Variation at complex disease resistance locus Rhg1

Corresponding Author
Andrew F. Bent:
Department of Plant Pathology
1630 Linden Dr, Madison, WI 53706, USA
(608) 265-3034
afbent@wisc.edu

Research Area: Genes, Development and Evolution
Distinct copy number, coding sequence and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode

Authors: David E. Cook¹⁺, Adam M. Bayless¹, Kai Wang², Xiaoli Guo¹⁺, Qijian Song³, Jiming Jiang², Andrew F. Bent¹⁺*

Affiliations:
¹Department of Plant Pathology, University of Wisconsin - Madison, Madison, WI 53706, USA.
²Department of Horticulture, University of Wisconsin - Madison, Madison, WI 53706, USA.
³Soybean Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA.

Summary:
A multi-gene locus in soybean conferring disease resistance has evolved three locus types defined by copy number, protein coding, expression, and DNA methylation differences.

Funding:
Funding was provided by awards from the United Soybean Board, Wisconsin Soybean Marketing Board, and Wisconsin Experiment Station Hatch funds to A.F.B., NSF grant DBI-0922703 to J.J., the Pioneer Fellowship in Plant Pathology awarded to D.E.C. by the American Phytopathological Society through a gift from Pioneer Hi-Bred, and a National Science Foundation Predoctoral Fellowship to A.M.B.
Present addresses:

† Laboratory of Phytopathology, Wageningen University, Wageningen, 6700EE, NL
‡ Division of Plant Sciences, University of Missouri - Columbia, Columbia, MO 65211, USA.

*To whom correspondence should be addressed. E-mail: afbent@wisc.edu
Abstract
Copy number variation of kilobase-scale genomic DNA segments, beyond presence/absence polymorphisms, can be an important driver of adaptive traits. 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 ten 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 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 multi-copy Rhg1 haplotypes that form two distinct groups. The tested high copy-number Rhg1 accessions, including PI 88788, contain a flexible number of copies (7 to 10) of the 31 kb Rhg1 repeat. The identified low copy-number Rhg1 group, including Peking and PI 437654, contains 3 copies of the Rhg1 repeat and a newly identified allele of Glyma18g02590 (predicted α-SNAP). There is strong evidence for a shared origin of the two resistance-conferring multi-copy Rhg1 groups and subsequent independent evolution. Differentially methylated DNA regions also were identified within Rhg1, that correlate with soybean cyst nematode resistance. The data provide insights into copy number variation of multi-gene segments, using as the example a disease resistance trait of high economic importance.
Introduction

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 have been characterized (Triglia et al., 1991; Labbe et al., 2007; Schmidt et al., 2010; Dassanayake et al., 2011; Heinberg et al., 2013), reviewed in (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; Hernando-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 both of the extent of methylation and other epigenetic marks throughout genomes, and of the plasticity of these marks (Schmitz et al., 2013; 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). Soybean cyst nematodes 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 US 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., 2005; Wrather and Koenning, 2009; Kim et al., 2010; Kim et al., 2011). The *Rhg1* (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., 2005; 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 amino acid transporter), *Glyma18g02590* (encoding a predicted α-SNAP 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 alternate allele of \textit{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 \textit{Rhg1}-mediated SCN resistance in commercially cultivated soybean varieties is PI 548402 (commonly and throughout this paper referred to as ‘Peking’). We found that the Peking \textit{Rhg1} contains three copies of the 31 kb region, but nucleotide sequences of the genes in Peking \textit{Rhg1} were not determined (Cook et al., 2012).

A well-documented epistasis occurs in Peking-derived SCN resistance, in which Peking \textit{Rhg1} has low efficacy relative to the \textit{Rhg1} from PI 88788, but only if Peking \textit{Rhg4} is not simultaneously present (Brucker et al., 2005; Liu et al., 2012). The responsible gene at \textit{Rhg4} was recently discovered to encode a serine hydroxymethyltransferase (SHMT) (Liu et al., 2012). Peking and PI 437654 (the source of the less used but commercially relevant ‘Hartwig’ or ‘CystX’ resistance), contain an \textit{Rhg4} allele whose product exhibits altered enzyme kinetics. Impacts of \textit{Rhg4} on SCN resistance are difficult to detect when deployed together with the high copy-number \textit{rhg1-b} from PI 88788 (Brucker et al., 2005). It is intriguing and of high economic relevance that SCN populations arise that partially overcome the resistance mediated by certain sources of \textit{Rhg1} while remaining sensitive to the resistance conferred by other \textit{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 \textit{Rhg1}, and in the potential to predict, discover and/or develop more effective versions of \textit{Rhg1}.

Here we use qPCR, fiber-FISH, whole-genome sequencing, and DNA methylation analyses to investigate the major SCN resistance locus \textit{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 SoyNAM (nested association mapping) project. We discovered
three classes of the \emph{Rhg1} locus that can be differentiated by gene dosage, copy number, and coding sequence. We also observed differential DNA methylation between resistant and susceptible \emph{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 multi-gene segment contributes to a defined adaptive trait.
Results

Commonly used sources for Rhg1 resistance possess either a low copy number or high copy number of Rhg1 repeats as compared to the wild-type single copy

To assess the natural variation present at Rhg1, beyond the previous determination that there are ten 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 (2 to 4 copies) and high-copy (>6 copies) (Fig. 1A). For lines estimated to contain >6 copies, qPCR produced variable results and unreliable absolute copy number estimates, possibly because it is difficult to reduce qPCR variation below ~50% (half of one PCR cycle) between replicate tissue samples. Copy number estimates based on qPCR did however 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 RNAseq 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 Jiang, 2012). Representative fiber-FISH images for soybean lines PI 90763, PI 89772, and PI 437654 shown in Figure 2B (top 3 panels) 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 ten copies in PI 209332 (Fig. 2B, bottom 3 panels). We had previously used fiber-FISH to determine that Fayette, a soybean variety containing a Rhg1 locus originally from PI 88788, carries ten 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 previously been sequenced (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.). Supplemental Table 1 and 2 provide details regarding the sequencing datasets. For the present study we focused on in-depth analysis of *Rhgl* on chromosome 18 and its paralogous locus on chromosome 11.

To initially uncover structural variation at *Rhgl*, we screened the SoyNAM genome sequence data sets by aligning Illumina reads to a portion of the Williams 82 reference genome corresponding to *Rhgl* on soybean chromosome 18 and similar loci (see Methods). This screen determined that 8 of 35 SoyNAM lines contain an estimated *Rhgl* copy number greater than 1, 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 use of a limited reference sequence region, Illumina sequencing reads were re-mapped 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 *Rhgl* 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 3 copies, while Cloud has 7 and PI 209332 has 10 (Fig. 2C). A soybean line derived from PI 88788 was previously estimated to carry ten copies of the *Rhgl* repeat, using read-depth analysis of whole genome sequence data (Cook et al., 2012). The majority of the soybean lines chosen for the SoyNAM study were found to carry a single copy of the *Rhgl* locus (Fig. 2C; Supplemental Table S3), and have not been reported to exhibit SCN resistance where information is publicly available from the Germplasm resources information network (GRIN) (USDA). Seven other SoyNAM lines contain 9 to 10 copies of the *Rhgl* locus, while one line contains an estimated 3 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 *Rhgl* is not a common phenomenon in *Glycine max* 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 $\alpha$-SNAP.

The whole genome sequence data of the SoyNAM and Hg Type Test lines were analyzed for Rhg1 nucleic acid and derived amino acid variation. 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 tool-kit (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, and was 260 ± 10 for the high-copy lines, while it was 23 ± 29 for the lines estimated to contain a single copy of Rhg1 (see Table I for full results). As is further described in a later section, 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 only 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 Glyma18g02610 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 which impact similar protein sites (Fig. 3A). The gene Glyma18g02590 encodes a predicted $\alpha$-SNAP; in other organisms these proteins have the canonical function of stimulating NSF 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
Rhgl 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 Rhgl haplotypes, including the previously reported sequence from Fayette, and a new allele was found in all tested lines carrying the low-copy Rhgl 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 5 or 6 amino acids, the residues that otherwise have the strongest consensus sequence across eukaryote α-SNAPs (Fig. 3C), including a substitution for the leucine 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 (NSF) 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 7 independent 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 Rhgl haplotypes (Supplemental Table S4). Interestingly, a small proportion of the cDNA clones sequenced from PI 88788 and Cloud (high-copy Rhgl 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 Rhgl repeats (described below). We did not detect any Williams 82-type Glyma18g02590 sequences in cDNAs from lines carrying the low-copy class Rhgl, 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 Rhgl lines, and this splice-isoform was not found in the high-copy or single-copy Rhgl lines (Supplemental Table S4; Fig. 3B).

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 PI
88788-derived resistance (Matsye et al., 2012). Our results from whole genome sequencing indicate, however, 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 paralog 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 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 non-hierarchical phylogenetic network using the NeighborNet algorithm in Splits-tree (Bryant and Moulton, 2004; Huson and Bryant, 2006). The network reveals a clear split between the *Rhg1* loci from SCN-resistant (right) and SCN-susceptible (left) lines (Fig. 4A). There is a further split in the multi-copy 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 common 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. The 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 variant sites not present in any tested SCN-susceptible genome were universally common within either the low-copy or the high-copy Rhg1 groups: 10 sites for low-copy and 7 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 (6), PI 88788 (0), PI 90763 (1), PI 437654 (0), PI 209332 (5), PI 89772 (0), and Cloud (1).

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 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 0.85-0.9 (Fig. 5A). The other 10-15% 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 1657025 is apparently in only 4 of the 7 copies in Cloud, and in only 5 or 6 of the 10 copies in PI 209332 and 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 1657807/1657816 and 1661264/1661293, Cloud and PI 209332 appear to carry only 5 and 7-8 copies respectively of the variant site while LD00-3309 appears to carry the variant site in all 9 non-Williams 82 repeats. Conversely, the set of polymorphisms at positions 1663007-1663250 are present in only 6-7 copies in LD00-3309, 8-9 copies in PI 209332, but are present in all 6 non-Williams 82 copies of Cloud (Supplemental Table S6). Inspection of
raw sequence data for these non-homogeneous 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 non-homogeneous variant sites, the sites are in intergenic regions at least one 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 to reference sequence at 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, 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; Brucker et al., 2005; Niblack et al.,
2006; Kim et al., 2010; Klink et al., 2011). To address this we analyzed data for soybean resistance to soybean cyst nematode, from greenhouse trials conducted by Alison Colgrove, Terry Niblack and colleagues as part of the Northern Regional Soybean Trial (Cary and Diers, 2010; Cary and Diers, 2011, 2012; Cary and Diers, 2013). The analysis included data from a total of 97 field populations collected from 2009 to 2012, including SCN field populations from 8-10 north central U.S. states and/or adjacent Canada provinces per year. The results from our analysis indicate that Cloud, which contains 7 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 9 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 role of low-copy and high-copy Glyma18g02590 amino acid polymorphisms in impacting resistance to SCN is 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 identify differentially methylated regions (DMR) 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 8 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 to or within the coding sequence of *Glyma18g02600*. We also used McrBC to analyze methylation at the *Rhgl*-adjacent but non-repeated genes *Glyma18g02570* and *Glyma18g02620* and did not observe DMRs (Fig. 7).

During preparation of this manuscript, a genome-wide methylome 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., 2013). LD is known to have SCN resistance derived from PI 437654 (low-copy *Rhgl* locus type), while LDX contains a single copy of *Rhgl* (Diers et al., 2010; Kim et al., 2011). To confirm our observations and gain single-base resolution for methylation, we highlighted and re-analyzed the Schmitz et al. data, focusing on *Rhgl*.

Consistent with the findings described above, our *Rhgl* copy number estimate was 2.93 for LD and 1.17 for LDX, with various LD x LDX F3-derived (and hence potentially heterozygous) progeny families giving a range of *Rhgl* copy number estimates between 1 and 3 (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-seq characterization (see Methods). Consistent with our present and previous findings (Fig. 1B and (Cook et al., 2012)), standardized RNA sequence read depth for non-infested plants, normalized to the susceptible LDX parent, showed elevated expression for the genes encoded within but not adjacent to the *Rhgl* repeat in LD and progeny 11272, but not progeny 11268 (Supplemental Fig. S3B). This is consistent with elevated *Rhgl* copy number as a significant cause of the elevated transcript levels.

DNA methylation levels were computed from the Schmitz et al. data in bins of 150 bp in the CG, CHG and CHH sequence context, in both parents and 27 progeny lines that had at least a 4x average sequencing depth. Consistent with our above findings of differential root DNA methylation in different *Rhgl* 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 \textit{Rhg1} (Fig. 8 and Supplemental Fig. S4). Data for the full set of lines can be seen in Figure S5 (see also methods). Consistent with the finding that methylation patterns are largely inherited based on the parental methylation pattern (Schmitz et al., 2013), for \textit{Rhg1} we observed high average levels of cytosine methylation (a characteristic of the LD parent that carries three \textit{Rhg1} copies) in the progeny that appeared homozygous for the three-copy \textit{Rhg1} haplotype, and lower average \textit{Rhg1} methylation (a characteristic of the LDX parent that carries one \textit{Rhg1} copy) in the progeny homozygous for single-copy \textit{Rhg1} haplotypes (Fig. 8B, 8D). Together, our findings and the data of Schmitz \textit{et al.} 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 \textit{Rhg1} resistance.
Discussion

Soybean cyst nematode is the most economically limiting pathogen for soybean, causing billions of dollars of yield losses annually in the United States alone (Wrather and Koenning, 2009). Major efforts in soybean breeding and biotechnology are focused on the incorporation of desirable Rhgl alleles, and on continued discovery of new and better sources of SCN resistance. We had previously determined that three very tightly linked genes at Rhgl contribute to SCN resistance, and that these genes reside on a 31 kb segment that is present in ten 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 Rhgl 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 Rhgl sources was not known. Here, we report the discovery of the structural, coding, and methylation differences present at Rhgl from a diverse population of soybean lines.

The identification in different soybean lines of 7, 9, and 10 copies of an Rhgl locus composed of highly similar sequences indicates that copy number at Rhgl is plastic, and malleable over the time scale of breeding cycles. This is evidenced by the discovery of 10 copies of Rhgl in Fayette, a line developed by backcrossing Williams 82 (single copy) to PI 88788 (9 copies) (Mikel et al., 2010). In contrast, all the sequenced SCN-resistant lines belonging to the low-copy Rhgl group contained 3 copies of nearly identical Rhgl repeats. It will be interesting to identify additional sources of SCN resistance to determine if the sequences in this Rhgl group can persist in greater than 3 copies. This information, coupled with the relationship between larger numbers of Rhgl repeats and increased resistance, suggests a new strategy to improve SCN resistance through addition of Rhgl copies.

There remains a need for improved assays that can inexpensively but accurately determine the copy number of Rhgl 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 a
two-fold coverage. Comparative genome hybridization (CGH) 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 to each other and to 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. Alternatively, 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., 2005; Liu et al., 2012). In the absence of Peking-type Rhg4, the 3 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., 2005; 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 the present study.

PI 437654 is recognized for its particularly high levels of resistance against diverse nematode populations (e.g., Colgrove and Niblack, 2008). However, we discovered 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 re-emphasizes the importance of identifying and cloning additional SCN resistance QTL 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 a ~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 within the repeats and between the resistant lines, which matches predictions for gene
duplicates fixed by positive selection for increased dosage and having a low rate of non-
synonymous to synonymous substitutions \( (K_N/K_S < 1) \) (Innan and Kondrashov, 2010). However, the presence of non-synonymous 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). It is also interesting to note 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. 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 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 significantly correlated with gene duplicates and silencing, suggesting a potential mechanism for the restoration of dosage imbalance (Chang and Liao, 2012). This mechanism has also been suggested to follow whole genome duplications, for example in soybean, where for a number of gene pairs, one copy of the paralogous pair was often found to have increased repressive methylation and decreased expression (Schmitz et al., 2013). 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 multi-copy \( Rhg1 \) genes might be even greater in the SCN-resistant genomes if there were not methylation. Although beyond the scope of the present study, recent identifications of dynamic methylation changes in \( Arabidopsis \) 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 two 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 previously described. To estimate *Rhg1* copy number qPCR reactions were run with using two separate primer pairs per sample. One set of primers previously described span the junction of repeated segmental *Rhg1* duplicates, which fail 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 five days post emergence. Plants were grown in a growth room in metromix at 24°C and 16 hrs 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 manufactures recommended procedures. Contaminating DNA was removed from the samples using Turbo DNAse following manufactures guidelines. To amplify cDNA from RNA, Biorad's iScript kit was used with 1ug of total RNA per reaction following manufactures recommended guidelines. qPCR was performed as previously published (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 *SKP16* for use in normalizing samples across plates (Cook et al., 2012).

The soybean 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 88788, PI 90763, PI 437654, PI 209332, PI 89772, PI 548316 (Cloud). The other line used and referenced in this work is Fayette, which was developed by crossing...
Williams(2) with PI 88788. Progeny from this cross were backcrossed with Willams(2) while selecting for SCN resistance.

**Transcript analysis**

To confirm the annotation of transcripts at *Rhg1*, rapid amplification of cDNA ends (RACE) PCR was performed for 3’ analysis of *Glyma18g02590* (Supplmental Table IV 87- 90) using the SMARTer RACE cDNA kit per manufacturer protocols (ClonTech, Mountain View, CA) and previously defined primers (Cook et al., 2012). Following RACE, PCR products were TA cloned into pCR8/GW/TOPO as previously mentioned. 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 previously detailed (Cook et al., 2012), and *Rhg1* repeat copy number findings are based on the maximum number of copies observed in at least ten 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 5 plants per sample totaling at least 3 grams of tissue to homogenize any somatic or possible intra plant 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 bp 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, PI89772, 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 HiSeq2500 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 kb to 3 kb. The mate-pairs libraries were constructed using the mate-pair library preparation kit from Illumina following manufactures protocols. All six libraries were sequenced in the forward and reverse direction on a single Illumina HiSeq 2000 lane generating sequencing lengths of 101 bp per direction. All samples were de-multiplexed using their respective adapter sequence and processed following Illumina's Cassava-1.8.2 pipeline to generate data in the fastq format used for downstream applications.

A manuscript is in preparation that will report the whole genome sequencing data for the lines in the SoyNAM project (Q. Song et al., manuscript in preparation). Briefly, each plant sample was paired-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 was downloaded from the Sequenced Read Archive (SRA) section of the National Center for Biotechnology Information (NCBI), stored under accession SRA009252 (Kim et al., 2010). Data from runs SRR020188, SRR020190 and SRR020182, SRR020184 were processed for analysis in this research.

**Short Read Genome Alignments**

*Rapid genome alignment for SoyNAM lines*

To rapidly estimate 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 1.89) corresponding to the *Rhg1* interval on chromosome 18 (1,581,000 – 1,714,000), and the homologous loci on chromosome 11 interval (37,361,000 – 37,456,000), chromosome 2 interval (47,705,000 - 47,855,000), chromosome 9 interval (45,995,000 – 46,345,000), and chromosome 14 position (4,240,265 – 4340,264). Paired-end reads were mapped
using default settings. Mapped reads were processed using Samtools (Li et al., 2009), and read depth was computed using the coverageBed program of BEDtools (Quinlan and Hall, 2010) over 1kb bins ranging from 1,600,000 to 1,694,000. Read depth was estimated by summing the number of reads corresponding to the region 5’ 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 standard error of the mean.

**Full Genome Alignment**

Illumina sequencing reads were aligned to the full Williams 82 reference genome (build 1.89; http://www.phytozome.net/cgi-bin/gbrowse/soybean/) using the program BWA (version 0.7.1) (Li and Durbin, 2009). Reads were mapped using the default settings of the aln function. Alignments were then paired using the sampe function. Alignments were further processed using the program Picard (version 1.83) to add read group information (AddOrReplaceReadGroups), mark PCR duplicates (MarkDuplicates), and merge alignments (MergeSamFiles) from separate sequencing reactions per genome. For the Hg Type Test data processing, PCR duplicates were marked at the lane level prior to merging the sequencing runs (McKenna et al., 2010).

**Sequence Variant Detection**

Sequence alignment files were processed for variant discovery using the Genome Analysis Tool-Kit (GATK) software package (version 2.4.9) (DePristo et al., 2011). The best practices were followed as described. Insertion and deletion sites were identified using the RealignerTargetCreator and set list of known INDELs. Because a known INDEL list is not publicly available for soybean, one was created following the GATK recommended guidelines. The list of known INDELs was created by selecting for concordance among high confident INDELS identified from the samples 4J105-34, LD00-3309, LG05-4292, and CL0J095-46 i.e., INDELs predicted with confidence from all 4 genomes was used as the list of knowns. Following the RealignerTargetCreator,
samples were re-aligned around INDEL sites using the `IndelRealigner` function with options: `--consensusDeterminationModel USE_READS --known INDELS --maxConsensuses 70 --LODThresholdForCleaning 0.5 --maxReadsForConsensuses 600 --maxReadsForRealignment 100000`. Following re-alignment, variants were called using the `UnifiedGenotyper` algorithm with options: `-stand_call_conf 20 -stand_emit_conf 15 -rf 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 estimates**

Read depth in the 1kb intervals was averaged over the two flanking intervals to determine average read depth of the region per re-sequenced genome, and used to determine the estimated copy number of the `Rhg1` locus and the flanking intervals. We used average read depth over 1kb intervals to estimate copy number from the whole genome re-sequencing data. The analyzed interval was (93kb) centered on the known 31kb `Rhg1` repeat with equally spaced flanking intervals. The average read depth in 1kb 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 `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 Soybean Cyst Nematode Test (Cary and Diers, 2010; Cary and Diers, 2011, 2012; Cary and Diers, 2013). In total, we analyzed data from greenhouse nematode trials conducted on the 7 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 (Team, 2009) with 1000 bootstrap replicates per genotype-nematode combination to estimate the variance. An ANOVA was computed using a linear mixed effect model (lmer) 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 T-values, 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 endonulease McrBC. McrBC digests DNA with methylated cytosines in a sequence independent manner while unmethylated DNA is unaffected. Restriction digestions were performed using 600-700ng of DNA and manufacture 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 37C for 90 minutes, and heat inactivated at 65C for 20 minutes. DNA was visualized in a 0.8% ethidium bromide stained gel to ensure DNA digestion. Both digested and control DNA samples were used for subsequent PCR using GoTaq flexi DNA polymerase (Promega, Madison WI). 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.
Data were downloaded from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) Series accession number GSE41753, previously deposited (Schmitz et al., 2013). These data were analyzed using custom scripts written in Java or Bash to compute the data, and results are presented in Figure 8, and Supplemental Figures S3, S4 and S5.

To estimate Rhgl copy number, sequence from GEO accession number GSE41753 (GEO number GMS1024005 through GMS1024008, GMS1134684, GMS1134698 through GMS1134700, GMS1134705, GMS1134706, GMS1134709, GMS1134712 through GMS1134714, GMS1134716, GMS1134718, GMS1134720, GMS1134722, GMS1134723, GMS1134729 through GMS1134732, GMS1134734, GMS1134736, GMS1134741, GMS1134744, GMS1134749 and GMS1134756) were analyzed. The total number of cytosine sequencing reads were summed over 1kb bins starting at position 1,600,225 counting till 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 1kb bins over the two 32kb intervals flanking Rhgl, which was used to normalize the read depth for each 1kb bin. Final copy number estimates of the three 32 kb intervals was calculated as the average normalized read depth over the respective 32kb interval, Supplemental Figure S3A.

To determine single base cytosine methylation at the Rhgl locus, sequences from GEO accession number GSE41753 were used for the corresponding groups: Parental Lines (GEO# GSM1024005 and GSM1024006); Single-copy Rhgl progeny (GEO# GSM1024007, GSM1134698 through GSM1134700, GSM1134709, GSM1134712, GSM1134714, GSM1134716, GSM1134720, GSM1134723, GSM1134729 through GSM1134731, GSM1134734, GSM1134741, and GSM1134749); Three-copy Rhgl progeny (GEO# GSM1024008, GSM1134684, GSM1134713, GSM1134732, GSM1134744, and GSM1134756). The total number of cytosine sequencing reads and the total number of cytosine sequencing reads supporting methylation were summed over 150bp bins starting at position 1,626,000 through 1,668,000 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 context. The data are represented in Figure 8 and Supplemental Figures S4 and S5.

To estimate expression of genes within and adjacent to the Rhg1 repeat, processed RNA-sequencing data were used to compare transcript levels across the 4 tested genotypes (GEO series GSE41753_RPKM supplementary file). To assess transcription differences, the reads per kilobase per million mapped sequence reads (RPKM) values from the 3 replicates of the single-copy Rhg1 parent LDX01-1-165 were first averaged. This number was used as a normalizer for the average of the RPKM of the three replicates for the other three lines tested.

Access to sequence data
The whole genome sequence data generated for this work has been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). The samples can be accessed through the BioProject PRJNA243933, or for each BioSample: Peking (SAMN02721112), PI90763 (SAMN02721113), PI437654 (SAMN02721114), PI209332 (SAMN02721115), PI89772 (SAMN02721116), and Cloud (SAMN02721117).

Acknowledgements
We thank Matthew Hudson and Tong Geon Lee for multiple conversations about this work and sharing of results prior to publications, and Guy 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.
References:


Brucker E, Carlson S, Wright E, Niblack T, Diers B (2005) Rhg1 alleles from soybean PI 437654 and PI 88788 respond differentially to isolates of Heterodera glycines in the greenhouse. Theoretical and Applied Genetics 111: 44-49


Klink VP, Hosseini P, Matsye PD, Alkharouf NW, Matthews BF (2011) Differences in gene expression amplitude overlie a conserved transcriptomic program occurring between the rapid and potent localized resistant reaction at the syncytium of the Glycine max genotype Peking (PI 548402) as compared to the prolonged and potent resistant reaction of PI 88788. Plant Molecular Biology 75: 141-165

Matson AL, Williams LF (1965) Evidence of a fourth gene for resistance to the soybean cyst nematode. Crop Sci. 5: 477


USDA A, National Genetic Resources Program (2014) Germplasm ressource information network (GRIN). In. National Germplasm Resources Laboratory, Beltsville, Maryland


www.plantphysiol.org on July 19, 2017 - Published by Downloaded from Copyright © 2014 American Society of Plant Biologists. All rights reserved.


Figure 1. Estimates of copy number and transcript abundance suggest different types of Rhg1 loci.

(A) Initial Rhg1 copy number estimates, obtained using qPCR to amplify genomic DNA, identify two groups of SCN-resistant lines: low-copy number, between 2 and 4 repeats (PI 90763, PI 89772, PI 437654) and high-copy number estimated to have greater than 7 repeats (Cloud, PI 209332, Fayette, PI 88788). Copy number is expressed as ratio of qPCR template abundance estimates for Rhg1 repeat junction and for a non-duplicated neighboring gene. (B) Transcript abundance, relative to SCN-susceptible Williams 82, indicates the presence of two expression groups of Rhg1 loci in SCN-resistant lines. Roots from lines identified in subsequent work as having 3 copies of the 31 kb Rhg1 repeat (Peking, PI 90763, PI 89772, PI 437654) exhibit lower transcript abundance than lines with 7, 9 or 10 Rhg1 copies (Cloud, PI 88788, PI 209332). The one complete copy of Glyma18g02570, located immediately outside of the Rhg1 repeat region (Cook et al. 2012 and present work), is expressed at a similar level across all the tested lines. Expression level of Glyma18g02600 is near the detection limit for qPCR.

Figure 2. Whole genome re-sequencing and fiber-FISH define the copy number of Rhg1 in Hg type lines.

(A) Diagram of red (11.5 kb) and green (25.3 kb) DNA probes used to detect Rhg1 repeats in fiber-FISH. Gene and bp numbers are from chromosome 18 of the soybean Williams 82 reference genome. (B) Representative fiber-FISH images collected from six Hg Type Test soybean lines. As previously documented for three soybean lines (Cook et al. 2012), the Rhg1 locus is present as multiple direct repeats on single DNA fibers. The present data indicate a copy number of 3 for PI 90763, PI 89772 and PI 437654, and copy numbers of 7, 9 and 10 for PI 548316 (Cloud), PI 88788 and PI 209332, respectively. The repeats are labeled for clarity for the representative fiber shown in box 88788 (PI 88788). White bar = 10 μm in each panel. (C) Rhg1 copy number for 30 soybean lines, based on whole genome sequence read depth analysis. Average read depth was determined for 1 kb bins across the Rhg1 repeat and for 30 kb on each side of the Rhg1 repeat region. Data
for the flanking single-copy regions from a given line were used to normalize the read depth data of 1kb bins within the Rhg1 repeat to determine copy number (mean ± standard error of the mean). (D) Copy numbers determined as in panel (C), but for the Rhg1 paralog locus on chromosome 11.

**Figure 3.** Resistant type Rhg1 classes encode unique α-SNAP alleles with polymorphisms in highly conserved residues localized at the C-terminus

(A) Structure of Glyma18g02590 from Williams82 modeled to the crystal structure of yeast α-SNAP (sec17p, PDB 1QQE). The Q203K substitution unique to high-copy Rhg1 encoded α-SNAPs is colored orange. The D208E substitution present only in low-copy Rhg1 α-SNAPs is shown in red. An alternative splice isoform detected in low-copy Rhg1 classes removes 12 residues from the full-length protein (displayed in yellow). Similar, but distinct polymorphisms in the final five - six C-terminal residues are found in both Rhg1 resistant-class α-SNAPs (Polymorphic residues not modeled; predicted to be unstructured). (B) Amino acid sequence of Glyma18g02590 from Williams82 (susceptible) aligned to predicted amino acid sequences of both high and low copy class Rhg1 α-SNAPs. Note that the low-copy splice isoform is predicted to exclude residues 209 – 220. No predicted amino acid polymorphisms in Glyma18g02590 from the sequenced SCN susceptible lines have been detected. (C) Logo displaying the consensus sequence for the final ten C-terminal residues of α-SNAP from eight diverse eukaryotes (H. sapiens, D. melanogaster, S. cerevisiae, C. elegans, D. rerio, B. taurus, A. thaliana, G. max). Strikingly, the unique C-terminal polymorphisms discovered in Rhg1 resistant type α-SNAPs occur at these five most highly conserved residues.

**Figure 4.** Network analysis and shared polymorphisms support three Rhg1 locus types and a single SCN-resistant-type progenitor

(A) Neighbor-net analysis indicates two distinct groups based on Rhg1 sequence, separating the SCN susceptible lines (left, single-copy Rhg1) and the SCN resistance
lines (right, multi-copy Rhg1). The resistant lines further split into two groups that correspond with Rhg1 copy number (see Figs. 1 and 2); lines containing 3 Rhg1 copies are noted in red and those containing ≥ 7 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 number 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 shown at correct x-axis position for reference (blue exons, black line introns, grey untranslated regions); gene name is above gene model (e.g., 2570 = Glyma18g02570).

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 homogeneous presence of same non-Williams 82 DNA sequence for variant sites in all copies (left quarter) or all but one copy (right three-quarters) of the Rhg1 repeat. X-axis: Location of DNA variant site (SNP or INDEL) within Rhg1 locus on soybean chromosome 18. Y-axis: 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 combined for the three Rhg1 high-copy class Hg Type Test soybean lines LD00-3309 (PI 88788), PI 209332 and Cloud; mean frequency and std. err. of mean for the three soybean lines are shown. Rhg1 locus gene models shown at correct x-axis position for reference (blue exons, black line introns, grey untranslated regions); gene name is above 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. Figure 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 indicates similarities in resistance profiles based on copy number.

Nematode development data obtained for the seven Hg Type Test SCN-resistant lines, for greenhouse assays conducted as part of the 2009-2012 Northern Regional SCN Tests. Data analyzed for 78 *H. glycines* nematode populations collected from 12 U.S. states and Canada provinces. Female index is the percentage of SCN cysts that developed on the resistant soybean line relative to the susceptible control soybean line. Boxes show median and 25%-75% range of data; whiskers extend to 10% and 90% of the data. For statistical analysis, variance was calculated by random replacement with 1000 bootstrap replicates for each line within a given nematode population (see methods). This calculated variance was used in a weighted ANOVA; soybean lines not sharing the same letter above whisker had significantly different means following Bonferroni correction for multiple testing, at a p-value < 0.001.

Figure 7. Differential *Rhg1* locus DNA methylation between SCN-resistant and SCN-susceptible lines, particularly in control regions upstream of SCN resistance genes.

(A) Representative gel images of PCR products from soybean root genomic DNA template treated with restriction endonuclease McrBC (+) or buffer only (-) prior to PCR. McrBC cleaves (G/A)mC sites containing methylcytosine, preventing PCR amplification of cleaved template strands so that PCR product abundance goes down with increasing levels of methylation. “Differential DNA Methylation” scored as positive if any soybean line differed from other lines in McrBC-sensitivity of the PCR product in two independent tests. Soybean lines are denoted as either resistant (R) or susceptible (s) to SCN. (B) Summary table for replicated McrBC study described in (A) with 23 PCR primer pairs used to assess DNA methylation within the *Rhg1* locus. The presence of methylation is listed as yes if both DNA samples showed reduced PCR amplification
following McrBC DNA treatment. Right column reports methylation differences between different soybean lines.

**Figure 8. DNA methylome sequence from three-copy and single-copy Rhg1 lines and their progeny further define differential methylation at Rhg1 SCN resistance genes.**

Levels of DNA methylation reported as proportion of methylated cytosines detected from bisulfite sequencing. Data are for 150bp bins represented by a single vertical line. *Rhg1* locus gene models are shown at the top of panel (A) at correct x-axis position along chromosome 18 shown below panel (D) for reference (blue exons, black line introns, grey untranslated regions); gene name above gene model (e.g., 2570 = Glyma18g02570).

(A) Levels of cytosine methylation for the sequence context CG, showing differential methylation of parental line LD (three-copies of *Rhg1*, red vertical lines above x-axis) relative to parental line LDX (single copy of *Rhg1*, black vertical lines below x-axis). The greatest differential methylation is present up and downstream of the *Glyma18g02580* open read frame (ORF), in the common promoter for *Glyma18g02580* and *Glyma18g02590*, and both up and downstream of the *Glyma18g02610* ORF, with more methylation in the three-copy *Rhg1* SCN-resistant line. (B) Average CG methylation in F3-derived progeny families of the cross between lines LD and LDX, either for all six progeny estimated to have an *Rhg1* copy number of 3 (red vertical lines above x-axis), or for all 16 progeny lines estimated to have an *Rhg1* copy number of 1 (black vertical lines below x-axis). Substantial similarities to the parental CG methylation patterns are evident. (C), (D) Levels of cytosine methylation for the sequence context CHG, where H can be either an A,T, or C. (C) Analysis similar to (A) except for CHG sequence context. The same regions identified as differentially methylated in (A) are again identified as hypermethylated. (D) Analysis similar to (B) except for CHG sequence context, using the same progeny as (B).
Table I. Summary statistics for DNA variant analysis at *Rhg1* from whole genome sequencing shows higher rates of polymorphism in SCN resistant lines.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Variant Class</th>
<th>Copy #</th>
<th>SNPs</th>
<th>IN.</th>
<th>DEL.</th>
<th>Total</th>
<th>Variant Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD00-3309</td>
<td></td>
<td>10.03</td>
<td>190</td>
<td>36</td>
<td>32</td>
<td>258</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG05-4292</td>
<td></td>
<td>9.23</td>
<td>195</td>
<td>38</td>
<td>34</td>
<td>267</td>
<td>Intron 6</td>
</tr>
<tr>
<td>4J105-34</td>
<td></td>
<td>9.63</td>
<td>196</td>
<td>35</td>
<td>32</td>
<td>263</td>
<td>Exon 6</td>
</tr>
<tr>
<td>CL03095-46</td>
<td></td>
<td>9.76</td>
<td>195</td>
<td>36</td>
<td>32</td>
<td>263</td>
<td>Intron 6</td>
</tr>
<tr>
<td>LD02-4485</td>
<td></td>
<td>9.79</td>
<td>189</td>
<td>36</td>
<td>33</td>
<td>258</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LD02-9050</td>
<td></td>
<td>9.56</td>
<td>177</td>
<td>32</td>
<td>28</td>
<td>237</td>
<td>Intron 6</td>
</tr>
<tr>
<td>Maverick</td>
<td></td>
<td>9.04</td>
<td>187</td>
<td>33</td>
<td>31</td>
<td>251</td>
<td>Exon 6</td>
</tr>
<tr>
<td>Cloud</td>
<td></td>
<td>7*</td>
<td>194</td>
<td>35</td>
<td>32</td>
<td>261</td>
<td>Intron 6</td>
</tr>
<tr>
<td>PI 209332</td>
<td></td>
<td>10*</td>
<td>197</td>
<td>35</td>
<td>33</td>
<td>265</td>
<td>Exon 6</td>
</tr>
<tr>
<td>PI 437654</td>
<td></td>
<td>3*</td>
<td>193</td>
<td>37</td>
<td>35</td>
<td>265</td>
<td>Intron 6</td>
</tr>
<tr>
<td>Peking</td>
<td></td>
<td>3*</td>
<td>200</td>
<td>38</td>
<td>33</td>
<td>271</td>
<td>Exon 6</td>
</tr>
<tr>
<td>PI 89772</td>
<td></td>
<td>3*</td>
<td>194</td>
<td>37</td>
<td>36</td>
<td>267</td>
<td>Intron 6</td>
</tr>
<tr>
<td>PI 90763</td>
<td></td>
<td>3*</td>
<td>192</td>
<td>37</td>
<td>34</td>
<td>263</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LD01-5907</td>
<td></td>
<td>2.95</td>
<td>146</td>
<td>18</td>
<td>17</td>
<td>181</td>
<td>Exon 6</td>
</tr>
<tr>
<td>NE3001</td>
<td></td>
<td>0.99</td>
<td>63</td>
<td>3</td>
<td>6</td>
<td>72</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG05-4317</td>
<td></td>
<td>1.01</td>
<td>72</td>
<td>4</td>
<td>5</td>
<td>81</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG94-1128</td>
<td></td>
<td>1.02</td>
<td>45</td>
<td>2</td>
<td>3</td>
<td>50</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG92-1255</td>
<td></td>
<td>1.02</td>
<td>46</td>
<td>6</td>
<td>3</td>
<td>55</td>
<td>Exon 6</td>
</tr>
<tr>
<td>PI 518751</td>
<td></td>
<td>1.07</td>
<td>44</td>
<td>2</td>
<td>3</td>
<td>49</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG94-1906</td>
<td></td>
<td>1.03</td>
<td>26</td>
<td>0</td>
<td>2</td>
<td>28</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG03-3191</td>
<td></td>
<td>0.98</td>
<td>28</td>
<td>0</td>
<td>2</td>
<td>28</td>
<td>Exon 6</td>
</tr>
<tr>
<td>CL03173-68</td>
<td></td>
<td>1.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>HS6-3976</td>
<td></td>
<td>1.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG04-4717</td>
<td></td>
<td>1.10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>PI 398881</td>
<td></td>
<td>1.09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>PI 427136</td>
<td></td>
<td>1.09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>PI 507681B</td>
<td></td>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG90-2550</td>
<td></td>
<td>1.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>Prohio</td>
<td></td>
<td>0.97</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td>LG98-1605</td>
<td></td>
<td>1.09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Exon 6</td>
</tr>
<tr>
<td><em>G.soja</em></td>
<td></td>
<td>1.10</td>
<td>42</td>
<td>3</td>
<td>4</td>
<td>49</td>
<td>Exon 6</td>
</tr>
</tbody>
</table>

Copy #: *Rhg1* copy number estimated from sequencing read depth, or, copy number estimates denoted with a * are from fiber-FISH. Variant Class: Whole genome sequencing (WGS) for 30 *Glycine max* soybean lines and one *Glycine soja* line were analyzed for DNA variants and classified as SNP (single nucleotide polymorphisms), IN. (insertion), or DEL. (deletion) relative to the Williams 82 reference genome. 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 number of variants in each type of genome region. Untranscribed region (UTR).
Table II. Amino acid polymorphisms for genes encoded within and adjacent to the Rhg1 repeat, from all analyzed soybean lines.

<table>
<thead>
<tr>
<th>Position</th>
<th>Low-Copy Lines</th>
<th>High-Copy Lines</th>
<th>NE3001</th>
<th>LG05-4317</th>
<th>LG94-1128</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glyma18g02570 (unknown function)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyma18g02580 (amino acid transporter)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1638989</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyma18g02590 (alpha-SNAP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1643208</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>V9I</td>
</tr>
<tr>
<td>1643225</td>
<td>D208E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1644965</td>
<td></td>
<td>E285Q</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1644968</td>
<td>D286Y</td>
<td>D286H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1644972</td>
<td>D287EV</td>
<td>D287EA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1644974</td>
<td>L288I</td>
<td>L288I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyma18g02600 (PLAC-8 family)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1647134</td>
<td></td>
<td></td>
<td>V23A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1647764</td>
<td>214 stop (Peking)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1648561</td>
<td></td>
<td></td>
<td></td>
<td>M480L</td>
<td>M480L</td>
</tr>
<tr>
<td>Glyma18g02610 (unknown function)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyma18g02620 (SEL-1-like)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Position: Chromosome 18 base-pair position relative to Williams 82 reference genome, with gene name (and putative gene product function in parentheses) above relevant bp positions. Remaining columns: soybean accessions. Low-copy lines: All five soybean lines estimated to contain 3 copies of Rhg1 have the same amino acid polymorphism and are represented by a single column. High-copy lines: All lines estimated to contain ≥ 7 copies of Rhg1 have the same amino acid polymorphism and are represented by a single column. Only three additional soybean lines contain amino acid polymorphisms for any of the six genes analyzed and are listed as individual columns. Amino acid polymorphisms are reported as the amino acid present in Williams 82, the amino acid position, and the resulting new amino acid discovered. Not shown in Table II: the soybean line Peking (but not the other low-copy lines) contains a single nucleotide deletion in some Rhg1 repeat copies that introduces a frame shift at amino acid 214 and results in a premature stop codon after amino acid 222 of Glyma18g02600 (eliminating 58% of predicted protein).
Figure 1. Estimates of copy number and transcript abundance suggest different types of *Rhg1* loci.

(A) Initial *Rhg1* copy number estimates, obtained using qPCR to amplify genomic DNA, identify two groups of SCN-resistant lines: low-copy number, between 2 and 4 repeats (PI 90763, PI 89772, PI 437654) and high-copy number estimated to have greater than 7 repeats (Cloud, PI 209332, Fayette, PI 88788). Copy number is expressed as ratio of qPCR template abundance estimates for *Rhg1* repeat junction and for a non-duplicated neighboring gene. (B) Transcript abundance, relative to SCN-susceptible Williams 82, indicates the presence of two expression groups of *Rhg1* loci in SCN-resistant lines. Roots from lines identified in subsequent work as having 3 copies of the 31 kb *Rhg1* repeat (Peking, PI 90763, PI 89772, PI 437654) exhibit lower transcript abundance than lines with 7, 9 or 10 *Rhg1* copies (Cloud, PI 88788, PI 209332). The one complete copy of *Glyma18g02570*, located immediately outside of the *Rhg1* repeat region (Cook et al. 2012 and present work), is expressed at a similar level across all the tested lines. Expression level of *Glyma18g02600* is near the detection limit for qPCR.
Figure 2. Whole genome re-sequencing and fiber-FISH define the copy number of Rhg1 in Hg type lines. (A) Diagram of red (11.5 kb) and green (25.3 kb) DNA probes used to detect Rhg1 repeats in fiber-FISH. Gene and bp numbers are from chromosome 18 of the soybean Williams 82 reference genome. (B) Representative fiber-FISH images collected from six Hg Type Test soybean lines. As previously documented for three soybean lines (Cook et al. 2012), the Rhg1 locus is present as multiple direct repeats on single DNA fibers. The present data indicate a copy number of 3 for PI 90763, PI 89772 and PI 437654, and copy numbers of 7, 9 and 10 for PI 548316 (Cloud), PI 88788 and PI 209332, respectively. The repeats are labeled for clarity for the representative fiber shown in box 88788 (PI 88788). White bar = 10 µm in each panel. (C) Rhg1 copy number for 30 soybean lines, based on whole genome sequence read depth analysis. Average read depth per kb was determined for 1 kb bins across the Rhg1 repeat and for 30 kb on each side of the Rhg1 repeat. Copy numbers for each line were determined for 1 kb regions from a given line were used to normalize the read depth data of 1kb bins within the Rhg1 repeat to determine copy number (mean ± standard error of the mean). (D) Copy numbers determined as in panel (C), but for the Rhg1 paralog locus on chromosome 11.
Figure 3. Resistant type Rhg1 classes encode unique α-SNAP alleles with polymorphisms in highly conserved residues localized at the C-terminus.

(A) Structure of Glyma18g02590 from Williams82 modeled to the crystal structure of yeast α-SNAP (sec17p, PDB 1QQE). The Q203K substitution unique to high-copy Rhg1 encoded α-SNAPs is colored orange. The D208E substitution present only in low-copy Rhg1 α-SNAPs is shown in red (red arrow points to this residue). An alternative splice isoform detected in low-copy Rhg1 classes removes 12 residues from the full-length protein (displayed in yellow). In both Rhg1 resistant-class α-SNAPs, similar but distinct polymorphisms in the final six C-terminal residues are present (these polymorphic residues not modeled; final residues predicted to be unstructured in PDB 1QQE). (B) Amino acid sequence of Glyma18g02590 from Williams82 (susceptible) aligned to predicted amino acid sequences of both high and low copy class Rhg1 α-SNAPs. Note that the low-copy splice isoform is predicted to exclude residues 209–220. No predicted amino acid polymorphisms in Glyma18g02590 from the sequenced SCN susceptible lines have been detected. (C) Logo displaying the consensus sequence for the final ten C-terminal residues of α-SNAP from eight diverse eukaryotes (H. sapiens, D. melanogaster, S. cerevisiae, C. elegans, D. rerio, B. taurus, A. thaliana, SCN-susceptible G. max). Strikingly, the unique C-terminal polymorphisms discovered in Rhg1 resistant type α-SNAPs locate in a 10-amino acid stretch of highly conserved residues.
Figure 4. Network analysis and shared polymorphisms support three Rhg1 locus types and a single SCN-resistant-type progenitor

(A) Neighbor-net analysis indicates two distinct groups based on Rhg1 sequence, separating the SCN susceptible lines (left, single-copy Rhg1) and the SCN resistance lines (right, multi-copy Rhg1). The resistant lines further split into two groups that correspond with Rhg1 copy number (see Figs. 1 and 2); lines containing 3 Rhg1 copies are noted in red and those containing 7 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 number 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 shown at correct x-axis position for reference (blue exons, black line introns, grey untranslated regions); gene name is above gene model (e.g., 2570 = Glyma18g02570).
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 homogeneous presence of same non-Williams 82 DNA sequence for variant sites in all copies (left quarter) or all but one copy (right three-quarters) of the Rhg1 repeat. X-axis: Location of DNA variant site (SNP or INDEL) within Rhg1 locus on soybean chromosome 18. Y-axis: 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 combined for the three Rhg1 high-copy class Hg Type Test soybean lines LD00-3309 (PI 88788), PI 209332 and Cloud; mean frequency and std. err. of mean for the three soybean lines are shown. Rhg1 locus gene models shown at correct x-axis position for reference (blue exons, black line introns, grey untranslated regions); gene name is above 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. Figure 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 indicates similarities in resistance profiles based on copy number.

Nematode development data obtained for the seven Hg Type Test SCN-resistant lines, for greenhouse assays conducted as part of the 2009-2012 Northern Regional SCN Tests. Data analyzed for 78 H. glycines nematode populations collected from 12 U.S. states and Canada provinces. Female index is the percentage of SCN cysts that developed on the resistant soybean line relative to the susceptible control soybean line. Boxes show median and 25%-75% range of data; whiskers extend to 10% and 90% of the data. For statistical analysis, variance was calculated by random replacement with 1000 bootstrap replicates for each line within a given nematode population (see methods). This calculated variance was used in a weighted ANOVA; soybean lines not sharing the same letter above whisker had significantly different means following Bonferroni correction for multiple testing.
Figure 7. Differential Rhg1 locus DNA methylation between SCN-resistant and SCN-susceptible lines, particularly in control regions upstream of SCN resistance genes. (A) Representative gel images of PCR products from soybean root genomic DNA template treated with restriction endonuclease McrBC (+) or buffer only (-) prior to PCR. McrBC cleaves (G/A)mC sites containing methylcytosine, preventing PCR amplification of cleaved template strands so that PCR product abundance goes down with increasing levels of methylation. “Differential DNA Methylation” scored as positive if any soybean line differed from other lines in McrBC-sensitivity of the PCR product in two independent tests. Soybean lines are denoted as either resistant (R) or susceptible (S) to SCN. (B) Summary table for replicated McrBC study described in (A) with 23 PCR primer pairs used to assess DNA methylation within the Rhg1 locus. The presence of methylation is listed as yes if both DNA samples showed reduced PCR amplification following McrBC DNA treatment. Right column reports methylation differences between different soybean lines.
Figure 8. DNA methylome sequence from three-copy and single-copy Rhg1 lines and their progeny further define differential methylation at Rhg1 SCN resistance genes.

Levels of DNA methylation reported as proportion of methylated cytosines detected from bisulfite sequencing. Data are for 150bp bins represented by a single vertical line. Rhg1 locus gene models are shown at the top of panel (A) at correct x-axis position along chromosome 18 shown below panel (D) for reference (blue exons, black line introns, grey untranslated regions); gene name above gene model (e.g., 2570 = Glyma18g02570). (A) Levels of cytosine methylation for the sequence context CG, showing differential methylation of parental line LD (three-copies of Rhg1, red vertical lines above x-axis) relative to parental line LDX (single copy of Rhg1, black vertical lines below x-axis). The greatest differential methylation is present up and downstream of the Glyma18g02580 open read frame (ORF), in the common promoter for Glyma18g02580 and Glyma18g02590, and both up and downstream of the Glyma18g02610 ORF, with more methylation in the three-copy Rhg1 SCN-resistant line. (B) Average CG methylation in F3-derived progeny families of the cross between lines LD and LDX, either for all six progeny estimated to have an Rhg1 copy number of 3 (red vertical lines above x-axis), or for all 16 progeny lines estimated to have an Rhg1 copy number of 1 (black vertical lines below x-axis). Substantial similarities to the parental CG methylation patterns are evident. (C), (D) Levels of cytosine methylation for the sequence context CHG, where H can be either an A, T, or C. (C) Analysis similar to (A) except for CHG sequence context. The same regions identified as differentially methylated in (A) are again identified as hypermethylated. (D) Analysis similar to (B) except for CHG sequence context, using the same progeny as (B).