Running head: NBS-LRR gene family in *Medicago truncatula*

Corresponding author: Dr. Steven Cannon
2021 Agronomy Hall
United States Department of Agriculture–ARS
Iowa State University
Ames, IA 50011
USA

Phone: 515-233-1904
Fax: 515-294-9359
Email: steven.cannon@ars.usda.gov

Journal research area: Genome Analysis
Identification and Characterization of NBS-LRR Genes in the Model Plant

Medicago truncatula

Carine Ameline-Torregrosa, Bing-Bing Wang, Majesta S. O’Bleness, Shweta Deshpande, Hongyan Zhu, Bruce Roe, Nevin D. Young and Steven B. Cannon*

Laboratoire des Interactions Plantes Microorganismes (LIPM), UMR CNRS-INRA 442-2594, Avenue de l’Agrobiopole, 31326 Castanet Tolosan, France (C.A.-T); Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, USA (C.A.-T., B.-B.W., N.D.Y); Department of Plant Biology (N.D.Y.); Advanced Center for Genome Technology, Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK 73109, USA (M.O’B., B.A.R.); Department of Plant & Soil Sciences, University of Kentucky, 230 KTRDC, Lexington, KY 40546-0236, USA (H.Z.); United States Department of Agriculture–Agricultural Research Service and Department of Agronomy, Iowa State University, Ames, IA 50011, USA (S.B.C)
Footnotes
This work was supported by the National Science Foundation (#0321664 and #0321460) to N.D.Y.

*Corresponding author: e-mail steven.cannon@ars.usda.gov; fax 515-294-9359
ABSTRACT

The NBS-LRR gene family accounts for the largest number of known disease resistance genes, and is one of the largest gene families in plant genomes. We have identified 333 non-redundant NBS-LRRs in the current *M. truncatula* draft genome (Mt1.0), likely representing 4-500 NBS-LRRs in the full genome, or roughly three times the number present in *Arabidopsis thaliana*. Although many characteristics of the gene family are similar to those described on other plant genomes, several evolutionary features are particularly pronounced in *M. truncatula*, including a high degree of clustering, evidence of significant numbers of ectopic translocations from clusters to other parts of the genome, a small number of more evolutionarily stable NBS-LRRs, and numerous truncations and fusions leading to novel domain compositions. The gene family clearly has had a large impact on the structure of the genome, both through ectopic translocations (potentially, a means of seeding new NBS-LRR clusters), and through two extraordinarily large superclusters. Chromosome 6 encodes approximately 34% of all TIR-NBS-LRRs, while chromosome 3 encodes approximately 40% of all CC-NBS-LRRs. Almost all atypical domain combinations are in the TIR-NBS-LRR subfamily, with many occurring within one genomic cluster. This analysis shows the gene family not only is important functionally and agronomically, but also plays a structural role in the genome.
INTRODUCTION

Plants have evolved sophisticated mechanisms to recognize and guard against pathogens. Interaction between hosts and pathogens triggers both localized and systemic resistance responses. Disease resistance frequently is governed by specific recognition between pathogen avirulence (Avr) genes and corresponding plant disease resistance (R) genes. This type of “gene for gene” interaction usually is accompanied by a hypersensitive response leading to the restriction of pathogen growth. In the past decade, R genes have been cloned from numerous plant species, conferring resistance to a wide range of plant pathogens including bacteria, fungi, oomycetes, viruses, and nematodes (Dangl and Jones, 2001; Meyers et al., 2003; DeYoung and Innes, 2006). However despite the wide range of pathogen taxa involved, R genes seem to encode a limited set of proteins consisting of conserved domains (reviewed in Dangl and Jones, 2001).

The largest class of R genes encodes proteins with a nucleotide-binding site (NBS) and a leucine-rich repeat (LRR) region. This domain architecture is consistent with a role in pathogen recognition and defense response signaling. The NBS domain contains several conserved motifs typically found in ATP- or GTP-binding proteins and also present in several structurally related regulators of animal apoptosis (Traut, 1994). In plant R proteins, the NBS region is a conserved domain that is responsible for the binding and the hydrolysis of ATP and GTP (Tameling et al., 2002). LRRs are typically involved in protein-protein interactions, and various studies indicate that the LRR motif is at least partly responsible for recognition specificity (Kobe and Deisenhofer, 1995; Leister and Katagiri, 2000). Between the NBS and LRR domains, the ARC domain has recently been identified to play a role in the recruitment of the LRR domain to the N terminal region and in the molecule inactive/active statement (Rairdan and Moffett, 2006). Some studies support a “guard hypothesis” model for R gene function, where NBS-LRR proteins guard plant targets (guardee) against pathogen effector proteins, and some R proteins have been shown to be activated upon interaction between pathogen virulence factors and guardee proteins (Axtell and Staskawicz, 2003; Mackey et al., 2003; Belkhadir et al., 2004).

The NBS-LRR family of R genes can be further divided into two subfamilies based on deduced N-terminal structural domains. One subfamily, termed TIR-NBS-LRR or TNL, encodes a domain with similarity to the intracellular signaling domains of the
**Drosophila** Toll and mammalian Interleukin (IL)-1 receptor, while the second, termed CC-NBS-LRR or CNL, codes for a putative Coiled-Coil domain in the N-terminal region. These two subfamilies can also be distinguished by the unique amino acid motifs found within the NBS domain itself (Meyers et al., 1999; Pan et al., 2000). Although the TIR domain interacts with effector molecules (Axtell and Staskawicz 2003; Burch-Smith et al. 2007), the TNL and CNL subfamilies seem to require different downstream factors, mediated primarily via the different N-terminal domains in these families. In genes characterized to date, the genes from the TNL and CNL subfamilies depend on EDS1- or NPR1-type signaling pathways, respectively (Aarts et al., 1998; Glazebrook, 1999; Peart et al., 2002; Hu et al., 2005; Wiermer et al., 2005). Nevertheless, the variety of domain arrangements in this large, diverse family suggests that members of the family will likely participate in a range of signaling pathways.

Conservation of the NBS domain has been used to study the genomic architecture of this gene family. *R* genes are unevenly distributed in plant genomes and many reside in local multigene clusters. The clustered distribution of *R* genes provides a reservoir of genetic variation from which new specificities can evolve. Mechanisms like duplication, unequal crossing-over, ectopic recombination, gene conversion, and diversifying selection have been proposed to contribute to the structure of *R* gene clusters and the evolution of resistance specificities (Michelmore and Meyers, 1998; Young, 2000; Sun et al., 2001). Moreover, the presence of conserved motifs primarily within the NBS domain has been used extensively to identify resistance genes homologues (RGHs) in model and crop species (Kanazin et al., 1996; Aarts et al., 1998; Penuela et al., 2002; Zhu et al., 2002; Ferrier-Cana et al., 2003; Yaish et al., 2004; Palomino et al., 2006). In species where *R* genes have been studied, this pattern of widespread multigene clusters is common (Young, 2000; Hulbert et al., 2001; Meyers et al., 2003). In *Arabidopsis thaliana*, where *R* genes have been studied in detail, 149 NBS-LRR-encoding genes plus 58 related genes lacking LRRs have been identified (Meyers et al., 1999; Meyers et al., 2003). Both CNL and TNL classes could be further subdivided based on specific motifs, intron position and genomic distribution. Interestingly, CNL genes are widely distributed in both monocot and dicot species, while TNLs appear to be restricted to dicot species (Meyers et al., 1999). A few “TN” genes (TNL but lacking an LRR domain) have been
identified in rice but differ greatly from typical TNL genes (Bai et al., 2002). Thus, it appears that the \( R \) gene arsenal of monocot and dicot species have significantly diverged during the evolution of these plant lineages.

*\( M. \) truncatula* is a self-fertile, annual and diploid plant that has been selected as a model legume (Barker et al., 1990; Cook, 1999). Previously, we reported on the identification of NBS encoding sequences in *\( M. \) truncatula* based on specific PCR amplification with primers designed from conserved regions of the NBS domain (Zhu et al., 2002). This earlier study identified 147 NBS-LRR sequences from *\( M. \) truncatula* (107 TNL and 40 CNL) (Zhu et al., 2002). More recently, *\( M. \) truncatula* has become the target for hierarchical genome sequencing based on bacterial artificial chromosome clones (BACs). This ongoing effort to sequence the *\( M. \) truncatula* genome now enables a much more extensive and detailed examination of the NBS-LRRs in this model legume.

In the public draft assembly, which is estimated to span approximately 60% of the euchromatic space of *\( M. \) truncatula*, we identified at least 333 NBS-LRR encoding genes in the A17 genotype that is now being sequenced. Here we report an analysis of the evolution and genomic organization of these genes in *\( M. \) truncatula* based on genomic sequence data from the first large-scale genome assembly of the ongoing sequencing project (www.medicago.org/genome).

**RESULTS**

**Genome assembly, gene prediction and nomenclature**

Genomic sequence consisted of the 1.0 draft genome assembly generated by the Medicago Genome Sequencing Consortium (MGSC; http://www.medicago.org/genome/downloads/Mt1/), with gene predictions from the International Medicago Genome Annotation Group (IMGAG) (Town, 2006). The assembly consists of 1,826 BAC clones, in 348 sequence contigs and 275 scaffolds, altogether comprising 186.2 Mbp non-redundant genome sequence. The MGSC estimates that the assembly covers 38-47% of the entire *\( M. \) truncatula* genome, and captures a larger proportion (55-58%) of genes (Mt1.0 fact sheet, http://www.medicago.org/genome/downloads/Mt1/Mt1.0.pdf).
The *M. truncatula* NBS-LRR genes in this study were identified from IMGAG annotated genes. Gene names (Supplemental Table 1) follow the IMGAG naming convention, as illustrated by gene AC148761_18.3: the characters before the underscore are the Genbank accession for the source BAC; the number after the underscore is the gene number within the BAC; and the number after the period is the version of this gene call. For convenience in tree figures, shorter, more informative names are used. Aliases to IMGAG names are provided in Supplemental Table 1. The format for short names is illustrated by Mt2g1873: genus and species in the first two characters, followed by chromosome number, then type (g = gene), then gene order in the pseudomolecule build (Mt1.0; http://www.medicago.org-genome/downloads/Mt1). These names are for use only in this paper; the persistent names (in Genbank and EMBL) use the IMGAG format.

**Identification of NBS-LRR genes in *M. truncatula***

We used similarity searches based on extended NBS-LRR domains (Materials and Methods) to identify NBS-LRR genes in the A17 ecotype genomic sequence. A total of 333 non-redundant sequences, consisting of 177 putative CNL and 156 TNL sequences, were used in subsequent analyses, as described in Materials and Methods (Supplemental Table 1). Thirty additional sequences that appear to be related to or derived from NBS-LRRs, but were too divergent for inclusion in phylogenetic analyses, are also shown in Supplemental Table 1. The 333 sequences included in phylogenetic analyses are distributed across all chromosomes of *M. truncatula* (Figure 1), with six (4 CNLs and 2 TNLs) located on still-unmapped BACs. Three chromosomes (3, 4 and 6) contain a disproportionately large number of NBS-LRRs (more than 54%).

As *M. truncatula* has not been fully sequenced, we also attempted to estimate how many NBS-LRR genes may be missing in the present study. We asked what proportion of a random set of expressed *M. truncatula* NBS genes are found in the Mt 1.0 assembly. We find that proportion to be approximately 196/294=2/3, as follows. For the “random set of expressed *M. truncatula* NBS genes” (the denominator), we took the *M. truncatula* Transcript Assemblies and singletons (“TA unigenes” from http://plantta.tigr.org, release 2) that 1) have a tblastn E-value of 1e-15 with some *M. truncatula* NBS gene, and 2) have higher similarity to an *Arabidopsis* NBS-LRR than to an *Arabidopsis* gene from any
other gene family. Applying only the first criterion, we find 661 NBS-like *M. truncatula* sequences among the 55,182 TA unigenes. Applying the second criterion lowers the number to 294 – because a large proportion of the initial candidates are actually more similar to some other *Arabidopsis* gene. Given the denominator (294), we then asked how many *M. truncatula* NBS genes have a nearly identical match to those Mt NBS sequences (196). Therefore, we estimate that the current genome sequence contains roughly two-thirds (196/294) of *M. truncatula* NBS-LRR genes. This is consistent with estimates that the Mt 1.0 assembly covers approximately two-thirds of the euchromatic space of *M. truncatula*.

**Genomic distribution and phylogenetic analysis**

Most NBS are physically clustered in the genome (Figure 1), with more than 54% of the NBS-LRR genes encoded on chromosomes 3, 4 and 6. Using a sliding window size of 100 kb, 79.8% of NBS domains occur in clusters of at least two genes, and 49.5% are in clusters of at least 5 genes. For this window size, the largest cluster (on chromosome 6) contains 14 genes. Using a sliding window size of 250 kb, 83.6% of NBS are in clusters of at least two genes, 68.9% are in clusters of at least 5 genes. For this window size, the largest cluster (on chromosome 3) contains 23 genes. Further relaxing these clustering criteria, a significant fraction of all *M. truncatula* NBS are in two very large, extended clusters: one at the north end of chromosome 3 containing 82 genes (73 CNL and 9 TNL) and extending across 55 BAC clones (including 11 unspanned gaps) and another at the south end of chromosome 6 containing 57 genes (all TNL) and extending across 34 BAC clones (including 10 unspanned gaps). Together, these two clusters contain a remarkable 39% of all NBS genes in this study.

Phylogenetic trees were constructed from 333 NBS sequences (177 CNLs and 156 TNLs). As previous studies have shown that phylogenies calculated from the NBS domain robustly distinguish the TNL and CNL subfamilies (Meyers et al., 1999), we constructed phylogenies separately for these subfamilies (CNL Figure 2A, B; TNL Figure 2C). NBS consensus sequences from TNL and CNL extracted from the previous alignments were used as outgroups to root the CNL and TNL trees respectively. To better visualize genomic location, the names of genes located on the same chromosome are
indicated by the same color, and by the short-name alias described above. We have divided the tree into clades on the basis of clade rooting with non-legume sequences from poplar and *Arabidopsis*. This results in 17 clades for CNL and ~8 for TNL clades (with uncertainty due to low bootstrap support for some clades).

Figures 1 and 2A-C provide overviews of the CNL and TNL subfamilies. Figure 1 shows gene position and subfamily on *M. truncatula* chromosome pseudomolecules. Figure 2 shows phylogenies, including chromosome of origin (by color and sequence name), clade age relative to *Arabidopsis thaliana* (‘Arabidopsis’ or ‘At’) or *Populus trichocarpa* (‘poplar’ or ‘Pt’) sequences (pink dots), relationship to putative internal genomic duplications (black arrows), gene relatedness and evolutionary rate (phylogenetic structure and branch length), and approximate expression levels and regulatory element counts (right). The chromosome of origin is informative in that it highlights local expansion of some sequence types, as well as changes in chromosome origin – suggesting either large-scale rearrangements or ectopic translocations. The approximate locations of *Arabidopsis* and poplar sequences, and the internal genomic duplications (black arrows) inferred from large-scale duplications within the genome, are informative in that they provide approximate relative age calibrations for each clade.

Most NBS derive from relatively recent gene duplications and for the most part they are highly similar to other NBS in the same genomic clusters (although some of the observed sequence similarity may also be the result of gene conversion). In Figures 2A-C, pink dots indicate approximate locations of coalescent points with poplar or *At* NBS clades. These points provide relative time references, showing which clades have probably expanded within legumes. The poplar and *Arabidopsis* coalescence points are reported together because, although poplar and *Arabidopsis* separated from the legumes at approximately 70-84 mya and 108-117 mya, respectively (Wikstrom et al., 2001, 2003; Sanderson et al., 2004; Tuskan et al., 2006), it was generally not possible to confidently distinguish these coalescences relative to legume NBS clades; see Supplementary Data, files Mt1NBS_repsEtcLjAtPt.nhx and Mt1NBS_repsEtcLjAtPt.fas.

Most legume NBS sequences are found at greater than 0.5 PAM units (accepted point mutations per site) from these coalescence points with non-legume species. By measuring the phylogenetic distance between *M. truncatula* NBS sequences, we can
assess how many have originated “recently.” In this context, a distance cutoff of 0.5 PAM units between *M. truncatula* sequences (or average distance of 0.25 PAM to the *M. truncatula-M. truncatula* coalescence point) is a reasonable indicator of “nearness” in that it is much shorter than the evolutionary distance to *Arabidopsis* or poplar sequences, and so represents gene duplications that most probably occurred within legumes. On average, each *M. truncatula* NBS is within 0.5 PAM of 9.2 other sequences, again indicating that many groups of sequences have high sequence similarity. This is evident in the trees in Figures 2A-C in the form of clades with many sequences and short branch lengths, such as the many similar sequences from chromosome 3 in Figure 2A.

The tree in Figure 2A-C is divided into 17 CNL and 8 TNL clades. These are “legume-specific,” if we define legume-specific to mean that each clade contains no sequences from *Arabidopsis* or poplar (using the coalescence points described above). Properly, this designation should be further qualified. More closely related non-legume species (for example, from Rosaceae) might still fall within these clades, and the absence of a poplar or *Arabidopsis* sequence could sometimes be due to gene loss. Gene conversion or homogenization might also foreshorten distances in some *M. truncatula* clades. Nevertheless, the genes in this large gene family trace to a small number of legume-specific progenitor sequences. It is also likely, therefore, that a much larger number of genes have arisen and been lost in this timeframe.

**Composition of clusters and evidence of transposing duplications**

Most legume-specific clades are dominated by sequences from one chromosome (and usually from one or a small number of genomic clusters), but many also contain small numbers of sequences from other chromosomes. Specifically, 14 CNL and 6 TNL legume-specific clades are “mixed” (with sequences from multiple chromosomes). This is 80% of the all clades. A mixed clade could arise in several ways: by chromosomal rearrangement (for example, breakage and fusion); by transposition; or by large-scale genomic duplication. In at least five clades, the origin of sequences from different chromosomes (or widely separated parts of one chromosome) can be traced to internal synteny in *M. truncatula* – the likely remnants of an early episode of polyploidy in the legumes (Cannon et al., 2006). Examples are shown in Figures 2A-C, with black two-headed arrows. For example, in 2C, clade TNL-1 contains a sub-clade with six sequences
from chromosome 4 and six from chromosome 6. These two clusters come from regions of synteny between chromosomes 4 and 6 (Cannon et al., 2006 and unpublished data).

While some mixed clades, including TNL-1 just described, can be traced to internal segmental duplications and others are probably cryptic remnants of duplications, no longer apparent after rearrangements, there are also other mixed clades and clusters that are best explained as ectopic translocations. Supplemental Table 1 indicates 29 such cases: instances in which a clade of closely related sequences from one genomic cluster, with one sequence occurring in a distant part of the genome. These instances can be thought of as having donor regions (a cluster of related sequences in one part of the genome) and acceptor regions (the location of the related gene outside of the “home cluster”). Examples of such clades are Figure 2A, CNL-3 (Mt5, Mt2, Mt7); Figure 2B, CNL-9 (Mt6, Mt3; Mt7, Mt5; Mt5, Mt1); Figure 2C, TNL-1 (Mt6, Mt2); TNL-4 (Mt6, Mt2; Mt6, Mt1).

Probable transpositions (donations) do not seem to target particular locations. They seem to occur throughout the genome and not just in NBS-rich regions. For example, there are seven donations to chromosome 1, but there are only 11 NBS-LRRs, in total, on chromosome 1. These are mostly unclustered (five donations occur as singletons, and the remainder occur in two clusters).

There are, however, some instances of apparent donations into existing clusters. Several examples are the TNL genes in the large CNL cluster on chromosome 3 (Mt6g1868 \(\not\sim\) Mt3g5125, Mt2g3436 \(\not\sim\) Mt3g4772), and the CNL genes in TNL clusters on chromosome 5 (Mt8g385 \(\not\sim\) Mt5g2441, Mt8g683 \(\not\sim\) Mt5g5043).

While some clusters contain genes that appear to have come from other regions of the genome, a broader phenomenon is physical clusters that include divergent sequences (regardless of the origin). There are 26 instances of TNL and CNL genes falling within 100 kb of one another (Supplementary Materials, file “MtNBS_clusters.xls”). In fact, five pairs of CNL and TNL genes fall within single BACs: Mt7g2240/AC169666_26.4 (CNL-4) and Mt7g2242/AC169666_23.4 (TNL-2); Mt5g3987/CT963106_1.4 (CNL-9) and Mt5g3991/CT963106_17.4 (TNL-8); Mt3g812/CT963074_17.3 (TNL-4) and Mt3g833/CT963074_13.3 (CNL-3); Mt3g796/CT963132_6.3 (TNL-4) and
Mt3g809/CT963132_9.3 (CNL-4); Mt3g654/CT967304_15.3 (CNL-2) and Mt3g655/CT967304_17.3 (TNL-8).

Domain analyses

Protein domains of the 333 NBS-encoding genes in this study were predicted using HMM searches against Pfam v. 20 (Eddy, 2001; Bateman et al., 2002), followed by correction of likely prediction errors (e.g. fusions with adjacent transposon proteins). Domain arrangements were divided into putative structural categories according to the nature, number and organization of their constituent domains, as listed in Supplemental Table 1 and counted in Table 1. Unusual domain arrangements are also shown in Figures 2A-C.

Comparisons between protein sequences within a single structural category revealed some likely inaccuracies in automated gene predictions and annotations. Probable mis-annotations were detected in 10 proteins (indicated with an asterisk in the “domains” column of Supplemental Table 1), and generally consisted of an additional exon in the C terminal region. Such exons include HSP70, reverse transcriptase, MMR-HSR (GTPase), RNase H and chaperone-associated domain. These are not included in the tally of “domain classes” in Table 1.

Pfam analyses could not identify the CC motif present in the N terminal region, even though previous studies have demonstrated that the presence or absence of this motif is correlated with specific signatures in the NBS domain (Meyers et al., 1999, 2002). We used these NBS signatures as the basis for classifying sequences in the CNL subfamily.

A majority of the proteins examined belong to the canonical classes described in the literature (Meyers et al., 1999; Meyers et al., 2003): CNLs or TNLs (Table 1, classes 2 and 8). A minority of genes, however, have less typical domain arrangements. Interestingly, almost all of these are in the TNL subfamily (discussed later).

In the CNL subfamily, the predominant “unusual” domain arrangement is a missing LRR; specifically, 25/177 (16%) lack the LRR (i.e. CN). Only one other unusual class is observed in the CNL is the result of a putative fusion in CU013515_1.4 / Mt5g1164, with the Rpw8 domain. The closest homolog of CU013515_1.4/Mt5g1164 in
*Arabidopsis* (At5g66910) displays the same domain structure. The *Rpw8* gene in *Arabidopsis* provides broad-spectrum resistance mildew resistance (Xiao et al., 2001). There are 6 proteins with *Rpw8* similarity in *Arabidopsis thaliana*; one of these (AT5G66910 = RPW8.1) is a C-terminal fusion with an NBS-LRR domain. Recent studies show that RPW8.1 from *A. thaliana* is absent in the *A. lyrata* genome (Orgil et al., 2007), presumably due to a deletion event (Xiao et al., 2001). Both the CU013515_1.4 / Mt5g1164 and At5g66910 proteins contain four LRR domains, and are reciprocal top matches to each other. Thus, it appears that both have been retained as single copy genes in their respective genomes over the ~100 million years since the last common ancestor of these plant lineages (supplementary file RPW8_At_Mt_alignments.doc).

In contrast to the CNL, the TNL subfamily is highly diverse in terms of domain arrangements. Only 86/156 (55%) are “typical” TNL. The second and third most common classes are TN (27/156 = 17%) and NL (25/156 = 16%). One of the most intriguing sets of atypical domain arrangements is within clade TNL-8 (bottom of Figure 2C), where a cluster of predicted peptides on chromosome 4 includes the structures TNTNL (5), TNLT (2), TNL (1), TTNL (1), N (1), NL (1). The sister clade, with genes from chromosomes 7 and 4 (and two unplaced BACs) contains one each of NT, NTNL, N, TNTNL, NT. That these sequences occur mainly on two chromosomes suggests that these unusual sequences have been maintained for some time, probably at least since polyploidy, early in the legumes (Pfeil et al. 2005, Schlueter et al. 2004, Cannon et al. 2006), with the majority of the sequences also expressed (Figure 2C). Comparisons with NBS sequences from *Arabidopsis*, poplar, and Lotus suggest that the *M. truncatula* chromosome 4 sequences have duplicated and diversified following the speciation with Lotus, meaning that the diverse domain arrangements occurred relatively recently within *M. truncatula* (supplementary files Mt1NBS_repsEtcLjAtPt.fas and Mt1NBS_repsEtcLjAtPt.pdf). Nevertheless, sequences from other legume and non-legume sequences from the broader clade (TNL-8) also contain unusual domain arrangements, although not the same as observed in *M. truncatula*. The homologs from Lotus include the structures TN, N, NL. The closest *Arabidopsis* match, At5g36930, which falls within this clade, has canonical structure TNL; the second-closest match (though outside this clade) is At3g25510, with atypical structure TNLTNL. Thus, this
clade seems to tolerate more domain rearrangement than other clades in the TNL subfamily, and certainly more than anywhere in the CNL subfamily.

An additional intriguing instance of a putative fusion in the TNL subfamily is a predicted protein with domains TNLTNL (AC126790_31.4 / Mt6g1826; Genbank ID ABE83302.2). Such a fusion would not be unprecedented, as at least one gene with similar structure is present in the current manually annotated Arabidopsis peptides (At3g25510; TIGR v.7). The Arabidopsis gene is not an ortholog, as it apparently results from an independent event. Rexamination of M. truncatula BAC AC126790.38 confirmed the initial prediction, and about a third of the sequence has 100% cDNA support (with ESTs CX538931 and CX524109). Therefore this gene structure and five exons are otherwise not unusual, and the 3123 nt of coding sequence occurs within the 4550 nt total gene region.

Motif analyses

Analyses of motifs within NBS domains reveal additional features. Since typical NBS domains often contain variable motifs that (NBS–A and –C, described in Meyers et al. 1999) while others are more conserved (NBS–B and –D), they can be used to distinguish TNL from CNL sequences. We identified short and highly conserved stretches within each domain configuration “class” (TNL, TNTNL, etc.) (Tables 2 and 3) using Pfam and HMM domain analysis (Eddy, 2001) and MEME motif analysis (Bailey and Elkan, 1995) on aligned domain subgroups. In most cases, the conserved sequences we observed have been described previously (Meyers et al., 1999; Meyers et al., 2002; Meyers et al., 2003; Tuskan et al., 2006). As expected, MEME analyses revealed that P-loop, Kin-2, and GLPL motifs all are conserved among NBS gene family members. In contrast, NBS-A and NBS-C from TNL proteins are less conserved and these motifs could not be identified at all in domain classes 6 and 8 (Tables 2 and 3) due to high levels of diversity in amino-acid sequences. We also examined proteins with doubled NBS domains (classes 5, 9, and 11; NTNL, TNTNL, TNLTNL) and found that the two domains within a single protein usually are dissimilar in motif structure (Table 1). In each case they have one truncated NBS domain, usually involving NBS-A and NBS-C motifs. Beyond differences in specific motifs, the two NBS domains found within a single gene
could also be distinguished by their overall amino-acid sequences. For example, Mt6g1826/AC126790_31.4, the only member of class 11, displays a notably high degree of difference between the two NBS-A motifs (Table 1). Examination of the two TIR domains within a single protein (classes 9, 10 and 11) does not reveal a high dissimilarity as such as observed in two adjacent NBS domains (not shown).

**In silico analysis of CNL and TNL gene expression using EST libraries**

To assess which genes in this study have expression support, we compared the predicted genes against available ESTs (231,765 ESTs from 55 libraries, from Genbank in April 2006). Because many NBS-LRR genes are similar to one-another, only top matches were considered, after applying a high match stringency of at least 95% of nucleotide identity between EST and genomic sequences. At this threshold, 168 NBS genes in this study have EST support, representing 50.5% of predicted genes in the study (indicated by EST matches in the right-hand column of Figures 2A-C, and Supplemental Table 1). Altogether, 530 EST matches were identified, with an average of 3.1 ESTs per expressed NBS gene. ESTs are approximately equally distributed in the CNL and TNL classes: 81 CNLs and 87 TNLs are represented by at least one EST. A majority of CNL and TNL genes display one or two ESTs, but 22 genes have at least 5 ESTs and one (AC135229_11.5/Mt8g3004) has 31 ESTs (Figs 2A-C and Supplemental Table 1). Among these relatively highly expressed genes, most are located on the lower arm of chromosome 6, within the supercluster described above. These genes are expressed in a wide range of libraries, including those constructed from various developmental stages, tissue types, and pathogen-challenged or non-challenged tissue.

Approximate expression patterns, judged by counts of EST matches, vary substantially between clades and even between highly similar genes within the same clade. For example, genes on most branches in the CNL tree in Figure 2A have lower expression than those in clade CNL-16 (bottom of Figure 2B). Within clade CNL-16, however, corresponding ESTs numbers range from 1 to 31.
**Pseudogenes**

There are 49 unique sequences in Mt1.0 with stop codons, identified using a CC or TIR consensus NBS sequences in a tblastn query against the Mt1.0 nucleotide chromosome assemblies, and filtered at E-value 1e-10. This is probably an underestimate, as either many pseudogene fragments may fall below this level of significance, or are sequences without stop codons in the region of this query, but not predicted among the IMGAG gene calls in this assembly. Nevertheless, the stated criteria give values for comparison. Pseudogene counts per chromosome are (for chromosomes 0-8) 1,2,2,4,15,6,7,5,7. Counts of CNL- and TNL-like pseudogenes are 22 and 27 (Supplemental Table 2).

We also have observed that 91.8% of all predicted NBS pseudogenes are within 100 kb of another predicted NBS gene. These pseudogenes are not, however, distributed on the chromosomes in the same way as predicted NBS genes without stop codons. More pseudogenes are found on Mt4 than would be expected (15 observed versus 6.1 expected), and fewer are found on Mt3 than expected (4 observed versus 13.5 expected). These differences are supported by a test for independence by chromosome, with a chi-squared value of 0.0033 (df=8). The differences by chromosome are primarily due to the excess of pseudogenes on Mt4 (13 of 15 of which are TNL-class) and the dearth of pseudogenes on Mt3 (most of which would have been expected to be CNL-class, following the pattern of distribution of predicted NBS genes). Further, all of the Mt4 TNL pseudogenes are near predicted TNL genes in the two large clusters of TNL genes with diverse domain arrangements (Fig. 1, Table 1, and Supplemental Table 1). Specifically, 9 of the 13 Mt4 TIR pseudogenes occur in the cluster that accounts for the majority of domain diversity in the TNL subfamily (five of nine TNL domain classes in Table 1; also Fig. 2C, bottom).

There also is evidence that some of the predicted pseudogenes may be expressed. Four of the 49 predicted pseudogenes match at least one EST at 99 to 100% identity over 58-89% of the genomic pseudogene length, and from 76-100% of the EST length (Supplemental Table 2). For example, TA38236_3880 (AL375406 AL375407) matches over 1403 nucleotides, with one mismatch, and neither genomic or EST sequence has
extended open reading frames; each has contains at least three stop codons in the 716 nt aligning region.

**In silico analysis of the promoter regions of the NBS-LRR genes**

We identified promoter sequences in 2 kb windows upstream of predicted NBS-LRR genes (Supplemental Table 1). Four regulatory elements implicated in either response to pathogens or plant stress were identified as being over-represented in the 2kb region upstream of NBS-LRRs. The regulatory elements were: WBOX cassettes, associated with the WRKY transcription factors (Dong et al., 2003); CBF and DRE boxes (Sakuma et al., 2006); and the GCC motif associated with the ERF-type transcription factors (Ohme-Takagi et al., 2000).

WBOX elements are the most numerous, averaging 8.6 for the CNL and 8.4 for the TNL subfamilies (Supplemental Table 1 and Figures 2A-C). In contrast, the average numbers of other element types, are 0.68 (CBF), 0.08 (GCC), 0.39 (DRE). Of the predicted NBS-LRR genes, 75% contain between 5-11 predicted WBOXs, with six WBOXs being the most common. The other three identified promoter regions are generally observed only once per upstream region, with just a few cases of multiple boxes predicted. A striking feature of counts of these regulatory elements is that they are distributed quite uniformly across the tree (Figs 2A-C). For example, average numbers of WBOXs calculated per major clade in this phylogeny (CNL1-17 and TNL1-8), the standard deviation is 2.15 on clade-wise averages of 9.0 WBOXs. Similarly, no clades show systematic excesses or deficits of the (less-numerous) CBF, GCC, or DRE box motifs. We see no clear evidence of a correlation between the arrangement of these promoter cassettes (WBOX, CBF, GCC and DRE) and in silico expression via EST counts.

**DISCUSSION**

Many aspects of the NBS-LRR disease resistance gene family have been extensively studied and described in other species. This study of NBS-LRRs in *M. truncatula* confirms many patterns observed in other plant species, but also clarifies some patterns and finds some features that differ at least quantitatively from those seen in other plants.
Analysis of overall localization, predicted domain structure, in silico gene expression, promoter regions and molecular evolution reveal a number of striking features: i) predominantly recently-derived sequences, with most having originated through local duplications; ii) evidence that NBS-LRR clusters, which in many cases dominate multi-megabase regions, have played an important role in genomic remodeling; iii) evidence of ectopic translocations of NBS-LRRs from many clusters to other parts of the genome; iv) surprisingly variable domain arrangements, primarily in the TNL subfamily; v) several novel domain combinations that appear to have originated and proliferated within the legumes; vi) dramatically varying expression patterns, with expression varying both between and within clades; vii) surprising uniformity of promoter regions across the gene family; viii) patterns of pseudogene distributions related to NBS gene distributions, but differing significantly between clusters.

**Genomic organization: clustered, donated singletons, or maintained singletons**

As is the case in other plant genomes, NBS genes predominantly are clustered physically in *M. truncatula*. This is clearly an outcome of the birth and death process that results from tandem duplication or contraction in a cluster. More intriguing, perhaps, are the exceptions. While most clusters are predominantly comprised of closely related genes, most clusters also include distantly related “strangers.” While most NBS are found in clusters, some exist as singletons, that in some cases, have close homologs elsewhere in the genome, but in other cases, appear to have been evolving independently. These exceptions have important implications for evolution of this family (and the genome), because although they are rare, they provide sources of novelty and change in the genome.

The pattern of clustered, related NBS sequences clearly is an outcome of the birth and death process that results from tandem duplication or contraction in a cluster (Michelmore and Meyers, 1998). In most such cases, the NBS genes all appear to be derived from local duplication events more recent than the split with poplar or *Arabidopsis*. Examples include clades CNL-2 and TNL-4. Although this pattern has been observed in other species described to date (Noel et al., 1999; Meyers et al., 2003; Monosi et al., 2004), the clusters on Mt-3 and Mt-6 are both exceptionally large and
relatively recent, highlighting the rapid turnover of most of this gene family. Expanded clusters account for a large proportion of NBS-LRRs in *M. truncatula*. Given a sliding window size of 100 kb, nearly 80% of all *M. truncatula* NBS genes reside in clusters. This compares to 61% of all NBS genes in clusters *Arabidopsis* (Meyers et al., 2003). Nearly 50% of *M. truncatula* NBS-LRRs lie in clusters of five or more, and the largest single cluster, on chromosome Mt-6, contains 14 members on just two BAC clones (AC148154 and AC127020).

Not only do *M. truncatula* NBS-LRRs tend to cluster, but many also lie in superclusters, such as the 82 NBS genes on the upper arm of chromosome 3 and the 57 NBS genes on the lower arm of chromosome 6. Interestingly, Mt-6 also is more transposon-dense than any other chromosome (Cannon et al. 2006 and unpublished data). Such an association between NBS and transposons has been observed before (Graham et al., 2002), but is not yet clear to what extent an association between NBS-LRR clusters and transposons is causal (in either direction) rather than merely associative.

The NBS genes encoded on the Mt-3 and Mt-6 superclusters represent more than 5% of all the genes, NBS and non-NBS, found on the upper arm of Mt-3 and the lower arm of Mt-6. At this scale, NBS superclusters probably played a central role in genomic remodeling during the evolution of these chromosome regions. Superclusters in *M. truncatula* resembles the situation in *Arabidopsis*, where 32 and 43 NBS-LRRs are found on chromosomes At-1 and At-5 respectively (Meyers et al., 1999; Meyers et al., 2003), and in rice, where more than 25% of all NBS genes are located on chromosome 11 (Monosi et al., 2004).

Although most clusters are predominantly composed of similar sequences, many clusters also contain some phylogenetically distant NBS genes. Indeed, 26 of 120 *M. truncatula* clusters include both TNL and CNL members. The presence of heterogeneous NBS clusters in *M. truncatula* resembles the situation in rice (Monosi et al., 2004) and *Arabidopsis*, where 10 of 40 clusters are phylogenetically mixed (Baumgarten et al., 2003; Meyers et al., 2003). At least some NBS clusters in *M. truncatula* are the result of large-scale segmental duplication events, as indicated by shared combinations of phylogenetic clades within multiple physical clusters and by their localization within regions of demonstrated intragenomic synteny (Cannon et al., 2006).
Among the minority of NBS-LRRs that are singletons, some of these are closely related to sequences elsewhere in the genome. Although this is a small proportion of all NBS genes in the genome, these genes may play the role of pioneers, seeding new regions of the genome with NBS-LRRs, and potentially establishing new locations for future clusters. Examples of singletons with related genes elsewhere are the three Mt-1 NBS-LRRs (pink) in clade TNL-4, Figure 2C, nested in a large clade of sequences primarily located on Mt-6 (blue). Approximately seven percent \( \frac{23}{333} = 6.9\% \) of NBS genes in this study have close homologs that appear to have come from clusters elsewhere in the genome.

The last class of NBS-LRRs are singletons which have no close relatives in the genome. Examples include the single-gene clade CNL-17 or the nine low-copy, unclustered genes in CNL-13 - CNL-15 (Figure 2B). In each case, there is a candidate ortholog from Lotus, poplar, or Arabidopsis, and in none of those cases are the orthologs in clusters in those genomes (Supplementary Data, files Mt1NBS_repsEtcLjAtPt3.*). Some NBS genes may remain as stable singletons simply by chance and because they are in stable regions of the genome. That is, singletons will tend to remain singletons. They are, by definition, not in clusters, which are inherently prone to expansion and contraction through unequal crossing over (Cooley et al., 2000; Kuang et al., 2004; Monosi et al., 2004). Alternatively, singletons may experience different kinds of selective pressures as they may be involved in more stable protein complexes or acting in long-term guard functions, similar to the physically and evolutionarily isolated Arabidopsis RPM1 gene (Mackey et al., 2003; Shao et al., 2003; Ashfield et al., 2004).

**Effects of whole-genome duplication on the gene family**

Several studies have described evidence of large-scale genomic duplication (possibly a “whole-genome duplication,” WGD) early in the evolution of the legumes (Schlueter et al., 2004; Cannon et al., 2006). For at least some of the “mixed” clades, the genes from multiple chromosomes can be mapped to larger internal genomic duplications (duplication blocks). Few such mappings were evident in comparisons to all M. truncatula duplication blocks because NBS-LRR clusters are intrinsically rapidly-evolving, and may eraze evidence of synteny; synteny within M. truncatula duplication
blocks is generally weak and degraded, suggesting significant gene loss and rearrangement in the genome following this early event; and the genome is not completely sequenced (approximately 60% of the euchromatin in the Mt1.0 draft release). Weak detection of *M. truncatula* duplication blocks may not be surprising, considering that estimates of the timing of the WGD place it quite early, at 85-55 Mya (Schlueter et al., 2004; Lavin et al., 2005). This pattern of high rates of loss of WGD evidence for the NBS-LRR family (relative to other large gene families) also has been described in *Arabidopsis* (Cannon et al., 2004).

**Promoter regions**

In an evaluation of the 2000 bp upstream of the NBS-LRR genes in *M. truncatula*, we found surprising uniformity in the numbers of four over-represented cis-elements. This uniformity was found across all clades examined, in both TNL and CNL subfamilies. At least within clusters, similar regulatory elements might be expected, if regulatory regions duplicate and undergo changes at rates similar to their associated genes. This would be consistent with the finding that tandemly duplicated genes in *Arabidopsis* have higher levels of conservation of *cis*-elements when compared to segmentally duplicated genes (Haberer et al., 2004). The finding that all upstream regions of NBS genes examined had at least one WBOX motif indicates that this motif is important for regulation of most, if not all NBS-LRR-family genes. Counts of WBOX elements are not significantly different between CC and TIR subfamilies, or between most clades (with the largest difference, interestingly, occurring in the most diverse clade in terms of domain composition, TNL-8). This similarity (at least for WBOX elements) does not necessarily imply that all *M. truncatula* NBS genes are under similar regulation in all respects, but does suggest that most have at least some regulatory features in common. In particular, pathogen elicitors and salicylic acid (SA) are rapid inducers of a large number of WRKY genes in various plants (Eulgem et al., 1999; Chen and Chen, 2000). In turn, WBOX motifs have been described upstream in the NPR1 gene (a positive regulator of inducible plant disease resistance) (Yu et al., 2001) and upstream of most *Arabidopsis* PR genes (Chen and Chen, 2002; Li et al., 2004). WRKY genes also activate NBS-LRRs in *Arabidopsis* and grape (Zheng et al., 2006; Marchive et al., 2007;
Zheng et al., 2007). The widespread presence of WBOX motifs upstream of so many of NBS genes, however, indicates that fine regulatory control must be due to less conserved, and presumably more variable factors.

**Domain structures and expression patterns**

*M. truncatula* NBS genes show diverse domain combinations, although almost all of the diversity exists in the TIR subfamily. The only variants in the CNL (apart from variation in LRR repeat number and a possible fusion with an RPW8 homolog) are CN (no LRR) and CNL (the canonical structure). In contrast, there are nine domain arrangements in the TIR subfamily: N, NL, NT, NTNL, TN, TNL, TNLT, TNLTNL, TTNL, TTNL.

The much greater domain diversity in the TNL subfamily compared with CNL might be explained in part by their exon-intron structure. The CNL proteins mostly are encoded by a single exon, unlike TNLs which usually are encoded by multiple exons (Meyers et al., 1999; Bai et al., 2002). NBS proteins lacking an LRR also occur in *Arabidopsis*, have been suggested to play a role as adapter proteins (Meyers et al., 2003). Apparent exon additions or fusions occur in other genomes, including WRKY-related domains and some metallopeptidases in *Arabidopsis* (Meyers et al., 2003), as well as the BED/DUF1544 domain in poplar (Tuskan et al., 2006). Meyers et al. (2002) described unusual domain arrangements in *Arabidopsis* chromosomes 4 and 2, and hypothesized that the TNTNL gene must have been a fusion of a TN and a TNL gene. Poplar also contains instances of the apparent TNLT, TN, TNL, TNLT, NL domain arrangements (Tuskan et al., 2006).

Intriguingly, much of the structural diversity in TNL genes exists in a small number of clusters, suggesting a linkage between physical organization in the genome and the origin of novelty in gene structure. All but one of the unusual TNL domain arrangements (TNLTNL) are found in a single clade (Figure 2B, TNL-8). Most of these sequences fall into two classes: singletons on Mt-8, Mt-5, Mt-7, Mt-3; and a cluster on Mt-4. Several pairs of most-similar sequences in this clade are singletons on Mt-5 and Mt-8, which show the largest amount of internal synteny when the *M. truncatula* genome is compared with itself (Cannon et al., 2006). A similar relationship exists between Mt-8
and Mt-4. Thus, it appears that several of these genes are related through a large-scale duplication, and have been maintained since that time. Although most of the genes in this clade are singletons, one cluster on Mt-4 contains highly diverse domain arrangements. It seems likely that recombination within this cluster has generated these arrangements -- at least several of which are viable genes, as 9 out of 11 have high-stringency EST matches.

In general, expression patterns (at least measured by counts of EST matches) are highly variable, and are not strongly associated with domain structure or sequence similarity. This is especially striking on chromosome Mt-6. Here, there are frequent instances of neighboring NBS genes differing significantly in both expression and structure. For example, a cluster of six TNL genes on Mt-6 (on a single BAC clone, AC126790) differ in EST counts ranging from 0 to 31. These same TNL genes also display four distinct domain combinations and differ in upstream WBOX counts, which range from 0 to 11.

**Pseudogenes**

An examination of pseudogenes supports rapid turnover of genes in this gene family, and identifies some particularly active clusters that have generated both large numbers of diverse new genes and pseudogenes. A relatively restrictive criterion for identifying pseudogenes finds 49, in comparison with the 333 predicted NBS genes reported here. This proportion is similar to that observed in the *Arabidopsis* TN and TIR-X subfamilies, which contain 47 genes and 4 pseudogenes, respectively (Meyers et al., 2002). It is possible that the ratio of pseudogenes to genes is relatively characteristic of a gene family, with families experiencing high rates of turnover also having large numbers of pseudogenes. The number of pseudogenes remaining at any given time would depend on the half-life of a pseudogene. The half-life of pseudogenes is thought to be relatively short, estimated at 8-9 million years in mouse and human (Sakai et al., 2007), and 14 million years in *Drosophila* (Petrov et al., 2000). Assuming similar rates in Mt, the observed *M. truncatula* pseudogenes would all have "died" recently in comparison to the timeframe of the legumes, which originated approximately 65 Mya (Sanderson et al., 2004).

It also is interesting to note that at least some of the pseudogenes may be expressed, and therefore not under neutral selection. Four of the predicted pseudogenes
have near-perfect (99-100% identity) support from ESTs. In one such case, the full 716 nt EST contig length matches the genomic pseudogene, and both contain at least three stop codons. Expressed pseudogenes have been observed to regulate the messenger-RNA stability of the corresponding homologous coding gene (Hirotune et al., 2003). Expressed NBS-LRR pseudogenes have been observed in pine (Liu and Ekramoddollah, 2003) and rice (Monosi et al., 2004).

Some mechanisms of gene turnover are suggested by the distribution of NBS pseudogenes in comparison to predicted NBS genes. Most (91.8%) of pseudogenes are found within 100 kb of predicted NBS genes, suggesting that most turnover occurs within clusters. However, there is clearly a greater rate of turnover in some clusters than others. A large excess of pseudogenes is present on Mt-4, with 15 observed versus 6.0 expected if the 50 pseudogenes were distributed as are the 333 predicted NBS genes. Further, most (10) of the pseudogenes on Mt-4 occur in the cluster that accounts for a large portion of domain diversity in the TNL subfamily (Table 1 and Fig. 2C, bottom). Thus, this large TNL cluster on Mt4 has generated unusual diversity, much of which has apparently been "discarded", but some of which retains and contains highly expressed genes (Supplemental Table 2). Just as the diverse Mt-4 cluster has generated a large share both of diverse genes and pseudogenes, other clusters have fewer pseudogenes than expected - specifically, Mt-3, with 4 observed versus 13.5 expected. These pseudogenes are in the CNL subfamily, and occur in the large CNL cluster on Mt-3. That cluster contains only two domain arrangements (domain classes 1 and 2 in Table 1), and is in this sense more "conservative" than the Mt-4 TIR clusters. A much smaller portion of these genes have clearly become pseudogenes, suggesting a clade expansion in which most genes have been accepted.

Conclusion

The NBS-LRR gene family remains, despite a great deal of work on many fronts, fascinating and surprising. There was little reason to suspect, prior to the sequencing of M. truncatula, that there would be three times as many NBS-LRRs in this genome as in Arabidopsis, or that they would dominate large parts of two chromosomes (Mt-3 and Mt-6). Similarly it was surprising to find such domain novelty, and to find that almost all the domain novelty exists in the TNL subfamily, and most of that within one genomic
cluster. Besides raising more intriguing questions (e.g. precisely how do NBS-LRR translocations occur, and are frequencies different between genomes?), these findings have direct practical agronomic implications. The dramatic pace of birth and death in the family is emphasized by the fact that the large majority of *M. truncatula* NBS-LRRs exist in cluster “nurseries,” and will not have one-to-one correspondences to NBS-LRRs in other species. A striking counterexample exists, however, for a minority of genes, which seem to follow a different, more stable evolutionary trajectory.

**MATERIALS AND METHODS**

**Identification of TNL and CNL sequences and pseudogenes in *M. truncatula***

We used the 1.0 draft genome assembly generated by the Medicago Genome Sequencing Consortium (MGSC; http://medicago.org genome/release1.0), with gene predictions from the International Medicago Genome Annotation Group (IMGAG) (Town, 2006). Candidate genes containing NBS domains were identified using blastp similarity (Altschul et al., 1997) at 1e-20 to the following consensus CNL and TNL consensus sequences from plant extended NBS domains (Cannon et al., 2002; Cannon et al., 2004):

**Consensus of CNLs used for blast search:**
```
erpsestiVGletmkewLldndvugvgyGMGGVGKTTLatqifNdfdvkgehFdrvWVvVSMkefn vekiqqdllEekLgLgdewrkteekaaeinLfqLegLkkfLLvLDDvWekevdldkgypfPdrenGskVvl fTTRsesavcmsgvdmxmevecLppeeAWEelFqkvfentlkspeielakevkkCGLPLALkVlG llacKrtvqEKraievllsslaaafsgmnesslplvLkISYdhnLppelKsCFLYcalfPFDylkelekLeyWia EGfidesegetaedvGyeyyleGvLveeLveedktdntrsetKvKMDVvREmALwiaesegfkeviiVra GylreipnvwntvrRmSmnneielLsdenpklrsLlqnsnh
```

**Consensus of TNLs used for blast search:**
```
RDFddlVGiEaHlekmlsLLcLdsdeVrMVGIwGPaGIKKTITARLFsqLSsFqlsaFmenlrgs yStrpaglDeYsmKLlQeqfLskLnlqDikhHLGVicELRkdqKVLlLlDLDVDdleQldALAK etqWFgpGSRlIIVTEDkpLkAgHlHrYuEvgsPekeALqICrAFqOnPDpGFeeLAreVtk LaGnPLLGLvlrVIGSSLRLGskseWedmLpRttsLDgklekvLrsYDgLhekDqALFhIAcIFN gekvdyVkaLadsnLDVrqGLKvLadKSLIhispldgtsieMHnLLqgLRelVrKQsidpKqRqFL vDaeeLCVDtmdTGTgsVIGlslDteieeelnleekAFegMrNLqFLriykkfrrdkg
```

Pseudogenes were identified using the same consensus CNL and TNL sequences, using a tblastn search (Altschul et al., 1997) against the Mt1.0 nucleotide chromosome assemblies, at E-value 1e-10. Matches were considered to be pseudogenes if tblastn
translations (relative to the consensus query sequences above) contained at least one stop codon.

Candidate NBS-LRR proteins were provisionally assigned to either the CNL or TNL groups on the basis of similarity, then were aligned to a Hidden Markov Model (HMM) calculated from a large collection of TNL and CNL extended NBS domains (Cannon et al., 2002; Cannon et al., 2004).

Consensus from NBS HMM used for whole-family NBS alignment:
GKTTLArVYNkiadhFeakcFlcvrefsvkhxlhklkqlxxxxxkei
kldnvleglsilkrLsgKKvLLVLDDVwneneQLeaLagglfwxxpGSRI
IITTRdkhvLsshgvrxxtYevegLneealeLFckkAFkgxxpvdpe
YeeigkkiVkyecGLPL

**Phylogenetic analyses**

Prior to phylogeny construction, sequences containing fewer than 75% of the HMM match-state residues were retained for subsequent analysis, and indels and poorly-aligning regions were removed by trimming regions outside the HMM match states. Also, although the IMGAG pseudomolecule assembly process removed most overlapping regions, some redundant sequence remains in the 1.0 draft in unfinished BAC clones. Phylogenies were calculated using parsimony and bootstrapped neighbor joining. Parsimony trees were calculated using protpars in the Phylip suite (Felsenstein, 2005). The input sequence order was jumbled five times, and a topology was calculated based on each data order. One most-parsimonious tree was chosen at random to serve as the basis for branch length calculations. Maximum likelihood branch lengths were calculated on the parsimony topologies using TreePuzzle 5.2 (Schmidt et al., 2002). The model of substitution was of Mueller-Vingron (2000), amino acid frequencies were calculated from the input trees, and rate heterogeneity was allowed with four Gamma rate categories. The neighbour joining calculation used the ClustalW implementation (Oliver et al., 2005), without Kimura distance correction, on the cleaned alignment from hmmalign, with 1000 bootstrap replicates.
Domain and motif predictions

Domains were predicted using hmmpfam (Eddy, 2001) comparisons to Pfam v20 (Bateman et al., 2002), with an initial E-value cutoff of 0.1. Predicted NBS-LRR protein sequences were compared to the Pfam v20 HMMs using HMMER 2.3.2. Predictions of motifs were made using MEME and MAST (Bailey and Elkan, 1995).

In silico expression analysis and estimation of NBS-LRR gene number

*Medicago truncatula* EST and cDNA sequences were downloaded from GenBank nucleotide database using query (txid3880[ORGN] AND "biomol mrna"[PROP]) for medicago (txid3880[ORGN] AND "biomol mrna"[PROP]). All EST/cDNA sequences were mapped to Mt1.0 BAC sequences by computer program GMAP (Wu and Watanabe, 2005). The alignments were processed and uploaded into MySQL database using ASIP pipeline (Wang and Brendel, 2006). We required >95% identity and 80% coverage for an EST/cDNA to be mapped. If one EST can be mapped to multiple genome location, only the location with best alignment score will be considered. By this method, we minimized the cross-mapping of ESTs among duplication genes.

To estimate the NBS-LRR gene number in EST collection, 55,182 Medicago Transcript Assemblies and singletons sequences (“TA unigenes”, release 2) were downloaded from http://plantta.tigr.org. Identified *Medicago truncatula* NBS-LRR protein sequences were used as query sequences to search against the TA unigenes by BLAST (Altschul et al., 1997), with an E-value threshold of 1e-15. All TA unigene hits were then searched against *Arabidopsis* proteins by blastx (1e-15). Only those TA unigenes with top match to *Arabidopsis* NBS-LRR genes were considered as candidates for expressed *Medicago truncatula* NBS-LRR genes. These expressed candidates were then searched against Mt1.0 BACs by blastn (1e-15) to find out the portion captured by the current *Medicago truncatula* genome.

Identification and analysis of the promoter regions

For each NBS predicted gene, the 2 kbp upstream regions were selected according to the position of the genes provided by the IMGAG annotation (Medicago Sequencing Resources) on the BAC sequences of *Medicago truncatula*. The extracted sequences were
screened against the PLACE database (Higo et al., 1999). Regulatory elements over-represented in the dataset and known to be involved in regulation during the resistance response and under stressed conditions were selected for further analysis (Jang et al., 2006). Among them, WBOX (sequence TGAC(C/T)), CBF (GTCGAC), DRE ((G/A)CCGAC) and GCC boxes were retained for further analysis.

**Comparisons to internal genomic duplications in Medicago**

For comparison of NBS-LRR gene duplications and large-scale duplications in the Medicago genome pseudomolecule build (Mt1.0; http://www.medicago.org/genome/downloads/Mt1). Syntenic regions were predicted using ncbi blastp self-comparisons at E-value 1e-10, then filtering to consider only the top reciprocal best hit between each chromosome pair, then synteny prediction using DiagHunter (Cannon et al., 2003) using the following parameters: compress_factor 2500, use_orientation, near_main_diag 30, min_diag_len 3, min_diag_qual 30, sensitivity 83. Duplications of NBS-LRR genes were compared against predicted synteny regions using OrthoParaMap (Cannon and Young, 2003) and manual evaluation.

**ACKNOWLEDGEMENTS**

Thanks to Roxanne Denny for lab assistance and to Xiaohong Wang, Jayprakash Vasdewani, and Ethalinda Cannon for bioinformatics assistance.

**LITERATURE CITED**


Belkhadir Y, Nimchuk Z, Hubert DA, Mackey D, Dangl JL (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


Traut TW (1994) The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites. Eur J Biochem 222: 9-19


FIGURES

Figure 1. Distribution of NBS-LRR encoded genes on the eight chromosomes of Medicago truncatula.
Red (TNL) and green (CNL) sequences on M. truncatula 1.0 draft chromosome assemblies. NBS consensus sequences, derived from TNL and CNL alignments, were used as blastp queries, and displayed using CViT-blast (www.medicago.org).
Pseudochromosomes are represented by the succession of BACs (vertical gray thick lines) separated by gaps (black horizontal lines). Not all gaps are visible at this resolution. Unmapped BACs are shown on the right as MtChr0. Detailed figure at much higher resolution, see Supplemental Figure “Fig1_NBS-25_lg.png”.

Figure 2. Phylogenetic tree of NBS domains of NBS-LRR predicted proteins from Medicago truncatula.
(A,B) Tree of CNL predicted proteins; (C) Tree of TNL predicted proteins
These are maximum parsimony trees with maximum likelihood branch lengths (Materials and Methods). Chromosomal origin of each gene is indicated with color and in the third character (Mt1 etc.) of each sequence name. For example, the large orange clades near the top of Figure 2A consists of sequences primarily from chromosome 3. Right-hand digits in each gene name indicate predicted gene order in this version of the draft genome pseudomolecule assembly. For stable public gene names, see Supplemental Table 1. Pink dots indicate approximate coalescence points for NBS sequences from Arabidopsis and/or poplar (Supplemental files Mt1NBS_repsEtcLjAtPt.*). Bootstrap values for important basal clades are indicated in blue. Black arrows indicate pairs of clades that can be mapped to an internal genomic duplication within M. truncatula. For example, in Figure 2A, arrows in the middle of the figure show sequences from M. truncatula
chromosomes 1 and 3 that both come from syntenic duplications blocks from those chromosomes. Columns on the right side show domain configurations (Supplemental Table 1 and Table 1), numbers of predicted LRR units, and predicted regulatory elements (Supplemental Table 1). Note: gene names shown in this figure are intended to simplify analysis, but are specific to this study only. For persistent gene names, please consult Supplemental Table 1.

**TABLES**

**Table 1: Domain combinations and motifs for Medicago NBS-LRR genes.**
This table summarizes counts of classes of predicted Pfam domain combinations, and motifs in alignments of these classes. Probable spurious C-terminal fusions were removed prior to summary ("domains" column in Supplemental Table 1 indicates edited domain predictions). Consensus motifs are shown as identified in Meyers et al. 1999. Pfam domains are indicated in column 1 (N: NBS; T: TIR; L: LRR).

**Supplemental Table 1. Predicted Medicago NBS-LRR genes and associated information.**
All predicted *M. truncatula* NBS-LRR genes described in this paper are listed here. Columns 1 and 2, short name displayed in trees and public IMGAG name for predicted gene; columns 3 and 4, gene order and numbered clades in trees (Figures 2A-C); column 5, closely related genes in which one appears to have translated to a distant location; column 6, predicted Pfam domains (domain abbreviations: L: LRR; N: NBS; T: TIR); column 7, counts of predicted LRR units; column 8, counts of matching ESTs in Genbank (as of April 2006); columns 9-12, counts of predicted regulatory elements within 2000 nt upstream of the NBS-LRR start codon (dashes indicate regions not assayed).

**Supplemental Table 2: Predicted Medicago NBS-LRR pseudogenes.**
This table lists the best predicted translation for 49 predicted NBS pseudogenes, and the corresponding near-identical EST matches (when present). Summary statistics are also shown for distributions of pseudogenes on chromosomal assemblies.
A.