CNL Disease Resistance Genes in Soybean and Their Evolutionary Divergence

Madhav P. Nepal and Benjamin V. Benson

Department of Biology and Microbiology, South Dakota State University, Brookings, SD, USA.

ABSTRACT: Disease resistance genes (R-genes) encode proteins involved in detecting pathogen attack and activating downstream defense molecules. Recent availability of soybean genome sequences makes it possible to examine the diversity of gene families including disease-resistant genes. The objectives of this study were to identify coiled-coil NBS-LRR (= CNL) R-genes in soybean, infer their evolutionary relationships, and assess structural as well as functional divergence of the R-genes. Profile hidden Markov models were used for sequence identification and model-based maximum likelihood was used for phylogenetic analysis, and variation in chromosomal positioning, gene clustering, and functional divergence were assessed. We identified 188 soybean CNL genes nested into four clades consistent to their orthologs in *Arabidopsis*. Gene clustering analysis revealed the presence of 41 gene clusters located on 13 different chromosomes. Analyses of the K_s-values and chromosomal positioning suggest duplication events occurring at varying timescales, and an extrapericentromeric positioning may have facilitated their rapid evolution. Each of the four CNL clades exhibited distinct patterns of gene expression. Phylogenetic analysis further supported the extrapericentromeric positioning effect on the divergence and retention of the CNL genes. The results are important for understanding the diversity and divergence of CNL genes in soybean, which would have implication in soybean crop improvement in future.

KEYWORDS: R-genes, evolutionary divergence, gene clustering, gene duplication, NBS-LRR, nucleotide binding site, soybean CNL genes

CITATION: Nepal and Benson. CNL Disease Resistance Genes in Soybean and Their Evolutionary Divergence. *Evolutionary Bioinformatics* 2015:11 49–63 doi: 10.4137/EBO.S21782.

RECEIVED: November 12, 2014. **RESUBMITTED:** February 01, 2015. **ACCEPTED FOR PUBLICATION:** February 01, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Original Research

FUNDING: Support for this research project came from the South Dakota Soybean Research and Promotion Council (SDSRPC), USDA-NIFA Hatch Project Fund to MN (Project No. H469–13) and South Dakota Agricultural Experiment Station (SDAES). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Madhav.Nepal@sdstate.edu

Background

Plants have evolved biotic stress sensory mechanisms that activate systemic and localized diseases-resistance responses.[1] A disease-resistance response occurs when an elicitor, either a microbe-associated molecular pattern (MAMP) or a damage-associated molecular pattern (DAMP),[2] activates the basal immune system in plants. Mechanism of action between disease-resistance genes (R-genes) and pathogen avirulence (Avr) genes was first described as the "Gene-for-Gene Model" by Harold Flor in 1971. This model describes resistance as a function of an individual R-gene protein for a single pathogenic elicitor. Most of the R-gene proteins contain nucleotide binding-site (NBS) and leucine-rich region (LRR) domains, which are triggered by elicitors produced by pathogens, and then send a systemic signal to activate plant defense responses.[3] Alternatively, the "Guard Model"[4] describes NBS-LRR proteins serving as guards of certain proteins that are targets of pathogen elicitors. The "Zig-zag Model"[4] describes the coevolution of pathogens and their prospective hosts: rapid adoption of one or more Avr proteins allows the pathogen to elude the host's basal immune system until the plant produces appropriate NBS-LRR proteins for enhanced detection of the Avr elicitors.[1] Rapid pathogen adaptation to the host defense system increases evolutionary pressure on the host at molecular level through gene duplication, unequal crossing over, ectopic recombination, gene conversion, and diversifying selection.[5–7]

Plant disease resistance genes have recently been classified into eight major families: 1) Nucleotide binding site (NBS)-leucine rich region(LRR)-Toll/interleukin-1-receptors (TIR) or TNL, 2) NBS-LRR-coiled coil (CC) or CNL, 3) LRR-transmembrane domain (TrD), 4) LRR-TrD-kinase, 5) TrD-CC, 6) LRR TrD protein degradation (proline-glycine-serine-threonine) (PEST), 7) TIR-NBS-LRR-nuclear localization signal (NLS) WRKY and 8) enzymatic R-genes.[8] Among these, CNL and TNL are two commonly occurring families, which are distinguished by the domain structure at the N-terminus of the R-protein.[1] The TNL genes are found only in eudicot plants, whereas CNL genes are found in both eudicots and monocots, making these genes suitable for studying evolutionary processes across plant species.[9] Both family members have several leucine-rich repeats (LxxLxLxx) at the C-terminus of their proteins. The LRR domain typically plays a role in protein–protein interactions either directly or

indirectly during a disease-resistance response, particularly while sensing the Avr molecules.[1]

Current understanding of the evolutionary process involving CNL genes is limited. Two methods are commonly described in the literature for studying CNL genes: 1) When complete genome sequences were not available, degenerate primers were used in polymerase chain reaction (PCR) targeting the highly conserved motifs of the NBS domain. 2) With complete genome sequences now available, bioinformatics approaches are commonly used to search for orthologs with the conserved NBS motifs in the published genomes.[10] Ashfield et al.[7] studied *Rpg1b* (resistance to *Pseudomonas glycinae 1b*) in *Phaseolus vulgaris* and *Glycine max*, and showed that the evolution of NBS-LRR genes was associated with a speciation event. Differences in these genes accumulated even at the subspecies level[7] through varied recombination rates coupled with retention or deletion of redundant regions.[11] Based on the neutral theory of molecular evolution, the rate of synonymous substitutions per synonymous sites (K_s) should parallel the mutation rate under the assumption that synonymous sites are not influenced by selection.[12] Functional partnering of CNL genes with TNL genes was exhibited by the NRG1 (N requirement gene 1), a CNL type protein requires N, a TNL type protein for the resistance to tobacco mosaic virus (TMV).[13] The tomato gene *Mi–1.2* and melon *Vat* gene confer resistance to nematodes and arthropods,[14] *Arabidopsis* RPS2 resists *Pseudomonas syringae* bacteria, and RPP8[15] and RPP13[16] resist a fungal pathogen *Hyaloperonospora arabidopsidis (Hpa)*.[17] In addition, the *Arabidopsis* RPS gene that confers resistance to *Phytophthora sojae* is shown to have specific protein interactions among R-genes products, other host proteins, and pathogen effectors.[18,19] The R-genes RPM1 (resistance to *Pseudomonas maculicola* 1) and RPS2 are reported to guard the RIN4 (RPM1-interacting 4) protein. RPM1 gene is induced to signal when RIN4 is phosphorylated by AvrRpm1 and AvrB, while RPS2 is triggered as a result of AvrRpt2's degradation of Rin4 in *Arabidopsis*.[20,21]

CNL gene diversity varies from species to species: 55 of them are reported in *Arabidopsis*, 177 in *Medicago*, 6 in papaya, and 370 in potato.[1] R-genes in the soybean (*Glycine max*) genome are yet to be identified, and are of particular interest because of their defense role against pathogens and potential role in symbiosis with *Rhizobia* for biological nitrogen fixation. Recent completion of the soybean genome-sequencing project[22] has allowed us to conduct genome-wide exploration of important genes such as CNL R-genes. The main objectives of this project were to identify the soybean CNL R-genes, infer their evolutionary relationships, and assess structural as well as functional divergence of the CNL genes. Since soybean is one of the most important crop species for protein feed and vegetable oil, identification and characterization of the CNL R-genes would have implication in creating a soybean race with more durable resistant genes.

Results

Soybean CNL genes and phylogenetic relationships. Altogether, 188 CNL genes were identified in the soybean genome. Phylogenetic relationships of soybean CNL genes are shown in Figure 1. Soybean CNL genes were nested into four major clades: 1) CNL-A with 14 members, 2) CNL-B with 37 members, 3) CNL-C with 135 members, and 4) CNL-D with 2 members. Although basal support for CNL-B and CNL-C was weak, there was a strong support for the crown group CNL-A (BS 97%) and CNL-D (BS 91%). The medial parts of CNL-C did not have strong bootstrap support either; however, many well-supported relationships were identified among the crown groups, for example, relationship between Glyma09g02401 and Glyma15g13290 had a strong bootstrap support (BS 99%). The fourth group, CNL-D, had strong bootstrap support for nearly every crown group.

Twenty putative conserved motifs obtained through MEME analysis are visualized in Figure 2 and the sequences are presented in Table 1. Seven of these conserved domains (ie, P-loop, RNBS-A, Kinase-2, RNBS-B, GLPL, RNBS-C, and RNBS-D) were present in 136 CNL proteins in *G. max* (Fig. 2). P-loop, Kinase-2, and GLPL motifs, however, were present in all CNL proteins. Glyma02g38743 was unique because it possessed P-loop, Kinase-2, and GLPL but lacked other motifs. The above-mentioned seven motifs identified in *G. max* were generally in the same order as in *A. thaliana*, although the RNBS-D motif in Glyma17g36400 and Glyma17g36420 appeared earlier in the sequences. The P-loop, Kinase-2, RNBS-B, and GLPL motifs showed a high level of conservation in *G. max* and *A. thaliana*, whereas the RNBS-A, RNBS-C, and RNBS-D motifs were more variable.

CNL gene clustering, K_s-values, and sequence divergence. Forty-one gene clusters were identified using a sliding window of 10 open reading frames (ORFs) (Table 2 and Fig. 3). The CNL gene clusters were assigned names based on chromosome location. The phylogenetic tree (Fig. 1) and clustering analysis (Fig. 3) indicated a single chromosomal domination of a clade.

Fifty-six percent of the CNL genes were located on 5 of 20 chromosomes, and CNL genes were completely absent from chromosome 10. An analysis of clustering placed 31 of the 41 clusters outside the pericentromeric region. A simple χ^2-test (p = 2.11E–5) showed that these clusters are primarily located outside the pericentromeric region. The results from the analysis of synonymous substitutions per synonymous site (K_s) of all 41 gene clusters are summarized in Supplementary File 1. The occurrence of tandem duplications at different time scales can be inferred from the directional decrease in the K_s-values (from 1.2727 to 0.044), with the oldest duplication event occurring in the gene cluster 19_4 (Glyma19g321800 has the highest K_s values; nested in clade C of the phylogenetic tree as shown in Fig. 1). Other gene members in the cluster

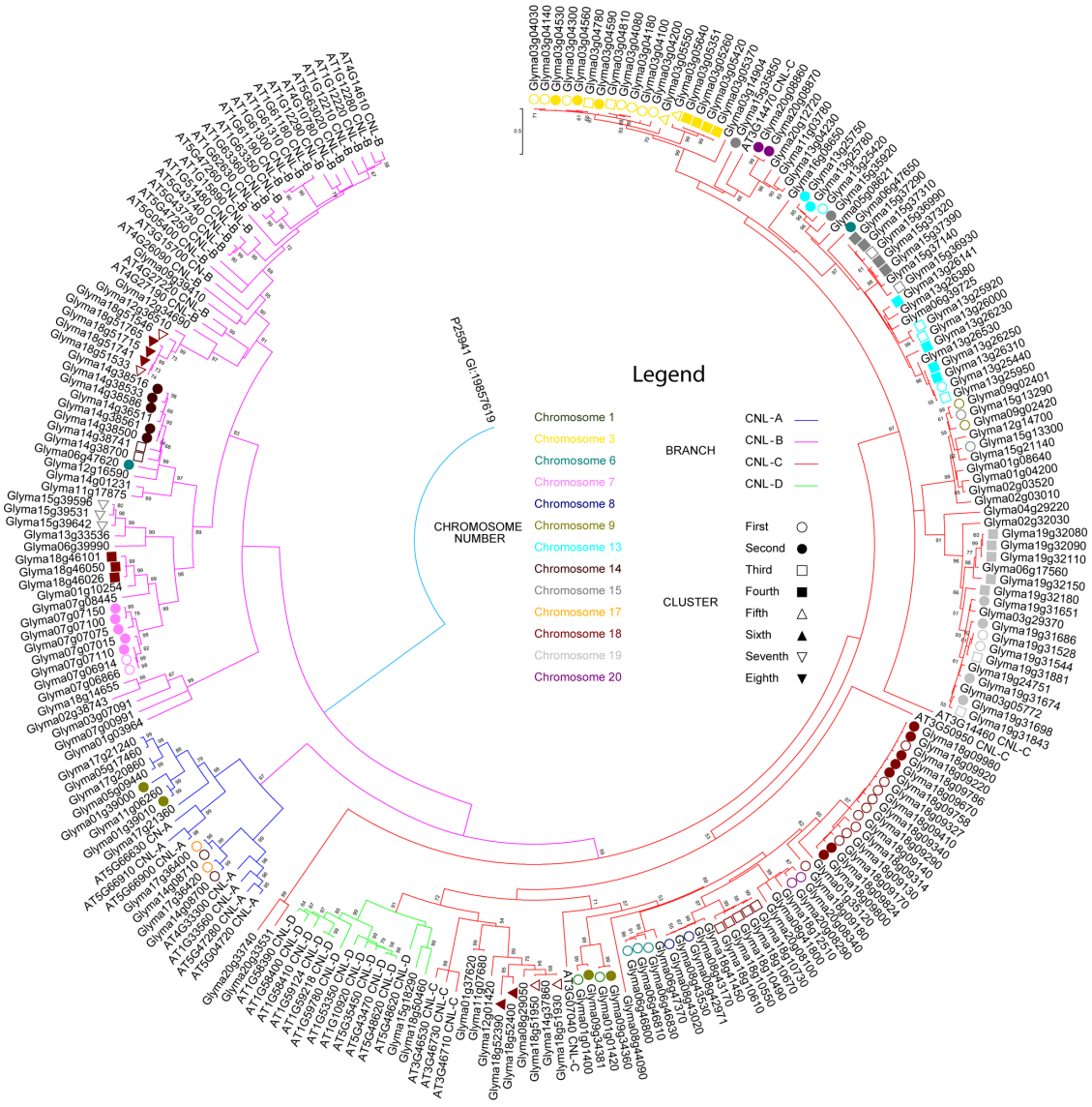

Figure 1. Maximum likelihood analysis of CNL *A. thaliana* orthologs in soybean genome.

Notes: JTT+G+I evolutionary model was used in the phylogenetic analysis. The values above the branches are the bootstrap support of 100 replicates. *Arabidopsis thaliana* (AT) and *Glycine max* (Glyma) accessions are tagged with their CNL identifier based on their phylogenetic placement. Clades are color-coded: CNL A–D in blue, purple, red, and green, respectively. Also included in the tree is the information on gene clustering: each shape indicates the order of appearance of a cluster (first to eighth represented by hollow circle, filled in circle, hollow square, filled in square, hollow triangle, filled in triangle, hollow upside down triangle, and filled in upside down triangle, respectively) on the chromosome. Each chromosome is represented by a color filled in the shape for the gene cluster.

19_4 may have arisen from consecutive tandem duplications, as evidenced from the decreasing K_s values within the cluster (Table 3). Figure 3 depicts the CNL gene locations and pericentromeric regions on the soybean chromosome pseudomolecules. Analysis of variance (ANOVA) showed a significant difference in the K_s values of the genes located within the pericentromeric region from those genes located outside ($P = 0.012$, $\alpha = 0.05$). However, the K_a (nonsynonymous substitutions per nonsynonymous site) values for the CNL genes within and outside the pericentromeric region showed no difference ($P = 0.260$, $\alpha = 0.05$). As expected, the majority of CNL gene clusters were located outside the pericentromeric region (Fig. 3; Table 2). Information on soybean CNL gene clusters on each chromosome is summarized in Table 4. Gene members in the cluster 7_1 had very low K_s-values (ie, 0.0037) suggesting the most recent duplication events. The average K_s-values of clusters 15_1 (0.219), 15_3 (0.228), and 13_2 (0.0511), 13_3 (0.169), and 13_4 (0.330) were near the suggested range for the recent duplication event of 13 MYA. The genes in cluster 19_4 were likely from tandem duplications, as summarized in Table 3. Supplementary File 1 includes the results from the K_s analysis.

Results from gene expression and structural variation analysis. Currently available expression data for soybean

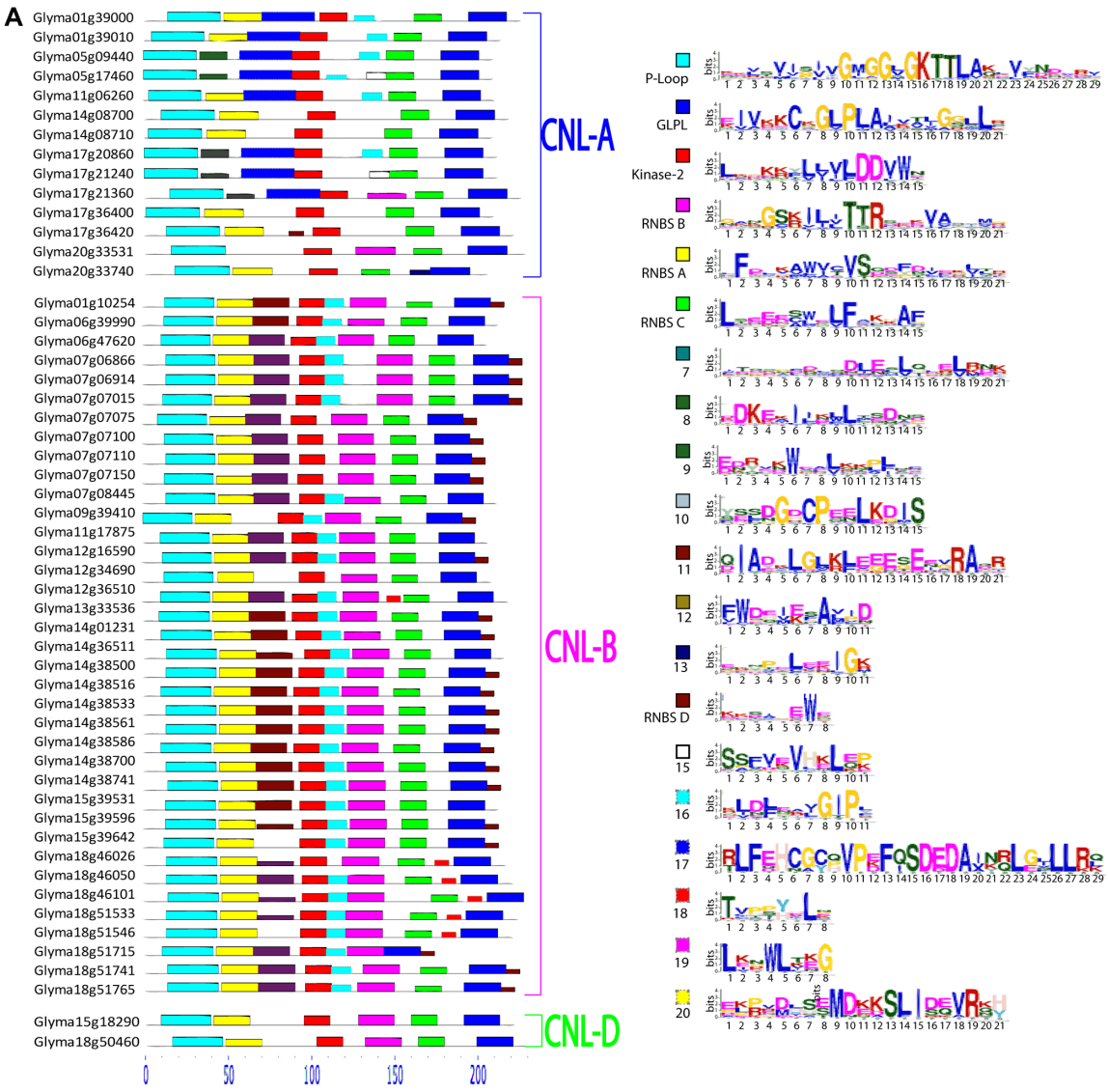

Figure 2. (*Continued*)

CNL genes are visualized as a heatmap in Figure 4 and data are presented in Supplementary File 2. Of the 188 CNL genes, 133 genes had uniquely mappable reads, ie, their expression profile would not be duplicated. Gene expression data were divided into quartiles: the upper quartile of the data (top 25% expression values) hereafter will be described as highly expressed. Highly expressed genes included 11, 6, 15, and 2 genes from CNL-A, CNL-B, CNL-C, and CNL-D, respectively. Available sets of expression data revealed that 11 of the 133 (8%) had zero expression in all tissues. CNL-A genes were among the most highly expressed genes ranging between 361 and 1845 reads (sum of expression in all tissues of q-PCR results). Within CNL-A, the gene paralogs Glyma17g36400 (479 reads) and Glyma14g08700 (361) were both low expressed, despite higher expression values for the related genes Glyma17g36420 (1147 reads) and Glyma14g08700 (756). The gene expression values for the CNL-B members ranged from zero (Glyma06g47620 and Glyma12g16590) to 1232 reads (Glyma07g07100). Low expression in Glyma06g47620 (CNL-B; zero reads) differed from its cluster mate Glyma06g47650 (CNL-C), which had a modest expression value. The expression value of CNL-C gene members ranged from zero (in nine genes) to 429 reads (in Glyma01g01400). The CNL gene clusters 18_1, 18_2, and 18_3 within the pericentromeric region had expression values ranging from zero to 30 reads with 3 exceptions of highly expressed genes (Glyma18g09920 with 38 reads, gene cluster 18_2; Glyma18g09180 with 90 reads, gene cluster 18_1; and Glyma18g09290 with 105 reads, gene cluster 18_1), as shown in Table 5. Finally, CNL-D had only two genes (Glyma15g18290 and Glyma18g50460) with expressions of 224 and 737 reads, respectively. Analysis of promoter regions showed that WBOX cassettes were present in the 2-kb upstream regions of 174 of 188 CNL genes (Fig. 4), with an average of three WBOX cassettes

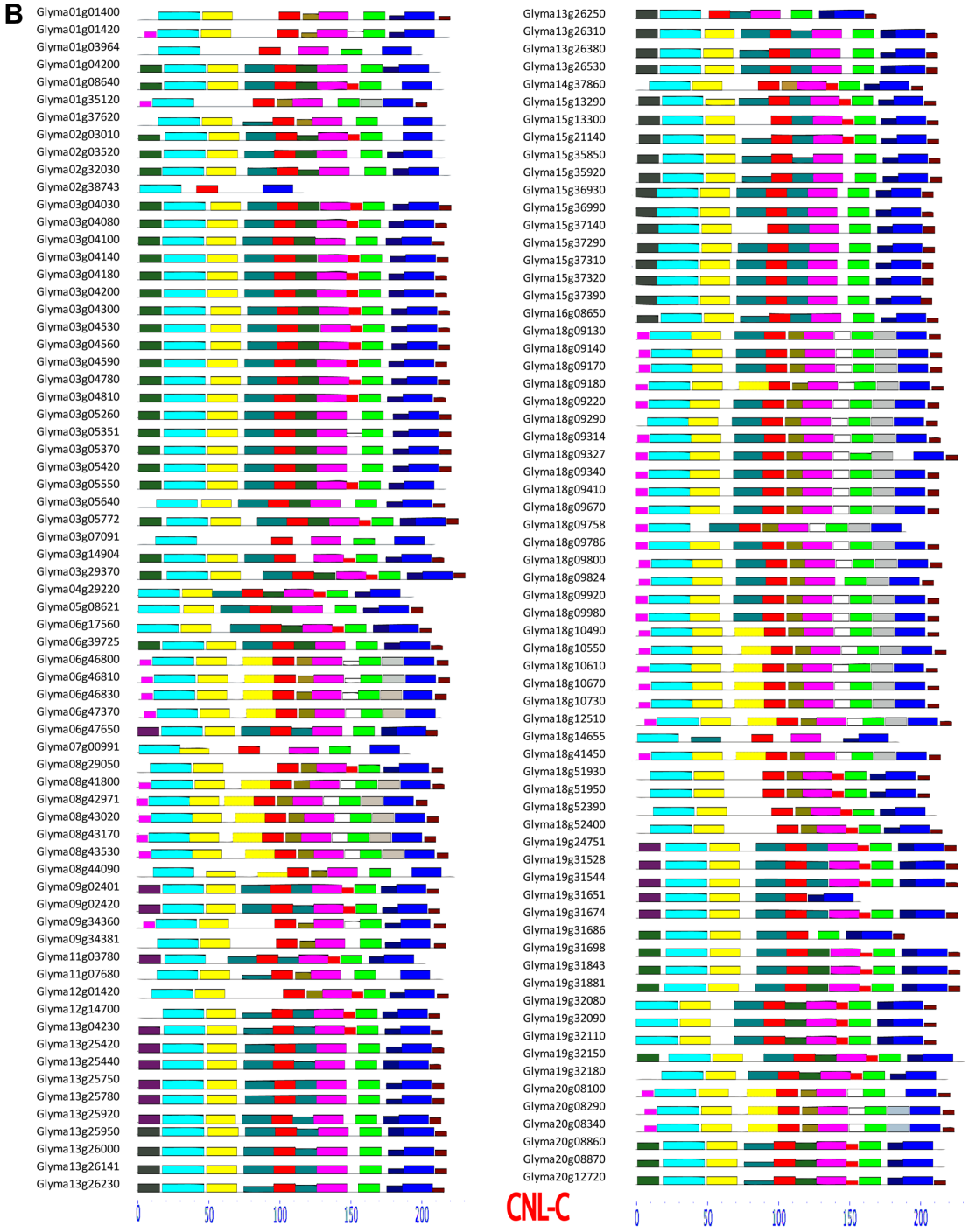

Figure 2. Conserved domains predicted by MEME analysis of soybean CNL genes.
Notes: Genes are divided into four groups (A–D) based on Figure 1. Analyzed NB-ARC regions span around 250 amino acids (~30 amino acids upstream of the P-loop to ~30 amino acids downstream of the GLPL motif). The search parameters were set to predict 20 unique motifs. (**A**) CNL-A,CNL-B,CNL-D MEME results, and the Weblogo legend for the MEME. (**B**) MEME results for CNL-C members.

per gene. For the 174 gene sequences with WBOX motif, the number of WBOX per sequences ranged from zero to nine; two WBOX motifs were most common, representing 30% of the sequences with the motif.

Figure 5 illustrates the intron–exon structure variation among soybean CNL genes and their orthologs in *Arabidopsis*. The number of exons of CNL-A, CNL-B, CNL-C, and CNL-D gene groups averaged 5.41, 5.22, 1.91, and 2 exons per gene, respectively. The number of exons in each of the CNL groups was similar to their *Arabidopsis* counterparts, except that in the members of CNL-B group, where soybean had an average of 5.22 exons per gene, which is higher than *Arabidopsis* with only 2 exons per gene. In this study, we observed a general trend where the number of exons and expression values were

Table 1. Conserved domains of soybean CNL genes as predicted by MEME analysis.

MOTIF ORDER	MOTIF ID	CONSENSUS MOTIF SEQUENCE	NUMBER
1	[19]	LKNWLTEG	38
1/3	[8]	HDKEMIINWLMSDNP	70/5
2	P-loop	NEVSVIPIVGMGGMGKTTLAQHVYNDPRV	188
3	RNBS-A	HFDCHAWVCVSQDFDIFQVQR	174
4	[7]	ITQQPCDMMDLEMLQNELRNK	96
4	[11]	QIAYMLGLKFEEESENGRAQR	33
4/12	RNBS-D	KHSAPEWE	2/145
4	[17]	RLFEHCGCQVPEFQSDEDAINRLGILLRQ	8
4	[20]	EPPHDHSEMDKKSLIDQVRQH	21
5	Kinase 2	LQGKRYLIVLDDVWN	188
6	[12]	FWDHMEFAMPD	51
6	[16]	YLDFNAIGIPY	36
6	[9]	EDYVNWEALQNPFNC	76
7	RNBS-B	GANGSRILITTRSEHVASYMQ	173
8	[15]	SSFVQVHKLQP	40
8/10	[18]	TVPPYHLP	49/3
9	RNBS-C	LTEEHCWELFCHHAF	184
10	[10]	YSSDGHCPEELKDIS	36
10	[13]	QCYPHCEEIGK	91
11	GLPL	EIVKKCKGLPLAIVTMGGMLH	188

Notes: The seven major motifs identified were P-loop, RNBS-A, RNBS-D, Kinase-2, RNBS-B, RNBS-C, and GLPL. Motif ID for the 13 previously unidentified motifs were assigned sequential numbers. The first column contains the relative location within the NB-ARC, and the last column contains the number of sequences with a match for the motif described. All motifs are presented in the order of their appearance.

Table 2. CNL gene clusters in *Glycine max* genome.

CLUSTER ID	NUMBER OF GENES	CLUSTER ID	NUMBER OF GENES	CLUSTER ID	NUMBER OF GENES	CLUSTER ID	NUMBER OF GENES
1_1	2	7_2	5	14_3	2	18_4	3
1_2	2	8_1	2	15_1	2	18_5	2
3_1	7	9_1	2	15_2	2	18_6	3
3_2	3	9_2	2	15_3	2	18_7	2
3_3	2	13_1	2	15_4	4	18_8	2
3_4	4	13_2	2	15_5	3	19_1	2
3_5	2	13_3	3	17_1	2	19_2	4
6_1	3	13_4	4	18_1	10	19_3	2
6_2	2	14_1	2	18_2	7	19_4	5
7_1	2	14_2	5	18_3	5	20_1	2
						20_2	2

Notes: Forty-one clusters containing 126 genes were identified using 10 open reading frame sliding window. Each cluster was assigned a cluster name based on its chromosomal position (chromosome number_ranked distance from the telomeric end of the short arm). The number of genes in a specific cluster is given in the column right to each cluster ID.

correlated: the gene members of CNL-A (with an average of 5.41 exons per gene) and CNL-B (5.2 exons per gene) had higher average gene expression values than the members of CNL-C and CNL-D with average of 1.9 and 2 exons per gene, respectively.

Discussion

Soybean CNL gene diversity compared to other plants. Several aspects of the NBS disease-resistance genes were previously studied in other plant species such as *Arabidopsis*,[24] *Brachypodium*,[25] *Lotus*,[5] and *Medicago*.[23] The

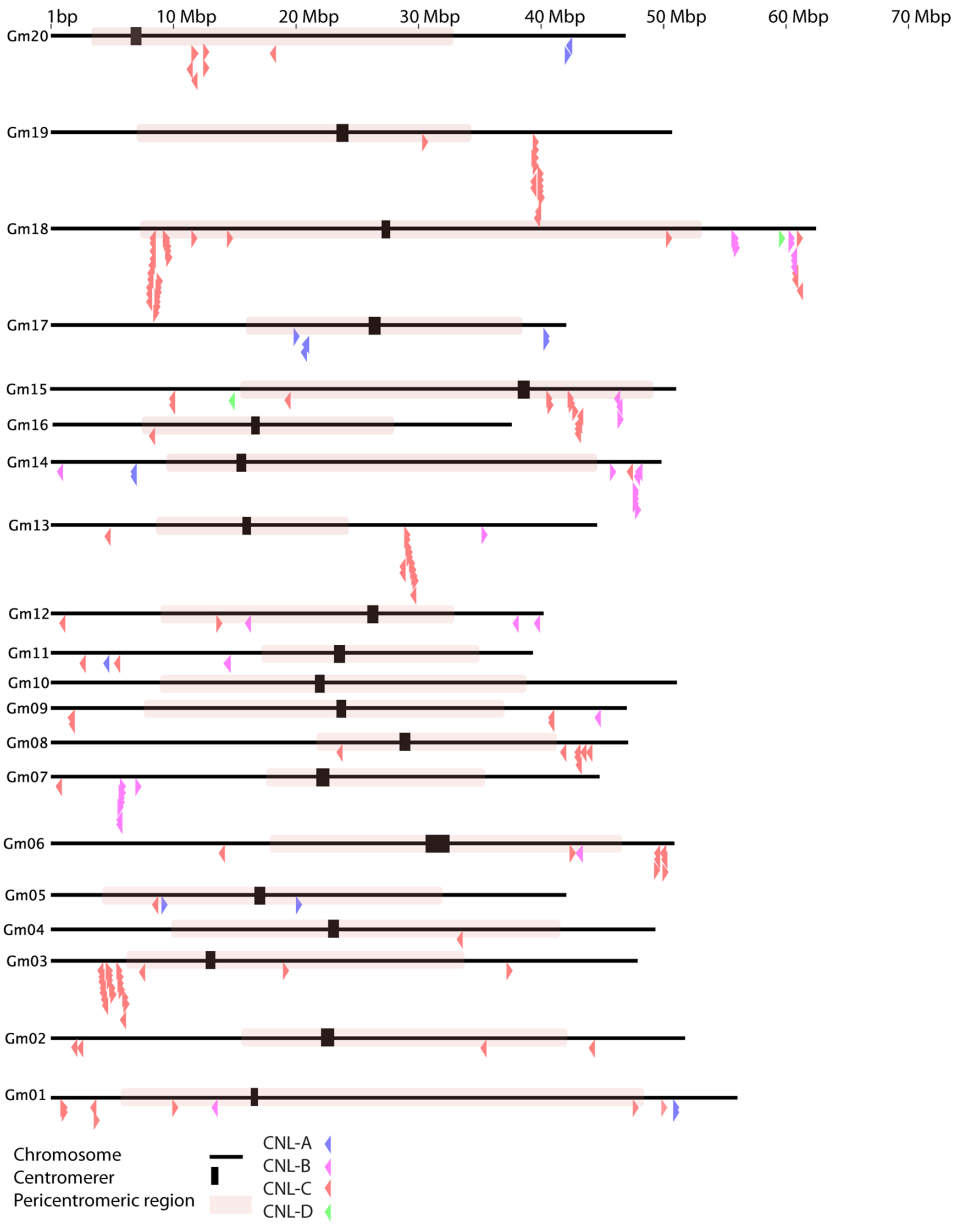

Figure 3. Chromosomal distribution of soybean CNL genes.
Notes: Each black line represents a chromosome, and each arrow indicates the location and orientation of a CNL gene. Gene groups CNL-A, CNL-B, CNL-C, and CNL-D are color-coded to blue, pink, red, and green, respectively. The black thick vertical bar indicates the centromere, and the red-shaded region is the predicted pericentromeric region.

188 CNL genes in soybean identified in the present study were similar to the 177 CNL genes in *Medicago* in number. The CNL gene number in soybean was higher than that in *Arabidopsis* (55) and papaya (6), but much fewer than that in potato (370). The NBS-encoding genes identified in soybean represented 0.35% of all the predicted proteins. As shown in Figure 1, the clade supports and tree topologies in soybean were similar to those previously reported in *Arabidopsis*.[26] The occurrence of a higher number of CNL genes in soybean than in *Arabidopsis* is attributable to polyploidization events that have increased the soybean's genome to have 3.1 copies of the majority of genes.[22] Because of their cost of fitness associated with expansion (ie, auto-activation of R-gene pathways cause the death of plants before they reproduce), and perhaps because of the "birth and death process" as discussed by Michaelmore,[27] the soybean genome has retained 3.4 copies of resistance genes similar to the expected number of gene copies in general. This increase in number comes primarily from an expansion of CNL-C members and many tandem duplications on chromosome 3 and 18. The total number of CNL genes in soybean, however, was much less than in *M. truncatula*, possibly due to genetic bottlenecks caused during the soybean domestication process. An alternate explanation would be the reduced pathogenic/parasitic environments in

Table 3. Pairwise comparisons of K_s-values for the members from the gene cluster 19_4.

GENE ACCESSIONS	19G32080	19G32090	19G32110	19G32150	19G32180
19G32080					
19G32090	0.044				
19G32110	0.1004	0.053			
19G32150	0.3917	0.4203	0.4395		
19G32180	1.1873	1.2347	1.2726	0.9312	

Notes: K_s-values are used to infer recent tandem duplications. The order of duplication is suggested by inverse of the K_s-value.[22] Glyma19g32180 and Glyma19g32080 are perhaps the oldest and youngest genes, respectively in the 19_4 cluster.

soybean, which has a longer domestication history compared to *M. truncatula*.[6]

Both NBS and LRR domains of R-proteins are vital for activating the defense signaling pathway against pathogen.[1] The NBS domain through its NTPase activity functions as a molecular switch for activating signal transduction. Some conserved motifs that can be distinguished in the NBS domain include GLPL, MHD, P-loop (Walker A or Kinase 1), Kinase-2 (Walker B), RNBS-A, RNBS-B, RNBS-C, and RNBS-D.[1] Our results showed that motifs surrounding the P-loop and those adjacent to the kinase-2 motif could be used to distinguish CNL-A group from the CNL-B group, and these two groups from the CNL-C and CNL-D groups, similar to previous findings in *Arabidopsis*.[24] Genome-wide analysis of *Brachypodium* disease-resistance genes also showed differences in these motifs flanking the P-loop and the kinase-2 motifs.[25]

Gene clustering and duplications. In the evolution of plants, gene duplications (tandem, segmental, or genome) have played important roles in the origin and maintenance of multiple gene families.[26] These gene duplications have contributed to the expansion of the NBS gene family in both eudicot and monocot lineages.[9] Plant gene clusters resulting from gene duplications experience heterogeneous rates of evolution: "fast" and "slow" evolving genes are termed "Type-I" and "Type-II" resistance genes,[28] respectively. Intraspecific nesting of the majority of CNL genes was abundant in both *Arabidopsis* and *Glycine*, suggesting that the occurrence of Type-I resistance genes evolved primarily through tandem duplications perhaps in response to rapidly evolving associated pathogens. On the other hand, there were a few cases where the CNL genes in *Arabidopsis* and *Glycine* were highly conserved (ie, AT4G26090 and Glyma09g39410 genes in CNL-B clade), suggesting the occurrence of Type II resistance genes.

Table 4. Chromosomal distribution of soybean CNL gene clusters.

CHROMOSOME	NUMBER OF GENES	PERCENTAGE (%)	CLUSTERS PER CHROMOSOME	GENES IN LARGEST CLUSTER	AVERAGE GENES/CLUSTER
1	10	5.32	2	2	2
2	4	2.13	N/A	N/A	N/A
3	22	11.70	5	7	3.6
4	1	0.53	N/A	N/A	N/A
5	3	1.60	N/A	N/A	N/A
6	9	4.79	2	3	2.5
7	9	4.79	2	5	3.5
8	7	3.72	1	2	2
9	5	2.66	2	2	2
10	0	0.00	N/A	N/A	N/A
11	4	2.13	N/A	N/A	N/A
12	5	2.66	N/A	N/A	N/A
13	15	7.98	4	3	2.75
14	12	6.38	3	5	3
15	16	8.51	5	4	2.6
16	1	0.53	N/A	N/A	N/A
17	5	2.66	1	2	2
18	38	20.21	8	10	4.25
19	14	7.45	4	5	3.25
20	8	4.26	2	2	2
Total	188	100	41	N/A	2.73

Note: Of the 188 CNL genes identified in the present study, the majority of them are located on seven different chromosomes.

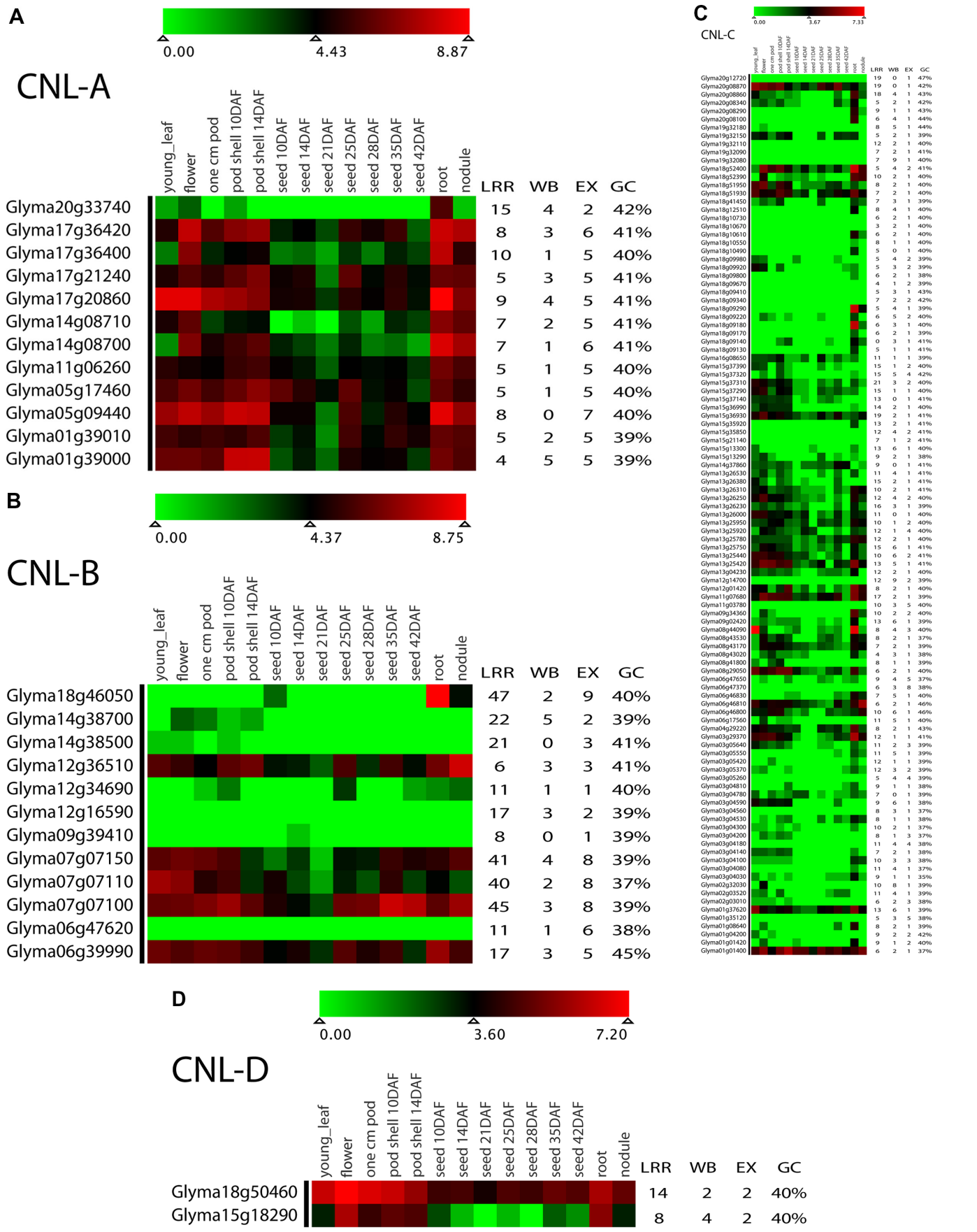

Figure 4. Expression profile for soybean CNL-A, CNL-B, CNL-C, and CNL-D genes visualized as heatmaps in panel A, B, C and D, respectively.
Notes: Heatmaps were constructed using $\log_2$-transformed data for the CNL genes in 14 tissue types shown at the top. The number of LRRs, WBOX (WB) regulatory factors, number of exons (EX), and the G + C content of the coding sequence are shown on the right.

Table 5. Expression values (number of reads) of the CNL genes located in the pericentromeric region of chromosome 18.

GENE ID	CLUSTER ID	GENE EXPRESSION VALUES (NUMBER OF READS)
Glyma18g09130	18_1	8
Glyma18g09140	18_1	30
Glyma18g09170	18_1	5
Glyma18g09180	18_1	90
Glyma18g09220	18_1	19
Glyma18g09290	18_1	105
Glyma18g09340	18_1	1
Glyma18g09410	18_1	0
Glyma18g09670	18_2	0
Glyma18g09800	18_2	4
Glyma18g09920	18_2	38
Glyma18g09980	18_2	16
Glyma18g10490	18_3	13
Glyma18g10550	18_3	3
Glyma18g10610	18_3	21
Glyma18g10670	18_3	0
Glyma18g10730	18_3	0

Phylogenetic analysis and gene clustering showed that all the members of each gene cluster were nested within the respective clade, consistent with the CNL genes in *Medicago truncatula*[23]; however, it was not the case of the 6_1 cluster, which is shared between CNL-A and CNL-C (CNL A-D; Fig. 1). The majority of crown groups in the present phylogenetic tree contained CNL members from the same chromosome, and usually from the same gene clusters, while a few crown groups contained CNL gene members from different chromosomes. These groups with CNL gene members from different chromosomes might have evolved by chromosomal rearrangement, genomic duplication, or transposition of chromosomal segments.[29] The presence of such heterogeneous subclades in *G. max* is consistent with the findings from previous studies in *Arabidopsis*[24] and *Medicago*.[23] Gene conversion could be one way in which such a clade might arise from a mismatch repair during recombination causing similarities among these homologous sequences.[9]

Physical clustering of plant R-genes have previously been reported in other plant species.[23,24] Meyers et al.[24] defined a gene cluster as two or more genes separated by fewer than nine non-CNL ORFs. In the present study, 126 genes (67.0%) adhered to this definition, forming 41 gene clusters, an average of three genes per cluster (Table 4). The largest gene cluster is located on chromosome 18 and contained 10 genes. These results on gene clustering and uneven chromosomal distribution of the CNL clusters are consistent with previous reports from other plant genomes.[23,24,30,31] Forty-one gene clusters were found to be unevenly distributed on chromosomes 1, 3, 6–9, 13–15, and 17–20. This is clearly an outcome of the tandem duplication and contraction in cluster size.[27] Such tandem duplications in R-gene clusters have been widely observed to produce small RNAs that could be used for chromatin modifications and transposable element insertion.[28] Also, the tandem duplications can be influenced by the pericentromeric regions,[11] where recombination rates are lower than in the euchromatic sites. In soybean, the CNL gene clustering seems to be independent of the TNL gene clustering, as revealed by a previous study by Kang et al,[6] who observed the largest TNL gene cluster of 34 members on chromosome 16. In soybean genome, 62 of 188 CNL genes did not form a gene cluster. Perhaps their nonclustered positioning in new regions on the chromosome plays an important role in establishing new locations for a future NBS-LRR gene clustering.[23–25]

An absolute age of the last genomic duplication can be inferred using K_s-values.[22] The K_s-value ranges in *G. max* were consistent with those predicted by Kang et al.[6] A region on the chromosome that underwent recent duplication contained Glyma01g01420 (1_1 gene cluster; Figure 1), which is nested with and syntenic to Glyma09g34360 (9_2 gene cluster), reinforcing the evidence of recent whole genome duplication as described in Schmutz et al.[22] CNL-C genes of Figure 1 also show evidence of ectopic translocation: genes from cluster 15_3 and 15_4 were nested with Glyma06g47650, located on chromosome 6, and these genes have high similarity to the region of recent duplication shared by chromosome 15 and 13 as identified as 18159398 by Kang et al.[6] Further reinforcing the suggested recent duplication event, phylogenetic analysis indicated that gene cluster 18_5 members Glyma18g51950 and Glyma18g51930 were nested with Glyma14g37860 and Glyma08g29050, which shared 96% and 77% sequence identity, respectively. This suggests that there are complex modes of duplication in which tandem duplicated genes may be moved to remote parts of the genome and maintained in the new place nearly as the original copy.[6]

Structural and functional divergence of the soybean CNL genes. Promoter elements are essential for recruiting transcriptional factors.[32] One of the promoter elements, the WBOX motif, was previously described upstream in the NPR1 gene (nonexpresser of PR genes[33]) and upstream of the majority of *Arabidopsis* R-genes.[34] Sequence variation in the WBOX regions in the 2-kb upstream region of the CNL genes suggests that different control mechanisms may be used between *Arabidopsis* and soybean genomes. Perhaps there are specific WRKY genes in *Arabidopsis* that may bind to different WBOX sequences than their counterparts in soybean. Differences in WBOX regions could also be reflected in the number of resistance genes present in these two genomes. The *Medicago* genome was first scanned in 2008 and again scanned in 2014 for resistance genes along with the analysis of WBOX promoter, resulting in 571 (NBS encoding R-genes).[35] Most of the 188 CNL genes in soybean contained this WBOX promoter regulatory element, averaging 2.77 WBOX cassettes

per gene, which is much lower than the 8.6 WBOX cassettes per gene reported in *Medicago truncatula* genome.[23] *Medicago* genome has more R-genes than *Glycine* while having a smaller genome and fewer duplication events; possibly, WBOX as a target for methylation is influenced by the punctuated evolution in resistance systems.[28] Punctuated evolution is a bout of increased rates of evolution. As described by Friedman,[28] these bouts are triggered when individual resistance genes in the plants that survived pathogen attacks had their genomic structure changed (ie, methylated), allowing increased tandem duplication and transposon activity expanding the gene family in the genome. The majority of the gene sequences (84.6%) had 1–5 WBOX motifs in contrast to that in *Medicago*, where 75% of the sequences had 6–11 WBOX motifs.[23] None of the CNL clades differed significantly in WBOX motifs ($P = 0.34$) relative to the expression values. Slight variation in the number of WBOX motifs alone does not seem to influence the expression level of the most highly expressed genes.

A scatter plot graph shown in Supplementary File 3 suggests a potential linear relationship between the number of introns and the expression values. A different analysis using one-way ANOVA suggested a significant relationship between the number of introns and the expression value ($P < 0.001$) of the CNL genes in soybeans, which is consistent with the results from a previous report.[36] Intron-mediated enhancement (IME), where an intron is located next to transcriptional start site, has been shown to increase the transcription from 2- to 10-fold (typically), with some extreme cases increasing to 100-fold.[36] These effects seem to decrease as the distance between the intron and 5' translational start site increases. And at a distance of 1 kb, the effect is abolished.[32,37] These patterns were witnessed in our exon/intron analysis of the soybean CNL genes, where the CNL-A and CNL-B gene members with more introns when located early in the sequences (Figs. 4 and 5) had high expression values, whereas the majority of the CNL-C genes with no introns had low or no expression at all (Supplementary File 2).

Conclusions

Systematic identification of resistance genes in soybeans is vital for understanding their roles in disease surveillance. In

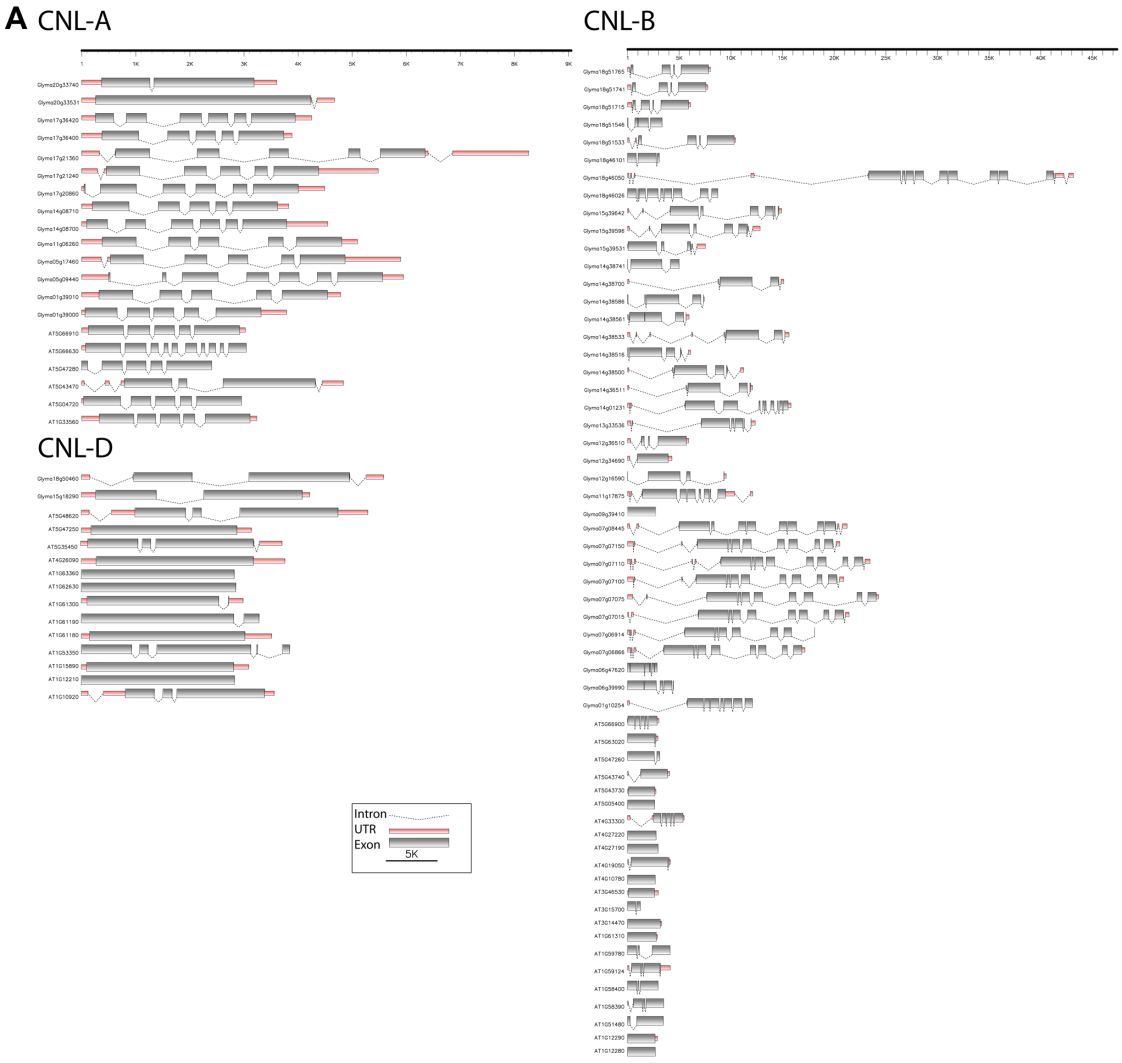

Figure 5. (*Continued*)

Figure 5. Exon-intron structures of *G. max* CNL genes and their orthologs in *A. thaliana*.
Notes: Each gene structure is shown in 5' to 3' orientation: a dashed line represents introns, a thin red line represents a UTR, and a gray box represents an exon. (**A**) CNL-A, CNL-B, and CNL-D members. (**B**) CNL-C members.

this study, altogether 188 CNL genes were identified. These genes were nested into four clades, and their evolutionary history indicated that they have evolved through tandem, segmental, or genomic duplications. These duplication events left 41 physically clustered CNL R-genes in the soybean genome. Of these 41 gene clusters, 31 clusters were located outside of the pericentromeric region. These genes outside of the pericentromeric region are allowed to freely recombine, as evidenced from the increased synonymous substitutions in the region. The presence of the majority of the CNL genes outside of the pericentromeric region would allow these genes to diversify in response to the rapidly evolving pathogens. Analysis of transcriptomic data showed differential expression patterns that ranged from nonexpression to high levels

of expression in nearly all examined tissues. These expression levels show evidence of functional divergence within the CNL R-genes. The advancement in the understanding of small RNA and methylation of the genome would, no doubt, reveal many confounding factors. Unraveling of up and downstream regulation of resistance genes and the constitutive expression of R-genes may allow for genetic modification of the plant, which is a cost-effective way to decrease yield loss by pathogen attack and control of the hypersensitive plant response.

Materials and Methods

Hidden Markov model (HMM) profiling, sequence identification, and motif analysis. Methods used in identification of CNL genes in *G. max* were similar to the methods previously used in *Arabidopsis*,[24] except for a few modifications. Fifty-three *Arabidopsis* CNL protein sequences identified by Meyers et al.[24] were obtained from the NIB-LRRS database (http://niblrrs.ucdavis.edu/, cross-linked to TAIR [The *Arabidopsis* Information Resource]),[38] and *G. max* protein sequences (version 1.0) were downloaded from Phytozome.net to create a local protein database for a HMM[39] profiling. During the HMM profiling, a model of protein domain was built from an alignment of known sequences with the domains expected of the protein. The model built from the alignment was used to scan a database of all known protein sequences of the species. For HMM[39] profiling, BLAST searched soybean protein sequences with the NBS motif along with the known *Arabidopsis* CNL protein sequences were aligned using the program ClustalW.[40] This aligned data matrix was used to create a *G. max*-specific HMM model following the method used in *Arabidopsis*.[24] This step was important to find the maximum number of candidate genes in the reiterative search process. The *G. max*-specific HMM profile was used to scan the complete set of the predicted *G. max* proteins, with a set threshold expectation value of 10^{-3}. Subsequently, the InterProScan database was searched using Geneious[41] in order to exclude the corresponding NBS proteins with the TIR motif. R-gene sequences are generally variable. Since they retain relatively higher level of conservation at the NB-ARC domain, only fully coding sequences of the functional NB-ARC were selected for evaluation.[1] MEME (multiple expectation maximization for motif elicitation) analysis[42] was used to confirm the presence of P-loop, Kinase-2, and GLPL motifs in the NBS domain of each of the selected sequences. The following criteria were used for MEME analysis: 1) the ideal motif width range was set to be between 6 and 50; 2) each search was set to identify a maximum of 20 motifs; and 3) default parameters were used for iterative cycles. The MEME visual output was sorted by CNL group, and CNL members were presented alphabetically.

Chromosomal locations of the NBS-LRR genes. Using the information on the start and end positions available at Phytozome.net, the CNL genes were located on their corresponding chromosomes. Centromere position was determined by identifying 91–92 nucleotide tandem repeats within the centromeric region, and the pericentric region was determined using recombination rate (predicted value zero or close to zero in contrast to regions with physical-to-genetic distance ratios of approximately 200 kb per 1cM, as suggested in a previous study.[11] Information about gene position, centromere, and pericentromeric region was used to annotate *G. max* CNL genes on each chromosome. The program Geneious[41] was used for graphic portrayal of *G. max* NBS-LRR gene positions, and clustering was defined using a 10 ORF window. The 10 ORF windows were used because, when larger windows of 25 ORF and 50 ORF were analyzed, the number of genes in the cluster showed little or no change similar to previously described in *Arabidopsis*.[24] A 10-ORF window is described as CNL genes, which are separated by no more than nine non-CNL ORFs. If additional genes were found within no more than eight non-CNL ORF separating these genes, then the new gene was added to the cluster. This search continued until more than nine non-CNL ORFs were found. Once the search was completed, a cluster name was assigned using the convention: chromosome number_ and number of cluster on the chromosome (eg, the first cluster on chromosome 18 was named 18_1, the second cluster was named 18_2, and so on). The 41 resulting clusters were then identified as inside or outside the pericentromeric region using this information. The X^2 -Test was used to confirm the significance of CNL cluster appearance executed using the program R (version 2.15.2, release 2012–10–26).[43]

Sequence alignment and phylogenetic analyses. Phylogenetic analyses were performed using MEGA (version 5.2.2).[44] Multiple alignments of the NBS amino acid sequences were performed using MUSCLE[45] with default settings. Maximum likelihood (ML) analysis using the best fit evolutionary model JTT+G+I (Jones–Taylor–Thornton with gamma distribution and invariant sites)[46] with the bootstrap support of 100 replicates was performed. The trees were rooted with *Streptomyces* accession (p25941) as out-group, which was also used in the analysis of CNL genes in *Arabidopsis*.[24]

Analysis of promoters, K_d/K_s values, and G + C content. The 2-kb upstream region for each predicted CNL gene was screened against the PLACE database.[47] Overrepresented regulatory elements known for their involvement in resistance responses under stress conditions were selected for further analysis. Among them, the WBOX (sequence TGAC[C/T]) associated with the WRKY transcription factor[34] was retained for further analysis. For each NBS containing protein, the amino acid motif (xxLxLxx) was searched downstream of the GLPL motif. For each CNL coding sequence, the percentage of guanine and cytosine (G + C) and K_a and K_s values were calculated using DnaSP (version 5.2 release 2012).[48] The interpretation of age of clusters based on pairwise K_s values followed Schmutz et al.[22] The approximate age of clusters based on pairwise K_s values were interpreted using a silent mutation rate of 5.17×10^{-3}, as reported previously.[22]

Structural variation and gene expression analysis. To gain insights into the expression profiles of the soybean CNL genes

in different tissues, RPKM normalized gene expression data were obtained from SoyBase.org and were $\log_2$-transformed for the MAYDAY heatmap visualization.[49] Intron/exon analysis was performed using information on genomic coordinates, orientation, and type of fragment available at Phytozome.net. The program FancyGene (http://bio.ieo.eu/fancygene) was used for visualization of the gene model. Gene mapping included positioning of the UTR (UnTranslated Region), exons, and introns. One-way ANOVA was used to compare the exon numbers and the distribution of introns using the program R.[43]

Acknowledgments

South Dakota Soybean Research and Promotion Council (SDSRPC) and South Dakota Agricultural Experiment Station (SDAES). Sarbottam Piya, Achal Neupane, Lukas Davison, Kenton MacArthur, Stacey Lindblom-Dreis, and Ethan Andersen contributed useful discussion on the manuscript. The results from this study were presented at the 2014 Annual Meeting of the Botanical Society of America, Boise, Idaho.

Author Contributions

Both authors made equal contributions. Carried out data mining, performed *in silico* and phylogenetic analyses, and drafted the manuscript: BVB. Conceived, designed and coordinated the project, supervised data analyses, and drafted the manuscript: MPN. Both authors reviewed and approved of the final manuscripts.

Supplementary Files

Supplementary File 1. K_s-values for the 41 CNL gene clusters in soybean.

The K_s-values were averaged by cluster and used in inferring the approximate age of the cluster following Schultz et al (2010). The cutoffs applied for K_s-values were from 0.04 to 0.4 and from 0.4 to 0.8 for the duplications of 13 MYA and 59 MYA, respectively.

Supplementary File 2. Expression values for the upper quartile (top 25% most expressed) soybean CNL genes preceded by the 11 genes that did not have any expression values.

Gene accessions are sorted in the order of expression level and then by the gene family.

Supplementary File 3. Scatter plot diagram comparing the number of introns with expression values in the 14 sampled tissues.

A regression line was drawn for each relationship between number of introns and expression values.

REFERENCES

1. Marone D, Russo MA, Laido G, De Leonardis AM, Mastrangelo AM. Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. *Int J Mol Sci*. 2013;14(4):7302–26.
2. Boller T, Felix G. A Renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annu Rev Plant Biol*. 2009;60:379–406.
3. Gao X, Chen X, Lin W, et al. Bifurcation of *Arabidopsis* NLR immune signaling via Ca2+-dependent protein kinases. *PLoS Pathog*. 2013;9(1):e1003127.
4. Jones JD, Dangl JL. The plant immune system. *Nature*. 2006;444(7117):323–9.
5. Li X, Cheng Y, Ma W, Zhao Y, Jiang H, Zhang M. Identification and characterization of NBS-encoding disease resistance genes in Lotus japonicus. *Plant Syst Evol*. 2010;289(1–2):101–10.
6. Kang YJ, Kim KH, Shim S, et al. Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean. *BMC Plant Biol*. 2012;12(1):139.
7. Ashfield T, Egan AN, Pfeil BE, et al. Evolution of a complex disease resistance gene cluster in diploid *Phaseolus* and tetraploid Glycine. *Plant Physiol*. 2012;159(1):336–54.
8. Gururani MA, Venkatesh J, Upadhyaya CP, Nookaraju A, Pandey SK, Park SW. Plant disease resistance genes: current status and future directions. *Physiol Mol Plant Pathol*. 2012;78:51–65.
9. Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J*. 1999;20(3):317–32.
10. Sanseverino W, Hermoso A, D'Alessandro R, et al. PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. *Nucleic Acids Res*. 2013;41(D1):D1167–71.
11. Du J, Tian Z, Sui Y, et al. Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. *Plant Cell*. 2012;24(1):21–32.
12. Kimura M. Evolutionary rate at the molecular level. *Nature*. 1968;217(5129):624–6.
13. Peart JR, Mestre P, Lu R, Malcuit I, Baulcombe DC. NRG1, a CC-NB-LRR protein, together with N, a TIR-NB-LRR protein, mediates resistance against tobacco mosaic virus. *Curr Biol*. 2005;15(10):968–73.
14. Smith CM, Clement SL. Molecular bases of plant resistance to arthropods. *Annu Rev Entomol*. 2012;57:309–28.
15. McDowell JM, Cuzick A, Can C, Beynon J, Dangl JL, Holub EB. Downy mildew (*Peronospora parasitica*) resistance genes in *Arabidopsis* vary in functional requirements for NDR1, EDS1, NPR1 and salicylic acid accumulation. *Plant J*. 2000;22(6):523–9.
16. Bittner-Eddy PD, Beynon JL. The *Arabidopsis* downy mildew resistance gene, RPP13-Nd, functions independently of NDR1 and EDS1 and does not require the accumulation of salicylic acid. *Mol Plant Microbe Interact*. 2001;14(3):416–21.
17. Mohr TJ, Mammarella ND, Hoff T, Woffenden BJ, Jelesko JG, McDowell JM. The *Arabidopsis* downy mildew resistance gene RPP8 is induced by pathogens and salicylic acid and is regulated by W box cis elements. *Mol Plant Microbe Interact*. 2010;23(10):1303–15.
18. Dorrance A, McClure S, DeSilva A. Pathogenic diversity of *Phytophthora sojae* in Ohio soybean fields. *Plant Dis*. 2003;87(2):139–46.
19. Costamilan LM, Clebsch CC, Soares RM, Seixas CDS, Godoy CV, Dorrance AE. Pathogenic diversity of *Phytophthora sojae* pathotypes from Brazil. *Eur J Plant Pathol*. 2013;135:845–53.
20. Mackey D, Belkhadir Y, Alonso JM, Ecker JR, Dangl JL. *Arabidopsis* RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance. *Cell*. 2003;112(3):379–89.
21. Liu J, Elmore JM, Lin Z-JD, Coaker G. A receptor-like cytoplasmic kinase phosphorylates the host target RIN4, leading to the activation of a plant innate immune receptor. *Cell Host Microbe*. 2011;9(2):137–46.
22. Schmutz J, Cannon SB, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. *Nature*. 2010;463(7278):178–83.
23. Ameline-Torregrosa C, Wang BB, O'Bleness MS, et al. Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant Medicago truncatula. *Plant Physiol*. 2008;146(1):5–21.
24. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell*. 2003;15(4):809–34.
25. Tan S, Wu S. Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*. *Comp Funct Genomics*. 2012;2012:1–12.
26. Cannon SB, Zhu H, Baumgarten AM, et al. Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. *J Mol Evol*. 2002;54(4):548–62.
27. Michelmore RW, Meyers BC. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res*. 1998;8(11):1113–30.
28. Friedman AR, Baker BJ. The evolution of resistance genes in multi-protein plant resistance systems. *Curr Opin Genet Dev*. 2007;17(6):493–9.
29. Young ND, Bharti AK. Genome-enabled insights into legume biology. *Annu Rev Plant Biol*. 2012;63:283–305.
30. Zhou T, Wang Y, Chen JQ, et al. Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol Genet Genomics*. 2004;271(4):402–15.

31. Kohler A, Rinaldi C, Duplessis S, et al. Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Mol Biol*. 2008;66(6):619–36.
32. Parra G, Bradnam K, Rose AB, Korf I. Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. *Nucleic Acids Res*. 2011;39(13):5328–37.
33. Yu D, Chen C, Chen Z. Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression. *Plant Cell*. 2001;13(7):1527–40.
34. Zheng Z, Mosher SL, Fan B, Klessig DF, Chen Z. Functional analysis of *Arabidopsis* WRKY25 transcription factor in plant defense against *Pseudomonas syringae*. *BMC Plant Biol*. 2007;7(1):2.
35. Shao ZQ, Zhang YM, Hang YY, et al. Long-term evolution of nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes: understandings gained from and beyond the legume family. *Plant Physiol*. 2014;166(1):217–34.
36. Rose AB. Intron-mediated regulation of gene expression. In: Reddy AN, Golovkin M, eds. *Nuclear pre-mRNA Processing in Plants*. Vol 326. Berlin, HD: Springer; 2008:277–90.
37. Rose AB. The effect of intron location on intron-mediated enhancement of gene expression in *Arabidopsis*. *Plant J*. 2004;40(5):744–51.
38. Lamesch P, Berardini TZ, Li D, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*. 2012;40(D1): D1202–10.
39. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39(suppl 2):W29–37.
40. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947–8.
41. Drummond AJ, Ashton B, Buxton S, et al. *Geneious* 5.6.5. 2011. Available at: http://www.geneious.com/.
42. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37(suppl 2):W202–8.
43. Team RC. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013. [ISBN 3–900051–07–0].
44. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28(10):2731–9.
45. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7.
46. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 1992;8(3):275–82.
47. Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res*. 1999;27(1):297–300.
48. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*. 2003;19(18):2496–7.
49. Battke F, Symons S, Nieselt K. Mayday – integrative analytics for expression data. *BMC Bioinformatics*. 2010;11:121.