Evolution and selection of *Rhg1*, a copy-number variant nematode-resistance locus

Tong Geon Lee, Indrajit Kumar, Brian W. Diers, Matthew E. Hudson

Supporting information

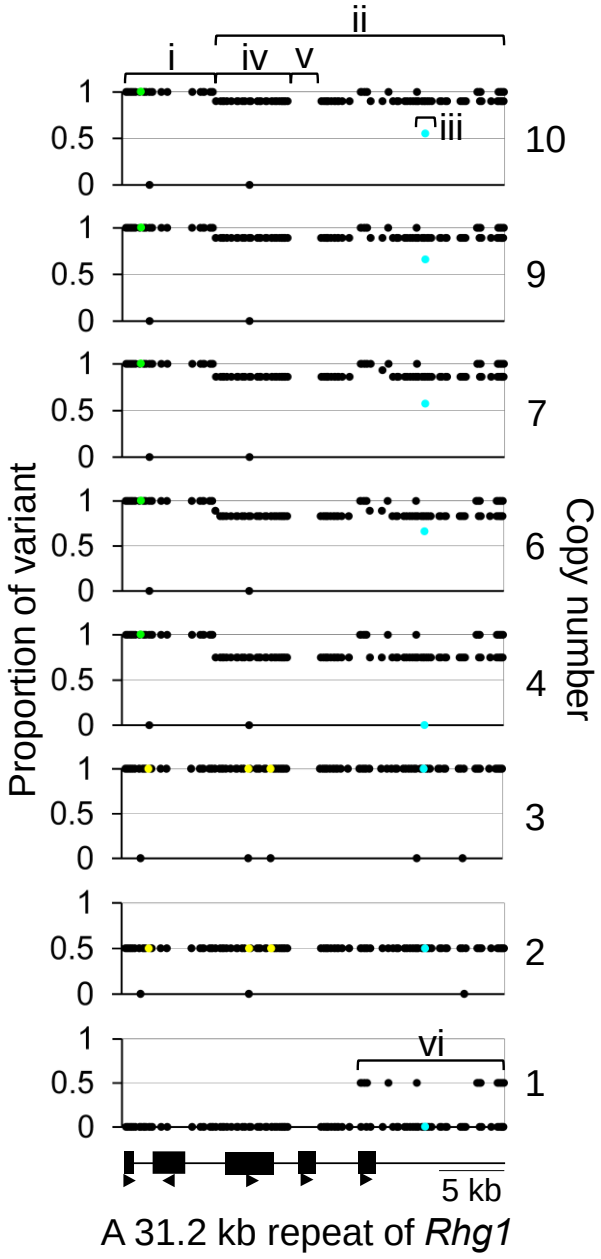
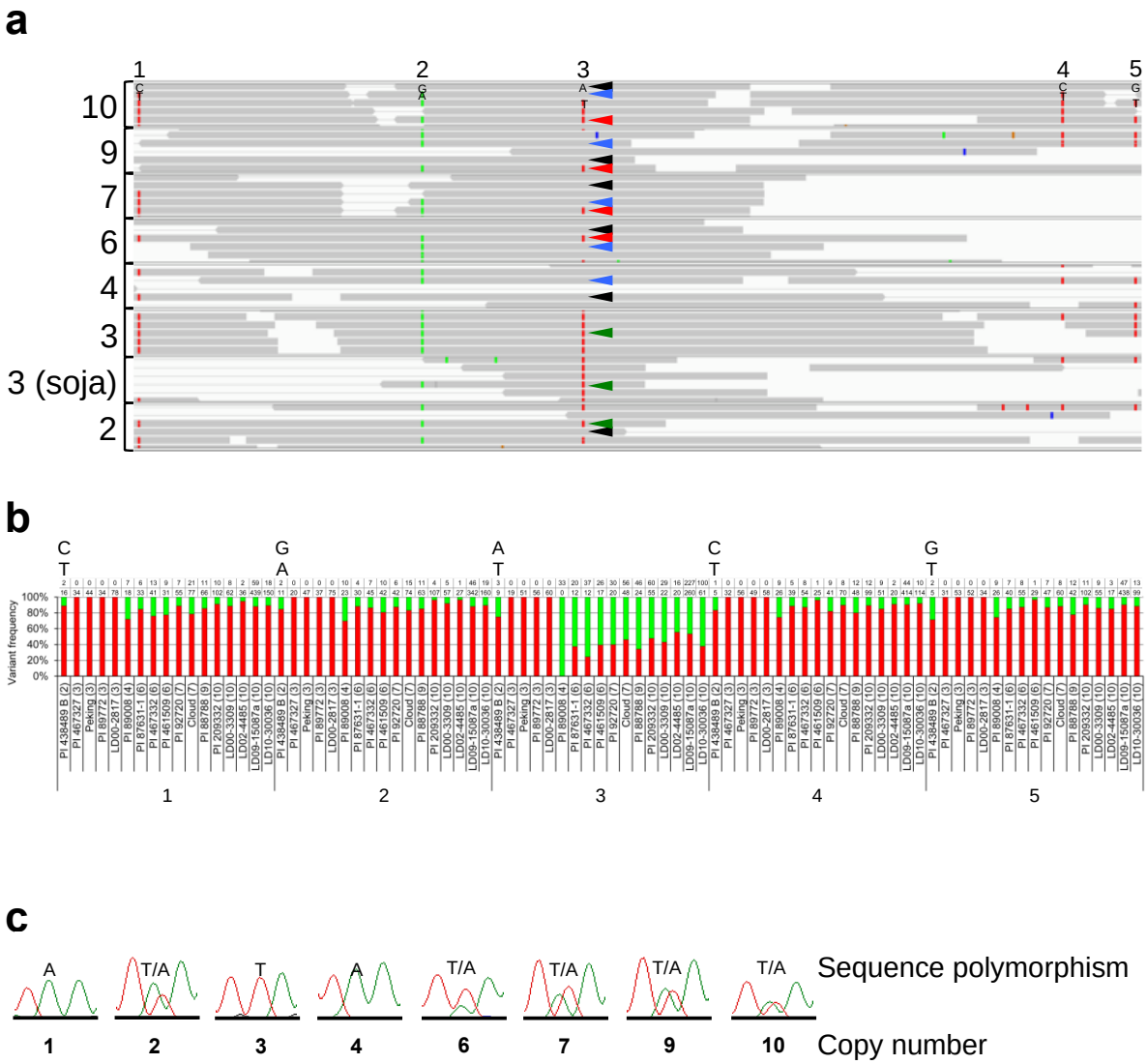


Fig. S1 Sequence variation within the *Rhg1* repeat. One hundred forty-nine single nucleotide variants are distributed in an *Rhg1* 31.2 kb repeat unit. The proportion of observed sequence variants in the reads, equivalent to an estimated probability of observing the variant ("Probability of variant") was depicted as separate graphs from top (10 copies) to bottom (single copy). A proportion of 1 represents 100% occurrence of a given

nucleotide in the reads aligning to that locus. ‘i’ marks a region where all resistant germplasm but the 2 copy variant shows 100% difference from the susceptible variant. The green dot represents a variant shared by all accessions with four or more copies. Region ‘ii’ shows probability consistent with one copy of the repeat in the lines with 4 or more copies having the same sequence as the reference, others being divergent. Yellow dots represent variants shared by two and three copy germplasm. ‘iii’ (blue dot) is a single position at 1,657,025 bp, where P varies from 0.5 to 1 depending on copy number. ‘iv’ marks the Glyma18g02590 gene encoding a predicted α-SNAP. ‘v’ marks a region with no SNPs. ‘vi’ marks the region with variants to the reference found in one single copy accession. Below the graph, gene models are shown from the Williams 82 reference genome; the final two exons of Glyma18g02570, Glyma18g02580, Glyma18g02590, Glyma18g02600, and Glyma18g02610 (from left to right).



Reconstruction of the sequences of individual repeat units. (a) Whole genome sequencing alignment view. Eight separate copy numbers are displayed in the Integrative Genomics Viewer (IGV) window. Copy number of *G. max* is denoted on the left. Copy number variation in *G. soja* is in parenthesis. Single nucleotide variants (SNVs) phased with sequences derived from the same DNA molecule are selectively noted with numbers (1 through 5). Positions of each SNV are as follows: 1; 1,656,898 bp on chromosome 18, 2; 1,656,979, 3; 1,657,025, 4; 1,657,162, and 5; 1,657,183. Representative reads with phased variants are labeled with colored arrows: black arrows; subtype W, green arrows; subtype P, blue arrows; subtype F_B, and red arrows; subtype F_A (b) Frequency of sequence variants from the Williams 82 reference at five SNV positions (Fig. 2a). Variant frequency at each position is displayed. The count observed for the two possible bases at each position are displayed above the corresponding graph. The *Rhg1* copy number of each germplasm is given in parentheses.

(c) Sanger sequencing used to confirm the presence of multiple sequences at each position. Multiple sequences at 1,657,025 bp on chromosome 18, which is equivalent to 24,801 bp within the repeat, were clearly detected from germplasm carrying two or more than two subtypes.

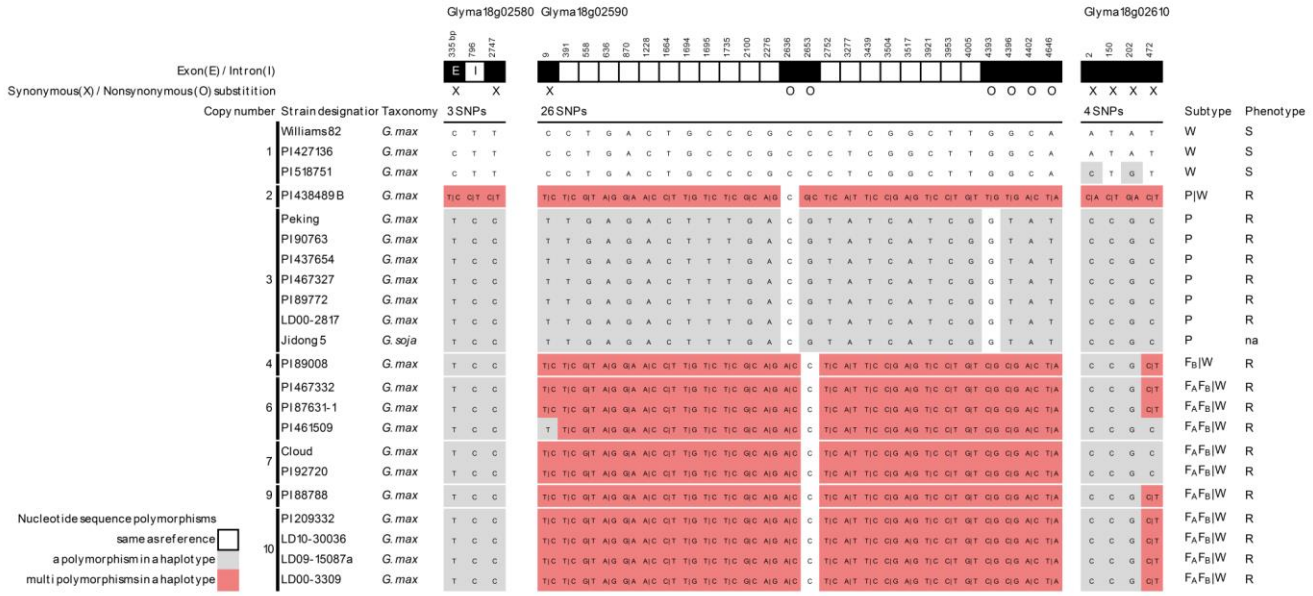


Fig. S3 Single nucleotide variations (SNVs) identified at three genes within the *Rhg1* repeat in 22 soybean germplasm accessions. Variant sites are numbered relative to the first nucleotide position of each gene. Variation at each position was colored grey (single genotype within an accession) or pale red (multiple genotypes within an accession). When multiple genotypes are observed within one accession, detected bases are divided by a vertical line. The repeat subtypes corresponding to each genotype are listed on the right. S and R indicate susceptible and resistant to SCN, respectively. ‘na’ means ‘data not available’.

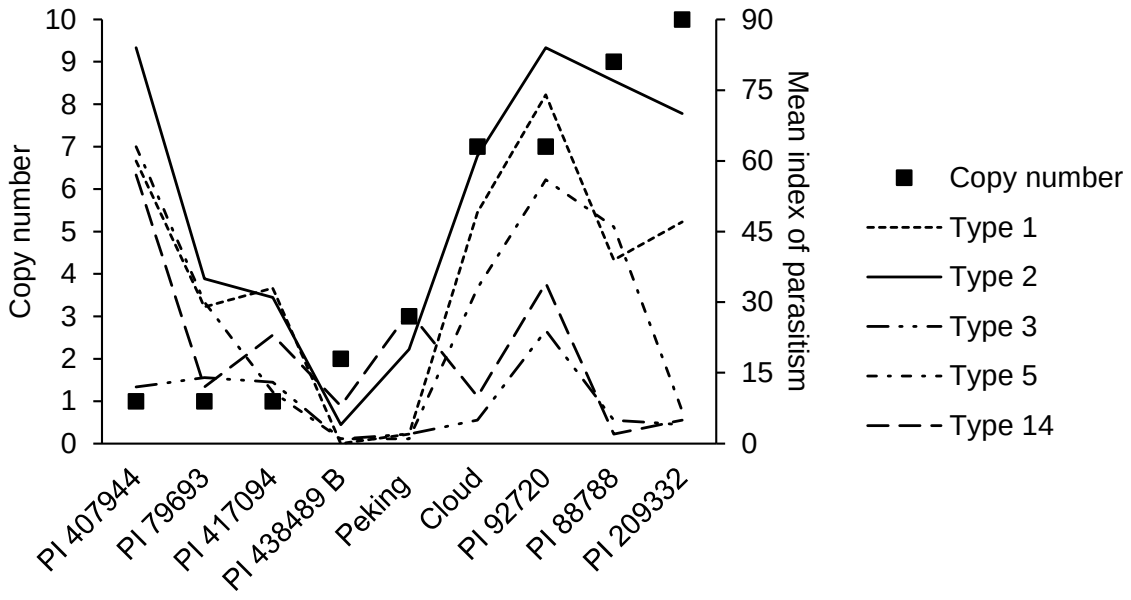


Fig. S4 Soybean cyst nematode (SCN) resistant germplasm with copy number confirmed by two methods, and their resistance reaction to SCN types. Nine SCN resistance germplasms with confirmed copy number at the *Rhg1* locus and resistance reactions to five commonly used SCN types are displayed. The mean index of parasitism was obtained from Diers *et al*. 1997.

Fig. S5 Sequence variants in the 400 kb region across the *Rhg1* locus, displayed according to copy number at *Rhg1*. Single nucleotide variants (SNVs) in coding DNA sequence (CDS) of genes in germplasm accessions are represented as colored blocks. Copy number in the accession is denoted on the left. S1, S2, and S3 are from Fig. 3b. 'tel' = direction of telomere.

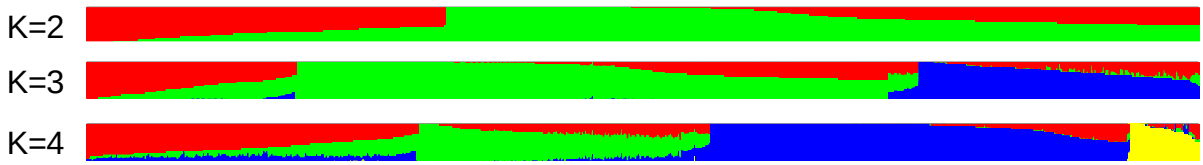


Fig. S6 Population structure estimation using *K*=2 through 4, using the same population and parameters as for **Fig. 5a**. Each individual is represented by a thin vertical line, which is partitioned into *K* colored segments that represent the individual's estimated membership fractions in *K* clusters.

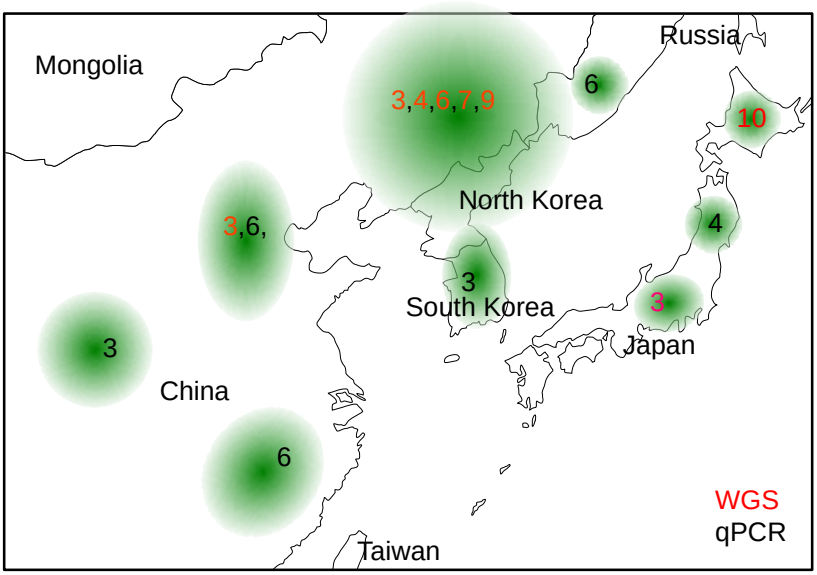


Fig. S7 Map of East Asian collection localities for SCN resistant soybean germplasm, showing copy number variation in the *Rhg1* locus. The thirty-eight resistant germplasm accessions with *Rhg1* copy number determined are distributed among diverse geographical regions. Germplasm collection coordinates (or city / province) were obtained from the National Plant Germplasm System (http://www.ars-grin.gov/npgs). Germplasm accessions without specific localities are as follows: twelve accessions from China, one from Japan, one from Korea, and

six with unknown origin. WGS: copy number data obtained by whole genome sequencing. qPCR: copy number data obtained by genomic qPCR.

Table S1 Presence/absence of repeat junction(s) and copy number estimation using whole genome sequencing (WGS) or genomic qPCR amplification of the gene Glyma18g02590 in *Rhg1*. (provided as a separate file)

Table S2 Sequence variants in the repeat junction: the sequence region that spans the centromere-proximal repeat and the adjoining non-duplicated region of the genome adjacent to *Rhg1*.

Supplemental Table S2. Sequence variants in the repeat junction: the sequence region that spans the centromere-proximal repeat and the adjoining non-duplicated region of the genome adjacent to *Rhg1*.

			Sequence variants* (-1, end of tandem repeat, 1663442 bp on chromosome 18)											
Strain designation	Cultivar name	CNV†	-198	-153	-149	-79~-80	+31	+38	+53	+63~+64	+86	+87	+105	+116
PI 209332		10	G	C	G	-	G	A	T	AATTTTTTGAATGGTGATAACGGCCAATAAT	A	T	T	A
PI 518674	Fayette, derivative of PI 88788	10	G	C	G	-	G	A	T	AATTTTTTGAATGGTGATAACGGCCAATAAT	A	T	T	A
LD09-15087a	breeding line; derivative of PI 88788	10	G	C	G	-	G	A	T	AATTTTTTGAATGGTGATAACGGCCAATAAT	A	T	T	A
PI 88788		9	G	C	G	-	G	A	T	AATTTTTTGAATGGTGATAACGGCCAATAAT	A	T	T	A
PI 548316	Cloud	7	G	C	G	-	G	A	T	AATTTTTTGAATGGTGATAACGGCCAATAAT	A	T	T	A
PI 87631-1		6	G	C	G	-	G	A	T	AATTTTTTGAATGGTGATAACGGCCAATAAT	A	T	T	A
PI 89008		4	G	C	G	-	G	A	T	AATTTTTTGAATGGTGATAACGGCCAATAAT	A	T	T	A
PI 548402	Peking	3	A	C	G	G‡	G	G	C	AATTTTTTGAATGGTGATAACGGCCAATAAT	G	C	C	G
PI 90763		3	A	C	G	G	G	G	C	AATTTTTTGAATGGTGATAACGGCCAATAAT	G	C	C	G
PI 437654		3	A	C	G	G	G	G	C	AATTTTTTGAATGGTGATAACGGCCAATAAT	G	C	C	G
PI 467327		3	A	C	G	G	G	G	C	AATTTTTTGAATGGTGATAACGGCCAATAAT	G	C	T	A
PI 89772		3	A	C	G	G	G	G	C	AATTTTTTGAATGGTGATAACGGCCAATAAT	G	C	C	G
PI 438489 B		2	G	A	A	G	A	G	C	----------	G	T	T	A
PI 518671	Williams 82	1	G	A	A	G	A	G	C	----------	G	T	T	A

*Variants were confirmed by Sanger sequencing method.

†Copy number variation (Fig. 1)

‡DNA sequence insertions based on Williams 82 genome assembly (www.phytozome.net).

Table S3 15,996 soybean germplasm accessions clustered by maximum parsimony phylogenetic analysis of the sequence region near the *Rhg1* allele, indicating germplasm accessions predicted to carry *Rhg1*. (provided as a separate file)

Table S4 DNA sequences of oligonucleotide primers used for assays.

Supplemental Table S4. DNA sequences of oligonucleotide primers used for assays.

ID	Primer	Sequence
1	1F	AGCCTGCTCCTCACAAATTCTTGC
2	9R-1	TCCTCTTGATCTCGTAGGAAAAGA
3	2590-forward	TGGAGTGGGCTGAATCTCTT
4	2590-reverse	ATGGAAGCAAGAGCAGCATT

Table S5 Summary of whole-genome shotgun sequencing data generated in this study from multiple-copy *Rhg1* germplasm accessions. (provided as a separate file)

References for Supporting information

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF (2012) Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science*, **338**, 1206–1209.

Diers BW, Skorupska HT, Rao-Arelli AP, Cianzio SR (1997) Genetic relationships among soybean plant introductions with resistance to soybean cyst nematodes. *Crop Science*, **37**, 1966–1972.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.