, fiber-FISH precisely determined the presence of seven Rhg1 copies in Cloud, nine copies in PI 88788, and 10 copies in PI 209332. We had previously used fiber-FISH to determine that Fayette, a soybean variety containing an Rhg1 locus originally from PI 88788, carries 10 copies of the Rhg1 repeat (Cook et al., 2012). Hence, the number of Rhg1 repeats varies not only between haplotype classes but also within the high-copy class and between lines with recent shared ancestry.

Read Depth Analysis from Whole-Genome Sequencing Identifies Rhg1 Copy Number and Predicts SCN Resistance

To further discover the nature of the diversity within soybean Rhg1, we performed whole-genome sequencing for six of the seven Hg Type Test soybean lines: Peking, PI 90763, PI 89772, PI 437654, PI 209332, and Cloud (Niblack et al., 2002). Derivatives of PI 88788 had been sequenced previously (Cook et al., 2012). In addition, we analyzed whole-genome shotgun sequence data for 35 diverse soybean lines, generated as part of the recently
initiated SoyNAM project, and analyzed previously published Illumina sequencing data from an undomesticated *Glycine soja* accession (Kim et al., 2010b). Supplemental Tables S1 and S2 provide details regarding the sequencing data sets. For this study, we focused on in-depth analysis of *Rhg1* on chromosome 18 and its paralogous locus on chromosome 11.

To initially uncover structural variation at *Rhg1*, we screened the SoyNAM genome sequence data sets by aligning Illumina reads to a portion of the Williams 82 reference genome corresponding to *Rhg1* on soybean chromosome 18 and similar loci (see “Materials and Methods”). This screen determined that eight of 35 SoyNAM lines contain an estimated *Rhg1* copy number greater than one, based on read depth across the known repeat and flanking regions (Supplemental Table S3). To further investigate the extent of copy number variation in this set of diverse soybean genomes and to eliminate possible mapping bias that might arise from the use of a limited reference sequence region, Illumina sequencing reads were remapped to the entire reference genome for 24 of the SoyNAM lines based on the results of the rapid alignment and sequencing depth. This provided more precise *Rhg1* copy number estimates based on read depth. Along with the SoyNAM lines, six Hg Type Test lines sequenced as part of this work and the available *G. soja* genome sequence were included for in-depth analysis. As shown in Figure 2C, the estimated copy numbers based on read depth for the six Hg Type Test lines are in agreement with the results from qPCR estimates and fiber-FISH. Lines Peking, PI 90763, PI 89772, and PI 437654 contain three copies, while Cloud has seven and PI 409332 has 10 (Fig. 2C). A soybean line derived from PI 209332 was previously estimated to carry 10 copies of the *Rhg1* locus, while one line contains an estimated three copies (Fig. 2C). These results are in agreement with pedigree information where it is publicly available. The above results indicate that increased copy number at *Rhg1* is not a common phenomenon in soybean accessions and likely can be traced to a limited number of parental lines. There is also no indication that structural variation has occurred at the paralogous locus on chromosome 11 (Fig. 2D).

**Sequence Analysis Reveals Extensive Rhg1 Locus DNA Sequence Variation, But Amino Acid Polymorphisms Are Only Present in the Predicted α-SNAP**

The whole-genome sequence data of the SoyNAM and Hg Type Test lines were analyzed for *Rhg1* nucleic acid and derived amino acid variations. Genomic DNA sequence variations, including single-nucleotide polymorphisms (SNPs) and small insertions and deletions relative to the Williams 82 reference genome, were identified using the Genome Analysis Toolkit (GATK) pipeline (McKenna et al., 2010; DePristo et al., 2011). A total of 409 DNA variant sites across the 31.2-kb *Rhg1* repeat interval (chromosome 18, bp 1,632,223–1,663,500) of the Williams 82 soybean reference genome, version 1.1) were identified in at least one of the 31 genomes. The average number of *Rhg1* variant sites per soybean line was 251 ± 40 (mean ± se) for the low-copy *Rhg1* lines, 260 ± 10 for the high-copy lines, and 23 ± 29 for the lines estimated to contain a single copy of *Rhg1* (Table I). As described below, within any single accession, the sequences of individual repeats were largely identical to the other repeats. Hence, there were approximately 250 polymorphisms per 31-kb repeat in the SCN-resistant genotypes but zero to 81 polymorphisms in the corresponding 31-kb *Rhg1* region of the investigated single-copy lines.

Despite the high number of sequence polymorphisms found within each *Rhg1* repeat in SCN-resistant lines, few affected protein-coding sequences. We did not detect any polymorphisms resulting in an altered amino acid sequence for Glyma18g02580 or Glyma18g02580 in any of the SCN-resistant lines. Curiously, in the derived amino acid sequences of Glyma18g02590, two SCN-resistant allele types were observed that carry distinct mutations but that impact similar protein sites (Fig. 3A). The gene Glyma18g02590 encodes a predicted α-SNAP; in other organisms, these proteins have the canonical function of stimulating N-ethylmaleimide-sensitive factor ATPase activity to assist the disassembly of SNARE components following vesicle-mediated transport (Morgan et al., 1994; Barnard et al., 1997; Rice and Brunger, 1999). Amino acid sequence alignment of the available 17 *Rhg1* single-copy soybean lines including the Williams 82 reference genome revealed an invariant primary sequence of Glyma18g02590. One type of alternative allele was found in all tested high-copy *Rhg1* haplotypes, including the previously reported sequence from Fayette, and a new allele was found in all tested lines carrying the low-copy *Rhg1* haplotype associated with SCN resistance (Fig. 3B). The novel alleles of α-SNAP found in SCN-resistant lines have amino acid polymorphisms changing the final five or six amino acids, the residues that otherwise have the strongest consensus sequence across eukaryote α-SNAPs (Fig. 3C), including a substitution for the Leu at the penultimate C-terminal amino acid. The presence of different amino acid substitutions at similar positions between the low- and high-copy-class 2590 alleles suggests a functional importance of these sites for SCN disease resistance. Mutations at these C-terminal residues are unexpected given previous findings that these residues are essential for stimulating N-ethylmaleimide-sensitive factor ATPase activity in other organisms (Barnard et al., 1996). Together, these findings suggest that the SCN resistance-associated Glyma18g02590 proteins may not possess classical α-SNAP functions and may instead...
promote SCN disease resistance through a novel mechanism.

For *Glyma18g02590*, we performed 3’ RACE and sequenced at least seven independent complementary DNA (cDNA) clones for each of the Hg Type Test lines and Williams 82. The novel (non-Williams 82) *Glyma18g02590* alleles predicted from genomic DNA sequences were present in cDNA from the respective lines carrying the low- or high-copy *Rhg1* haplotypes (Supplemental Table S4). Interestingly, a small proportion of the cDNA clones sequenced from PI 88788 and Cloud (high-copy *Rhg1* lines) contained Williams 82-type *Glyma18g02590* sequences, consistent with the identification of a single Williams 82-type genomic DNA sequence in one of the copies of the 31-kb *Rhg1* repeats (described below). We did not detect any Williams 82-type *Glyma18g02590* sequences in cDNAs from lines carrying the low-copy-class *Rhg1*, again consistent with the genomic DNA sequence data. However, a splice isoform of the *Glyma18g02590* cDNA was identified in all of the tested low-copy *Rhg1* lines, and this splice isoform was not found in the high-copy or single-copy *Rhg1* lines (Fig. 3B; Supplemental Table S4).

A naturally occurring truncated allele encoding a predicted α-SNAP was recently implicated in SCN disease resistance derived from Peking and PI 437654 but not in PI 88788-derived resistance (Matsye et al., 2012). Our results from whole-genome sequencing, however, indicate that the sequence encoding that truncated α-SNAP is not encoded by a *Glyma18g02590* gene at *Rhg1* on chromosome 18 but rather by *Glyma11g35820*, the paralogue of *Glyma18g02590* on chromosome 11 (Supplemental Fig. S1; Supplemental Table S5). The SNPs at *Glyma11g35820* responsible for encoding the truncated allele were also identified in the high-copy *Rhg1* SCN-resistant lines Cloud and LG05-4292.

Another *Rhg1* sequence polymorphism was identified in the Peking genome: a nucleotide deletion in the

| Genotype      | Variant Class | Variant Location |
|---------------|---------------|------------------|
| Copy No.      | SNPs          | Insertion        | Deletion | Total | Exon | Intron | Intergenic | UTR |
| LD00-3309     | 10.03         | 190              | 36       | 32    | 258  | 6      | 24         | 220 8 |
| LG05-4292     | 9.23          | 195              | 38       | 34    | 267  | 6      | 25         | 228 8 |
| G. soja       | 1.10          | 42               | 3        | 4     | 49   | 1      | 3          | 45   0 |

### Table 1. Summary statistics for DNA variant analysis at *Rhg1* from whole-genome sequencing shows higher rates of polymorphism in SCN-resistant lines

Copy No., *Rhg1* copy number estimated from sequencing read depth or from fiber-FISH (asterisks). Variant Class, Whole-genome sequencing for 30 soybean lines and one *G. soja* line was analyzed for DNA variants and classified as SNPs, insertion, or deletion relative to the Williams 82 reference genome. The total number of DNA variants of each type across the 31-kb *Rhg1* sequence are reported. Total, Sum of the SNP, insertion, and deletion variants. Variant Location columns report numbers of variants in each type of genome region. UTR, Untranscribed region.
second exon of the Glyma18g02600 coding sequence (Table II) observed as a heterozygous deletion (see below). Translation of the resulting mRNA results in a stop codon eight codons downstream of the deletion, truncating the predicted protein by 314 amino acids (removing 58% of the wild-type protein sequence).

Resistance-Conferring Rhg1 Loci Developed from a Common Source But Underwent Copy Number Expansion in Distinct Lineages

To further explore the evolutionary history of the Rhg1 locus, DNA sequence variation sites in a diverse set of soybean lines were used to construct a nonhierarchical phylogenetic network using the NeighborNet algorithm in SplitsTree (Bryant and Moulton, 2004; Huson and Bryant, 2006). The network reveals a clear split between the Rhg1 loci from SCN-resistant and SCN-susceptible lines (Fig. 4A). There is a further split in the multicopy clade, separating the low- and high-copy Rhg1 groups from each other (Fig. 4A). A common origin of the high-copy and low-copy Rhg1 repeats was suggested by the identity of their repeat-junction sequences (Cook et al., 2012) and is now further supported by the high number of DNA sequence variant sites shared by the two groups but absent in single-copy Rhg1 lines (Fig. 4B). In total, 147 DNA variant sites not detected in the single-copy Rhg1 SCN-susceptible lines are commonly to all of the sequenced high-copy and low-copy Hg Type Test lines. This is 75% of the 197 DNA variant sites present in at least one Hg Type Test line but not present in any of the examined SCN-susceptible lines. These data suggest that a common progenitor had accumulated the 147 DNA variant sites prior to subsequent divergence of the two copy number groups. In support of subsequent divergence of the low-copy lines from the high-copy lines, a small number of DNA variants not present in any tested SCN-susceptible genome were universally common within either the low-copy or the high-copy Rhg1 group: 10 sites for low copy and seven for high copy (Supplemental Fig. S2). Even more recent divergence is highlighted by the presence of a small number of DNA variants unique to a single tested genotype: Peking (six), PI 88788 (zero), PI 90763 (one), PI 437654 (zero), PI 209332 (five), PI 89772 (zero), and Cloud (one).

The degrees of similarity between Rhg1 repeats within any single genome or within a copy number group can be analyzed by the frequency of variant sequence relative to reference sequence from the
Table II. Amino acid polymorphisms for genes encoded within and adjacent to the Rhg1 repeat, from all analyzed soybean lines

| Position | Low-Copy Lines | High-Copy Lines | NE3001 | LG05-4317 | LC04-1128 |
|----------|----------------|-----------------|--------|-----------|-----------|
| Glyma18g02570 (unknown function) | No polymorphisms | | | | |
| Glyma18g02580 (α-SNAP) | Q203K | | | | |
| Glyma18g02590 | D208E | E285Q | V9I | | |
| Glyma18g02600 | D286Y | D287EA | | | |
| Glyma18g02610 | L288I | L288I | | | |
| Glyma18g02620 | | | | | |

whole-genome sequence data sets. Within the high-copy genomes of Cloud, PI 209332, and LD00-3309 (PI 88788 derivative), most of the variant sites on the right three-quarters of the interval as shown in Figure 5 have a sequence frequency of roughly 0.85 to 0.9 (Fig. 5A). The other 10% to 15% of sequence reads at these positions match the Williams 82 reference sequence, suggesting that roughly three-quarters of one of the Rhg1 repeats in the high-copy Rhg1 accessions contains Williams 82-type sequence. This is consistent with the Glyma18g02590 cDNA data described above (Supplemental Table S4). Most variant sites across the left one-quarter of the Rhg1 repeat (Fig. 5A) are invariant for the alternate sequence, indicating its presence in all copies.

A small number of DNA variant sites do not follow the above trend and indicate the development and propagation of variant sequences in a smaller number of the total copies. Specifically, the DNA variant at chromosome 18 base pair position 1,657,025 is apparently found in only four of the seven copies in Cloud and in only six to seven copies in LD00-3309, suggesting as one possibility the emergence of this DNA polymorphism in one repeat at an intermediate stage of copy number expansion of the locus (Supplemental Table S6). However, propagation of the repeats apparently was not symmetric between genomes, because (for example), at positions 1,657,007/1,657,016 and 1,661,264/1,661,293, Cloud and PI 209332 appear to carry only five and seven to eight copies, respectively, of the variant site while LD00-3309 appears to carry the variant site in all nine non-Williams 82 repeats. Conversely, the set of polymorphisms at positions 1,663,007 to 1,663,250 are present in only six to seven copies in LD00-3309, eight to nine copies in PI 209332, but in all six non-Williams 82 copies of Cloud (Supplemental Table S6). Inspection of raw sequence data for these nonhomogenous variant sites suggests that they are valid sequence calls rather than data-processing errors and suggests unequal propagation of specific copies during evolution of the locus. Although we cannot rule out phenotypic selection among the high-copy Rhg1 soybean lines for revertants that carry more copies of the Williams 82 reference sequence at these nonhomogenous variant sites, the sites are in intergenic regions at least 1 kb away from known transcription start sites. Hence, it may be more parsimonious to assume that they are neutral sites that reflect the source of progenitor repeats that were utilized during Rhg1 repeat expansion.

Analysis of the low-copy Rhg1 lines (Peking, PI 89772, PI 90763, and PI 437654) shows a different pattern of repeat expansion and may partially account for well-established functional differences between the high-copy (PI 88788-type) and low-copy (Peking-type) Rhg1 loci. The frequency of variant sequence at reference sequence polymorphic sites in all Rhg1 low-copy lines is nearly 1 (i.e., mostly uniform across the 31.2-kb repeat region; Fig. 5B). This suggests that the low-copy lines experienced copy number expansion from a single shared progenitor and/or homogenization across the repeats by gene conversion or other mechanisms after at least some repeats had already formed. Loss of repeats carrying divergent copies may have also occurred. This in-depth analysis of sequencing frequencies shows that not only are the two resistance groups diverging for Rhg1 copy number but the sequence composition of the repeats is also following different evolutionary paths.
Variation in Soybean Resistance to Diverse Nematode Populations Supports the High-Copy and Low-Copy Rhg1 Groupings and Suggests a Relationship between Copy Number and Resistance

Previous research has described differences in SCN resistance between Peking-, PI 437654-, and PI 88788-derived soybean sources, measured in terms of genetics, cell biology, nematode development, and nematode race specificity or Hg type specificity, but the causes for these observations have remained elusive (Arelli and Webb, 1996; Mahalingam and Skorupska, 1996; Kim et al., 1998, 2010c; Brucker et al., 2005a; Niblack et al., 2006; Klink et al., 2011). To address this, we analyzed data for soybean resistance to SCN from greenhouse trials conducted by Alison Colgrove, Terry Niblack, and colleagues as part of the Northern Regional SCN Tests (Cary and Diers, 2010, 2011, 2012, 2013). The analysis included data from a total of 97 field populations collected from 2009 to 2012, including SCN field populations from eight to 10 north central U.S. states and/or adjacent Canadian provinces per year. The results from our analysis indicate that Cloud, which contains seven copies of Rhg1, was significantly less resistant than the other lines tested (Fig. 6). The other two lines in the high-copy Rhg1 class, PI 88788 and PI 209332, which contain nine and 10 copies, respectively, form a statistically significantly more resistant cluster than Cloud, suggesting that higher Rhg1 copy number may increase SCN resistance. Because Cloud, PI 88788, and PI 209332 lines are not isogenic at other loci, this comparison is only suggestive. The low-copy Rhg1 lines are significantly more resistant to diverse SCN populations, but since these lines carry an SCN resistance-conferring allele of Rhg4, a simple comparison of the relative contributions or efficacies of Rhg1 loci between low-copy and high-copy lines is obfuscated. Moreover, the roles of low-copy and high-copy Glyma18g02590 amino acid polymorphisms in impacting resistance to SCN are unknown.

Rhg1 Loci from Different Sources Contain Differentially Methylated Regions That Correlate with SCN Resistance

In addition to determining the genome structure and nucleic acid variation present at the Rhg1 locus from different sources, we investigated potential differences in DNA methylation states. In a broad survey of root DNA methylation patterns at Rhg1, we used DNA methylation-sensitive restriction enzymes coupled with PCR to

Figure 4. Network analysis and shared polymorphisms support three Rhg1 locus types and a single SCN-resistant-type progenitor. A, NeighborNet analysis indicates two distinct groups based on Rhg1 sequence, separating the SCN-susceptible lines (left, single-copy Rhg1) and the SCN-resistant lines (right, multicopy Rhg1). The resistant lines further split into two groups that correspond with Rhg1 copy number (Figs. 1 and 2); lines containing three Rhg1 copies are noted in red and those containing seven or more copies are noted in blue. The four coding genes within the Rhg1 repeat, including 200 bp of sequence upstream of the start codon, were used for analysis. B, DNA variant sites present in all seven Hg Type Test soybean lines (low-copy and high-copy Rhg1, SCN resistant) but absent from all sequenced SCN-susceptible single-copy Rhg1 lines. Vertical red lines show locations of these 148 DNA variant sites, which are 75% of all of the 197 SNP or INDEL DNA variant sites present in at least one Hg Type Test line but not present in any of the examined SCN-susceptible lines. Rhg1 locus gene models are shown at correct x axis positions for reference (blue, exons; black line, introns; and gray, untranslated regions); gene names are given above the gene model (e.g. 2570 = Glyma18g02570).
to or within the coding sequence of Glyma18g02600. We also used McrBC to analyze methylation at the Rhg1-adjacent but nonrepeated genes Glyma18g02570 and Glyma18g02620 and did not observe DMRs (Fig. 7).

During the preparation of this article, a genome-wide methyleome study was published in which whole-genome bisulfite sequencing was performed for soybean lines LDX01-1-165 (referred to here as LDX), LD00-2817P (referred to here as LD), and progeny from their cross (Schmitz et al., 2013a). LD is known to have SCN resistance derived from PI 437654 (low-copy Rhg1 locus type), while LDX contains a single copy of Rhg1 (Diers et al., 2010; Kim et al., 2011). To confirm our observations and gain single-base resolution for methylation, we highlighted and reanalyzed the data of Schmitz et al. (2013a), focusing on Rhg1.

Consistent with the findings described above, our Rhg1 copy number estimate was 2.93 for LD and 1.17 for LDX, with various LD × LDX F3-derived (and hence potentially heterozygous) progeny families giving a range of Rhg1 copy number estimates between one and three (Supplemental Fig. S3A). We were also able to estimate transcript abundance for the two parents along with the two F3-derived progeny families that were subjected to RNA sequencing characterization (see “Materials and Methods”). Consistent

identify differentially methylated regions (DMRs) between SCN-resistant and SCN-susceptible genotypes. The enzyme McrBC restricts DNA at sites of methylated cytosines of the sequence (G/A)mC and does not restrict unmethylated DNA (Sutherland et al., 1992). Hence, genomic DNA digestion by McrBC followed by PCR will not produce a product if the PCR product spans methylated cytosines. Using a total of 23 primer pairs, we discovered eight DMRs between SCN-susceptible genomes (carrying a single-copy Rhg1 locus) and SCN-resistant genomes (carrying low- or high-copy Rhg1 loci; Fig. 7). Hypermethylated DMRs were detected in SCN-resistant lines in the shared promoter for genes Glyma18g02580 and Glyma18g02590 and within and flanking the coding sequence of Glyma18g02610. We did not observe DMRs in the gene body of Glyma18g02580, nor did we observe substantial methylation or DMRs adjacent

Figure 5. The frequency of DNA variant sites across Rhg1 repeats reveals heterogeneity between repeats in high-copy but not low-copy Rhg1-containing lines. A, Nearly homogenous presence of the same non-Williams 82 DNA sequence for variant sites in all copies (left one-quarter) or all but one copy (right three-quarters) of the Rhg1 repeat. The x axis shows the locations of DNA variant sites (SNP or INDEL) within the Rhg1 locus on soybean chromosome 18, and the y axis shows the proportion of all DNA sequence reads with variant (high-copy-type) sequence rather than the reference Williams 82-type (single-copy Rhg1) sequence at the designated DNA variant site. Data were combined for the three Rhg1 high-copy-class Hg Type Test soybean lines LD00-3309 (PI 88788), PI 209332, and Cloud; mean frequency and ± for the three soybean lines are shown. Rhg1 locus gene models are shown at correct x axis positions for reference (blue, exons; black line, introns; and gray, untranslated regions); gene names are given above the gene model (e.g. 2570 = Glyma18g02570). B, Near identity of the three repeats in Rhg1 low-copy lines and absence of a Williams 82-type segment. Details are as in A except showing combined data for the four Rhg1 low-copy-class Hg Type Test soybean lines Peking, PI 90763, PI 437654, and PI 89772.

Figure 6. Nematode resistance data from 78 diverse SCN populations indicate similarities in resistance profiles based on copy number. Nematode development data were obtained for the seven Hg Type Test SCN-resistant lines for greenhouse assays conducted as part of the 2009 to 2012 Northern Regional SCN Tests. The data analyzed for 78 SCN populations were collected from 12 U.S. states and Canadian provinces. Female index is the percentage of SCN cysts that developed to or within the coding sequence of Glyma18g02600. We also used McrBC to analyze methylation at the Rhg1-adjacent but nonrepeated genes Glyma18g02570 and Glyma18g02620 and did not observe DMRs (Fig. 7).

During the preparation of this article, a genome-wide methyleome study was published in which whole-genome bisulfite sequencing was performed for soybean lines LDX01-1-165 (referred to here as LDX), LD00-2817P (referred to here as LD), and progeny from their cross (Schmitz et al., 2013a). LD is known to have SCN resistance derived from PI 437654 (low-copy Rhg1 locus type), while LDX contains a single copy of Rhg1 (Diers et al., 2010; Kim et al., 2011). To confirm our observations and gain single-base resolution for methylation, we highlighted and reanalyzed the data of Schmitz et al. (2013a), focusing on Rhg1.

Consistent with the findings described above, our Rhg1 copy number estimate was 2.93 for LD and 1.17 for LDX, with various LD × LDX F3-derived (and hence potentially heterozygous) progeny families giving a range of Rhg1 copy number estimates between one and three (Supplemental Fig. S3A). We were also able to estimate transcript abundance for the two parents along with the two F3-derived progeny families that were subjected to RNA sequencing characterization (see “Materials and Methods”). Consistent
with our present and previous findings (Fig. 1B; Cook et al., 2012), standardized RNA sequence read depth for noninfested plants, normalized to the susceptible LDX parent, showed elevated expression for the genes encoded within but not adjacent to the \( \text{Rhg1} \) repeat in LD and progeny 11272 but not progeny 11268 (Supplemental Fig. S3B). This is consistent with elevated \( \text{Rhg1} \) copy number as a significant cause of the elevated transcript levels.

DNA methylation levels were computed from the data of Schmitz et al. (2013a) in bins of 150 bp in the CG, CHG, and CHH sequence contexts in both parents and 27 progeny lines that had at least 4× average sequencing depth. Consistent with our above findings of differential root DNA methylation in different \( \text{Rhg1} \) copy number groups, we observed differential hypermethylated DNA in all three sequence contexts at the same regions in lines estimated to contain multiple copies of \( \text{Rhg1} \) repeat in LD and progeny 11272 but not progeny 11268 (Supplemental Fig. S3B). This is consistent with elevated \( \text{Rhg1} \) copy number as a significant cause of the elevated transcript levels.

DNA methylation levels were computed from the data of Schmitz et al. (2013a) in bins of 150 bp in the CG, CHG, and CHH sequence contexts in both parents and 27 progeny lines that had at least 4× average sequencing depth. Consistent with our above findings of differential root DNA methylation in different \( \text{Rhg1} \) copy number groups, we observed differential hypermethylated DNA in all three sequence contexts at the same regions in lines estimated to contain multiple copies of \( \text{Rhg1} \) (Fig. 8; Supplemental Fig. S4). Data for the full set of lines can be seen in Supplemental Figure S5 (see “Materials and Methods”). Consistent with the finding that methylation patterns are largely inherited based on the parental methylation pattern (Schmitz et al., 2013a), for \( \text{Rhg1} \) we observed high average levels of cytosine methylation

(a characteristic of the LD parent that carries three \( \text{Rhg1} \) copies) in the progeny that appeared homozygous for the three-copy \( \text{Rhg1} \) haplotype (Fig. 8, B and D). Lower average \( \text{Rhg1} \) methylation (a characteristic of the LDX parent that carries one \( \text{Rhg1} \) copy) was observed in the progeny homozygous for single-copy \( \text{Rhg1} \) haplotypes (Fig. 8, B and D). Together, our findings and the data of Schmitz et al. (2013a) describe, in detail, across tissue types and different sources of SCN resistance, stably inherited hypermethylated DNA regions at the resistance-conferring alleles of the genes shown to mediate \( \text{Rhg1} \) resistance.

**DISCUSSION**

SCN is the most economically limiting pathogen for soybean, causing billions of dollars of yield losses annually in the United States alone (Wrather and Koening, 2009). Major efforts in soybean breeding and biotechnology are focused on the incorporation of desirable \( \text{Rhg1} \) alleles and on the continued discovery of new and better sources of SCN resistance. We had previously determined that three very tightly linked genes at \( \text{Rhg1} \) contribute to SCN resistance and that these genes reside on a
A segment that is present in 10 copies in a common SCN-resistant variety along with an altered amino acid sequence for one of the genes (Cook et al., 2012). However, the extent of Rhg1 structural variation present in a broader set of soybean germplasm, the presence of alternate coding alleles and their expression levels, and the relatedness of different Rhg1 sources were not known. Here, we report the discovery of the structural, coding, and methylation differences present at Rhg1 from a diverse population of soybean lines.

The identification in different soybean lines of seven, nine, and 10 copies of an Rhg1 locus composed of highly similar sequences indicates that copy number at Rhg1 is plastic and malleable over the time scale of breeding cycles. This is evidenced by the discovery of 10 copies of Rhg1 in Fayette, a line developed by backcrossing Williams 82 (single copy) to PI 88788 (nine copies; Mikel et al., 2010). In contrast, all the sequenced SCN-resistant lines belonging to the low-copy Rhg1 group contained three copies of nearly identical Rhg1 repeats. It will be
interesting to identify additional sources of SCN resistance to determine if the sequences in this Rhg1 group can persist in greater than three copies. This information, coupled with the relationship between larger numbers of Rhg1 repeats and increased resistance, suggests a new strategy to improve SCN resistance through the addition of Rhg1 copies.

There remains a need for improved assays that can inexpensively but accurately determine the copy number of Rhg1 or other high-copy-number loci that confer adaptive traits (Curtis et al., 2012; Maron et al., 2013; Stebbing et al., 2013). We initially utilized qPCR with genomic DNA templates for this purpose but found it challenging to obtain precise results for copy numbers above approximately four. Fiber-FISH provided definitive data, and whole-genome sequencing provided accurate estimates of higher copy number regions as long as the genome-wide read depth exceeded approximately 2-fold coverage. Comparative genome hybridization methods can also be used (Roberts et al., 2012). However, these relatively complex procedures are not likely to be useful, for example, in a plant breeding germplasm screen that seeks to identify rare individuals or infrequent recombination events carrying usefully elevated copy numbers.

Biochemical characterization of the wild-type (Williams 82-type), low-copy, and high-copy versions of the Glyma18g02590 α-SNAP alleles also is needed to determine what, if any, altered functions they have compared with each other and with canonical α-SNAP functions. We speculate that while the genomes containing the PI 88788-type α-SNAP have apparently benefited from an increase in Rhg1 copy number, the genomes with the Peking-type α-SNAP may have remained at three copies because of selection against an unknown negative impact of the Peking-type full-length α-SNAP. Rhg1 copy number in these genomes may also be affected by the shorter splice isoform of the Glyma18g02590 α-SNAP that was only detected in the low-copy Rhg1 lines. Alternately, the loss of a wild-type (Williams 82-like) α-SNAP coding sequence in the three-copy genomes may have limited expansion of the locus. It is also possible that interactions with a specific Rhg4 allele may favor the Rhg1 locus configurations found in the low-copy Rhg1 haplotypes.

The identification of the different copy numbers at Rhg1 also suggests a hypothesis regarding the relatively ineffectual nature of low-copy Rhg1 in the absence of the resistance-conferring Rhg4 allele (Brucker et al., 2005a; Liu et al., 2012). In the absence of Peking-type Rhg4, the three copies of Rhg1 now known to be present in low-copy lines such as Peking have been shown to be more resistant to SCN infection than single-copy Rhg1 lines, suggesting that this Rhg1 can function independently of resistance-associated Rhg4 alleles (Brucker et al., 2005a; Liu et al., 2012). This raises the possibility that Rhg4 combined with the high-copy Rhg1 may provide a broader spectrum SCN resistance, while the Peking-type Rhg1 resistance could possibly be improved by increasing the copy number or expression level. Also, stacked deployment of both types of Rhg1 in single soybean lines could attenuate the development of virulent nematode populations. This type of research is increasingly important given the slow but ongoing erosion of the widely deployed PI 88788-derived resistance (Niblack et al., 2008; Tylka et al., 2012).

Our data help to explain the overlaps observed by many SCN-resistance specialists when comparing different soybean accessions with regard to their spectrum of resistance to a range of different SCN populations. For example, the resistance spectra of the Hg Type Test lines PI 88788, PI 209332, and Cloud (PI 548316) correlate highly, as do those of Peking (PI 548402), PI 90763, PI 89772, and PI 438489B (Colgrove and Niblack, 2008). Those two groupings match the Rhg1 DNA sequence, copy number, and α-SNAP groups discovered in this study.

PI 437654 is recognized for its particularly high levels of resistance against diverse nematode populations (Colgrove and Niblack, 2008). However, we discovered the near identity of PI 437654 Rhg1 copy number and sequence to other, less broadly resistant Rhg1 low-copy soybean lines. Although Rhg1 makes one of the strongest contributions to PI 437654-derived resistance (Webb, 2012), the present finding reemphasizes the importance of identifying and cloning additional SCN resistance quantitative trait loci from PI 437654 (Wu et al., 2009).

Current models for evolution by gene duplication are often applied to single gene duplicates. A fascinating and unusual element of Rhg1 is that gene copy number selection occurred, and research hypotheses are being tested, for an approximately 30-kb block of four genes that encode completely dissimilar proteins, three of which have been shown to contribute (Cook et al., 2012) to the phenotype that apparently has driven selection. Determining the exact course of evolution of the Rhg1 locus is difficult, but our data strongly suggest that the repeats in the low-copy and high-copy class have a common origin. It is not clear if the common Rhg1-resistant progenitor diverged from susceptible lines prior to duplication or if the divergence occurred after duplication. Either scenario could account for the highly similar sequence and the identical repeat junction found between low- and high-copy Rhg1 lines if repeat homogenization or gene conversion has played a role in the evolution of the Rhg1 locus and caused the high sequence identity between repeats within single plant lines.

Our data suggest that multiple evolutionary forces could have differentially affected the different genes in the repeat. Two of the proteins encoded at Rhg1 (Glyma18g02580 and Glyma18g02610) have identical derived amino acid sequences among the repeats and between the resistant and susceptible lines, which matches predictions for gene duplicates fixed by positive selection for increased dosage and having a low rate of nonsynonymous to synonymous substitutions ($K_{\text{a}}/K_{\text{s}} < 1$; Innan and Kondrashov, 2010). However, the presence of nonsynonymous substitutions in Glyma18g02590 in both the low- and high-copy Rhg1 lines, caused by different
nucleotide polymorphisms, suggests a different evolutionary course; the duplication and divergence scenario that is applicable to many gene duplicates (Ohno, 1970). The identification of a premature stop codon in one copy of Glyma18g02600 in Peking, despite the highly similar SCN resistance between Peking and the other resistant lines in the low-copy class, is also interesting. This provides further evidence that Glyma18g02600 is not required for full Rhg1-mediated resistance and could be the first glimpse of pseudogenation (Lynch and Conery, 2000). Hence, the different genes in the Rhg1 repeat apparently represent different evolutionary trajectories.

The identification of Rhg1 DNA regions that exhibit differential methylation between SCN-resistant and SCN-susceptible accessions adds an additional layer of complexity to the control of phenotype expression at Rhg1 and probably to Rhg1 locus evolution. The observation of highly similar gene duplicates in the genomes of many organisms has led to the hypothesis that decreased expression of duplicate gene copies is a mechanism to maintain normal physiology following gene duplication (Qian et al., 2010). In recent work on mammalian gene duplicates, increased DNA methylation of promoter regions has been shown to follow whole-genome duplications, for example, in Syrian hamster (Schmitz et al., 2013a). Our observations for Rhg1 may seem to be the opposite of this, because in SCN-resistant lines with multiple Rhg1 copies, hypermethylation is observed at genes that exhibit increased transcript abundance. However, expression of the multicopy Rhg1 genes might be even greater in the SCN-resistant genomes if there were not methylation. Although beyond the scope of this study, recent identifications of dynamic methylation changes in Arabidopsis (Arabidopsis thaliana) following biotic stress (Dowen et al., 2012; Yu et al., 2013) suggest the hypothesis that the differentially methylated cytosine regions found upstream of Glyma18g02580, Glyma18g02590, and Glyma18g02610 could result in lower constitutive expression and increased expression of these genes following nematode infection. Future experiments to test this hypothesis may reveal further mechanisms that provide increased fitness and thereby impact the evolution of gene copy number variation.

MATERIALS AND METHODS

Estimating Copy Number and Transcript Abundance

To estimate the number of Rhg1 copies present in the Hg Type Test lines, we collected tissue for DNA extraction from 2-week-old plants grown in Metro Mix at 26°C. Leaf tissue was collected and flash frozen in liquid nitrogen, and DNA extraction was performed as described previously. To estimate Rhg1 copy number, qPCR was run using two separate primer pairs per sample. One set of primers described previously spanned the junction of repeated segmental Rhg1 duplicates, which failed to amplify a product in genomes with the wild-type single copy of the locus. A second primer pair used in a separate reaction amplified a product corresponding to a DNA interval from the gene Glyma18g02620, which is adjacent to but not present in the Rhg1 repeat. The ratio of the two products was used to determine the number of Rhg1 repeats.

To quantify the relative transcript abundance for the genes within and adjacent to the Rhg1 repeat interval, tissue was collected from the roots of plants 5 d after emergence. Plants were grown in a growth room in Metro Mix at 24°C and 16 h of light. The entire root-soil mass was removed from the pot, quickly immersed in water to remove excess soil, and flash frozen in liquid nitrogen. RNA was extracted using Trizol following the manufacturer’s recommended procedure. Contaminating DNA was removed from the samples using Turbo DNase following the manufacturer’s guidelines. To amplify cDNA from RNA, Bio-Rad’s iScript kit was used with 1 μg of total RNA per reaction following the manufacturer’s recommended guidelines. qPCR was performed as described previously (Cook et al., 2012). Briefly, primer pairs corresponding to transcripts of Glyma18g02570, Glyma18g02580, Glyma18g02590, Glyma18g02600, and Glyma18g02610 were used to amplify products for each sample in duplicate technical replicates. A product was also amplified from each sample corresponding to transcript of gene SKP1/ASK-interacting protein16 for use in normalizing samples across plates (Cook et al., 2012).

The soybean (Glycine max) lines previously defined to make up the Hg Type Test nematode test were chosen for analysis (Niblack et al., 2002). The lines are PI 548402 (Peking), PI 88778, PI 90763, PI 437654, PI 209332, PI 89772, and PI 548316 (Cloud). The other line used and referenced in this work is Fayette, which was developed by crossing Williams(2) with PI 88778. Progeny from this cross were backcrossed with Williams(2) while selecting for SCN resistance.

Transcript Analysis

To confirm the annotation of transcripts at Rhg1, RACE PCR was performed for 3’ analysis of Glyma18g02590 (Supplemental Table S4) using the SMARTer RACE CDNA kit per the manufacturer’s protocols (Clontech) and previously defined primers (Cook et al., 2012). Following RACE, PCR products were TA cloned into pCR8/GW/TOPO as mentioned previously. Randomly chosen colonies were sequenced to confirm the 3’ ends of individual transcripts.

Fiber-FISH

Fiber-FISH experiments were carried out using the same methods and probes as detailed previously (Cook et al., 2012), and Rhg1 repeat copy number findings are based on the maximum number of copies observed in at least 10 separate probe-hybridizing DNA fibers for a given plant genotype.

Whole-Genome Sequencing

Whole-genome sequencing was performed for lines Peking (PI 548402), PI 90763, PI 437654, PI 209332, PI 89772, and Cloud (PI 548316). Tissue was collected from at least five plants per sample totaling at least 3 g of tissue to homogenize any somatic or possible intraplant DNA variants. DNA was extracted following previously published protocols (Swaminathan et al., 2007). Two separate DNA libraries were constructed for each sample. For construction of the paired-end library, DNA was randomly sheared, separated, and enriched for DNA fragments ranging from 200 to 400 bp in length. Adapter-sequence was added to the ends of each sample for bar coding following Illumina guidelines. Paired-end libraries for samples PI 209332, PI 89772, and Cloud were sequenced on a single Illumina HiSeq 2000 lane, producing reads of 101 bp sequenced from both ends of the fragment. Paired-end libraries for samples Peking, PI 90763, and PI 437654 were sequenced on Illumina’s HiSeq 2500 using the rapid sequencing run, producing sequence of 101 bp in both the forward and reverse directions. A separate library was also constructed for each sample using larger insert sizes, known as a mate-pair library. DNA for each sample was randomly sheared, separated, and collected, ranging in size from 2 to 3 kb. The mate-pair libraries were constructed using the mate-pair library preparation kit from Illumina following the manufacturer’s protocols. All six libraries were sequenced in the forward and reverse directions on a single Illumina HiSeq 2000 lane, generating sequencing lengths of 101 bp per direction. All samples were demultiplexed using their respective adapter sequences and processed following Illumina’s Cassava-1.8.2 pipeline to generate data in the fastq format used for downstream applications.

An article is in preparation that will report the whole-genome sequencing data for the lines in the SoyNAM project (Q. Song, B.W. Diers, and P. Cregan,
unpublished data). Briefly, each plant sample was pair-end sequenced on an Illumina HiSeq 2000, producing reads 151 bp in length in each direction. DNA insert sizes from the samples were 300 bp.

Previously sequenced Glycine soja data were downloaded from the Sequenced Read Archive section of the National Center for Biotechnology Information, stored under accession number SRA009252 (Kim et al., 2010b). Data from runs SRK020188, SRK020190, SRK020182, and SRK020184 were processed for analysis in this research.

**Short-Read Genome Alignments**

**Rapid Genome Alignment for SoyNAM Lines**

To rapidly estimate the copy number of the *Rhg1* interval in the SoyNAM, reads were aligned to a limited reference using the program Bowtie2 (Langmead and Salzberg, 2012). The reference for mapping was created using the Bowtie2 build indexer function with input sequence corresponding to the Williams 82 reference genome (version 1.1, assembly B73). Alignments were further processed using the program Picard (version 1.95). Mapped reads were processed using the program BWA (version 0.7.17; Li and Durbin, 2009), and read depth was computed using the coverageBed program of BEDtools (Quinlan and Hall, 2010) for each 1-kb interval from 1,600,000 to 1,694,000. Depth was estimated by mapping the number of reads corresponding to the region 5'-3' of the *Rhg1* repeat (1,600,000-1,631,999), the *Rhg1* repeat (1,632,000-1,663,999), and the 3' region (1,664,000-1,694,000). Copy number was estimated using both flanking regions, computed as the ratio of read depth corresponding to the *Rhg1* interval divided by the total reads in the flanking interval. Read depth was reported as the average of these two ratios along the 1-kb intervals.

**Full Genome Alignment**

Illumina sequencing reads were aligned to the full Williams 82 reference genome (build 189; http://www.phytozome.net/cgi-bin/show_genome?soybean) using the program BWA (version 0.7.1; Li and Durbin, 2009). Reads were mapped using the default settings of the *alt* function. Alignments were then paired using the *sam* function. Alignments were further processed using the program Picard (version 1.83) to add read group information (AddOrReplaceReadGroups), mark PCR duplicates and merge alignments (MergeSamFiles), and process for analysis in this research.

**Sequence Variant Detection**

Sequence alignment files were processed for variant discovery using the GATK software package (version 2.4.9; DePristo et al., 2011). The best practices were followed as described. Insertion and deletion (INDEL) sites were identified using the RealignerTargetCreator and a set list of known INDELS. Because a known INDEL list is not publicly available for soybean, one created following the GATK recommended guidelines. The list of known INDELS was created by selecting for concordance among high-confidence INDELS identified from the samples. 4105-34, LD00-3309, LG05-4292, and CLJ0995/46 (i.e., INDELS predicted with confidence from all four genomes were used as the list of known INDELS). Following the RealignerTargetCreator, samples were realigned around INDEL sites using the IndelRealigner function with the following options: --consensusDeterminationModel USE_READS, --known INDELS, --maxConsensuses 70, --LODThresholdForCleaning 0.5, --maxReadsForConsensuses 600, --maxReadsForRealignment 100000. Following realignment, variants were called using the UnifiedGenotyper algorithm with the following options: --stand call_conf 20, --stand emit_conf 15, --BadCigar, --A VariantType, --glm BOTH. To remove false variants, a filter was applied to remove variants not sequenced at least three times and having a quality score greater than 50. Variant files were annotated with the program SnpEff as documented (Cingolani et al., 2012).

**Copy Number Estimation**

Read depth in the 1-kb intervals was averaged over the two flanking intervals to determine average read depth of the region per resequenced genome and used to determine the estimated copy number of the *Rhg1* locus and the flanking intervals. We used average read depth over 1-kb intervals to estimate copy number from the whole-genome resequencing data. The analyzed interval was 93 kb, centered on the known 31-kb *Rhg1* repeat with equally spaced flanking intervals. The average read depth in 1-kb bins was determined for the flanking *Rhg1* regions and used to normalize read depth across bins. Final copy number estimates were made by averaging the normalized read depth across the three 32-kb intervals.

**Network Analysis**

To determine the *Rhg1* sequence relationships between soybean lines, we performed multiple sequence alignment using ClustalW2. The open reading frames for the genes Glyma18g02580, Glyma18g02590, Glyma18g02600, and Glyma18g02610 including 200 bp of upstream promoter sequence were concatenated and aligned. The alignment was used in SplitsTree (version 4.13.1) to construct a sequence network (Huson and Bryant, 2006). The analysis pipeline included Uncorrect P for distances and NeighborNet for network construction. Parsimony-Uninformative sites were excluded from the network.

**Analysis of Nematode Resistance**

To determine the relationship between nematode resistance and lines containing different copy numbers of *Rhg1*, we analyzed data collected as part of the Northern Regional SCN Tests (Cary and Diers, 2010, 2011, 2012, 2013). In total, we analyzed data from greenhouse nematode trials conducted on the seven *Hg*-type soybean lines and the susceptible control line Lee for 78 SCN field populations. Six plants per genotype were tested against the 78 different nematode populations. To more accurately estimate the variance for female index, we performed random replacement using the software R (R Development Core Team, 2009) with 1,000 bootstrap replicates per genotype-nematode combination. An ANOVA was computed using a linear mixed-effect model with bootstrap variances used to weight observations, expressed as the inverse of the variance. Residuals were checked for normality. P values were calculated using the generated test statistics, and a Bonferroni correction was applied to account for false positives resulting from multiple testing.

**Methylation Analysis**

**Restriction Enzyme-Based Methylation Discovery**

Locus-specific DNA methylation was analyzed using the methylation-specific endonuclease McrBC. *McrBC* digests DNA with methylated cytosines in a sequence-independent manner, while unmethylated DNA is unaffected. Restriction digestions were performed using 600 to 700 ng of DNA and manufacturer protocols. Adding the same amount of DNA to the reaction buffer with no restriction enzyme was used to set up control reactions. Samples with and without the restriction enzyme were incubated at 37°C for 90 min and then inactivated at 65°C for 20 min. DNA was visualized on a 0.8% ethidium bromide-stained gel to ensure DNA digestion. Both digested and control DNA samples were used for subsequent PCR using GoTaq flexible DNA polymerase (Promega). For *McrBC*-treated DNA, PCR primers that spanned methylated DNA did not produce the intended product following PCR because the template DNA was digested by *McrBC*. DNA that was not methylated or not treated with the enzyme yielded a product of the expected size.

**Computational Methylation Analysis**

Data were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) with series accession number GSE41753, deposited previously (Schmitz et al., 2013a). These data were analyzed using custom scripts written in Java or Bash to compute the data, and the results are presented in Figure 8 and Supplemental Figures S3, S4, and S5. To estimate *Rhg1* copy number, sequences from GEO accession number GSE41753 (GEO nos. GSM1134705, GSM1134706, GSM1134707, GSM1134708, GSM1134709, GSM1134712, GSM1134714, GSM1134716, GSM1134718, GSM1134720, GSM1134722, GSM1134723, GSM1134724, GSM1134725, GSM1134734, GSM1134736, GSM1134741, GSM1134743, GSM1134744, and GSM1134756) were analyzed. The total number of cytosine sequencing reads was summed over 1-kb bins starting at position 1,600,225 and counting to the end of bin 1,696,224, for a total of 96 bins. Average sequencing coverage in the region was calculated by averaging the number of cytosine reads in the 1-kb bins over the two 32-kb
intervals flanking Rhg1, which was used to normalize the read depth for each 1-kb bin. Final copy number estimates of the three 32-kb intervals were calculated as average normalized read depth over each respective 32-kb interval (Supplemental Fig. S3A).

To determine single-base cytosine methylation at the Rhg1 locus, sequences from GEO accession number GSE41753 were used for the corresponding groups: parental lines (GEO nos. GSM1024007; GSM134698–GSM134700; GSM134709; GSM134712, GSM134714, GSM134716, GSM134720, GSM134723, GSM134729–GSM134731, GSM134734, GSM134741, and GSM134749); and three-copy Rhg1 progeny (GEO nos. GSM1024008, GSM134694, GSM134713, GSM134732, GSM134744, and GSM134756). The total number of cytosine sequencing reads and the total number of cytosine sequencing reads supporting methylation were summed over 150-bp bins starting at position 1,626,000 and counting to the end of bin 1,666,000, for a total of 280 bins. For each bin, the methylation level was computed by dividing the total number of cytosine reads supporting methylation by the total number of cytosines sequenced. Methylation levels were computed in the CG, CHG, and CHH sequence contexts. The data are represented in Figure 8 and Supplemental Figures S4 and S5.

To assess transcription differences, the reads per kilobase per million mapped (RPKM) summary statistics were used to compare transcript levels across the four tested genotypes (GEO series GSE41753_RPKM supplementary file). The samples can be accessed through the BioProject PRJNA243933 or for each Biosample: Feking (SAMN02721112), PI 90763 (SAMN02721113), PI 437654 (SAMN02721114), PI 209332 (SAMN02721115), PI 89772 (SAMN02721116), and Cloud (SAMN02721117).

Supplemental Data

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

Supplemental Figure S1. Nucleic acid sequence alignment of alpha-SNAP alleles and homolog.

Supplemental Figure S2. Positions of DNA variant sites in low- and high-copy locus types.

Supplemental Figure S3. Estimated Rhg1 copy number and expression data from a recombinant population.

Supplemental Figure S4. Differential CHH DNA methylation.

Supplemental Figure S5. Differential DNA methylation for two parents and 27 progeny at Rhg1.

Supplemental Table S1. SoyNAM sequencing summary statistics.

Supplemental Table S2. Hg-Type sequencing summary statistics.

Supplemental Table S3. Estimated Rhg1 copy number of SoyNAM lines using rapid mapping.

Supplemental Table S4. Summary of alpha-SNAP allele expression from cDNA sequencing.

Supplemental Table S5. Amino acid polymorphisms at Rhg1 paraloeus locus.

Supplemental Table S6. Sequencing frequency at Rhg1 DNA variant sites in high-copy loci.

ACKNOWLEDGMENTS

We thank Matthew Hudson and Tong Geon Lee for multiple conversations about this work and for sharing results prior to publication. We also thank Coy Plunkett, Marie Adams, John Alliet, Xiao-yu Liu, and the University of Wisconsin-Madison Biotechnology DNA Sequencing and Bioinformatics Resource Center for education and support, Nicholas Keuler for advice on statistical analysis of nematode resistance data, and the Henry Nguyen laboratory for sharing the cost of obtaining whole-genome sequence data.

Received January 15, 2014; accepted March 19, 2014; published April 14, 2014.

LITERATURE CITED

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Robertson-Lowe C, Marshall AJ, Petretto E, et al (2006) Copy number polymorphism in Fgfr3 predisposes to glomerulonephritis in rats and humans. Nature 440: 851–855

Arelle APR, Webb DM (1996) Molecular genetic diversity among soybean plant introductions with resistance to Heterodera glycines. Curr Sci 71: 230–233

Barnard RJ, Morgan A, Burgyne RD (1996) Domains of alpha-SNAP required for the stimulation of exocytosis and for N-ethylmaleimide-sensitive fusion protein (NSF) binding and activation. Mol Biol Cell 7: 693–701

Baud R, Morin M, Burgyne RD (1997) Stimulation of NSF ATPase activity by alpha-SNAP is required for SNARE complex disassembly and exocytosis. J Cell Biol 139: 875–883

Bass C, Field LM (2011) Gene amplification and insecticide resistance. Pest Manag Sci 67: 886–890

Brucker E, Carlson S, Wright E, Niblack T, Diens B (2005a) Rhg1 alleles from soybean PI 437654 and PI 89788 respond differentially to isolates of Heterodera glycines in the greenhouse. Theor Appl Genet 111: 44–49

Brucker E, Niblack T, Kopisch-Obuch F, Diens B (2005b) The effect of rhg1 on reproduction of Heterodera glycines in the field and greenhouse and associated effects on agronomic traits. Crop Sci 45: 1721–1727

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21: 255–265

Caldwell BR, Brim CA, Ross JP (1960) Inheritance of resistance of soybean to the cyst nematode Heterodera glycines. Agron J 52: 635–636

Cary T, Diens BW (2010b) 2009 Northern regional soybean cyst nematode tests. University of Illinois, Urbana. http://crops.cici.illinois.edu/research/ SCN-tests (May 2014)

Cary T, Diens BW (2011) 2010 Northern regional soybean cyst nematode tests. University of Illinois, Urbana. http://crops.cici.illinois.edu/research/ SCN-tests (May 2014)

Cary T, Diens BW (2012) 2011 Northern regional soybean cyst nematode tests. University of Illinois, Urbana. http://crops.cici.illinois.edu/research/ SCN-tests (May 2014)

Cary T, Diens BW (2013) 2012 Northern regional soybean cyst nematode tests. University of Illinois, Urbana. http://crops.cici.illinois.edu/research/ SCN-tests (May 2014)

Chang AY, Liu BY (2012) DNA methylation rebalances gene dosage after mammalian gene duplications. Mol Biol Evol 29: 133–144

Chen Z (2007) Genetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol 58: 377–406

Chinglopi P, Platts A, Wang I, Coon M, Nguyen T, Wang L, Land SJ, Lu Y, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80–92

Colgrove AL, Niblack TL (2008) Correlation of female indices from virulence assays on indrel lines and field populations of Heterodera glycines. J Nematol 40: 24–45

Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9: 938–950

Concibido VC, Diens BW, Arelle PR (2004) A decade of QTL mapping for cyst nematode resistance in soybean. Crop Sci 44: 1121–1131

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, et al (2012) Copy number variation of multiple genes at Rhg1 mediates resistance in soybean. Science 338: 1206–1209

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiva S, Yuan Y, et al (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486: 346–352

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, Yun DJ, Bressan RA, Zhu JK, Bohnert H, et al (2014) The genome of the extramphile crucifer Thellungiella parvula. Nat Genet 46: 913–918

DeBolt S (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. Genome Biol Evol 2: 441–453

Demuth JP, Hahn MW (2009) The life and death of gene families. Bioessays 31: 29–39
identifying significant copy number aberrations from array comparative genomic hybridisation data. Adv Bioinformatics 2012: 876976

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, Bogwitz MR, Martin J, Daborn PJ, Goddard ME, Batheram P, et al (2010) Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cytpg1 locus. PLoS Genet 6: e1000998

Schmitz RJ, Hey, Valdés-López O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G, et al (2013a) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, et al (2013b) Patterns of population epigenetic diversity. Nature 495: 193–198

Severin AJ, Woody JL, Bokon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, et al (2010) RNA-Seq atlas of Glycine max: a guide to the soybean transcriptome. BMC Plant Biol 10: 160

Sharma SB (1998) The Cyst Nematodes. Kluwer Academic Publishers, Dordrecht, The Netherlands

Steemans P, Hérissé AL, Melvin J, Miller MA, Paris F, Verniers J, Wellmann CH (2009) Origin and radiation of the earliest vascular land plants. Science 324: 353

Sutherland E, Coe L, Raleigh EA (1992) McrBC: a multisubunit GTP-dependent restriction endonuclease. J Mol Biol 225: 327–348

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

Tang YC, Amon A (2013) Gene copy-number alterations: a cost-benefit analysis. Cell 152: 394–405

Triglia T, Foote SJ, Kemp DJ, Cowman AF (1991) Amplification of the multidrug resistance gene pimdr1 in Plasmodium falciparum has arisen as multiple independent events. Mol Cell Biol 11: 5244–5250

Tylka GL, Gebhart GD, Marnett CC, Mullaney MP (2012) Evaluation of soybean varieties resistant to soybean cyst nematode in Iowa in 2012. Iowa State University Extension and Outreach. http://www.plantpath.iastate.edu/tylkalab/files/2012%20ISU%20VT%20report_0.pdf (May 2014)

USDA (2014) Germplasm resources information network (GRIN). National Germplasm Resources Laboratory, Beltsville, MD. http://www.ars-grin.gov/npgs/index.html (May 2014)

Walling JG, Jiang JM (2012) DNA and chromatin fiber-based plant cyto-genetics. In Plant Cytogenetics: Genome Structure and Chromosome Function, Vol 4. Springer, New York, pp 121–130

Webb DM (November 1, 2012) Quantitative trait loci associated with soybean cyst nematode resistance and uses thereof. US Patent No. 20120279953 A1

Webb DM, Ballazar BM, Rao-Arelli AP, Schupp J, Clayton K, Keim P, Beavis WD (1995) Genetic mapping of soybean cyst nematode race-3 resistance loci in the soybean PI 437.654. Theor Appl Genet 91: 574–581

Wraa ER, Koening SR (2009) Effects of diseases on soybean yields in the United States 1996 to 2007. Plant Health Progress. http://dx.doi.org/10.1094/PHP-2009-0401-01-RS (May 2014)

Wu HJ, Zhang ZH, Wang JY, Oh DH, Dassanayake M, Liu BH, Huang QF, Sun HX, Xia R, Wu YR, et al (2012) Insights into salt tolerance from the genome of Thellungiella salsuginea. Proc Natl Acad Sci USA 109: 12219–12224

Wu XL, Blake S, Sleper DA, Shannon JG, Cregan P, Nguyen HT (2009) QTL, additive and epistatic effects for SCN resistance in PI 437654. Theor Appl Genet 118: 1093–1105

Young LD (1996) Yield loss in soybean caused by Heterodera glycines. J Nematol 28: 604–607

Yu A, Lepère G, Jay F, Wang JY, Bapaume L, Wang Y, Abraham AL, Penterman J, Fischer RL, Voinnet O, et al (2013) Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense. Proc Natl Acad Sci USA 110: 2289–2294

Ziller MJ, Gu H, Müller F, Dunaghey J, Tsai LTY, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein RE, et al (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature 500: 477–481