Genome-Wide Analysis of NBS-LRR Genes in Sorghum Genome Revealed Several Events Contributing to NBS-LRR Gene Evolution in Grass Species

Libertas Academica
FREEDOM TO RESEARCH

Xiping Yang[1] and Jianping Wang[1–3]

[1]Agronomy Department, University of Florida, Gainesville, FL, USA. [2]Plant Molecular and Cellular Biology Program, Genetics Institute, University of Florida, Gainesville, FL, USA. [3]FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China.

ABSTRACT: The nucleotide-binding site (NBS)–leucine-rich repeat (LRR) gene family is crucially important for offering resistance to pathogens. To explore evolutionary conservation and variability of NBS-LRR genes across grass species, we identified 88, 107, 24, and 44 full-length NBS-LRR genes in sorghum, rice, maize, and *Brachypodium*, respectively. A comprehensive analysis was performed on classification, genome organization, evolution, expression, and regulation of these NBS-LRR genes using sorghum as a representative of grass species. In general, the full-length NBS-LRR genes are highly clustered and duplicated in sorghum genome mainly due to local duplications. NBS-LRR genes have basal expression levels and are highly potentially targeted by miRNA. The number of NBS-LRR genes in the four grass species is positively correlated with the gene clustering rate. The results provided a valuable genomic resource and insights for functional and evolutionary studies of NBS-LRR genes in grass species.

KEYWORDS: disease resistance gene, NBS-LRR, local duplication, evolution, grass species

CITATION: Yang and Wang. Genome-Wide Analysis of NBS-LRR Genes in Sorghum Genome Revealed Several Events Contributing to NBS-LRR Gene Evolution in Grass Species. *Evolutionary Bioinformatics* 2016:12 9–21 doi: 10.4137/EBO.S36433.

TYPE: Original Research

RECEIVED: October 12, 2015. **RESUBMITTED:** December 02, 2015. **ACCEPTED FOR PUBLICATION:** December 04, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 3792 words, excluding any confidential comments to the academic editor.

FUNDING: This study is financially supported by the Florida Sugar Cane League. 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: wangjp@ufl.edu

Introduction

Plants have developed sophisticated mechanisms to recognize and guard against pathogens, including pathogen-associated molecular pattern-triggered immunity (PTI) and effector-triggered immunity (ETI).[1] PTI is based on the actions of a basal immune system, which can be activated by generic signals of a pathogen, such as bacterial flagellins, lipopolysaccharides, and elongation factors. ETI is based on the actions of an adaptive immune system, which has a specific recognition of plant disease resistance (*R*) genes and pathogen avirulence (*Avr*) genes.[2,3] The gene-for-gene interaction of the ETI is usually accompanied by a hypersensitive response leading to the restriction of pathogen growth. However, during the never-ending battle of plants and pathogens, a host species often loses the ETI resistance due to its pathogen races evolving new virulent genes.[4] Therefore, identification of *R* genes in plants is a critical step toward developing resistant varieties and investigating corresponding resistant mechanisms.

To date, >100 *R* genes have been cloned from different plant species, and at least five different classes of *R* genes have been identified based on protein domain organization. The most dominant class of *R* gene contains a nucleotide-binding site (NBS) and a leucine-rich repeat (LRR) domain.[5–7] For example, in a survey of 22 isolated *R* genes in rice, 21 of them had NBS-LRR domain.[8] It was estimated that rice genome contains ~500 NBS-LRR genes, though a significant portion of them had partial sequences or were pseudogenes.[8] This large gene family is involved in the recognition of diverse plant pathogens through gene-for-gene interaction and plays significant roles in plant disease resistance. The NBS domains are involved in signaling and include several highly conserved and strictly ordered motifs. In contrast, the LRR domains typically evolve binding specificities and are adaptable under diversifying selection.[3]

The NBS-LRR genes are unevenly distributed in plant genomes and most of them reside in clusters. There are two types of NBS-LRR gene clusters: monophyletic cluster and mixed cluster. The monophyletic cluster contains genes with close relationship and high sequence similarity, while the mixed cluster contains clustered genes with diverged relationship and low sequence similarity. The distribution of *R* genes in a cluster may provide a reservoir of genetic variation from which new specificities can evolve. Mechanisms such as gene duplication, unequal crossing over, ectopic recombination, gene conversion, and diversifying selection have been proposed to contribute to the diverse structure of *R* genes and to the evolution of resistance specificities.[2,5,9] Presence/absence variation polymorphisms of *R* genes have been found between species.[1] Such rapid copy number evolution has been postulated to be driven by gene loss or expansion within a species through repeated cycles of duplication, divergence, and eventual loss by pseudogene formation or deletion in response to diverse pathogens.[10] Variable numbers of NBS-LRR genes, ranging from 50 to 600, have been identified mostly through sequence similarity from different plant species such as papaya, soybean, and rice.[3] However, few studies have investigated and compared NBS-LRR genes across grass species at the whole genome level, specifically with a focus on the full-length NBS-LRR genes and their evolving events within and across genomes. The full-length NBS-LRR genes can be identified with high confidence. We hypothesized that the structural and evolutionary characterization of the potentially functional full-length NBS-LRR genes can generate conserved evolutionary events within and across genomes. With the availability of fully sequenced and annotated genome sequences of cereal crops, including sorghum, rice, maize, and *Brachypodium*, we have an opportunity to explore the evolutionary conservation and variability of this major component in the plant immune system across grass species.

The objectives of this study are to (1) identify the full-length NBS-LRR genes from grass species using InterProScan, an integrated protein domain recognition method; (2) perform a comprehensive analysis on classification, genome organization, evolution, expression, and regulation of these NBS-LRR genes using sorghum as a representative of grass species; and (3) conduct a comparative analysis of NBS-LRR gene number, cluster, duplication, and targeting site of miRNA across four grass species. This comprehensive bioinformatics analysis of the full-length NBS-LRR genes will shed light on how NBS-LRR genes evolve and will provide foundation for *R* gene isolation to facilitate breeders for crop improvement.

Methods

Identification, classification, and genomic cluster of full-length NBS-LRR genes. Genome assemblies and predicted gene models for sorghum (v2.1),[11] maize (v6a),[12] rice (*Oryza sativa* subsp. *japonica*; Release 7),[13] and *Brachypodium distachyon* (*B. distachyon*) (v2.1)[14] were obtained from JGI Genome Portal (http://www.phytozome.net). To identify NBS-LRR genes in the four species, predicted protein sequences were analyzed using InterProScan 5 standalone version released in 2013, including all 13 integrated databases (CATH-Gene3D, HAMAP, PANTHER, PIRSF, PRINTS, PROSITE patterns, PROSITE profiles, Pfam, PfamB, ProDom, SMART, SUPPERFAMILY, and TIGRFAMs).[15] The output of each sequence was manually checked for conserved domains, and genes with both NBS and LRR domains in full length were characterized as NBS-LRR genes. The physical locations of sorghum NBS-LRR genes were mapped to 10 chromosomes based on the information on sorghum GFF file. The distribution map of NBS-LRR genes was generated using MapDraw.[16] NBS-LRR genes were classified based on the combination of protein domains. Complete sequences of gene model of all identified NBS-LRR genes from four grass species were used for discovering simple sequence repeats (SSRs). Both SSR detection and primer design were performed using a combined script of MISA and Primer3.[17]

Linked NBS-LRR genes were grouped into clusters when they were interrupted by less than eight open reading frames not encoding NBS-LRR proteins following the same definition of NBS-LRR gene cluster described previously.[18] Monophyletic clusters were composed solely of genes belonging to monophyletic clades based on phylogenetic analysis.[5] If the linked NBS-LRR genes belong to different clades in the phylogenetic tree, they are considered as a mixed cluster. The expected number (μ) of mixed clusters in whole sorghum genome was calculated using the following formula[18]: $\mu = (33{,}032 - n)p(1 - q^{n-1})$ with $n = 10$, $p = 88/33{,}032$, and $q = 32{,}944/33{,}032$, where 33,032 is the total number of genes in sorghum genome with 88 NBS-LRR genes.

Gene duplication and gene conversion. Duplication events of NBS-LRR genes across whole sorghum genome were identified by conducting BLASTP search of NBS-LRR gene sequences against the annotated sorghum protein sequences. The hits of NBS-LRR genes were filtered using the following threshold: expectation (E) value $\leq 10^{-40}$, cumulative identity percentage (CIP) $\geq$60%, and cumulative alignment length percentage (CALP) $\geq$70%.[1] The survived hits were considered as the duplicated events of NBS-LRR genes in sorghum genome.

Coding DNA sequences (CDSs) of each cluster and duplication family were retrieved from sorghum CDSs and aligned in each cluster and family, respectively. Alignments were analyzed using GENECONV version 1.81a[19] according to the protocol outlined by Drouin[20] to detect gene conversion. Alignments containing only two sequences were analyzed using the *include_monosites* and *g2* options, whereas alignments containing more than two sequences were analyzed using only the g2 option.

Phylogenetic analysis of NBS-LRR genes and exon–intron configuration. A phylogenetic tree was constructed using the Molecular Evolutionary Genetics Analysis software version 6.0 (MEGA 6.0)[21] following the method described by Hall.[22] NBS domains coevolved with other protein domains, including the N terminal and LRR regions, which may contain unique information on a subgroup of NBS-LRR genes. To cover all sequence information, we used the complete protein sequence of NBS-LRR genes to construct a maximum likelihood phylogenetic tree, including the chromosome of origin (by sequence name), intron–exon configurations, gene classification, and expression sequence tag (EST) representatives. In brief, protein sequences of all NBS-LRR genes in sorghum were aligned using MUSCLE with default settings.[23] After the alignment, the best substitution model was chosen using the feature *Find Best DNA/Protein Models* (*ML*) of MEGA 6.0. Then, the maximum likelihood method and the best model $L + G$ were applied to construct a phylogenetic tree with bootstrapping 1000 replicates. Maximum likelihood branch lengths were calculated for each NBS-LRR protein using MEGA 6.0.

The exon/intron positions and phases of the NBS-LRR genes in sorghum were extracted from sorghum gene GFF file. Intron phases are classified based on Sharp's description[24]: phase-0 introns lie between two codons, phase-1 introns interrupt a codon between the first and the second nucleotides in the codon, and phase-2 introns interrupt a codon between the second and the third nucleotides in the codon. The exon/intron structures were obtained using the online Gene Structure Display Server (http://gsds.cbi.pku.edu.cn).

Expression analysis of sorghum NBS-LRR genes. Representations of EST for each sorghum NBS-LRR gene were studied by searching the NCBI EST database using the predicted cDNA sequence of each NBS-LRR gene. As of January 3, 2015, this database contained a total of 199,401 sorghum EST sequences, which included all EST sequences from multiple tissues of sorghum updated in the database by different researchers. All sorghum ESTs with the best match to NBS-LRR genes and sequence identity >80% were counted as representations.[25]

About 105 M 100 bp single-end reads (FASTQ files) of sorghum and *Bipolaris sorghicola* transcriptome were obtained from NCBI from the research done by Yazawa et al.[26] In this experiment, the transcriptomes of the disease-resistant sorghum cultivar SIL-05 and *B. sorghicola* (BC-24) in infected leaves at early stages of infection (12 and 24 hours postinoculation) were mixed and sequenced by Illumina mRNA-Seq technology. The analysis of RNA-seq data followed cufflink pipeline.[27] Briefly, the reads were trimmed using Trimmomatic version 3.2[28] and aligned to the sorghum reference genome of the BTx623 (v2.1)[11] using TopHat version 2.0.9[29] with a GFF file supplied and default values for the rest of the parameters. Cufflinks version 2.2.1 was used to predict transcripts by piling up the aligned reads. The values of reads per kilobase per million mapped reads (RPKM) were calculated as: RPKM = (10^9 × number of reads mapped to a gene)/(total mapped reads in the experiment × gene length) for the annotated transcripts. Differentially expressed genes included these genes with significantly different expression levels between control and 12 hpi or 12 and 24 hpi.

miRNA databases of each species were used to predict miRNA-targeting sites of NBS-LRR genes. There were 241, 713, 321, and 464 miRNAs preloaded in psRNATarget (http://plantgrn.noble.org/psRNATarget/) for sorghum, rice, maize, and *Brachypodium*, respectively.[30] The prediction was conducted using psRNATarget,[31] a tool designed for plant miRNA analysis.

Identification of paralogs/orthologs and synteny analysis. Paralog and ortholog groups were identified by using BLASTP with the same criteria as those in the gene duplication study (E value $\leq 10^{-40}$, CIP $\geq$60%, and CALP $\geq$70%). The paralog groups were recognized as homologous NBS-LRR gene groups among all the NBS-LRR genes identified in sorghum genome, while the ortholog groups were discovered among the NBS-LRR genes from four grass species. cDNA sequences of each paralogous group and orthologous group were subjected to multiple sequence alignments, and calculation of the number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number of synonymous substitutions per synonymous site (Ks) was done in MEGA 6.0.[21]

For each gene pair of orthologs, 100 flanking genes (50 genes from each end) surrounding orthologous gene pairs were selected to study syntenic relationships. If flanking pairs with E value $\leq 10^{-10}$ were observed, then these genes were considered as syntenic genes in this region.[1]

Results

Identification, classification, and genomic cluster distribution of NBS-LRR genes. A total of 88 full-length NBS-LRR genes (Supplementary Table 1) were identified from sorghum genome by using InterProScan 5.[15] The number of these NBS-LRR genes on each chromosome varied from 1 on chromosome 4 to 21 on chromosome 5 (Fig. 1), indicating an uneven distribution of NBS-LRR genes in sorghum. Based on the combinations of different protein domains, including coiled-coil (CC), NBS, LRR, and X (other domains except CC, NBS, and LRR), the 88 NBS-LRR sorghum genes were classified into four types: NBS-LRR (NL), CC-NBS-LRR (CNL), X-NBS-LRR (XNL), and X-CC-NBS-LRR (XCNL) (Table 1). The majority of these genes were classified as genes with two domains, NBS-LRR (46.6%), and as genes with three domains, CC-NBS-LRR (37.5%). For each type, the mean and range of CDS length were calculated. On average, NBS-LRR gene CDSs contained 3475 nucleotides, which was much longer than that of the remaining genes in sorghum (1202 nucleotides).

Out of the 88 NBS-LRR genes in sorghum, a total of 43 (48.9%) were localized in 16 clusters (Table 2), each containing two to four genes. Of these 16 clusters, 11 were in monophyletic clusters containing 31 genes and 5 were in mixed clusters containing 12 genes. The cluster sizes ranged from 9.4 to 799.0 kb, with an average length of 128.4 kb. The monophyletic clusters had a smaller mean cluster size (92.9 Kb) compared to the mean cluster size of mixed clusters (206.5 kb) (Table 2). The expected number of mixed clusters in whole sorghum genome was 2.1 as calculated according to the formula in the method, which was far smaller than the five we identified. To further understand the composition of genes involved in the cluster, all genes (Supplementary Table 2A) in the 16 clusters (85 genes, including NBS-LRR genes and other genes within a cluster), were analyzed using InterProScan. The results (Supplementary Table 2B) showed a high density of *R*-like genes in clusters.

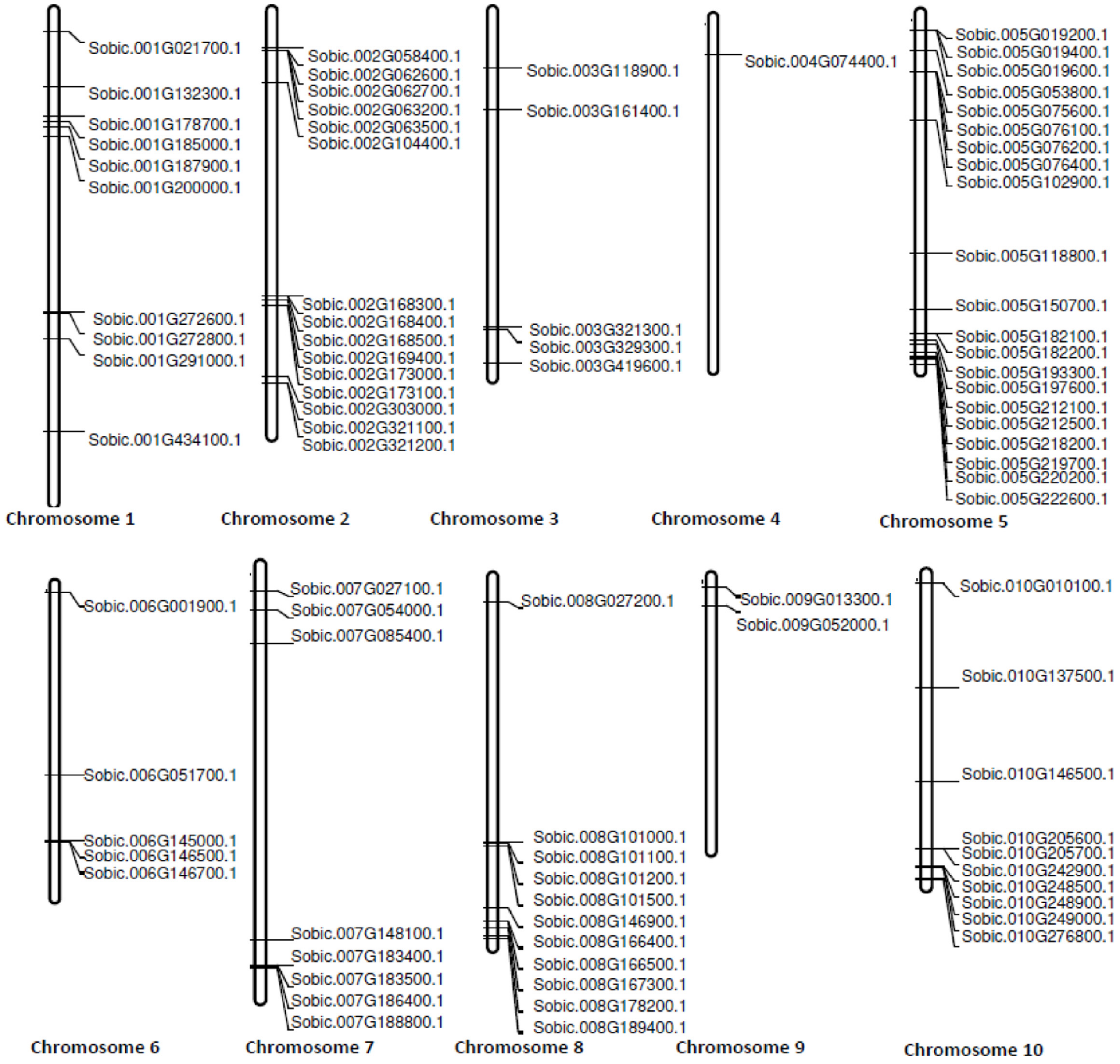

Figure 1. Physical location of NBS-LRR genes on sorghum chromosomes.

Table 1. Classification of NBS-LRR genes based on domains in sorghum and their CDS lengths.

TYPES	LETTER CODE	NUMBER OF GENES	CDS LENGTH IN RANGE (bp)	MEAN LENGTH OF CDS (bp)
NBS-LRR	NL	41	2,109–4,059	3,223
CC-NBS-LRR	CNL	33	2,673–7,182	3,700
X-NBS-LRR	XNL	9	2,829–5,019	4,280
X-CC-NBS-LRR	XCNL	5	1,536–3,312	2,606
Total		88	1,536–7,182	3,475

Duplication and gene conversion of NBS-LRR genes. In sorghum, 39 out of the 88 full-length NBS-LRR genes were duplicated in the whole genome (Table 3). The duplication rate of NBS-LRR genes (60.2%) was higher than that of the remaining genes (31.7%) in sorghum genome, indicating that NBS-LRR genes might be under evolution pressure to increase its number by duplication. There were 35 duplicated genes identified as NBS-LRR-like genes, which were not characterized

Table 2. Cluster summary of NBS-LRR genes in sorghum.

CATEGORY	NO. OF CLUSTERS	NO. OF GENES (%)	RANGE OF CLUSTER SIZE (MEAN) (Kb)	AVERAGE SEQUENCE IDENTITY (%)
All clusters	16	43 (48.9%)	9.4–799.0 (128.4)	
Mono-cluster	11	31 (35.2%)	9.9–484.5 (92.9)	65.2
Mixed-cluster	5	12 (13.6%)	9.4–799.0 (206.5)	31.2
Not clustered		45 (51.1%)		
Total genes		88		

Table 3. Duplication of NBS-LRR genes in sorghum, rice, maize, and *Brachypodium* genomes.

CATEGORY	SINGLE NBS-LRR GENES	DUPLICATED NBS-LRR GENES	NBS-LRR DERIVED GENES	NBS-LRR DUPLICATION RATE
Sorghum	49	39	35	60.2%
Rice	72	35	27	46.3%
Maize	19	5	11	45.7%
Brachypodium	28	16	39	66.3%

as NBS-LRR genes due to the lack of either NBS domain or LRR domain. These NBS-LRR-like genes shared a high amino acid sequence identity (75.9%) with NBS-LRR genes. Most of these NBS-LRR-like genes (31) had NBS domains, and only one of them had an LRR domain, suggesting that NBS domains are highly conserved after duplication.

Of the 43 NBS-LRR genes, 26 (60.5%) in clusters contained gene conversion with 64.5% in the monophyletic cluster and 50% in the mixed cluster (Table 4 and Supplementary Fig. 1). Compared to the mixed clusters, more gene conversions were detected and longer conversion tract sizes were present in monophyletic clusters. Monophyletic clusters contained NBS-LRR genes with higher sequence similarity and shorter cluster sizes than mixed clusters, which may increase the detection rate of gene conversion and the size of conversion tracts, reflecting at a rate of 64.5% of the genes (20) containing at least one gene conversion event and mean size of conversion tract as 378.0 bp (Table 4). This was consistent with the results of gene conversion study on paralogs of NBS-LRR genes, with 66.7% (16) of the genes containing at least one gene conversion event and a mean size of conversion tract of 310.1 bp.

Phylogenetic analysis of sorghum NBS-LRR genes and exon–intron configurations. A maximum likelihood phylogenetic tree of 88 NBS-LRR genes was constructed to illustrate the evolutionary relationship of these NBS-LRR genes (Fig. 2A). The tree showed a mixture of the four types of NBS-LRR genes: NL with NBS-LRR domains, CNL with CC-NBS-LRR domains, XNL with X-NBS-LRR domains, and XCNL with X-CC-NBS-LRR domains. This phylogenetic mixture suggested that coevolving or exchange of genetic information may happen among the four different types of NBS-LRR genes. Most of the NBS-LRR genes on the same chromosome were grouped in the same clades with a few exceptions (Fig. 2A), indicating a more recent local duplication than ectopic duplication. The NBS-LRR genes from the same clades tended to have similar exon–intron configuration, suggesting high conservation of exon–intron configuration during evolution. The average rate of amino acid substitution showed an ascending order with the domain number increasing in the four types of NBS-LRR genes (Fig. 2B). Since the number of mutations per amino acid in a protein increases almost linearly with evolutionary time,[32] the XCNL-type genes may be divergent from their ancestor *R* genes earlier than other NBS-LRR genes, and the NL type was the youngest type. In the four types of NBS-LRR genes, as the domain number increased, the percentage of genes with introns also increased (Fig. 2C). The NL type of NBS-LRR genes contained the lowest percentage of genes with introns and 10.3% of those were genes with phase-0 intron, while the XCNL-type genes contained the highest percentage of genes with introns and 50% of those were genes with phase-0 introns (Fig. 2C).

Table 4. Gene conversion events of NBS-LRR genes in sorghum genome.

TYPE	AFFECTED CLUSTERS (%)	AFFECTED GENES (%)	GENE CONVERSION EVENTS	MEAN (MEDIAN) SIZE OF CONVERSION TRACTS (bp)
All cluster	9 (56.2%)	26 (60.5%)	28	323.1 (54)
Mono-cluster	6 (50%)	20 (64.5%)	23	378.0 (64.5)
Mixed-cluster	3 (60%)	6 (50.0%)	5	81.6 (14)
Paralogs	NA	16 (66.7%)	18	310.1 (92.5)

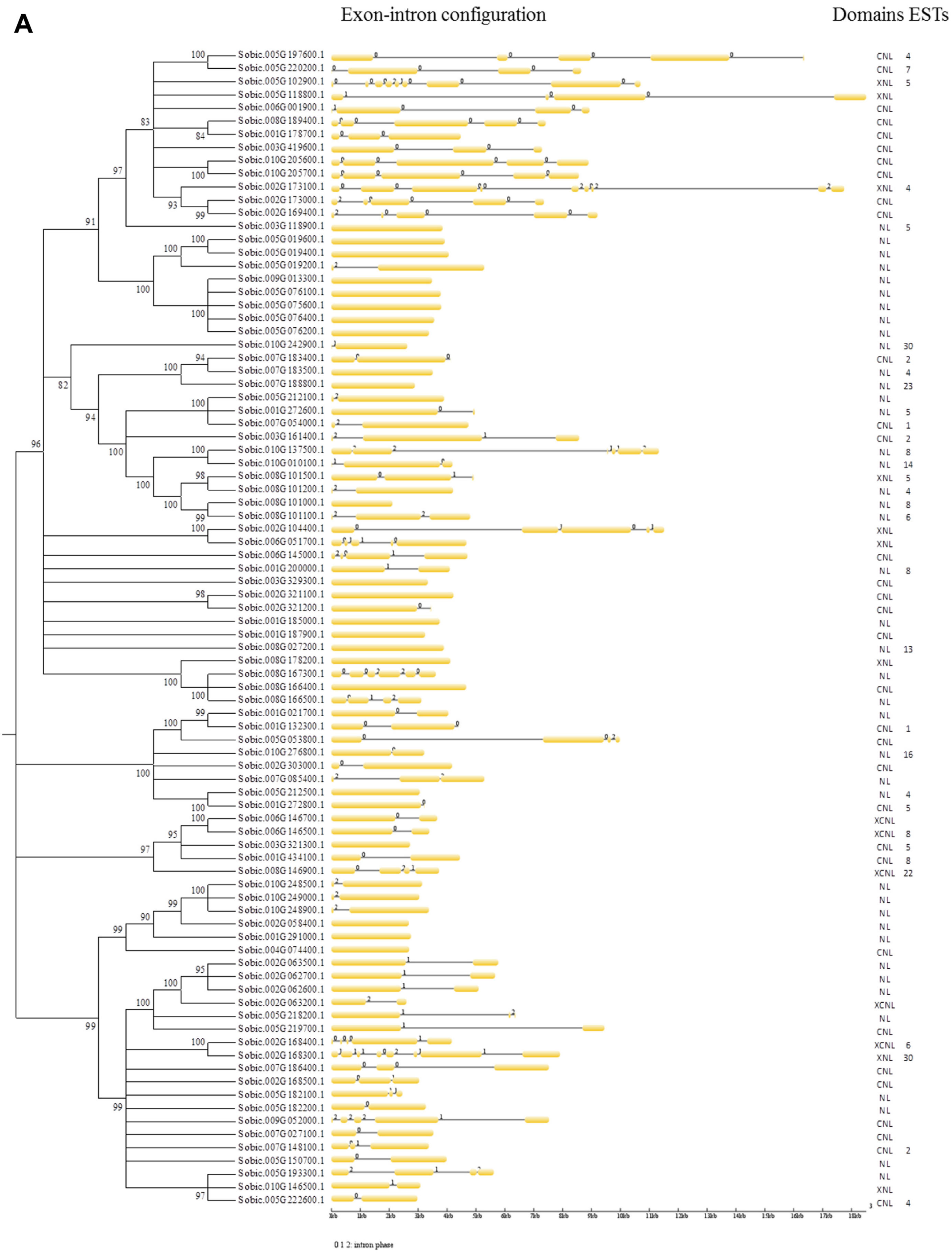

Figure 2. (*Continued*)

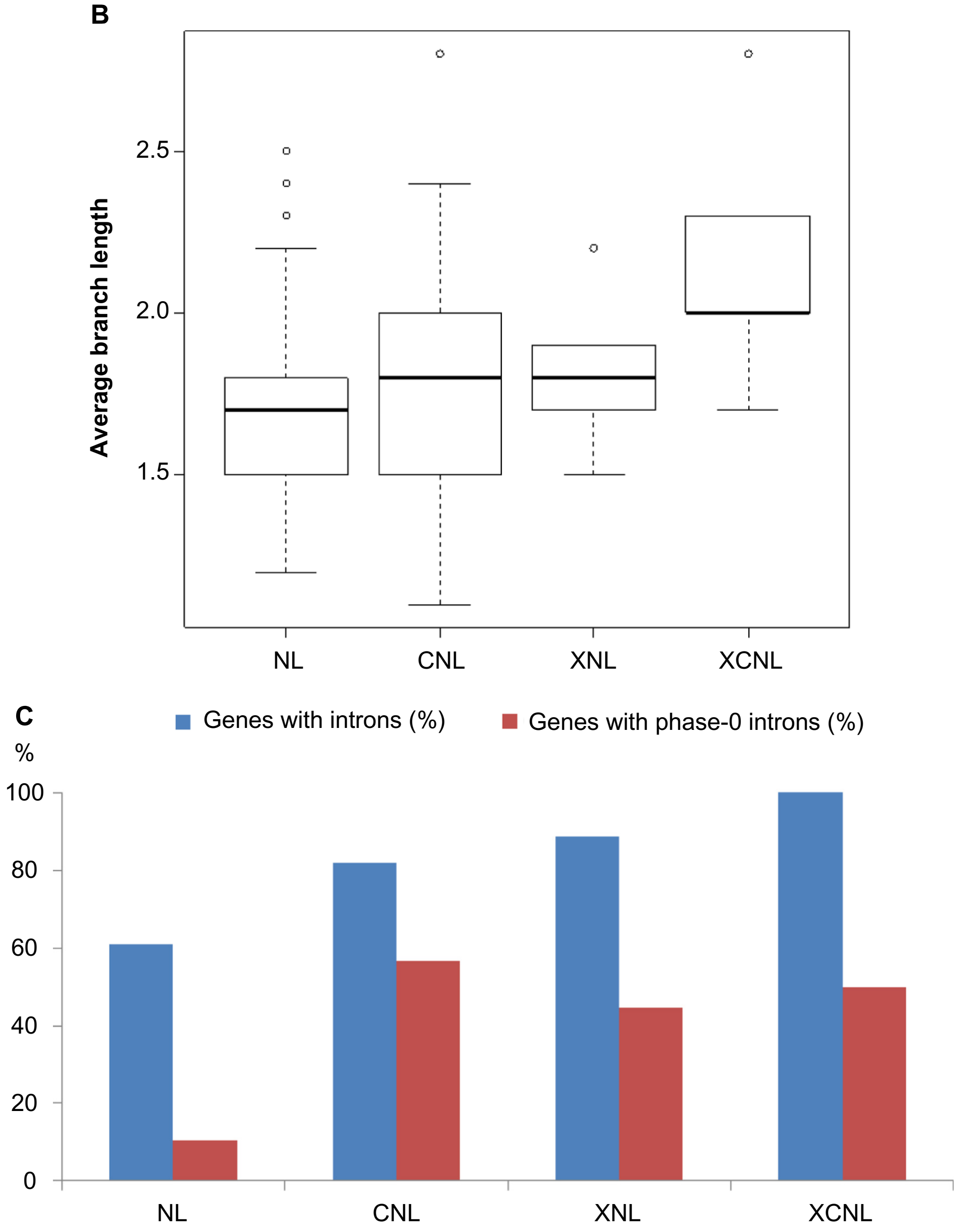

Figure 2. (**A**) A phylogenetic tree of NBS-LRR genes was constructed using MEGA 6.0. Branch numbers represent percentage of bootstrap values in 1000 bootstrapping replicates, and the scale indicates branch length. Chromosomal location of genes is included in its name (005G in Sobic.005G197600.1 showed that this gene is located on chromosome 5). Introns and exons are drawn to scale with the full encoding regions of their respective genes. Boxes indicate the exon, and lines indicate the intron. 0, phase-0 intron; 1, phase-1 intron; 2, phase-2 intron. Information on domains and number of supporting ESTs for each gene is also shown in the last two columns. (**B**) The average branch length of the four types of NBS-LRR genes. (**C**) The percentage of genes with introns and the percentage of genes with phase-0 introns in the four types of NBS-LRR genes.

Expression analysis of sorghum NBS-LRR genes. Expression of the 88 NBS-LRR genes in sorghum was evaluated by identifying the sorghum EST hits in EST GenBank with ~200,000 entries, using BLAST program. Only 32 NBS-LRR genes (36.3%) had ESTs detected with an average of 8.4 ESTs for each gene. Twenty-eight of the 33 expressed NBS-LRR genes had four or more EST representatives per gene. Fifty-five NBS-LRR genes were not detected in this depth of EST sampling.

The expression of the 88 NBS-LRR genes in sorghum was further evaluated by aligning 104 million high-quality reads of a set of RNA-seq data[26] with the genomic sequences of the 88 genes. Sixty-eight of the 88 NBS-LRR genes had an FPKM value of ≥0.05, indicating certain level of expression. The average FPKM value of the 88 NBS-LRR genes in sorghum was 5.4, which is much lower than the expression levels at the whole genome level (17.6), suggesting relatively low basal expression of NBS-LRR genes. Compared to the control, 553 sorghum genes were differentially expressed, and one of them was an NBS-LRR gene (Sobic.002G104400.1) (Fig. 3). The expression of this NBS-LRR gene was significantly reduced at 12 hpi

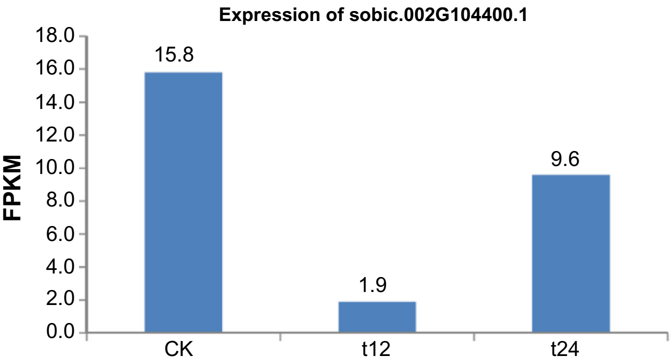

Figure 3. Gene expression (in FPKM) plot showing differences of one NBS-LRR gene across three conditions. CK, control, no inoculation; t12, 12 hours postinoculation; t24, 24 hours postinoculation.

and then it increased but with a lower level than the control at 24 hpi, suggesting the critical function of this gene for the host during the interaction with this pathogen. Further analysis indicated that this gene encoded an RPP13-like protein and its ortholog in rice contained WRKY- and DNA-binding domains. Further analysis is needed to validate the function of this NBS-LRR gene in disease resistance.

Out of the 88 NBS-LRR genes, 6.8% were identified as targets of miRNA by using the miRNA of sorghum, which were significantly higher than the remaining genes in the genome with the targeted percentage of 2.2% (Fig. 4 and Supplementary Table 3).

Identification of paralogs/orthologs and syntenic relationships. In total, 24 NBS-LRR genes belonging to eight paralogous groups were identified from the 88 NBS-LRR genes (Supplementary Table 4). These eight groups have an average of 72.7% protein sequence identity. Twenty-one NBS-LRR paralogs (87.5%) are located in clusters. In order to study the divergence of these paralogs, Ka and Ks were calculated for CDSs of NBS domain, LRR domain, and complete gene sequence (Fig. 5A). On average, the Ka values of complete gene sequence, NBS domain, and LRR domain were 0.18, 0.15, and 0.51, respectively. The Ka values were not significantly different between complete gene sequence and NBS domain, except for the LRR domain, as revealed by the *t*-test ($P < 0.05$). The Ks values from the three resources were similar (0.51, 0.52, and 0.46 respectively). The Ka and Ks values showed that NBS domains were highly conserved, while LRR domains had high diversity. The average Ka/Ks ratios of complete CDSs and NBS domain were much smaller (0.39 and 0.32, respectively) than the average Ka/Ks ratio of LRR (0.69). Considering that more than half of LRR domain sequences could not be used for Ka/Ks calculation due to remarkable mismatching, the actual ratio of Ka/Ks for LRR domain would be higher than 0.69. The Ka/Ks ratios of NBS and LRR domains indicated strong purifying selection on NBS domains, which were more conserved than LRR domains.

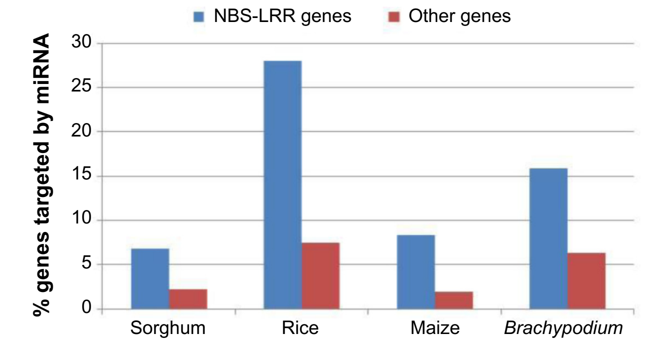

Figure 4. Percentage of genes targeted by miRNA of each species in NBS-LRR and the other genes (not NBS-LRR genes) in the genomes of the four grass species. There were 241, 713, 321, and 464 miRNAs preloaded in psRNATarget for sorghum, rice, maize, and *Brachypodium*, respectively, used for this analysis.

In total, four ortholog groups of NBS-LRR genes (with an average of 81.1% protein sequence identity) (Supplementary Table 5) were identified among sorghum, rice, maize, and *Brachypodium*. The Ka and Ks values were calculated using complete CDSs. The average Ka and Ks values between orthologs of sorghum and maize were 0.04 and 0.17, respectively, which were significantly smaller than the values of other groups, indicating the closest relationship between sorghum and maize (Fig. 5B). This was consistent with the phylogenetic analysis of the 16 orthologs (Fig. 6), in which orthologs from sorghum and maize were grouped in the same clades. The average Ka/Ks ratios of all orthologous groups was 0.24 (Fig. 5B), indicating that NBS-LRR genes may be under purifying selection after speciation in grass species.

Comparative analysis of NBS-LRR genes in grass species. In sorghum, rice, maize, and *Brachypodium*, we identified 326, 498, 163, and 354 genes with NBS domain, respectively. However, only ~20% of them were characterized as NBS-LRR genes harboring both NBS and LRR domains. The genes with full-length coding sequences of NBS-LRR, including 88, 107, 24, and 44 in sorghum, rice, maize, and *Brachypodium*, respectively (Supplementary Table 1), were used for further comparative analyses. Genome-wide comparison of NBS-LRR genes in gene number, density, duplication, targets by miRNA of NBS-LRR genes, and SSRs in NBS-LRR genes showed that sorghum and rice shared the same NBS-LRR gene density, similar duplication rate, and SSR density (Table 5). We observed a strong positive correlation between the total number of NBS-LRR genes and the percentage of NBS-LRR genes in clusters ($R^2 = 0.93$). In rice, the number of clustered genes, number of clusters, and genes targeted by miRNA were higher than those of sorghum. In *Brachypodium*, the NBS-LRR gene duplication rate, targeted rate by miRNA, and SSR density were higher than those in sorghum, while the gene density, percentage of clustered genes, and clusters were much lower than those in sorghum. Compared to other grass species, maize was unique with a low NBS-LRR gene density, no cluster, a low duplication

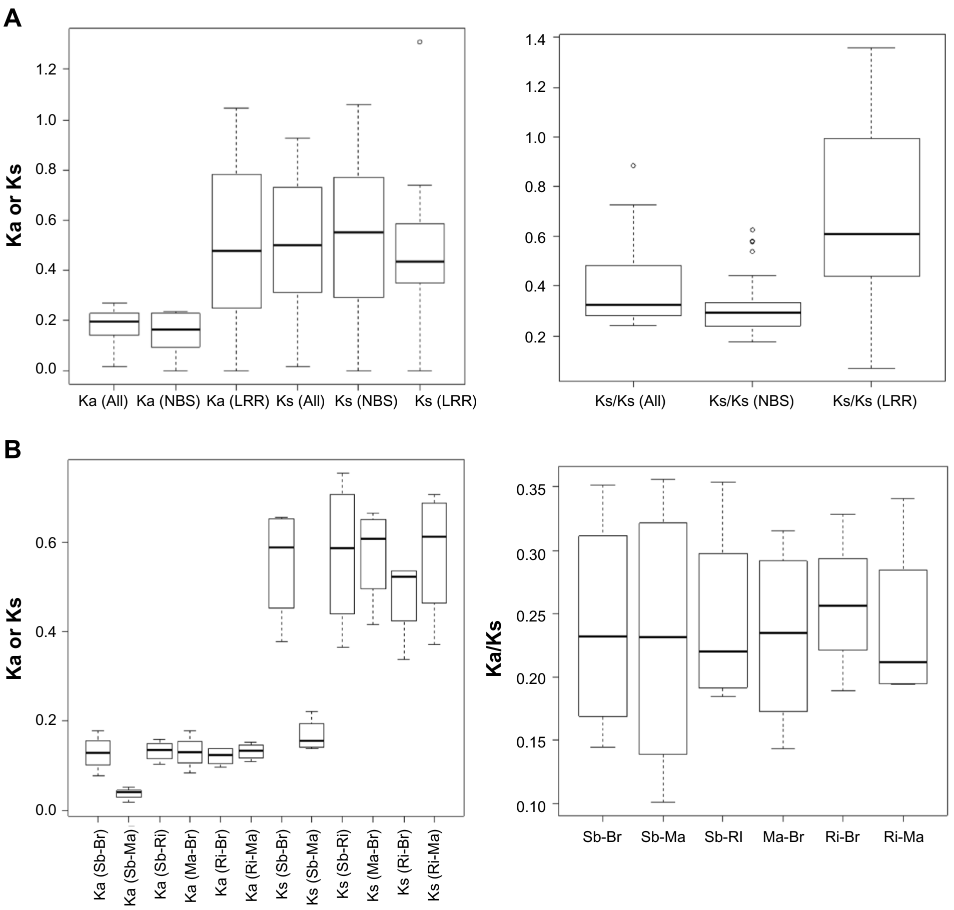

Figure 5. Divergence of NBS-LRR genes: (**A**) divergence of different classes of NBS-LRR genes in sorghum and (**B**) divergence of NBS-LRR genes in grass species.
Abbreviations: Sb, sorghum; Br, *Brachypodium*; Ma, maize; Ri, rice.

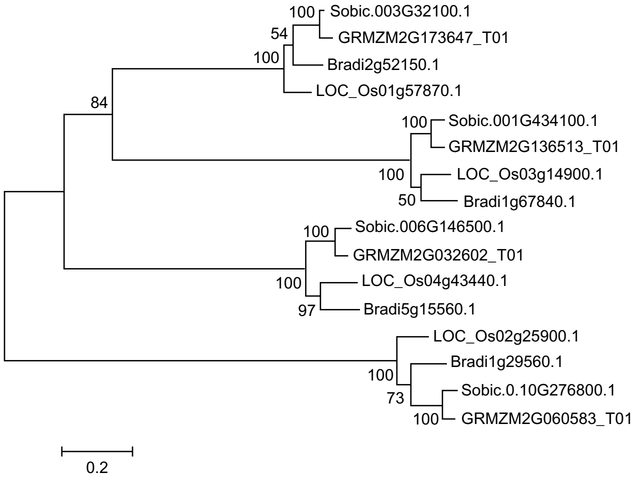

Figure 6. Phylogenetic tree of 16 orthologs in grass species.

rate, and a low SSR density. The percentage of other genes (not NBS-LRR genes) targeted by miRNA was significantly lower than that of NBS-LRR genes in grass species (Fig. 4 and Supplementary Table 3).

To understand gene conservation around regions of four orthologs, the syntenic relationships of the ortholog regions surrounding the NBS-LRR gene pairs were investigated. Among the four grass species, sorghum and *Brachypodium* shared a high level of syntenic relationship, as 63 out of 100 sorghum genes had syntenic genes in the corresponding region of *Brachypodium* genome (Table 6 and Fig. 7). *Brachypodium* shared relatively low level of syntenic relationships with maize. Out of 100 *Brachypodium* genes, 38 had orthologs in the corresponding region of maize genome. Syntenic relationships of the ortholog 4 showed that the corresponding region in rice genome was completely unaligned with the other three species, except the NBS-LRR gene. Further investigation revealed that the region in rice around ortholog 4 is on chromosome 6 instead of chromosome 4, where the rice NBS-LRR gene ortholog 4 is located. The NBS-LRR gene of ortholog 4 in rice may have been translocated to chromosome 4 from chromosome 6 during evolution.

Discussion

Identification of NBS-LRR genes based on a protein domain recognition method. In this study, we used

Table 5. NBS-LRR genes and their clustering, duplication, miRNA targeting, and SSRs in four grass species.

	SORGHUM	RICE	MAIZE	*BRACHYPODIUM*
No. of NBS-LRR genes (Density[1])	88 (2.7)	107 (2.7)	24 (0.4)	44 (1.4)
No. (%) of NBS-LRR genes in No. of clusters	43 (48.9%) (16)	49 (55.1%) (21)	0 (0) (0)	12 (27.3%) (6)
No. of NBS-LRR paralogs (%)	24 (27.3%)	25 (23.4%)	2 (8.3%)	9 (20.5%)
Targeted by miRNA (%)	6 (6.8%)	30 (28%)	2 (8.3)	7 (15.9%)
No. of SSRs (Density[2])	94 (1.91)	106 (1.96)	15 (1.67)	31 (1.82)

Notes: [1]The density of NBS-LRR genes is calculated as: 1000 × the total number of NBS-LRR genes in each species/the total number of genes in each species. [2]The density of SSRs is calculated as: the total number of SSRs in each species/the number of NBS-LRRs with SSRs in each species.

Table 6. The number of syntenic genes identified out of 100 flanking genes surrounding four NBS-LRR gene groups of orthologs.

	Sb-Br	Sb-Ri	Sb-Ma	Br-Ri	Br-Ma	Ri-Ma
Ortholog 1	69–67	66–64	49–57	63–65	42–47	43–49
Ortholog 2	74–75	49–55	50–53	51–52	40–41	38–38
Ortholog 3	59–49	69–63	49–58	50–53	36–49	36–49
Ortholog 4	49–49	1–1	40–50	1–1	33–41	1–1

Note: Sb-Br, Sb-Ri, Sb-Ma, Br-Ri, Br-Ma, and Ri-Ma represent the number of genes in the former species that had corresponding syntenic genes in the latter species.
Abbreviations: Sb, sorghum; Br, *Brachypodium*; Ma, maize; Ri, rice.

InterProScan, a software packet with an integration of different protein domain recognition methods and a comprehensive database, to identify full-length NBS-LRR genes, a large gene family in plant genomes involved in diverse pathogen recognition. The numbers of NBS-LRR genes in this study were much less than the numbers reported before, such as 160 genes in rice, 175 in *Brachypodium*, and 228 in sorghum, from studies that relied on BLAST or hidden Markov models (HMMs).[10,33] For BLAST or HMM methods, a large number of representative start sequences are required because BLAST hits are chosen based on sequence similarity with query

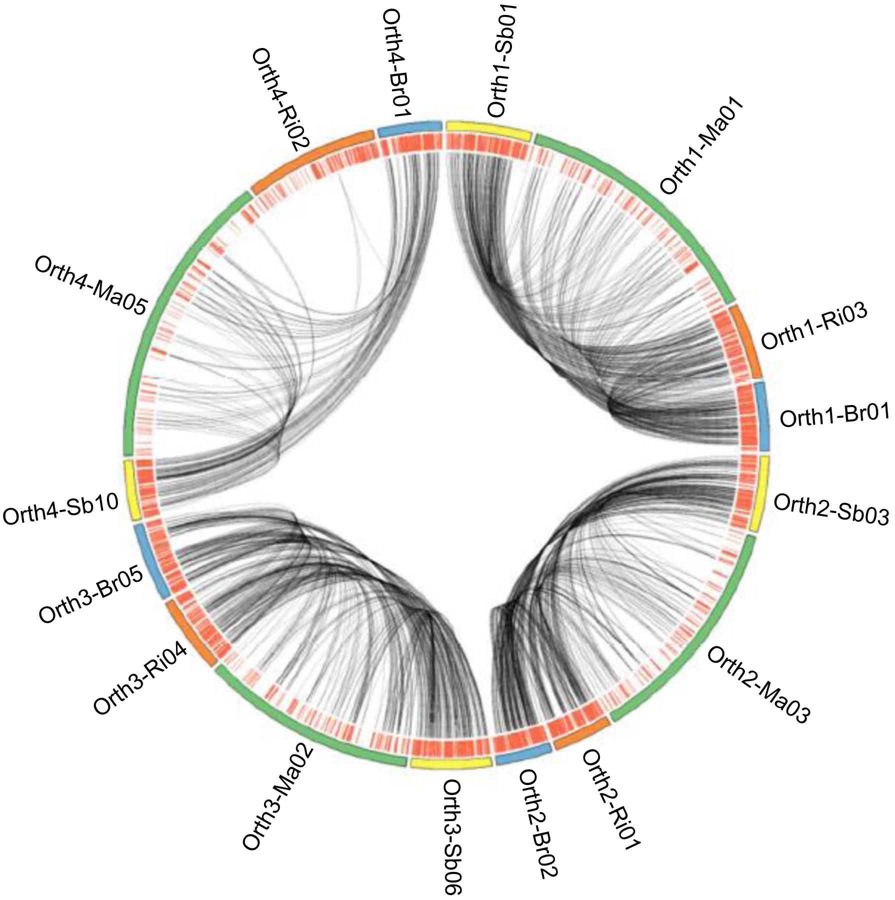

Figure 7. Syntenic relationships surrounding NBS-LRR orthologs in grass species.
Notes: Flanking genome regions and 100 flanking genes of four orthologs are represented by color bars and lines, respectively. Four colors are used: green, maize; yellow, sorghum; blue, *Brachypodium*; orange, rice. Syntenic genes in each ortholog are linked by black lines.

sequences and HMMs need to be trained on seed sequences. The parameter stringency for good hits in BLAST and selection of proper HMMs generated from the seed sequences could affect the number of identified genes. As a consequence, some NBS-LRR-like genes, genes of partial sequences, pseudogenes with nonsense point mutations, or small insertions/deletions causing frameshift mutations were identified through this approach. In contrast, InterProScan analyzes candidate protein sequences against 13 compressive databases and combines different protein signature recognition methods into one resource.[15] Domains or motifs will be checked by multiple recognition methods, such as weight matrices and fingerprinting method, and compared with 13 databases, thus increasing the chances of correct domain identification. Therefore, the stringency of InterProScan is relatively high. In addition, in this method, only genes characterized confidently with full coding sequences of both NBS and LRR domains were considered as NBS-LRR genes, instead of partial sequences containing either NBS or LRR domain. Therefore, the results from the analysis of InterProScan in this study were more conserved than the methods used in other studies.

The accurate set of NBS-LRR genes is the foundation for further structure, distribution, and evolutionary analyses. To ensure that the number of NBS-LRR genes identified in this study is not a significant underestimation due to systemic and bioinformatics errors, we compared the 88 NBS-LRR genes identified from the sorghum genome with the more than 200 NBS-LRR genes reported previously by Mace.[10] We were able to retrieve the sequences of 194 NBS-LRR genes reported in Mace's article. The InterProScan indicated that only 53 of them were NBS-LRR genes, while 140 genes contained only NBS domain or just the LRR domain and one gene contained neither of the two domains. Furthermore, 10 of the 140 NBS genes were randomly chosen and manually checked using Pfam (http://pfam.xfam.org/), which is a publicly accepted method to determine protein domains. The results from Pfam confirmed that none of those 10 genes contained both of the domains. In addition, we compared the 44 NBS-LRR genes identified from the genome of *Brachypodium* using InterProScan with a previous study,[33] in which 120 NBS-LRR genes (including 126 transcripts) were identified. Similarly, only 37 of them were NBS-LRR genes, and the remaining 83 genes contained NBS domain, only missing LRR domain. Therefore, the partial genes (missing one domain or with incomplete domain) were considered as NBS-LRR genes in the study by Tan and Wu.[33] The comparison with previous studies indicated that genes with a high tendency toward becoming partial genes were categorized as NBS-LRR genes in previous reports, while the InterProScan package in this study identified only the full-length NBS-LRR genes.

Plasticity of NBS-LRR genes in clusters. Although the number of NBS-LRR genes in clusters varies depending on species, NBS-LRR gene clustering is a common feature in plant genomes. For example, nearly 50% of NBS-LRR genes reside in clusters in *Medicago truncatula*,[2] and ~60% genes reside in clusters in *Arabidopsis lyrata* and *Arabidopsis thaliana*.[9] In our analysis, about 50% of the NBS-LRR genes of sorghum and rice resided in clusters. In *Brachypodium*, 27.3% of the NBS-LRR genes were clustered. The total number of NBS-LRR genes in grass species is positively correlated with the percentage of NBS-LRR genes in clusters and the duplication rate, indicating that the local gene duplication contributes to the increase and diversification of NBS-LRR genes. Maize contained the lowest number of NBS-LRR genes, which may be due to the fact that after one recent whole genome duplication (WGD) in maize, duplicated *R* genes were deleted to return to a singleton status, but local duplication has not established yet.[1] The different number of NBS-LRR genes in grass species may reflect the potential environmental disease pressures. Compared to maize, rice may encounter more pathogen attacks since it mainly grows in waterlogged areas with high humidity, which promotes the emergence of pathogens.

The clusters provide a reservoir of genetic variation of NBS-LRR genes through mechanisms such as duplication, gene conversion, and diversifying selection.[2] Of the 24 sorghum NBS-LRR parologs, 21 were in clusters, and 27 out of 43 genes in clusters were duplicated. Local duplication in clusters could be the major evolutionary mechanism of NBS-LRR gene expansion and could potentially determine the number of varieties of NBS-LRR genes in grass species. The majority of the genes in clusters were affected by gene conversion, especially for monophyletic clusters with 64.5% affected genes. LRR domain was involved in the recognition of pathogen ligands and is usually highly variable, while NBS domain was involved in signaling and included highly conserved and strictly ordered motifs.[10] The Ka and Ks values in this study showed that the diversity of NBS-LRR genes mainly came from LRR domains, while the NBS domain might be under purifying selection.

Compared to mixed clusters, monophyletic clusters contained NBS-LRR genes with a high sequence similarity and small cluster size, indicating that the two types of clusters may have different originating mechanisms. Monophyletic clusters might result from local duplication of NBS-LRR genes on the chromosome, while mixed clusters might result from ectopic recombination, in which heterozygous NBS-LRR genes combined with a physical cluster. In addition, the number of mixed clusters (5) is greater than expected at a genome-wide level (2.1), indicating that the genes in mixed clusters promote certain functions of NBS-LRR genes. In fact, an emerging theme was reported that two NBS-LRR genes functioned together to mediate disease resistance and many of them were in genomic clusters (eg, *RPP2A-RPP2B*, *RRS1-RPS4*, *Lr10-RGA2*, *Pikm1-TS-Pikm2-TS*, and *pi5-1-pi5-2*).[34] For those NBS-LRR genes in pairs, both NBS-LRR proteins are required for disease resistance. In sorghum, we identified five mixed clusters.

Regulation mechanisms of NBS-LRR gene expression. Most of the previously investigated species (except papaya with a relatively even distribution) showed varying number of NBS-LRR genes and an uneven distribution of genes on chromosomes.[3] Grass species, such as sorghum, rice, maize, and *Brachypodium*, underwent one shared WGD, and maize experienced a recent extra-WGD five million years ago.[1] After WGD, *R* genes are subjected to biased deletion from duplicated chromosomes and return to a singleton status. However, local duplication could compensate the uneven distribution of *R* genes during evolution.[1] This is consistent with our results. In our analysis, sorghum and rice, with a high duplication rate, contained high number of NBS-LRR genes, while maize and *Brachypodium*, with a low duplication rate, contained smaller number of NBS-LRR genes.

miRNA plays an important role in RNA silencing and posttranscriptional regulation of gene expression. For example, many transcription factors and development-related genes have been reported as targets of these regulatory small RNAs. The dosage effect of NBS-LRR genes could be balanced through miRNA regulation as more NBS-LRR genes were predicted to be targets of miRNA in these species with high NBS-LRR duplication. Our results showed that there was a higher percentage of NBS-LRR genes targeted by miRNA than the remaining genes in the genomes of the four grass species. miRNA might serve as a regulator, controlling the expression levels of NBS-LRR genes. When no pathogen attacks, the host plant would regulate the expression of NBS-LRR genes at a low level to reduce its possible detrimental impact, especially for those NBS-LRR genes with high dosage. Targeting sizes were frequently located at NBS or other conserved domains (63.2%). Expression of NBS-LRR genes needs to be regulated to limit their metabolic costs and detrimental effects.[3]

Conclusion

In summary, we have identified and characterized full-length NBS-LRR genes in sorghum, rice, maize, and *Brachypodium*. Local duplication, mainly in clusters, is the major source for the expansion of NBS-LRR genes. Clusters also provide a reservoir for NBS-LRR genes to evolve. Basal expression levels and high chances to be targeted by miRNA indicated that the expression of NBS-LRR genes was under control to reduce detrimental effects.

Acknowledgments

The authors thank Ze Peng, Dev Paudel, Johnny Molestina, and Yang Zhao at the Agronomy Department, University of Florida, for reviewing this article.

Author Contributions

Conceived the study: JW, XY. Analyzed the data: XY. Wrote the article: XY, JW. Both authors read and approved the final article.

Supplementary Materials

Supplementary Table S1. Full length NBS-LRR genes in grass species.

Supplementary Table S2A. 85 genes in 16 NBS-LRR clusters of sorghum.

Supplementary Table S2B. 58 R-like genes in 16 NBS-LRR clusters of sorghum.

Supplementary Table S3. NBS-LRR genes from grass species targeted by miRNA.

Supplementary Table S4. Twenty four NBS-LRR genes belonging to eight paralogous groups in sorghum.

Supplementary Table S5. Four ortholog groups of NBS-LRR genes in grass species.

Supplementay Figure 1. Lengths (bp) of gene conversion tracts.in mono-cluster, mixed-cluster, and paralogs.

REFERENCES

1. Zhang R, Murat F, Pont C, et al. Paleo-evolutionary plasticity of plant disease resistance genes. *BMC Genomics*. 2014;15:187.
2. 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;146:5–21.
3. Marone D, Russo MA, Laido G, et al. Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. *Int J Mol Sci*. 2013;14:7302–26.
4. Yu J, Tehrim S, Zhang F, et al. Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana*. *BMC Genomics*. 2014;15:3.
5. Meyers BC, Kozik A, Griego A, et al. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell*. 2003;15:809–34.
6. Caplan J, Padmanabhan M, Dinesh-Kumar SP. Plant NB-LRR immune receptors: from recognition to transcriptional reprogramming. *Cell Host Microbe*. 2008;3:126–35.
7. 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:D1167–71.
8. Luo S, Zhang Y, Hu Q, et al. Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. *Plant Physiol*. 2012;159:197–210.
9. Guo YL, Fitz J, Schneeberger K, et al. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*. *Plant Physiol*. 2011;157:757–69.
10. Mace E, Tai S, Innes D, et al. The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes. *BMC Plant Biol*. 2014;14:253.
11. Paterson AH, Bowers JE, Bruggmann R, et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*. 2009;457:551–6.
12. Schnable PS, Ware D, Fulton RS, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326:1112–5.
13. Ouyang S, Zhu W, Hamilton J, et al. The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res*. 2007;35:D883–7.
14. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*. 2010;463:763–8.
15. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30:1236–40.
16. Liu RH, Meng JL. MapDraw: a microsoft excel macro for drawing genetic linkage maps based on given genetic linkage data. *Yi Chuan*. 2003;25:317–21.
17. You FM, Huo N, Gu YQ, et al. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*. 2008;9:253.
18. Richly E, Kurth J, Leister D. Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. *Mol Biol Evol*. 2002;19:76–84.
19. Sawyer S. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 1989;6:526–38.
20. Drouin G. Characterization of the gene conversions between the multigene family members of the yeast genome. *J Mol Evol*. 2002;55:14–23.
21. Tamura K, Stecher G, Peterson D, et al. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 2013;30:2725–9.
22. Hall BG. Building phylogenetic trees from molecular data with MEGA. *Mol Biol Evol*. 2013;30:1229–35.
23. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004;5:113.
24. Sharp PA. Speculations on RNA splicing. *Cell*. 1981;23:643–6.
25. Tan X, Meyers BC, Kozik A, et al. Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in *Arabidopsis*. *BMC Plant Biol*. 2007;7:56.

26. Yazawa T, Kawahigashi H, Matsumoto T, et al. Simultaneous transcriptome analysis of *Sorghum* and *Bipolaris sorghicola* by using RNA-seq in combination with de novo transcriptome assembly. *PLoS One*. 2013;8:e62460.
27. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7:562–78.
28. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
29. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25:1105–11.
30. Meyers BC, Axtell MJ, Bartel B, et al. Criteria for annotation of plant MicroRNAs. *Plant Cell*. 2008;20:3186–90.
31. Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res*. 2011;39:W155–9.
32. Nei M. *Mutation-Driven Evolution*. 1st ed. Oxford: Oxford University Press; 2013.
33. Tan S, Wu S. Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*. *Comp Funct Genomics*. 2012;2012:418208.
34. Eitas TK, Dangl JL. NB-LRR proteins: pairs, pieces, perception, partners, and pathways. *Curr Opin Plant Biol*. 2010;13:472–7.