Genetics and Molecular Research

Genome-wide identification and evolutionary analysis of nucleotide-binding site-encoding resistance genes in *Lotus japonicus* (Fabaceae)

H. Song*, P.F. Wang*, T.T. Li, H. Xia, S.Z. Zhao, L. Hou and C.Z. Zhao

Bio-Tech Research Center, Shandong Academy of Agricultural Sciences, Shandong Provincial Key Laboratory of Crop Genetic Improvement, Ecology and Physiology, Jinan, China

*These authors contributed equally to this study.
Corresponding author: C.Z. Zhao
E-mail: zhaochuanzhi@gmail.com

Genet. Mol. Res. 14 (4): 16024-16040 (2015)
Received August 10, 2015
Accepted October 22, 2015
Published December 7, 2015
DOI http://dx.doi.org/10.4238/2015.December.7.16

ABSTRACT. Nucleotide-binding site (NBS) disease resistance genes play a crucial role in plant defense responses against pathogens and insect pests. Many NBS-encoding genes have been detected in *Lotus japonicus*, an important forage crop in many parts of the world. However, most NBS genes identified so far in *L. japonicus* were only partial sequences. We identified 45 full-length NBS-encoding genes in the *L. japonicus* genome, and analyzed gene duplications, motifs, and the molecular phylogeny to further understand the NBS gene family. We found that gene duplication events rarely occur in *L. japonicus* NBS-encoding (LjNBS) genes. In addition, LjNBS genes were subjected to selection pressure, and codon usage bias was evident. We tested for purifying selection (specifically in the CC-NBS-LRR and TIR-NBS-LRR groups), and found strong purifying selection in the TIR-domain-containing sequences, indicating that the CC-NBS-LRR group is more likely to undergo expansion than the TIR-NBS-

LRR group. Moreover, our results showed that both selection and mutation contributed to LjNBS codon usage bias, but mutational bias was the major influence on codon usage.

Key words: *Lotus japonicus*; Nucleotide-binding site disease resistance genes; Gene duplication; Selection pressure; Codon usage bias

INTRODUCTION

Plants have evolved various mechanisms to protect themselves from infections by diverse microorganisms such as bacteria, fungi, and viruses (Dangl and Jones, 2001). The plant immune response system is characterized by gene-for-gene interactions between a host disease resistance (*R*) gene and a pathogen avirulence (*Avr*) gene (Jones and Dangl, 2006). Plant *R* genes can be grouped into at least five categories based on the structure and function of the encoded proteins (Dangl and Jones, 2001). The largest category of *R* genes is the nucleotide-binding site (NBS)-encoding protein group (Jones and Dangl, 2006). Previous studies have shown that NBS-encoding proteins play a critical role in plant disease resistance to diverse pathogens, but the detailed mechanisms that control these interactions are not well understood.

NBS-encoding genes have been identified in numerous plants, including *Arabidopsis thaliana* (Meyers et al., 2003), *Brachypodium distachyon* (Tan and Wu, 2012), *Brassica rapa* (Mun et al., 2009), *Glycine max* (Zhang et al., 2011b), and *Medicago truncatula* (Song and Nan, 2014). The structure of NBS-encoding proteins is characterized by three domains: a coiled-coil (CC) or a Toll/mammalian interleukin-1 receptor (TIR) domain at the N-terminal, an NBS, and a leucine-rich repeat (LRR) domain at the C-terminal (Dodds and Rathjen, 2010). The TIR domain is found in many dicot species, but is rarely present in monocots. In contrast, the CC domain is found in both monocot and dicot species (Mun et al. 2009). Statistical analysis of the *M. truncatula* genome revealed that 39.0% of NBS-encoding genes are non-TIR type, 36.0% of NBS-encoding genes are TIR-type, and 25.0% are NBS/NBS-LRR-type genes (Song and Nan, 2014).

Lotus japonicus is an important forage crop, and it is widely planted in many parts of the world. It has been used extensively as a model legume due to its short life cycle, self-fertility, and relatively simple diploid genome (Sato et al., 2008). After the release of the *L. japonicus* draft genome (build 2.5) sequence in 2008 (Sato et al., 2008), Li et al. (2010) identified 158 NBS-encoding genes, including 10 full-length and 148 incomplete sequences. These sequences were used to analyze evolutionary relationships and gene family structure (Li et al., 2010). One potential problem with this analysis was that the partial sequences might have introduced error into the phylogenetic and structural analyses. In the present study, 206 NBS-encoding genes were identified from the *L. japonicus* genome (build 2.5), including 45 full-length NBS-encoding genes and 161 incomplete sequences. We analyzed phylogenetic relationships of the gene family, conserved motifs, gene structure, gene duplications, codon usage bias, and selection pressure on the full-length NBS-encoding genes. Results of the full-length gene analyses may provide more reliable information for future studies on NBS-encoding genes. Therefore, our main goals were 1) to test if the 45 full-length NBS-encoding gene sequences result in different conclusions about this gene family than the previous study that included partial sequences; and 2) to further explore the evolutionary processes affecting the *L. japonicus* NBS-encoding genes, based on codon usage bias and selection pressure.

MATERIAL AND METHODS

Database search and sequence retrieval

The *L. japonicus* genome sequence (build 2.5) was downloaded from http://www.kazusa.or.jp/lotus/. The NBS-encoding gene sequences were identified using an iterative process. First, a hidden Markov model (HMM) profile of the NBS domain (PF00931) was downloaded from the Pfam database (http://pfam.sanger.ac.uk/) (Finn et al., 2006) and used for the identification of NBS-encoding amino acids from the *L. japonicus* genome using a local BLASTp program (P value = 10^{-2}). To ensure that the maximum number of NBS-encoding sequences could be detected, another HMM profile was generated from alignments of *A. thaliana* NBS-encoding sequences downloaded from http://niblrrs.ucdavis.edu. Secondly, a manual reannotation was performed using the Pfam database to identify TIR, NBS, and LRR domains, and COILS (http://www.ch.embnet.org) was used to specifically detect CC domains.

Phylogenetic analysis and gene structure

Multiple-sequence alignments of the NBS amino acid domain sequences were performed with the Clustal X 1.83 software (Thompson et al., 1997). Unrooted neighbor-joining trees were constructed with MEGA 4.0 (Tamura et al., 2007), and were subjected to a bootstrap analysis with 1000 iterations. In addition, we constructed a phylogenetic tree using TIR-NBS-LRR (TNL) and CC-NBS-LRR (CNL) amino acid sequences from *M. truncatula* (Song and Nan, 2014) and *L. japonicus* using the same methods.

The Gene Structure Display Server (GSDS) program was used to illustrate exon-intron structure for individual NBS-encoding genes by comparing cDNA sequences with their corresponding genomic DNA sequences.

Identification of conserved motifs, gene duplication, and chromosomal location

The MEME 4.9 online program was used to elucidate motifs in *L. japonicus* NBS (LjNBS)-encoding amino acid sequences. The MEME package was run with the following parameters: any number of repetitions; an optimum motif width between 6 and 200 residues; and a maximum of 20 motifs. Structural motif annotation was performed using the Pfam and COILS databases.

It is well known that NBS-encoding resistance genes are subject to gene duplication (Zhou et al., 2004). Using the local BLASTn (P value = 10^{-10}) program, each NBS-encoding gene sequence was used as a query for comparisons against 45 full-length NBS-encoding genes in the *L. japonicus* genome. The BLAST outputs were imported into the MCScanX software (Wang et al., 2013), and NBS-encoding genes were classified into various types of duplications, including segmental, tandem, proximal, retrotransposed, DNA-based transposed, and dispersed (under a default criterion).

In order to determine the physical locations of 45 NBS-encoding genes in *L. japonicus* chromosomes, we used each NBS-encoding gene as a query on the local BLASTn (P value = 10^{-10}) program for comparison with the whole genome, thereby confirming the initiation point of each gene. The MapInspect software was used to draw the location images of the LjNBS-encoding genes (http://mapinspect.software.informer.com/).

Selection pressure in TNL and CNL gene sequences

The resulting amino acid alignments were used to guide the alignment of coding sequences using the PAL2NAL program. The Ks (synonymous) and Ka (nonsynonymous) values for CNL and TNL genes were calculated using the F3 × F4 maximum likelihood model in the CODEML module of PAML (Yang, 2007). Generally, Ka/Ks = 1, >1, and <1 indicate neutral, positive, and purifying selection, respectively.

In order to detect whether amino acids underwent selective pressure at the protein level, we examined selected nodes from the CNL and TNL phylogenetic trees. We calculated variation in Ka/Ks among sites by employing a likelihood ratio test (LRT) between M0 *vs* M3 and M7 *vs* M8 models. Nodes were considered to have undergone positive selection if they satisfied the following criteria: 1) Ka/Ks >1 under M8; 2) sites identified to be under positive selection by Bayes Empirical Bayes analysis; and 3) a statistically significant LRT (Song et al., 2014).

Synonymous codon usage bias in *L. japonicus* NBS-encoding genes

To avoid sampling bias, the CDS were filtered based on the following criteria: 1) full-length CDS shorter than 300 bp were excluded from this analysis; 2) the presence of a start codon beginning and a stop codon ending in each CDS was required.

The G + C content was determined for all codons in the entire gene sequence and for the first and second codon positions (GC1 and GC2) using EMBOSS (http://emboss.bioinformatics.nl/). The GC12 value was calculated as the mean of GC1 and GC2, and it was used for neutrality analyses (Kawabe and Miyashita, 2003).

The Codon W 1.4 program (http://codonw.sourceforge.net) was used to assess a range of statistics related to codon usage bias: the ENC (effective number of codons), RSCU (relative synonymous codon usage), A3 (frequency of adenine at the third positions of codons), T3 (frequency of thymine at the third positions of codons), G3 (frequency of guanine at the third positions of codons), C3 (frequency of cytosine at the third positions of codons), and GC3s (G+C frequency at the third positions of codons). Among these statistics, A3, T3, G3, and C3 were used in a Parity Rule 2 (PR2) bias plot analysis (Sueoka, 1999a). PR2 is an intrastrand rule where A = T and G = C are expected if there is no bias in mutation and selection between the two complementary DNA strands (Sueoka, 1999a). To examine the influence of GC content on codon usage, an ENC-plot was calculated according to the equation described by Wright (1990). The (ENC expected - ENC observed) / ENC expected plot allows us to exclude the effect of differences in GC content caused by neutral mutations (Kawabe and Miyashita, 2003).

To obtain the estimated optimal codon usage for LjNBS, the codon usage of the highest 5.0% and lowest 5.0% of the ENC values was used as estimates of high and low expression data, respectively. To calculate an optimal codon usage value, the RSCU values must satisfy the following criteria: 1) the RSCU value is greater than 1.0 in the highly expressed gene data; 2) the RSCU value is less than 1.0 in gene data with low expression rates; and 3) ΔRSCU (comparison of RSCU values between two expression values) is larger than 0.3 (Zhang et al., 2011a).

RESULTS

Database search and sequence retrieval

We identified 206 NBS-encoding genes in the *L. japonicus* genome. Among these, 45

full-length NBS-encoding gene sequences were used for further analyses. These full-length NBS-encoding genes were divided into three groups, non-TIR (18 sequences), TIR (20), and NBS/NBS-LRR type (7), according to the presence or absence of specific domains (Table 1). Non-TIR-type genes could be placed into two different subgroups: CC-NBS (CN) and CNL. The TIR-type genes could be further divided into three subgroups: NBS-TIR (NT), TIR-NBS (TN), and TNL. Finally, the NBS/NBS-LRR-type genes could be divided into two types: NBS (N) and NBS-LRR (NL). We found eight potential pseudogenes among the 45 full-length NBS-encoding genes that contained either a premature stop codon or a frame shift mutation (**Table S1**). These eight potential pseudogenes were excluded from further analysis because of the high possibility that they were non-functional.

Table 1. Number of predicted NBS-encoding genes in *Lotus japonicus*.

Predicted	Letter code	*Lotus japonicus*
Non-TIR-NBS-encoding genes		
CC-NBS	CN	3
CC-NBS-LRR	CNL	15
TIR-NBS-encoding genes		
NBS-TIR	NT	3
TIR-NBS	TN	4
TIR-NBS-LRR	TNL	13
NBS-encoding/ NBS-encoding-LRR genes		
NBS	N	3
NBS-LRR	NL	4
Total regular NBS genes		45

Phylogenetic analysis of the NBS-encoding gene family

Phylogenetic analyses of NBS domains have been used to distinguish different groups of *R* genes (Pan et al., 2000). We constructed a composite phylogenetic tree for all candidate genes using the NBS domain sequence (Figure 1A). We detected distinct clades of TIR-type and non-TIR-type sequences that received high support from bootstrap values. These clades reflect ancient differentiation of NBS-encoding genes into two major groups. Within the TIR-type genes, sequences were divided into two major clades, TIR-1 and TIR-2 (Figure 1A). TIR-1 contained 15 sequences, including 14 TIR-type sequences and one NBS/NBS-LRR-type sequence. TIR-2 contained three TIR-type sequences and one NBS/NBS-LRR-type sequence. Interestingly, all of the TNL or TN genes were found in clade TIR-1, and the NT genes were restricted to clade TIR-2, suggesting a different origin between TNL or TN and NT genes. The non-TIR clade contained 13 non-TIR-type sequences and three NBS/NBS-LRR-type sequences. In addition, two NBS/NBS-LRR-type sequences were not closely related to the other sequences in the tree.

We constructed a phylogenetic tree of *M. truncatula* (MtNBS) and LjNBS-encoding genes to identify orthologs and paralogs with strong bootstrap values (100) (Figure 2, **Figure S1** and **S2**). We detected four paralogs (two CNL and two TNL sequences) and three orthologs (two CNL and one TNL) based on phylogenetic relationships (Li et al., 2003). Orthologous genes tend to have similar structures and functions, and these three orthologous gene pairs may be useful for future transgenic research in leguminous plants.

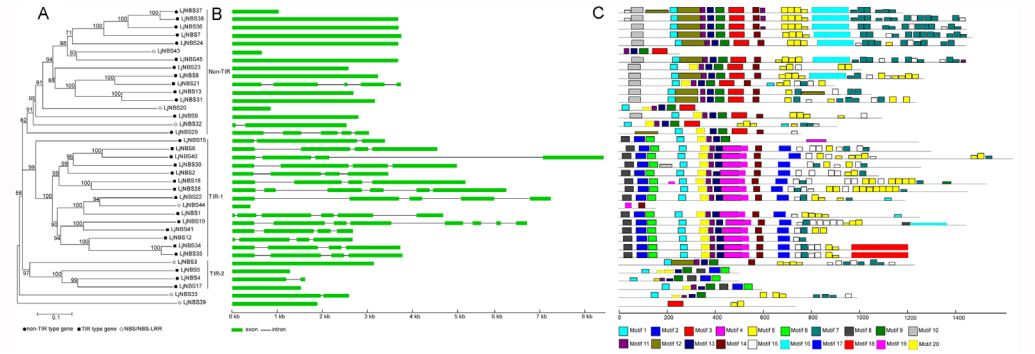

Figure 1. Phylogenetic relationships, gene structure, and motifs of NBS-encoding genes in the *Lotus japonicus* genome. **A.** The phylogenetic tree was generated with MEGA 4.0 using the neighbor-joining method with 1000 bootstrap replicates. **B.** Exon/intron structure of NBS-encoding genes from *L. japonicus*. **C.** Distribution of 20 putative conserved motifs.

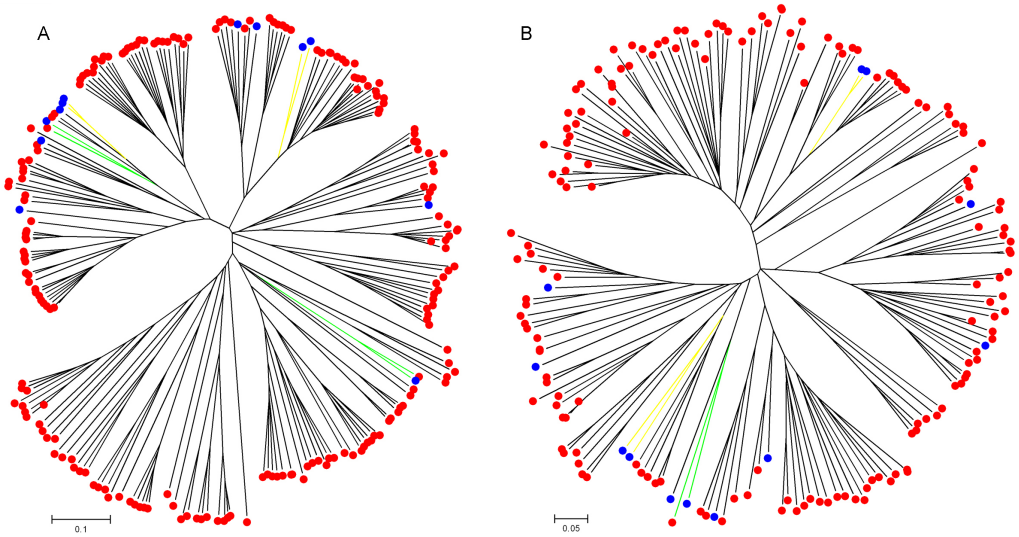

Figure 2. Phylogenetic comparison of *Lotus japonicus* (blue circle) and *Medicago truncatula* (red circle) NBS-encoding genes. The phylogenetic tree was generated with MEGA 4.0 using the neighbor-joining method with 1000 bootstrap replicates. **A.** Phylogenetic tree based on the whole NBS-encoding sequence from *L. japonicus* (blue circle) and *M. truncatula* (red circle) CC-NBS-LRR genes. **B.** Phylogenetic tree based on the whole NBS-encoding sequence from *L. japonicus* (blue circle) and *M. truncatula* (red circle) TIR-NBS-LRR genes. Yellow and cyan lines indicate paralogs and orthologs, respectively.

Analysis of gene structure, conserved motifs, and gene duplication

In order to understand the structural diversity of NBS-encoding genes, we analyzed the exon-intron structure of LjNBS genes. Sequences in the same clade exhibit a similar number of exons and introns, but in some subclades, the lengths of exons and introns varied greatly (e.g., LjNBS 6 and LjNBS 40; Figure 1B). Gene structure appeared to be more variable in some

subclades than in others. For example, LjNBS 22 contained five exons and four introns, while LjNBS 44 had only one exon (Figure 1B). In addition, most non-TIR-type genes (11 of 13) lacked an intron, except LjNBS 21 and LjNBS 29, whereas most TIR-type NBS-encoding genes (15 of 17) contained both introns and exons. Similarly, CN/CNLs contained fewer introns (0-4) than TN/TNLs (3-7), as previously found in *Arabidopsis* (Ronquist and Huelsenbeck, 2003) and *Populus* (Kohler et al., 2008) for CNLs and TNLs.

NBS-encoding proteins were characterized by identifying the TIR/CC, NBS, and LRR domains (Meyers et al., 2002). The MEME analysis found that TIR- and non-TIR-type genes had different conservative motifs, especially in the N-terminal regions. The CC motifs of CNL proteins were classified into four types in a previous study (Tan and Wu, 2012). However, we identified only one type of CC motif in non-TIR-type protein sequences (Motif 10; Figure 1C and **Figure S3**). Four TIR motifs were found in NBS-encoding protein sequences (Meyers et al., 2003), whereas we found only three types of TIR motifs in the TIR-type sequences. We found that TIR motifs could be more variable than CC motifs. However, this result is not consistent with previous studies in *A. thaliana* (Meyers et al., 2003) and *M. truncatula* (Song and Nan, 2014) where CC motifs were more variable than TIR motifs.

Eight major NBS motifs were identified in *A. thaliana* (Meyers et al., 2003), including P-loop, RNBS-A, Kinase 2, RNBS-B, RNBS-C, GLPL, RNBS-D, and MHD. The sequences of these motifs differ in CNL and TNL proteins, and we found all of the consensus sequence motifs in the CNL and TNL sequences (Figure 1C, **Figure S3**). The RNBS-A and Kinase 2 motifs were found in close proximity, and together they constituted a single motif in the present study (CNL: motif 12; TNL: motif 20). Notably, there was a clear difference between the MHDV motif in *A. thaliana* (Meyers et al., 2003) and the MHDL motif in the NBS domains of *L. japonicus* (**Figure S3**). Furthermore, the P-loop, RNBS-B, RNBS-C, and MHDL motifs shared similar sequences in CNL and TNL proteins in the present study, suggesting that RNBS-A, Kinase, GLPL, and RNBS-D are conserved.

The LRR motif is considered an important component of disease resistance (Belkhadir et al., 2004). The precise pattern of LRR varies, but the basic pattern is conserved as LxxxLxxLxxLxxLxLxxC (or T, S) xx (Bella et al., 2008). The conserved LxL and/or LxxL elements are the core of an LRR motif (Zhou et al., 2004). We identified three typical LRR motifs in the CNL and TNL sequences in *L. japonicus* NBS genes (motif 5, 7, and 15; Figure 1C and **Figure S3**).

Gene duplication events may originate as tandem or segmental duplications (Cannon et al., 2004). We detected four genes involved in tandem duplication events in *L. japonicus* (Figure 3). Tandem duplication events of NBS-encoding genes occurred on chromosome 6 of *L. japonicus*. The percentage of duplicated NBS-encoding genes was significantly lower (10.8%) than those found in *A. thaliana* (46.6%), *O. sativa* (53.7%), or *M. truncatula* (51.0%; Table 2), suggesting that the expansion of NBS-encoding genes in *L. japonicus* occurred in ways other than gene duplication.

Chromosomal location

The NBS-encoding genes sampled in this study were unevenly distributed throughout the six *L. japonicus* chromosomes (Figure 3). In the currently released sequences, some NBS-encoding genes have no precise location information, so unmapped NBS-encoding genes are temporarily mapped onto fictional chromosome 0. In addition, chromosome 2 contains the most NBS-encoding genes (13), while chromosome 4 contains three NBS-encoding genes (excluding chromosome 0).

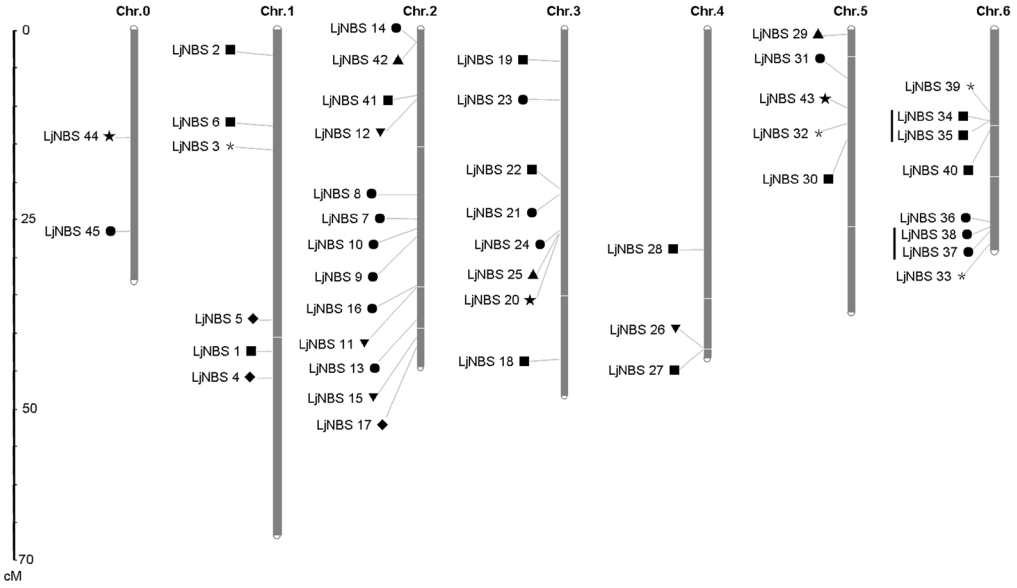

Figure 3. Chromosomal locations of *Lotus japonicus* NBS-encoding genes. The chromosome numbers are shown at the top of each chromosome (chromosome; gray bars). The names on the left side of each chromosome correspond to the approximate location of each LjNBS gene. The markers next to the gene names represent the groups to which each LjNBS gene belongs (circles: CNL; triangles: CN; stars: N; lazenges: NT; squares: TNL; inverted triangles: TN; asterisks: NL). The black lines to the left of the LjNBS gene names indicate the duplicated genes. Unmapped LjNBS genes are shown on chromosome 0.

Table 2. Organization of NBS-encoding genes in four plant genomes.

Organization	*Lotus japonicus*	*Arabidopsis*[a]	*Oryza sativa*[b]	*Medicago truncatula*[c]
Single-genes	40	93	240	240
Multi-genes	4	81	279	250
Percentage of multi-genes	10.8	46.6	53.7	51.0

[a]Data from Meyers et al. (2003); [b]data from Zhou et al. (2004); [c]data from Song and Nan (2014).

There was no obvious difference between the distributions of the TIR-type and non-TIR-type genes on the chromosomes (Figure 3). However, chromosome 2 contributes roughly eight (44.4%) of all non-TIR-type genes, while chromosomes 1 and 4 only contain TIR-type genes. The result is incongruent with the NBS-encoding genes in *M. truncatula* where chromosome 6 encodes approximately 44.5% of all TIR-type genes, while chromosome 3 encodes approximately 27.5% of all non-TIR-type genes (Song and Nan, 2014).

Selection pressure in TNL and CNL gene sequences

To avoid unreliable Ka/Ks values, we removed values (Ks < 0.005) when calculating the CNL and TNL sequences for selection pressure. We found that the Ks values of CNL and TNL sequences were within a range of 1.0-3.0 (Figure 4). However, the number of Ka/Ks values greater than 0.4 were greater in CNL gene pairs than in TNL gene pairs. Meanwhile, Ks values ranged from

1.5-2.5 and 1.0-1.5 in CNL and TNL genes, respectively. The average value of Ka/Ks (0.35) in CNL genes was nearly equality that in TNL genes (0.32).

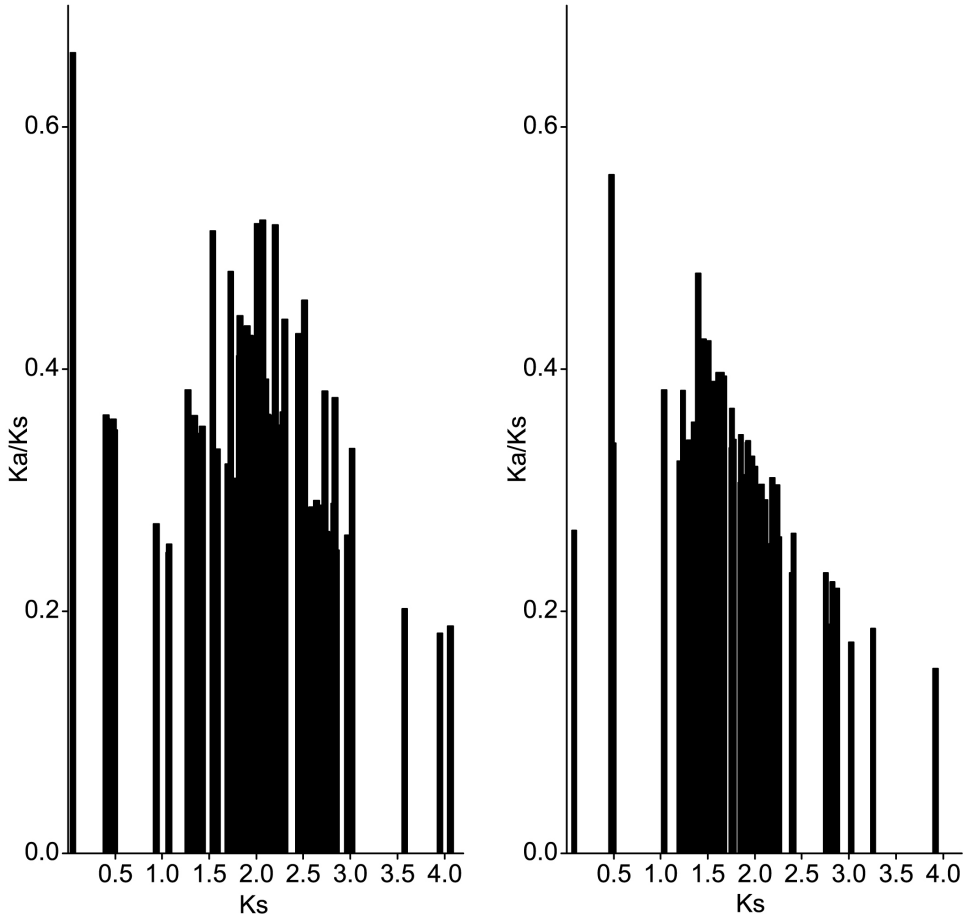

Figure 4. Relationship between Ka/Ks and Ks of CC-NBS-LRR (left) and TIR-NBS-LRR (right) genes.

Table 3 illustrates the selection pressure acting on CNL and TNL genes. Five nodes of the phylogenetic tree (**Figure S4**, nodes 1-4, 7) were under positive selection, including 18 sites under positive selection in CNL genes. Two nodes of the phylogenetic tree (**Figure S4** node 3, 4) and 10 positively selected sites were composed of TNL genes, based on the positive selection criteria. Therefore, TNL genes might experience more purifying selection pressure than CNL genes during evolutionary or selective constraints.

Synonymous codon usage bias in NBS-encoding genes

The GC content in three codon positions was investigated using the EMBOSS online program. The GC1 value (0.48) was higher than that of GC3 (0.38) and GC2 (0.35), and these

results showed that the GC content of these positions differed. The average GC content of all codons was 0.40, indicating that AT content in LjNBS genes was higher than GC content. Neutrality plots (GC12 *vs* GC3s) were used to analyze the relationships among the three codon positions. The results showed that GC content in the LjNBS genes had a wide range of GC3s values (0.29-0.54). We detected a significant positive correlation (r = 0.5884, P < 0.001) between GC12 and GC3s (Figure 5), indicating that GC mutational bias leads to similar GC content at all codon positions.

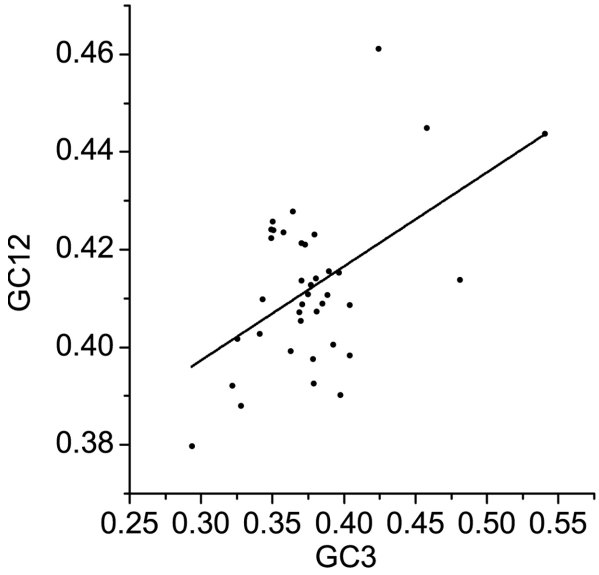

Figure 5. Neutrality plots (GC12 *vs* GC3s). Regression line: y = 0.19253X^2 + 0.33958.

An ENC value smaller than 35 was regarded as strong codon usage bias (RoyChoudhury and Mukherjee, 2010). The ENC value was positively correlated with the GC3s value of each gene (r = 0.5884, P < 0.001; data not shown), suggesting that LjNBS genes with lower GC3s and ENC values had a strong bias. The ENC-plot (GC3s *vs* ENC) indicated that the difference in ENC is related to the difference in GC content (Wright, 1990). The continuous curve indicated no difference in selection pressure between ENC and GC3s based on codon bias. In the present study, all LjNBS genes fell below the high ENC standard curve (>35) (Figure 6), suggesting that other factors independent of nucleotide composition have affected codon usage bias (Richly et al., 2002). A frequency distribution of the ENC ratio (Figure 7) showed that most LjNBS genes have a 0.05-0.2 ENC ratio, which suggests that their ENC values were smaller than expected. These results also indicated that LjNBS codon use could be predicated from GC3s, with the exception of other factors affecting codon usage bias. These findings show that mutational bias played a role in shaping codon usage in these *L. japonicus* genes.

We detected optimal codon usage based on RSCU values in high and low expression data (**Table S2**). Nine codons were considered the optimal codons to encode isoleucine (Ile), valine (Val), histidine (His), asparagine (Asn), lysine (Lys), threonine (Thr), arginine (Arg), and glycine (Gly). The codons ended with G or C (with the exception of Thr, ending with A), suggesting that codon usage bias in LjNBS genes favored G or C at the third position.

Natural selection can be detected if GC or AT are disproportionately found at the third codon; however, GC or AT are proportionally used in mutational bias at the third codon position (Wright, 1990). In order to detect whether these biased codon choices were widespread in all LjNBS genes, we constructed a PR2 bias plot between G and C content and between A and T content (Figure 8). The results showed that GC or AT were used disproportionately, with G and T used more frequently than C and A at the third position of all LjNBS genes.

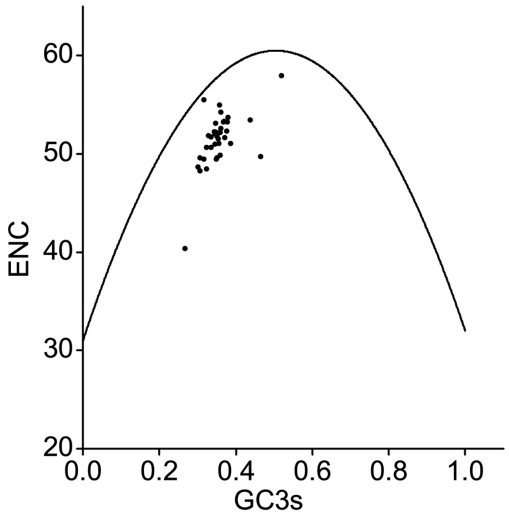

Figure 6. Relationship between the effective number of codons (ENC) and GC3s. The continuous curve between ENC and GC3 suggests that no selection pressure acts on codon bias.

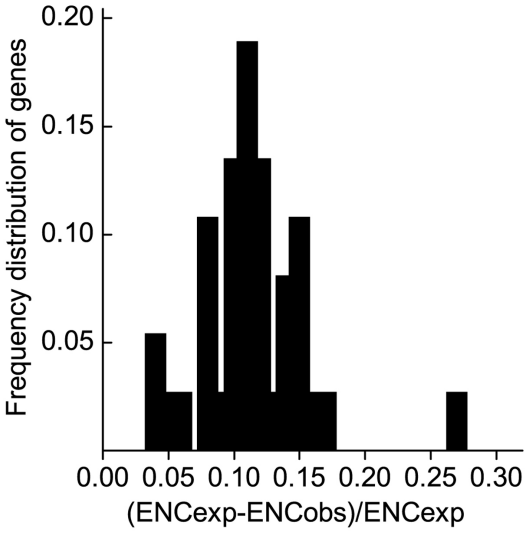

Figure 7. Frequency distribution of the effective number of codons (ENC) ratio.

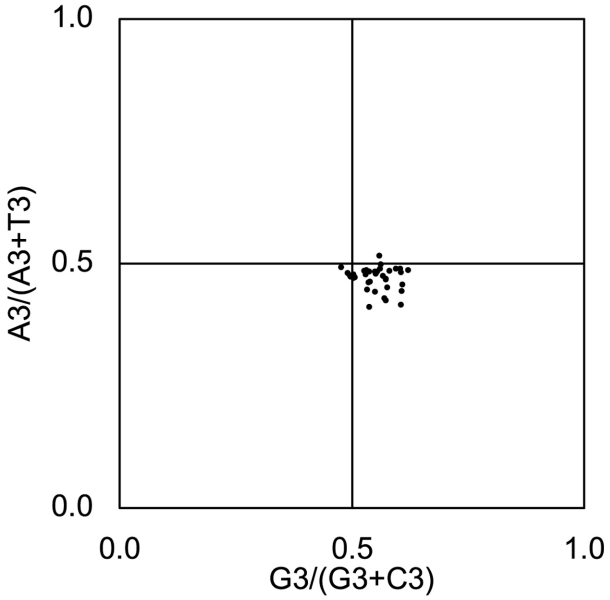

Figure 8. PR2-bias plot of the third codon position.

DISCUSSION

NBS-encoding genes have been detected in both monocotyledonous and dicotyledonous model plants, and previous studies have shown that NBS-encoding genes play a crucial role in plant defense against a variety of pathogens and pests (Dangl and Jones, 2001). Although characterization of LjNBS genes has been reported, our results provide novel information on the structure of LjNBS genes and associated selective pressure. We identified 206 NBS-encoding gene sequences in the *L. japonicus* draft genome using an HMM profile that was generated from alignments of *A. thaliana* NBS-encoding sequences and the Pfam database. However, Li et al. (2010) detected 158 NBS-encoding gene sequences using an HMM profile of the NBS domain from the Pfam database. Our results identified eight motifs in NBS domain regions according to MEME 4.9. However, the previous study (Li et al., 2010) only identified four motifs in NBS domain regions using several incomplete amino acid sequences, which may have led to the loss of some motifs. We found a reasonable pattern of evolutionary relationships between the LjNBS genes. Most researchers have identified two clades of NBS genes in dicots, TIR-type and non-TIR-type genes (Liu et al., 2012). However, research in monocots has been unable to resolve relationships among NBS domain sequences, and has yielded a star-like topology (Cheng et al., 2012). Our results indicated a typical dicot pattern, whereas the tree in the previous study (Li et al., 2010) was unresolved. We suspect that the incomplete amino acid sequences used in the previous study caused statistical problems during the phylogenetic analysis. According to our gene duplication criteria (see Material and Methods), if short sequences were included in the analysis, more gene duplication events could be detected. Therefore, it is reasonable that the previous study (Li et al., 2010) detected more gene duplication events (45) than our study (4).

Gene duplication events in *L. japonicus* NBS-encoding genes

Previous research has shown that a large number of NBS-encoding genes are produced by gene duplication, including tandem and segmental duplications (Cannon et al., 2004). However, the full-sequence LjNBS genes contain only four multi-genes derived from tandem duplication events on chromosome 6. Compared with other plants, the number of duplication events found in LjNBS genes was significantly lower than that in *A. thaliana* (81), *O. sativa* (279), and *M. truncatula* (250) (Table 2). Was this lower proportion caused by the use of fewer NBS-encoding genes in the present study? We compared the ratio of NBS-encoding genes in different genomes, and NBS-encoding genes accounted for 0.54% (206 of 30,799) of the *L. japonicus* genome sequences. This ratio is significantly lower than the ratio of NBS-encoding genes in *G. max* (0.92%) (Zhang et al., 2011b) and *M. truncatula* (0.99%) (Song and Nan, 2014). However, the current draft of the *L. japonicus* genome sequences cover only about 91.30% of the gene space (Sato et al., 2008). In addition, the ratio of duplicated genes in *G. max* (Zhang et al., 2011b) and *M. truncatula* (Song and Nan, 2014) NBS-encoding genes accounted for 77.00 and 83.00% of the number of genes, respectively. We further analyzed gene duplication events using the 206 LjNBS genes, and only found 45 NBS-encoding genes (22.0%) involved in gene duplication events. A smaller proportion of LjNBS gene duplication events were not associated with the use of LjNBS genes with full sequences only.

Previous studies indicated that duplication-derived genes could succumb to several different fates. They may be stably maintained in genomes, deleted, or they may become non-functional due to birth-and-death mechanisms during evolution (Nei and Rooney, 2005). A large number of duplicated NBS-encoding genes might suddenly disappear during a selective sweep of the genome (Mun et al., 2009). Therefore, the observed low frequency of duplication events in LjNBS genes could be explained by either fewer duplications or rapid deletion of the copies after duplications.

Selection pressure in *L. japonicus* NBS-encoding genes

In the present study, we determined the selection pressure of *L. japonicus* CNL and TNL genes. Generally, large Ks/Ka ratios suggest the existence of birth-and-death mechanisms, where similar proteins are retained due to purifying selection (Piontkivska et al., 2002). In *L. japonicus*, the average CNL and TNL Ks values were greater than Ka values, which suggests that CNL and TNL genes have evolved through birth-and-death mechanisms, and it partially explains why there have been fewer gene duplication events.

We calculated the Ka/Ks values of nodes on the phylogenetic tree that should also be subject to purifying selection (**Figure S4** and Table 3). Positive selection was detected at seven nodes (CNL: 5 and TNL: 2), and most of these were located deep within the tree (CNL: node 1-4 and TNL: node 3), except for node 7 in the CNL clade and node 4 in the TNL clade (Table 3). This result could be caused by a few highly divergent sequences, especially in the N- and C-terminal regions. The multiple-sequence alignment shows that CNL and TNL sequences share few amino acid residues. Node 7 in the CNL and node 4 in the TNL clade are ancestral to 11 and 8 genes, respectively, but only 1 and 3 positively selected sites, respectively, were detected at these two nodes. This suggests that when the sequence number increases, the number of positively selected sites decreases.

Table 3. Likelihood ratio test results of *Lotus japonicus* NBS-encoding genes.

Node[a]	Ka/Ks under M0[b]	2ΔlnL M3 *vs* M0	2ΔlnL M8 *vs* M7	M8 estimates[c]	No. of positively selected sites[d]
CNL					
1	0.45	120.57**	29.40*	Ka/Ks>1 (p = 0.02, q = 0.01)	5
2	0.35	221.61**	37.40**	Ka/Ks>1 (p = 0.43, q = 0.69)	4
3	0.29	394.41**	51.77**	Ka/Ks>1 (p = 0.73, q = 1.72)	4
4	0.27	482.44**	49.87**	Ka/Ks>1 (p = 0.83, q = 2.00)	4
5	0.29	289.38**	17.19**	Ka/Ks>1 (p = 0.72, q = 1.32)	0
6	0.31	162.98**	9.12*	Ka/Ks>1 (p = 0.49, q = 0.84)	0
7	0.24	622.50**	18.20**	Ka/Ks>1 (p = 0.90, q = 2.13)	1
TNL					
1	0.35	280.88**	19.90**	Ka/Ks>1 (p = 0.81, q = 1.07)	0
2	0.32	332.40**	40.25**	Ka/Ks>1 (p = 1.04, q = 2.58)	0
3	0.38	304.52**	60.53**	Ka/Ks>1 (p = 0.73, q = 1.34)	7
4	0.33	817.20**	33.69**	Ka/Ks>1 (p = 0.76, q = 1.12)	3
5	0.23	112.85**	0.00	Ka/Ks>1 (p = 0.81, q = 1.71)	0

*P < 0.05 and ** P < 0.01 (χ^2 test); [a]node number from the phylogenetic tree; [b]Ka/Ks is the average ratio over sites under a codon model with one ratio; [c]p and q are the parameters of the beta distribution under model M8; [d]the number of amino acid sites estimated to have undergone positive selection under model M8.

Codon usage bias in *L. japonicus* NBS-encoding genes

A number of factors have been proposed to explain the mechanism of codon usage bias. Natural selection and mutation are two typical and recognized hypotheses (Duret, 2002). Natural selection occurs in highly expressed genes, such as translation elongation factors and ribosomal proteins, to ensure efficient and/or accurate translation (Hershberg and Petrov, 2008). The mutation hypothesis proposes that codon usage bias between different organisms is caused by phylogenetically constrained GC content (Hershberg and Petrov, 2008). Both mutational pressure and selective forces are involved in codon usage bias in many organisms (Hershberg and Petrov, 2008). In this scenario, the selection-mutation-drift theory has been proposed as the model under which codons are used in the genome (Duret, 2002). This model proposes that selection favors the optimal codon over minor codons, while mutational pressure and genetic drift allow minor codons to persist.

GC content may be one of the most important factors during the evolution of genomic structures (Bellgard et al., 2001), and we investigated GC content in LjNBS genes. The average GC3s value was 0.38, indicating a high frequency of AT in LjNBS gene sequences at the third position. Reduced selection against mutations in plant genomes could lead to a wide distribution bias of GC3s values (Liu et al., 2012). In LjNBS genes, the GC3s values ranged from 0.29 to 0.54, suggesting that LjNBS genes underwent mutation pressure or selection against mutation. A neutrality plot (GC12 *vs* GC3s) revealed the relationship between mutation and selection bias. If the points were distributed along the diagonal line, the results would indicate that the genes underwent neutral mutation. In contrast, if the points were distributed on the parallel lines of abscissa, the results would indicate that the genes were completely non-neutral. In other words, the regression coefficient (slope) provided a measure of relative neutrality of GC12 to GC3s. The extent of the slope smaller than unity indicated that the extent of GC12 neutrality was less than that of GC3s (Sueoka, 1999b). In the present study, the significant positive correlations between GC12 and GC3s indicated that LjNBS underwent neutral mutation.

Our PR2 plot results showed that LjNBS codons were used disproportionately and that most codons ended with G and T, which indicates the involvement of natural selection. Combined

with the neutrality plot, our results support the hypotheses that both mutation and selection contribute to LjNBS codon usage bias, and that mutational bias could be considered the major factor shaping codon usage.

Nine optimal codons were identified in LjNBS gene sequences. Among them, most codons ended in G and C. Interestingly, LjNBS optimal codons were AT-rich. These findings are in agreement with studies of *Drosophila* and *Caenorhabditis* optimal codons (Hershberg and Petrov, 2008). However, we are not sure why this is the case, and to our knowledge, no hypotheses have been proposed to explain this phenomenon.

CONCLUSIONS

We compared the results of our study, which used only full-length LjNBS sequences, to a previous study that used partial and full-length LjNBS sequences. We found that we obtained different results than the previous study, and our results were consistent with what has been found in other plants. Our results indicate that the incomplete NBS-encoding sequences used in the previous study may have led to incorrect conclusions. In addition, we observed a low frequency of duplication events in LjNBS genes, and we speculated that the reason for this is either fewer duplications or rapid deletions of the duplicated genes. Moreover, our analysis of LjNBS genes suffered due to selection pressure and codon usage bias. The results suggest that strong purifying selection occurred in the TNL and CNL sequences. Moreover, the results showed that both mutation and selection contributed to LjNBS codon usage bias, and mutational bias could be considered the main factor shaping codon usage.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#31101427), the Initial Special Research for "973" Program of China (#2012CB126313), grants from the Ministry of Science and Technology of China (#2013AA102602, #2012BAD33B07, #2011BAD35B04), and the Shandong Province Germplasm Innovation and Utilization Project, Young Talents Training Program of Shandong Academy of Agricultural Sciences.

Supplementary material

REFERENCES

Belkhadir Y, Nimchuk Z, Hubert DA, Mackey D, et al. (2004). *Arabidopsis* RIN4 negatively regulates disease resistance mediated by RPS2 and RPM1 downstream or independent of the NDR1 signal modulator and is not required for the virulence functions of bacterial type III effectors AvrRpt2 or AvrRpm1. *Plant Cell* 16: 2822-2835.

Bella J, Hindle KL, McEwan PA and Lovell SC (2008). The leucine-rich repeat structure. *Cell Mol. Life Sci*. 277: 519-527.

Bellgard M, Schibeci D, Trifonov E and Gojobori T (2001). Early detection of G+C differences in bacterial species inferred from the comparative analysis of the two completely sequenced *Helicobacter pylori* strains. *J. Mol. Evol*. 53: 465-468.

Cannon SB, Mitra A, Baumgarten A, Young ND, et al. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol*. 4: 10.

Cheng Y, Li X, Jiang H, Ma W, et al. (2012). Systematic analysis and comparison of nucleotide-binding site disease resistance genes in maize. *FEBS J*. 279: 2431-2443.

Dangl JL and Jones JD (2001). Plant pathogens and integrated defence responses to infection. *Nature* 411: 826-833.

Dodds PN and Rathjen JP (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet*. 11: 539-548.

Duret L (2002). Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev*. 12: 640-649.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, et al. (2006). Pfam:clan, web tools and services. *Nucleic Acids Res*. 34: 247-251.

Hershberg R and Petrov DA (2008). Selection on codon bias. *Annu. Rev. Genet*. 42: 287-299.

Jones JD and Dangl JL (2006). The plant immune system. *Nature* 444: 323-329.

Kawabe A and Miyashita NT (2003). Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet. Syst*. 78: 343-352.

Kohler A, Rinaldi C, Duplessis S, Baucher M, et al. (2008). Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Mol. Biol*. 66: 619-636.

Li L, Stoeckert CJJ and Roos DS (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13: 2178-2189.

Li X, Cheng Y, Ma W, Zhao Y, et al. (2010). Identification and characterization of NBS-encoding disease resistance genes in *Lotus japonicus*. *Plant Syst. Evol*. 289: 101-110.

Liu H, Huang Y, Du X, Chen Z, et al. (2012). Patterns of synonymous codon usage bias in the model grass *Brachypodium distachyon*. *Genet. Mol. Res*. 11: 4695-4706.

Meyers BC, Morgante M and Michelmore RW (2002). TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *Plant J*. 32: 77-92.

Meyers BC, Kozik A, Griego A, Kuang HH, et al. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 15: 809-834.

Mun JH, Yu HJ, Park S and Park BS (2009). Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa*. *Mol. Genet. Genomics* 282: 617-631.

Nei M and Rooney AP (2005). Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet*. 39: 121-152.

Pan Q, Wendel J and Fluhr R (2000). Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J. Mol. Evol*. 50: 203-213.

Piontkivska H, Rooney AP and Nei M (2002). Purifying selection and birth-and-death evolution in the histone H4 gene family. *Mol. Bio. Evol*. 19: 689-697.

Richly E, Kurth J and Leister D (2002). Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. *Mol. Bio. Evol*. 19: 76-84.

Ronquist F and Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572-1574.

RoyChoudhury S and Mukherjee D (2010). A detailed comparative analysis on the overall codon usage pattern in herpesviruses. *Virus Res*. 148: 31-43.

Sato S, Nakamura Y, Kaneko T, Asamizu E, et al. (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Res*. 15: 227-239.

Song H and Nan Z (2014). Genome-wide analysis of nucleotide-binding site disease resistance genes in *Medicago truncatula*. *Chin. Sci. Bull*. 59: 1129-1138.

Song H, Wang P, Nan Z and Wang X (2014). The WRKY transcription factor genes in *Lotus japonicus*. *Int. J. Genomics* 2014: 420128.

Sueoka N (1999a). Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene* 238: 53-58.

Sueoka N (1999b). Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A=T and G=C. *J. Mol. Evol*. 49: 49-62.

Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Bio. Evol*. 24: 1596-1599.

Tan S and Wu S (2012). Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*. *Comp. Funct. Genomics* 2012: 418208.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, et al. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 25: 4876-4882.

Wang Y, Li J and Paterson AH (2013). *MCScanX-transposed*: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics* 29: 1458-1460.

Wright F (1990). The 'effective number of codons' used in a gene. *Gene* 87: 23-29.

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Bio. Evol*. 24: 1586-1591.

Zhang L, Guo Y, Luo L, Wang YP, et al. (2011a). Analysis of nuclear gene codon bias on soybean genome and transcriptome. *Acta Agron. Sin*. 37: 965-974.

Zhang X, Feng Y, Cheng H, Tian D, et al. (2011b). Relative evolutionary rates of NBS-encoding genes revealed by soybean segmental duplication. *Mol. Genet. Genomics* 285: 79-90.

Zhou T, Wang Y, Chen JQ, Araki H, et al. (2004). Genome-wide identification of NBS genes in *japonica* rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol. Genet. Genomics* 271: 402-415.