

D. J. Bertioli · S. C. M. Leal-Bertioli · M. B. Lion
V. L. Santos · G. Pappas Jr · S. B. Cannon
P. M. Guimarães

A large scale analysis of resistance gene homologues in *Arachis*

Received: 21 February 2003 / Accepted: 23 June 2003 / Published online: 19 August 2003
© Springer-Verlag 2003

Abstract *Arachis hypogaea* L., commonly known as the peanut or groundnut, is an important and widespread food legume. Because the crop has a narrow genetic base, genetic diversity in *A. hypogaea* is low and it lacks sources of resistance to many pests and diseases. In contrast, wild diploid *Arachis* species are genetically diverse and are rich sources of disease resistance genes. The majority of known plant disease resistance genes encode proteins with a nucleotide binding site domain (NBS). In this study, degenerate PCR primers designed to bind to DNA regions encoding conserved motifs within this domain were used to amplify NBS-encoding regions from *Arachis* spp. The *Arachis* spp. used were *A. hypogaea* var. Tatu and wild species that are known to be sources of disease resistance: *A. cardenasii*, *A. duranensis*, *A. stenosperma* and *A. simpsonii*. A total of 78 complete NBS-encoding regions were isolated, of which 63 had uninterrupted ORFs. Phylogenetic analysis of the *Arachis* NBS sequences derived in this study and other NBS sequences from *Arabidopsis thaliana*, *Medicago trunculata*, *Glycine max*, *Lotus japonicus* and

Phaseolus vulgaris that are available in public databases. This analysis indicates that most *Arachis* NBS sequences fall within legume-specific clades, some of which appear to have undergone extensive copy number expansions in the legumes. In addition, NBS motifs from *A. thaliana* and legumes were characterized. Differences in the TIR and non-TIR motifs were identified. The likely effect of these differences on the amplification of NBS-encoding sequences by PCR is discussed.

Keywords *Arachis* · Resistance genes · Nucleotide-binding site · Legume · Peanut

Communicated by M.-A. Grandbastien

Electronic Supplementary Material Supplementary material is available for this article if you access the article at <http://dx.doi.org/10.1007/s00438-003-0893-4>. A link in the frame on the left on that page takes you directly to the supplementary material.

D. J. Bertioli (✉) · V. L. Santos · G. Pappas Jr
Universidade Católica de Brasília,
Pós Graduação Campus II, SGAN 916,
DF CEP 70.790-160, Brasília, Brazil
E-mail: david@cenargen.embrapa.br
Tel.: +55-61-3405550 Ext 129

S. C. M. Leal-Bertioli · M. B. Lion · P. M. Guimarães
EMBRAPA Recursos Genéticos e Biotecnologia,
Parque Estação Biológica, Final W5 Norte,
CP 02372, DF CEP 70.770-900, Brasília, Brazil

S. B. Cannon
Plant Biology Department,
University of Minnesota,
St. Paul, MN 55108, USA

Introduction

Arachis hypogaea L., commonly known as peanut or groundnut, is an important and widespread food legume with a large tetraploid genome ($1C=1.74 \times 10^9$ bp; Bennet and Smith 1976), compared to that of *Arabidopsis thaliana* ($1C=1.25 \times 10^8$ bp; The *Arabidopsis* Genome Initiative 2000). Although morphologically diverse, genetic diversity in *A. hypogaea* is low and it lacks sources of resistance to many pests and pathogens (Kochert et al. 1996). This is possibly because *A. hypogaea* has its origin in a single allotetraploidization event in a hybrid between two wild diploid species. The resulting plant, containing two distinct genomes, would have been reproductively isolated from its wild relatives. Therefore, all land races of peanuts may be derived from a single plant (Kochert et al. 1991). Wild diploid *Arachis* species, which are native to South America, are, on the other hand, genetically diverse and rich in sources of disease resistance (Halward et al. 1992; Galgano et al. 1997). Resistances from some wild diploid *Arachis* spp. can be transferred to *A. hypogaea* through complex crosses and introgression (Simpson 2001).

In the course of evolution, plants have developed various defense mechanisms to protect themselves against diseases and parasites. These defenses include: pre-existing structural defenses, such as thick cuticle and

leaf hairs; the production of inhibitors, such as phenolic compounds, tannins and lectins; and enzymes such as glucanases and chitinases (Bowles 1990; Nicholson and Hammerschmidt 1992). Another class of defense mechanism involves the specific recognition of pathogens by the host plant. Among the cellular events that characterize this type of resistance are an oxidative burst, cell wall strengthening, induction of defense gene expression, and rapid cell death at the site of infection (Ryals et al 1994; De Wit 1995; Agrios 1997; Heath 2000). Ongoing work in many labs is clarifying the structure and mode of action of the resistance genes (R-genes) involved in this type of response.

A number of R-genes that control disease resistance of this type have been identified in the model plant *A. thaliana* (eg. Simonich and Innes 1995; Gassmann et al. 1999; Deslandes et al. 2002) and in several other plant species (eg. Lawrence et al. 1995; Thomas et al. 1997; Milligan et al. 1998; Wang et al. 1999). R-genes can be divided into families based upon homologous domains in their protein products. Many of these resistance proteins are characterised by the presence of so-called LRR and NB-ARC domains (Meyers et al. 1999). The LRR (leucine rich repeat, Kobe and Deisenhofer 1994) domain appears to be responsible primarily for elicitor recognition (Jones and Jones 1997) or, in an alternative model, act as “guards” of cellular machinery that are susceptible to attack by pathogens (Dangl and Jones 2001). The NB-ARC domain is proposed to act in signal transduction pathways that operate in response to pathogen attack, and is involved in programmed cell death in distantly related organisms. It is named after Nucleotide Binding, human APAF-1, Plant Resistance Genes and Caenorhabditis elegans CED-4 (van der Biezen and Jones 1998; Peso et al. 2000). In plants, the only function so far associated with the NB-ARC is in disease resistance, which is often manifested by a hypersensitive response involving programmed cell death. For plant genes it has become more common to refer to the NB-ARC as the Nucleotide Binding Site domain (NBS). Although the term NBS has been widely used for some time, nucleotide binding by this domain has only recently been demonstrated (Tameling et al. 2002).

The NBSs can be divided into two classes, TIR and non-TIR, on the basis of the presence or absence in the complete protein of an N-terminal region homologous to the *Toll* and Interleukin Receptor-like regions (TIR) (Meyers et al. 1999; Pan et al. 2000; Young 2000). The non-TIR proteins often contain a coiled-coil motif, with a subset of these coding for a leucine zipper structure (LZ). The coiled-coil or LZ domain potentially plays a role in the interaction of R-proteins with molecules downstream in the signal transduction pathway.

The NBS contains a number of amino acid motifs, most notably the P-loop, kinase-2 and GLPL motifs, which are present in both TIR and non-TIR NBSs (Meyers et al. 1999). Using degenerate PCR primers designed from these domain sequences, Leister et al. (1996), Kanazin et al. (1996) and Yu et al. (1996)

developed a targeted technique for isolating R-gene homologues (RGHs, also called by some authors R-gene analogs or RGAs) in potato and soybean. Since these publications, a series of studies have been conducted using the same strategy. Large numbers of relatively diverse NBS-encoding regions are typically amplified when this degenerate PCR approach is used. Cloned NBS-encoding regions have been shown to be genetically linked to known R-genes, or indeed to be fragments of known R-genes themselves (Kanazin et al. 1996; Yu et al. 1996; Aarts et al. 1998; Collins et al. 1998, 1999, 2001; Shen et al. 1998; Hayes and Saghai-Marooof 2000; Donald et al. 2002; Peñuela et al. 2002).

In this study we used degenerate primers designed to bind to regions encoding NBS motifs to isolate RGHs from a number of *Arachis* species known to be sources of resistances to diseases and pests. The use of primers based on the P-loop and GLPL motifs resulted in a bias towards the isolation of TIR-NBS sequences. Therefore, we also employed a targeted PCR approach using a primer based on the RNBS-D motif that is specific to non-TIR-NBSs (Peñuela et al. 2002).

To investigate the basis of the apparent bias in PCRs using P-loop and GLPL primers, we characterized these motifs from all NBS sequences available for *A. thaliana* and legumes. Differences in the P-loop and GLPL motifs of TIR and non-TIR proteins were identified, and an electronic simulation of PCR was used to model the likely effect of these motif differences on the amplification of NBS-encoding sequences.

A phylogenetic study was carried out on the *Arachis* spp. NBS sequences obtained in this study and non-redundant protein sequences of NBSs from *A. thaliana*, *Medicago trunculata*, *Glycine max*, *Lotus japonicus* and *Phaseolus vulgaris*. The evolution of the NBS and the representativeness of legume NBS datasets are discussed.

Materials and methods

Plant material

Arachis spp. seeds were kindly supplied by Dr. José Valls, curator of the Germplasm Bank at EMBRAPA Recursos Genéticos e Biotecnologia. Seeds were germinated as described by Nelson et al. (1989) and plants were maintained in the greenhouse. The species used for this study were all from the taxonomic section *Arachis* (accession numbers in parentheses): *A. cardenasii* (GKP10017), *A. duranensis* (V14167 and K7988), *A. hypogaea* var. Tatu, *A. stenoperma* (V7762 and V10309) and *A. simpponii* (V13710).

PCR amplification, cloning and sequencing

Young leaves were harvested before expansion and frozen immediately in liquid nitrogen. DNAs were extracted by the CTAB method (Rogers and Bendich 1988). Primers P1B-fwd, P3D-rev, P3A-rev, P1A-fwd and P3D-rev for PCR (Table 1) were designed using a protein alignment of the following sequences (database accession numbers are given in parentheses): *L6* rust R-gene from *Linum usitatissimum* (gi 7488901), R-gene *N* against tobacco

Table 1 Primer sequences used for amplification of *Arachis* spp. genomic DNA

Primer	Motif	Motif sequence	Primer sequence ^a
P1a-fwd	P-loop	GM[PG]G[IVS]GKTT	GGIATGCCIGGGIIIIGGIAARACIAC
P1b-fwd	P-loop	GM[PG]G[IVS]GKTT	GGIATGGGGIGGGIIIIGGIAARACIAC
P3a-rev	GLPL	GLPL[TAV][LAV][KND]	AIITYIRIYYIAGIGGYAAICC
P3d-rev	GLPL	GLPL[TAV][LAV][KND]	AIITYIRIIRYYAAIGGIAGICC
LM638	P-loop	GGVGKTT	GGIGGIGTIGGIAAIACIAC
RNBS-D-rev	RNBS-D	CFLYCALFP	GGRAAIARISHRCARTAIIVIRAARC

^aI, inosine. Codes for degenerate positions are: R, A/G; Y, C/T; S, G/C; H, A/C; V, A/C

mosaic virus from *Nicotiana glutinosa* (gi 1086263), gene *NL* 25 from *Solanum tuberosum* mRNA (gi 3947733), gene *RPS5* of *A. thaliana* for resistance to *Pseudomonas syringae* (gi 15221252), R-gene *Mi-1* against nematodes and aphids from *Lycopersicon esculentum* (gi 7489037) and gene *Rpp* 8 of *A. thaliana* (gi 17064876). The sequence of primer RNBS-D-rev was kindly provided by Nevin Young. Primer LM638 was developed by Kanazin et al. (1996).

Six out of the nine possible primer combinations were used: P1B-fwd with P3D-rev, P1B-fwd with P3A-rev, P1A-fwd with P3D-rev, LM638 with P3A-rev, LM638 with P3D-rev, and LM638 with RNBS-D-rev. The positions of the primers are shown in Fig. 1. PCRs were performed in a 25- μ l volume containing each dNTP at 200 μ M, each primer at 1 μ M, 1 U of *Taq* DNA polymerase, 2.5 μ l of 10 \times PCR buffer (Gibco-BRL), and 10–100 ng of genomic DNA. All PCR mixtures were overlaid with mineral oil and incubated on a MJ PT-100 thermocycler (MJ Research). The conditions used for PCR were 92 C for 5 min, 35 cycles of 92 C for 1 min, 48 C for 1 min, and 72 C for 1 min, with a final extension at 72 C for 5 min.

PCR products were visualized on a 1% agarose gel by staining with ethidium bromide. Bands of the appropriate sizes were excised from the gel and purified using a QIAquick gel extraction column (Qiagen, Valencia, Calif.). In some cases, total PCR products were prepared for cloning by size-exclusion spin-chromatography in Sepharose. Purified PCR products were cloned into a plasmid vector using a pGEM-T-Easy cloning kit (Promega, Madison, Wis.). *Escherichia coli* was transformed by electroporation. Bacteria were plated onto LB medium containing ampicillin, X-Gal and IPTG, and recombinant plasmids were chosen by blue/white selection. Selected colonies were grown in LB with ampicillin for the purification of plasmid DNA. Plasmid DNA was purified by alkaline lysis (Sambrook et al. 1989) and sequenced on an Applied Biosystems 377 using either the DYEnamic Terminator Cycle Sequencing Kit (Amersham Pharmacia Biotech) or the ABI BigDye Terminator Sequencing Kit according to the manufacturer's instructions. Selected clones were sequenced in both orientations. To estimate the rate of mutation caused by PCR, the contig containing the largest number of independent clones (95) in the Staden database was selected (see below for information in the Staden database). The total number of bases sequenced in independent clones and the number of bases that differed from the consensus were calculated.

Production of contigs

First, traces representing sequences similar to R-genes were separated from those that did not or were of poor quality. The PERL

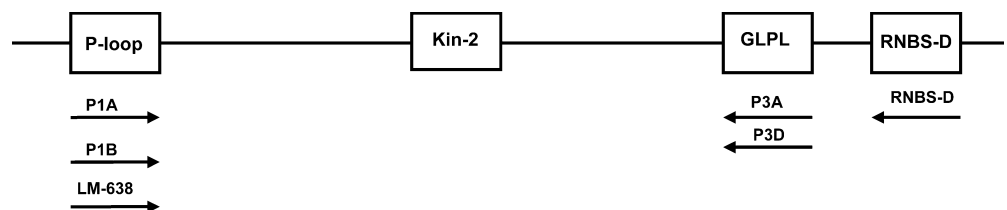
script (<http://www.perl.com/>) phredPhrap (<http://www.phrap.org/>) was used to read the chromatograms and produce a FASTA sequence multifile with vector masked. This was used as the input file for a BLASTX sequence similarity search (Altschul et al. 1997) against a local database consisting of the *A. thaliana* R-genes and their homologues listed at <http://niblrns.ucdavis.edu/> (Meyers et al. 2002). A PERL script was then used to read the BLAST output and move chromatograms with significant e-values into a "hits" folder and the others into a folder "non-hits".

The "hit" chromatograms were entered into a Staden database (Staden 1996). Default parameters were generally used although higher stringency values were used during assembly, with maximum pads per read of 5 and maximum percent mismatch of 1. Data were then edited, contigs split and contigs joined using the "find internal joins" function where necessary. For processing, a reading representing an individual clone was joined to a contig if it had up to three base differences. However, even a single base difference was considered to be sufficient to keep two contigs separate if the base differences were supported by multiple clones. For the addition of new batches of chromatograms to the database the "assemble independently" function allowed the addition of new data without the undesirable joining of very similar contigs in the older data. Old and newly formed contigs were then joined using the "find internal joins" function. Primer sequences were removed manually. Contigs were considered complete when both primers had been removed and if the sequence was of high quality. Contigs were exported and complete ones were translated. Only sequences with contiguous ORFs were used for the major part of the comparative analysis.

Construction of the *G. max*, *M. truncatula*, *P. vulgaris* and *L. japonicus* dataset

Prior to searches of public databases, two hidden Markov models (HMMs) were made using 46 of the *A. thaliana* non-TIR NBS-LRR protein sequences, and 99 of the *A. thaliana* TIR NBS-LRR sequences. Alignments were made using T-Coffee (Notredame et al. 2000), and HMMs were made using HMMER (<http://hmmer.wustl.edu/>; Eddy 1998). The HMMs were used to generate consensus sequences for these two gene subfamilies. We then used the consensus sequences as TBLASTN queries (Altschul et al. 1997) against the following public nucleotide sequence databases: (1) the TIGR June 2002 Gene Indices for *Glycine max* and *G. soja*, *Medicago truncatula* and *Lotus japonicus* (containing CAP4 EST contigs and EST singletons) and (2) the Genbank nucleotide NR database from July 28, 2002. Caveats should be mentioned

Fig. 1 Schematic representation of some NBS motifs (not to scale). The arrows show the positions of the primers used for this work



regarding the use of EST sequences. EST sequences and EST contigs are inherently prone to sequence errors, which may include misassembly (combination of non-identical alleles or separate gene sequences or chimera generation), mis-sequencing, inclusion of clone-artifact chimeras, or partial reads. The *Glycine* Gene Indices predominantly consist of sequences from *G. max* (with approximately 90 EST libraries), but also include five libraries constructed from *G. soja*. Our sequence processing methods should minimize some errors of these types. Specifically, use of translated sequences that are good matches to consensus protein sequences from relatively highly-conserved domains reduces the likelihood of the inclusion of widely divergent chimeras in these regions. Use of protein rather than nucleotide sequences should reduce the inclusion of additional alleles or gene variants that are due to synonymous site changes. Nevertheless, counts from EST sequences and EST contigs should be treated as approximate.

All sequences were checked for redundancy, translation quality and completeness using PERL scripts and also manually. Sequences from the databases that appeared to be incomplete between (but not including) the P-loop and GLPL motifs were removed, as were sequences with stop codons or ambiguities. Regions outside the NBS and redundant sequences were also removed.

The *A. thaliana* dataset

All putative protein sequences from *A. thaliana* (Schoof et al. 2002), retrieved from <http://mips.gsf.de/> were searched using a HMM (Eddy 1998) of the NB-ARC. Proteins with significant matches were used to conduct a search of the Pfam database (<http://Pfam.wustl.edu/>; Bateman et al. 2002).

Sequence analysis and phylogenetic reconstructions

Arachis and *Arabidopsis* sequences were initially aligned using ClustalX (Thompson et al. 1997), with default parameters. For the purpose of constructing the alignments, consensus P-loop GKTT and GLPL sequences were added to the *Arachis* sequences. Alignment files were edited and interpreted using Jalview (<http://www.ebi.ac.uk/~michele/jalview/download.html>). To identify insertion-deletion (indel) regions, we modelled the alignment using a hidden Markov model, and then realigned the sequences to the model. We used HMMER (Eddy 1998) with stringent parameters for alignment "match states" (archpri=0.7 and gap-max=0.3). This assigns all remaining residues in an alignment to "insertion states" or "deletion states," and these indel sites were removed prior to tree construction. All alignments (ClustalX and hmalign alignments with and without indel sites removed) are available in Electronic Supplementary Material.

Sequence Logos were made using <http://weblogo.berkeley.edu/> (Schneider and Stephens 1990).

Phylogenetic trees were constructed using both maximum parsimony and neighbor-joining (NJ) techniques. These gave generally similar tree topologies, and are available in Electronic Supplementary Material. Topologies generated by each of these methods were then used as the basis for computing maximum likelihood branch lengths. Parsimony trees were calculated using the 'protpars' program in the Phylip suite (ver. 3c; available from the author, J. Felsenstein, Department of Genetics, University of Washington, Seattle). This produced a single most-parsimonious tree, which was fed to the Tree-Puzzle program (Schmidt et al. 2002) for calculation of maximum likelihood branch lengths. The model of substitution was that of Adachi and Hasegawa (1996), amino acid frequencies were calculated from the input trees, and rate heterogeneity was allowed with 8 Gamma rate categories.

Results

PCR amplification, cloning and sequencing

Almost all PCRs yielded a major band of the expected size upon agarose gel electrophoresis. This band was, in the case of P-loop and GLPL primers, about 500 bp long, and in the case of LM638 and RNBS-D-rev, about 700 bp in length (Fig. 1, Table 1, Meyers et al. 1999). However, sometimes a number of other, non-specific bands were also amplified. In the former case, direct cloning of PCR products worked well, in the latter, purification of the band of the expected size from the agarose gel gave the best results.

The primers used had a high proportion of inosine bases. When transformed into bacteria, we found that these were converted into guanosines. This had the effect of turning regions of the P1a-fwd and P1b-fwd primers into poly-G tracts. Our tests showed that the DYEnamic Terminator Cycle Sequencing Kit (Amersham Pharmacia Biotech) was better at sequencing through these regions than the ABI BigDye kit. In the contig with the largest number of clones, a total of 37,219 bases were sequenced with 26 "PCR mutations". This gives a PCR mutation rate of 7×10^{-4} errors per base cloned.

Production of contigs

A total of 1333 sequences from more than 1000 independent colonies were generated. Of these, 664 were NBS encoding sequences and 669 non-NBS encoding sequences. The proportions of NBS and non-NBS encoding sequences varied with each cloning; most notably the primer combination LM638 and RNBS-D-rev gave a lower percentage of NBS-encoding sequences. However, this was compensated for by the fact that the sequences isolated using these primers encoded non-TIR type NBSs, which made up only a small proportion of the sequences generated by the other primer combinations. Analyses of the sequences which did not encode NBS sequences showed that a large proportion of them were retroelement sequences.

Processing of the 664 sequence traces in the Staden database gave a total of 127 contigs. The average length of read used was 423 bp and the average number of reading characters per consensus character was 4. This coverage considerably reduces the number of PCR errors that are incorporated into contigs. Of the total 127 contigs, 78 were complete and of high quality between the primer sequences used for amplification. Of these complete sequences, only 63 had contiguous ORFs. Of the other 15 contigs that presented stops in all reading frames, one was represented by only a single clone (S5_A_190P) and may be a PCR mutant, but the other 14 were isolated independently several times, and probably represent fragments of pseudogenes.

Most complete sequences had the expected sizes of about 500 bp or 700 bp. However, a few were anomalous. One sequence, named C8_XY_340, isolated using P-loop and GLPL primer combinations, in two independent PCRs from *A. cardenasii*, was 1018 bp long (Genbank Accession No. AY157783). The GLPL motif sequence was the rare variant GSPL; perhaps because of this, the GLPL-based primer did not bind to this site, but to a 3'-distal site, explaining the larger and unexpected size of this product. Two additional sequences (C8_X_301P and C8_Y_309P), also isolated from *A. cardenasii*, were shorter (126 and 290 bp, respectively), and, since they had multiple stops, were apparently amplified from pseudogenes.

Sequence analysis and phylogenetic reconstructions based on data from *Arachis* spp.

Conceptual translations of the 63 NBS-encoding sequences from *Arachis* spp. gave 61 non-redundant protein sequences. These sequences were very divergent, and clearly formed two groups in phylogenetic analyses (Fig. 2). In the largest group, with 45 sequences, the kinase-2 motifs can be characterized by the absence of a tryptophan residue. This missing tryptophan is typical of TIR-NBS sequences (Meyers et al. 1999). The smaller group had a total of 16 sequences, 14 of which had tryptophan residues in the kinase-2 motif, while two had glycine residues. Searches of Genbank found five *A. hypogaea* NBS sequences with contiguous ORFs. All these sequences were TIR-NBSs, and these were also included in the analyses. For Genbank Accession Nos., see Table 2.

Phylogenetic analysis of sequences from *G. max*, *M. truncatula*, *P. vulgaris* and *L. japonicus*

In total, 37 *G. max*, 43 *M. truncatula*, 21 *P. vulgaris* and 2 *L. japonicus* non-redundant NBS protein sequences that were unambiguous and complete between P-loop and GLPL motifs were obtained from public databases. These sequences grouped clearly into TIR-NBS and non-TIR-NBS like the sequences from *Arachis* spp. These sequences were used in the phylogenetic analysis presented in Fig. 2.

Phylogenetic analysis of sequences from *A. thaliana*

A domain search against the 25458 putative *A. thaliana* proteins with the program HMMer (<http://hmmer.wustl.edu/>; Eddy 1998) yielded 157 significant homologies to the NB-ARC (This compares to 149 reported by The *Arabidopsis* Genome Initiative 2000). These 157 NB-ARC containing proteins were used for a second HMMer search for all Pfam-defined domains. As expected, numerous LRR and TIR domains were also found in the NB-ARC containing proteins. In addition,

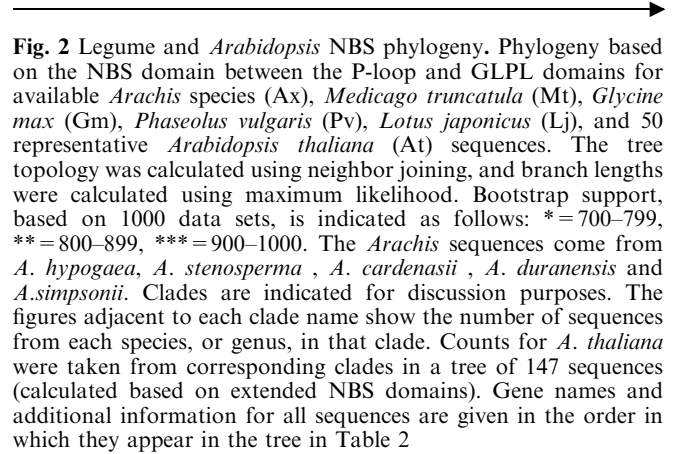


Fig. 2 Legume and *Arabidopsis* NBS phylogeny. Phylogeny based on the NBS domain between the P-loop and GLPL domains for available *Arachis* species (Ax), *Medicago truncatula* (Mt), *Glycine max* (Gm), *Phaseolus vulgaris* (Pv), *Lotus japonicus* (Lj), and 50 representative *Arabidopsis thaliana* (At) sequences. The tree topology was calculated using neighbor joining, and branch lengths were calculated using maximum likelihood. Bootstrap support, based on 1000 data sets, is indicated as follows: *=700–799, **=800–899, ***=900–1000. The *Arachis* sequences come from *A. hypogaea*, *A. stenosperma*, *A. cardenasii*, *A. duranensis* and *A. simpsonii*. Clades are indicated for discussion purposes. The figures adjacent to each clade name show the number of sequences from each species, or genus, in that clade. Counts for *A. thaliana* were taken from corresponding clades in a tree of 147 sequences (calculated based on extended NBS domains). Gene names and additional information for all sequences are given in the order in which they appear in the tree in Table 2

there were significant homologies to other domains: one zinc finger FYVE domain, two WRKY DNA binding domains, and one Mob1 phocein domain, which is a highly conserved domain present in human and yeast that is involved in mitotic checkpoint regulation. These extra domains may reflect as yet uncharacterised diversity in the function of NB-ARC domains in plants.

Multiple alignments using ClustalX (Thompson et al. 1997) of these NB-ARC regions from *A. thaliana* showed that the data were very divergent and formed two main groups and two small outlying groups (data not shown). The largest group, with a total of 101 sequences, consisted almost entirely of TIR-associated NB-ARCs, and the other large group comprised 52 sequences of NB-ARCs not associated with TIR domains. One of the smaller outlying groups consisted of three NB-ARCs not associated with any other domain detected by HMMER and the other group of the single NB-ARC sequence associated with a Mob1 phocein domain.

There were differences in the motifs between TIR-NBSs and non-TIR-NBSs, most notably in the kinase-2 motif. In the non-TIR-type sequences this typically had a characteristic tryptophan which was absent in TIR-NBS sequences (Meyers et al. 1999). The P-loop and GLPL motifs also differed. In almost all non-TIR-NBS sequences, the P-loop and GLPL sequences were GKTT and GLPL, respectively. On the other hand, in the TIR-NBS sequences, deviations from these patterns were common. This greater variability of TIR-NBS motifs compared to non-TIR-NBS motifs contrasts with the greater global variability of non-TIR-NBSs compared to TIR-NBSs (Cannon et al. 2002; also apparent in Fig. 2). To illustrate this difference in motifs, Sequence Logos (<http://weblogo.berkeley.edu/>; Schneider and Stephens 1990) were made from the P-loop, kinase-2 and GLPL motifs of TIR-NBSs and non-TIR-NBSs. For comparison, Sequence Logos were also made for all legume P-loop, kinase-2 and GLPL motifs available in the protein NR database (Fig. 3).

These differences in the P-loop and GLPL motifs between TIR and non-TIR NBSs could cause bias in the amplification of NBS-encoding sequences using redundant primers, resulting in the over- or under-representation of

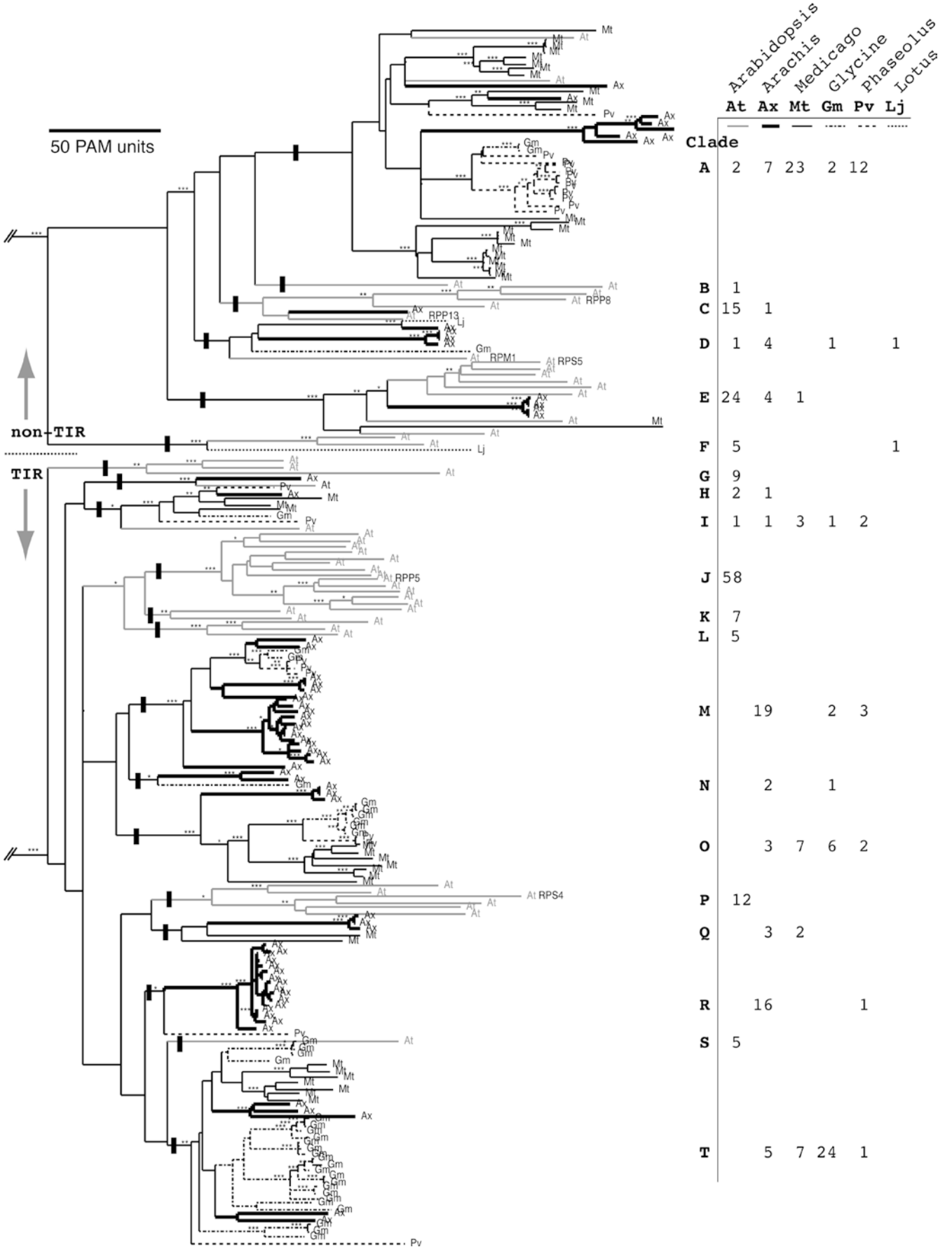


Table 2 Key to sequences in the tree shown in Fig. 2

Clade ^a	Tree ID ^b Internal or other ID ^c	Acc. No. ^d	Clade ^a	Tree ID ^b Internal or other ID ^c	Acc. No.	Clade ^a	Tree ID ^b Internal or other ID ^c	Acc. No.										
A1	Mt	EST610980	BQ165111	D3	Ax	C8_Y_121	Gm	AY157788	M3	Gm	gi1663546	R2	Ax	U55808	R2	Ax	gi21465186	AF520429
A2	At	A13g14460	NP188064	D4	Ax	C8_QY_454	Gm	AY157799	M4	Gm	gi21654931	R3	Ax	AY048864	R3	Ax	S4_A_164	AY157813
A3	Mt	gi12897952	AZ773511	D5	Ax	C8_QY_289	Pv	AY157803	M5	Pv	gi13937097	R4	Ax	AF363803	R4	Ax	M_A_21	AY157769
A4	Mt	gi12897950	AZ773510	D6	Gm	gi21654929	Pv	AY048863	M6	Pv	gi13937093	R5	Ax	AF363801	R5	Ax	TC8_TA_474	AY157801
A5	Mt	gi12897954	AZ773512	D7	At	A13g07040	Pv	NP187360	M7	Pv	gi13937095	R6	Ax	AF363802	R6	Ax	M_A_80	AY157770
A6	Mt	gi12897948	AZ773509	E1	At	A14g10780_RPSS	Ax	NP192816	M8	Ax	S5S1_A_412	R7	Ax	AY157946	R7	Ax	M_A_100	AY157768
A7	Mt	gi12897946	AZ773508	E2	At	A1g12280	Ax	NP172692	M9	Ax	T_A_43	R8	Ax	AY157821	R8	Ax	M_A_48	AY157771
A8	Mt	NF007H08ST	AW688464	E3	At	A15g63020	Ax	NP201107	M10	Ax	T_A_16	R9	Ax	AY157822	R9	Ax	S4_A_26	AY157816
A9	Mt	gi12897968	AZ773519	E4	At	A1g61180	Ax	NP176313	M11	Ax	S1_A_36	R10	Ax	AY157808	R10	Ax	S3S4_A_19	AY157815
A10	At	A13g14470	NP188065	E5	At	A15g47260	Ax	NP199537	M12	Ax	D6_A_424	R11	Ax	AY157807	R11	Ax	S4_A_165	AY157814
A11	Ax	C8_V_253	AY157798	E6	At	A15g47250	Ax	NP199536	M13	Ax	D7_A_12	R12	Ax	AY157767	R12	Ax	D7_A_391	AY157765
A12	Mt	gi12897960	AZ773515	E7	Ax	C8_V_431	Ax	AY157805	M14	Ax	S5_A_366	R13	Ax	AY157939	R13	Ax	T_A_24	AY157825
A13	Ax	C8_XY_340	AY157783	E8	Ax	C8_V_434	Ax	AY157804	M15	Ax	S5_A_195	R14	Ax	AY157490	R14	Ax	TD7_A_112	AY157766
A14	Mt	gi12897933	AZ773502	E9	Ax	C8_Y_265	Ax	AY157790	M16	Ax	C8_X_304	R15	Ax	AY157787	R15	Ax	SIIM7D7_A_23	AY157812
A15	Mt	gi12897931	AZ773501	E10	Ax	C8_V_414	Ax	AY157797	M17	Ax	C8_T_481	R16	Ax	AY157792	R16	Ax	C8_X_269	AY157778
A16	Pv	gi13937087	AF363798	E11	At	A14g26090	Ax	NP194339	M18	Ax	C8_A_212	R17	Pv	AY157786	R17	Pv	gi13937089	AF363799
A17	Ax	C8_V_504	AY157794	E12	Mt	EST487879	Mt	BG586114	M19	Ax	S3S5D7_A_409	S1	At	AY157763	S1	At	At1g72890	NP177432
A18	Ax	C8_Y_282	AY157779	E13	At	A14g27190	Ax	NP194449	M20	Ax	S1C8_A_30	T1	At	AY157773	T1	Gm	gi1663538	U55804
A19	Ax	C8_Y_240	AY157776	F1	At	A15g04720	Ax	NP196092	M21	Ax	S1C8_A_168	T2	Gm	AY157809	T2	Gm	NP332002	AF325685
A20	Ax	C8_V_441	AY157810	F2	At	A14g33300	Ax	NP195056	M22	Ax	T_A_104	T3	Gm	AY157823	T3	Gm	gi13194661	AF325685
A21	Ax	NP332000	AF060192	G1	Lj	AV420577	Ax	AV420577	M23	Ax	T_A_51	T4	Gm	AY157820	T4	Gm	gi13194663	AF325685
A22	Gm	gi13111696	AF060192	G2	At	A15g45050	Ax	NP199318	M24	Ax	C8_XY_263	T5	Mt	AY157780	T5	Mt	gi12897888	AZ773480
A23	Gm	gi13111696	AF060192	G2	At	A14g36140	Ax	NP195337	N1	Ax	S5_A_384	T6	Mt	AY157942	T6	Mt	gi12897886	AZ773479
A24	Pv	gi3493148	AF084026	G3	At	A14g12020	Ax	NP192939	N2	Ax	S4C8_AX_402	T7	Mt	AY157818	T7	Mt	gi12897894	AZ773483
A25	Pv	gi4234951	AF098969	H1	Ax	gi21465184	Gm	AF520427	N3	Gm	gi1663552	T8	Mt	U55811	T8	Mt	EST313460	AW559555
A26	Pv	gi8925788	AF143557	H2	At	A1g27170	At	NP174037	O1	Ax	S1_A_37	T9	Mt	AY157811	T9	Mt	gi11958677	AW696339
A27	Pv	gi14348630	AF306503	I1	Pv	gi8925782	At	AF143553	O2	Ax	S5_A_375	T10	Mt	AY157945	T10	Mt	gi11900724	BF636566
A28	Pv	gi8925786	AF143556	I2	Ax	gi21465188	Ax	AF520431	O3	Ax	C8_A_172	T11	Mt	AY157774	T11	Mt	EST484813	BG583063
A29	Pv	gi3228062	AF072168	I3	Mt	gi13607107	Gm	BG588967	O4	Gm	gi17974505	T12	Ax	AF305392	T12	Ax	C8_X_294	AY157781
A30	Pv	gi14348624	AF306504	I4	Mt	gi20285339	Gm	BQ148280	O5	Gm	gi1663544	T13	Ax	U55807	T13	Ax	S5_A_346	AY157943
A31	Pv	gi14348624	AF306503	I5	Mt	NF041H05NR	Gm	AW686734	O6	Gm	NP418868	T14	Ax	AF305389	T14	Ax	S5_A_399	AY157944
A32	Pv	gi3328060	AF072167	I6	Gm	gi1663548	Gm	U55809	O7	Gm	gi17974500	T15	Gm	AF305389	T15	Gm	gi1663550	U55810
A33	Pv	gi3328060	AF072167	I7	Pv	gi13937083	Gm	AF363796	O8	Gm	gi17978524	T16	Gm	AY008380	T16	Gm	NP423942	AF327903
A34	Pv	gi8925779	AF143551	I8	At	A15g36930	Gm	NP198509	O9	Gm	gi1663542	T17	Gm	U55806	T17	Gm	gi18033110	AF327903
A35	Mt	gi12897958	AZ773514	J1	At	A13g44630	Pv	NP190049	O10	Pv	gi13937099	T18	Gm	AF363804	T18	Gm	gi9965107	AF175398
A36	Mt	gi12897966	AZ773518	J2	At	A12g14080	Pv	NP179024	O11	Pv	gi13937085	T19	Gm	AF363797	T19	Gm	NP233579	AF175388
A37	Mt	gi12897962	AZ773516	J3	At	A15g18350	Mt	NP197336	O12	Mt	gi12897884	T20	Gm	AZ773478	T20	Gm	gi9965102	AF175388
A38	Mt	gi12897941	AZ773506	J4	At	A15g41740	Mt	NP198998	O13	Mt	gi12897880	T21	Gm	AZ773476	T21	Gm	gi1663536	U55803
A39	Mt	gi12897937	AZ773504	J5	At	A14g14370	Mt	NP193173	O14	Mt	gi12897882	T22	Gm	AZ773477	T22	Gm	NP332001	AF325684
A40	Mt	gi12897939	AZ773505	J6	At	A15g18360	Mt	NP197337	O15	Mt	gi12897892	T23	Gm	AZ773482	T23	Gm	gi1663540	U55805
A41	Mt	gi12897920	AZ773496	J7	At	A15g17970	Mt	NP197298	O16	Mt	EST508976	T24	Gm	BG647357	T24	Gm	gi13194659	AF325684
A42	Mt	gi12897922	AZ773500	J8	At	A15g49140	Mt	NP199725	O17	Mt	gi12897890	T25	Gm	AZ773481	T25	Gm	gi9858475	AF175394
A43	Mt	gi12897922	AZ773497	J9	At	A14g16990	Mt	NP193432	O18	Mt	EST333758	T26	Gm	AW774607	T26	Gm	NP232875	AF175394
A44	Mt	NF007H12ST	BG444758	J10	At	A14g16950_RPP5	P1	NP193428	P1	At	A14g19510	T27	Gm	NP193686	T27	Gm	gi17978525	AY008381
A45	Mt	gi12897924	AZ773498	J11	At	A15g51630	P2	NP199976	P2	At	A14g12010	T28	Gm	NP192938	T28	Gm	gi17974498	AF305388
A46	Mt	gi12897964	AZ773517	J12	At	A15g46510	P3	NP199463	P3	At	A15g45250_RPS4	T29	Gm	NP199338	T29	Gm	NP232874	AF175395
B1	At	A13g50950	NP190664	J13	At	A15g22690	P4	NP197661	P4	At	A15g17380	T30	Gm	NP197290	T30	Gm	gi9965104	AF175389
C1	At	At1g59780	NP176187	J14	At	A15g46450	P5	NP199457	P5	At	A15g45230	T31	Gm	NP199336	T31	Gm	gi18033508	AF345652
C2	At	At1g58400	NP176136	K1	At	A15g17680	P6	NP197270	P6	At	A15g44870	T32	Ax	NP199300	T32	Ax	S5_A_378	AY157941

C3	At	A15g43470_RPP8	NP199160	K2	At	At1g72840	NP177427	Q1	Ax	C8_X_305	AY157785	T33	Ax	gi21465189	AF520432
C4	At	At1g50180	NP175437	L1	At	At1g72850	NP177428	Q2	Ax	D6C8S4_AX_400	AY157817	T34	Gm	NP303214	AF322632
C5	Ax	C8_Y_260	AY157782	L2	At	At1g17610	NP173204	Q3	Ax	S5C8_AYX_370	AY157784	T35	Gm	gi12056927	AF322632
C6	At	A15g46530_RPP13	NP190237	L3	At	A15g48780	NP199689	Q4	Mt	gi11905913	BF641755	T36	Gm	gi12056929	AF322633
D1	Lj	AV409895	AV409895	M1	Ax	T_A_44	AY157819	Q5	Mt	EST589704	BM779129	T37	Pv	gi13937091	AF363800
D2	Ax	C8_Y_57	AY157777	M2	Ax	gi21465187	AF520430	R1	Ax	T_A_41	AY157824				

^aSequences are listed in tree order. Clade identification and position

^bThe Tree ID indicating species or genus

^cInternal or other ID (consisting of *Arachis* sequence contig name for sequences identified in this work, GenBank gi, or EST ID). *Arachis* contig names contain information in the three fields separated by underscores. The first field indicates the species from which the sequences were identified: C, *Arachis cardenasii*; D, *A. duranensis*; T, *A. hypogaea* var. Tatu; S, *A. stenoperma*; and M, *A. simpsonii* (the numbers that follow letter codes were generated for laboratory use). The second field gives the primer combination used to amplify the sequences within the contig: A, P1B-fwd with P3D-rev; Y, P1B-fwd with P3A-rev; X, P1A-fwd with P3D-rev; Q, LM638 with P3D-rev; T, LM638 with P3A-rev; and V, LM638 with RNBS-D-rev. The third field contains the contig number generated by the Staden software

^dGenbank ID

certain sub-classes of NBSs. To simulate this, we searched all *A. thaliana* NB-ARC regions for motifs defined as: P-loop [AG]xxxxGK[ST] and GLPL [GLS][LNGF][PR][LI]. From the entire set of 157 NB-ARCs, only 53% (83 sequences) had both these motifs. Out of the 52 Non-TIR-NBSs, 92% (48 sequences) contained these motifs. In contrast, out of a total of 101 TIR-NBS sequences, only 33% had these motifs (33 sequences).

Phylogenetic reconstructions based on all sequences from legumes and *A. thaliana*

The phylogeny in Fig. 2 is a reconstruction of possible evolutionary relationships among NBSs from *A. thaliana*, *Arachis* spp., *M. truncatula*, *G. max*, *P. vulgaris* and *L. japonicus*. In order to keep the number of sequences within reasonable limits, only 50 representative *A. thaliana* sequences are shown. These were chosen based on their phylogenetic positions in a tree of all *A. thaliana* sequences.

For purposes of discussion, clades have been identified. We tried to choose deeply-rooted clades that had both good bootstrap support and relatively long branch lengths, although not all clades are well supported. Phylogenies making use of other taxa and more sites flanking the NBS suggest that some clades are better supported than is indicated in this data set. Clade A, for example, corresponds to the moderately well-supported clade N1.1 in Cannon et al. (2002), and contains the same make-up of two *A. thaliana* sequences and *M. truncatula* sequences (differing, of course, by the addition of *Arachis* sequences).

Of the 20 indicated clades, 14 are in the TIR-NBS subfamily and six are in the non-TIR-NBS subfamily. Of the six non-TIR clades, five contain both legume and *A. thaliana* sequences, and the remaining clade contains a single *A. thaliana* sequence. In contrast, of the 14 TIR-NBS clades, only two contain both legume and *A. thaliana* sequences, while six clades are legume specific and six are *A. thaliana* specific. In general, bootstrap support is lower for clades containing both *A. thaliana* and legume sequences.

The phylogenetic results indicate several clades (e.g. A, O, T) that contain many representatives from all well-sampled legume genera. These may be indications of sequence types that have expanded in the legumes. In contrast, other clades (e.g., E and I) are much more sparsely represented and may contain sequences that have been maintained at low copy number, or perhaps have been lost from some legume genera (though the absences could easily be an artifact of sampling).

Discussion

The genome sequence of *A. thaliana* is now essentially complete and that of *Oryza sativa* is also nearing

completion. These plants have been chosen for large-scale sequencing partly because of their unusually small genomes. Crop plants with large genomes are unlikely to be sequenced in the foreseeable future; therefore one of the challenges for genomics is to use the information available from completely sequenced organisms to further the understanding and manipulation of non-model species.

One way to do this is to identify genes or gene families with particularly important characteristics (candidate genes) in model organisms, and then clone and characterize homologues from non-model organisms. Peanut and its wild relatives are clear examples of plants that will not have their genomes sequenced in the foreseeable future. Therefore, we are following a candidate gene strategy to study disease resistance in *Arachis* spp. We have amplified R-gene NBS homologues from *Arachis* spp., and have used information from model plants and public databases to help provide a context for the *Arachis* sequences.

Sequence comparisons of known R-genes and their homologues have allowed the construction of primers for the P-loop and GLPL motifs, and the amplification of NBS encoding regions. In legumes, primers for P-loop and GLPL motifs seem to amplify many more TIR-NBS- than non-TIR-NBSs encoding regions (Yu et al. 1996). Our work with *Arachis* spp. showed this same bias: using P-loop and GLPL primers, we generated 55 NBS encoding sequences with contiguous ORFs, but only 10 were of the non-TIR type. Since TIR-NBS- and non-TIR-NBS encoding genes generally do not seem to occur together in the same clusters (Richly et al. 2002), candidate gene approaches could be handicapped by this bias in PCR. Therefore, we used a non-TIR-NBS specific primer (RNBS-D-rev; Peñuela et al. 2002) to isolate six extra non-TIR-NBS encoding sequences, providing a better representation of NBS sequences.

We were interested in how the bias towards TIR-NBS encoding sequences could occur in PCR. To study this, we analyzed the complete set of NBS sequences from *A. thaliana*, and showed that P-loop and GLPL motifs are more variable in TIR-NBSs than in non-TIR-NBSs (Fig. 3). From this it follows that primers designed using consensus motifs are likely to bind a smaller proportion of TIR-NBS than non-TIR-NBS sequences. This will lead to an under-representation of TIR-NBS sequences amplified by PCR. Surprisingly, this bias was exactly the opposite of that observed in legumes. To investigate this further, legume motif sequences available from GenBank and from this study were also used to make Sequence Logos (Fig. 3). Although the number of legume P-loop and GLPL motifs identified was small, the characteristics of these motifs in legumes seemed to be generally similar to those from *A. thaliana* (Fig. 3). Therefore, our preferred hypothesis for the observed bias in PCR is that there are more TIR-NBS than non-TIR-NBS encoding sequences in legume genomes, so the difficulties encountered in amplifying non-TIR-NBS

encoding sequences from legumes can be explained by competition in PCR (Bertioli et al. 1995).

In order to investigate the diversity of the NBSs in *Arachis* spp., we made a phylogenetic study of the derived protein sequences. For context, we chose to use NBS datasets from *A. thaliana* (because it is the most complete) and from *M. truncatula*, *G. max*, *P. vulgaris*, and *L. japonicus* because of their importance in legume studies.

For the comparative analysis of sequences, only NBS protein regions with unambiguous sequences that were complete between the P-loop and GLPL motifs were used. Searches of public databases yielded 43 *M. truncatula*, 37 *G. max*, 21 *P. vulgaris* and 2 *L. japonicus* sequences. This compares with the 61 non-redundant sequences obtained from *Arachis* spp. in this study and the 5 *A. hypogaea* sequences in GenBank. Of course, although these counts in no way estimate actual NBS domain numbers in these species, they do crudely indicate sampling levels in this analysis. A search of the set of putative proteins in *A. thaliana* with the HMMer model of the NB-ARC gave 157 significant homologies. The combined set of all legume and *A. thaliana* NBS sequences would have been too large for reasonable phylogenetic tree calculations or visualizations. Therefore, based on preliminary neighbor-joining trees of *A. thaliana* and legume sequences, we chose 50 representative *A. thaliana* sequences for use in the final tree (Fig. 2). These were made up of phenotypically defined *A. thaliana* genes, as well as members of all phylogenetically basal sequences and members of all the remaining more derived clades.

The legume-Arabidopsis evolutionary tree shows several important features

Firstly, most legume sequences fall within legume-specific clades. This is consistent with other studies that have found that many major sequence clades in this gene family are family-specific (Meyers et al. 1999; Pan et al. 2000; Cannon et al. 2002). This implies significant birth and/or death of particular sequence types following the split between Brassicaceae and Fabaceae (Michelmore and Meyers 1998). For example, clades O and T are well represented in *M. truncatula*, *G. max*, *P. vulgaris*, and *Arachis* spp., but have no representatives from *A. thaliana*. Similarly, Clade A contains two *A. thaliana* sequences, but many more legume sequences (44). Because the *A. thaliana* genome has been completely sequenced we can confidently infer that, in these clades, the number of R-gene sequences has expanded in legumes, or that loss of these sequence types has occurred in *A. thaliana*, or both. There seems no reason to believe that the expansion of legume R-gene copy number in, e.g., clade A is due to whole-genome expansions in the legumes, because high copy numbers are not seen uniformly across the gene family, and because there are precedents for dramatic expansion of particular lineages through local gene duplications (Meyers et al. 1999; Michelmore and Meyers 1998; Noel et al. 1999; Peñuela et al. 2002; Richley et al. 2002; Young 2000).



Fig. 3 Sequence logos. Sequence logos (Schneider and Stephens 1990) are depicted for P-loop, kinase-2 and GLPL motifs for non-TIR and TIR NBSs from *A. thaliana* and legumes. The *A. thaliana* sequences used were the 153 that could be separated clearly into TIR-NBSs and non-TIR-NBSs. For legumes, motif sequences from legume NB-ARC sequences in Genbank and the *Arachis* sequences isolated in this study were used. P-loop and GLPL sequences that were probably derived from PCR primers were excluded; hence, for the legumes, the number of sequences contributing to each Logo is variable, as indicated below each Sequence Logo. Although the numbers of sequences contributing to P-loop and GLPL Logos are limited for the legume sequences, the degree of similarity between *A. thaliana* and legume motifs is striking. The lines terminated with 3' indicate the regions usually used for primer design. The arrows indicate amino acid residues which are less conserved in TIR than in non-TIR motifs. This is likely to contribute to the bias in PCR towards the isolation of non-TIR sequences

Secondly, most *A. thaliana* sequences fall within *Arabidopsis*-specific clades. For instance, clade J, which contains the phenotypically defined *RPP5* gene, contains 58 close homologues in *A. thaliana* and none from any of the legume species in this study. This clade has strong bootstrap support and is deeply separated from all other clades. It must be borne in mind, however, that this observation of “no legume sequences in clade J” (or any other clade) depends on the assumption of sufficient and sufficiently unbiased sampling from the legumes.

Thirdly, while some clades have expanded or contracted dramatically within a family, other clades seem to be more stable. Clade I, which contains small numbers of sequences from *A. thaliana* and from the four well-sampled legume species (*Arachis* spp., *M. truncatula*, *G. max* and, to a lesser degree, *P. vulgaris*), may show this type of stability. Similarly, clade D contains one *A. thaliana* sequence (which happens to be the *RPM1* R-gene), and possible orthologues from *G. max* and *L. japonicus*, and four orthologue candidates from *A. cardenasii*.

The largest legume clades are represented by members from all of the well-sampled legumes. It is

interesting to note that *Arachis* spp. sequences are present in more clades (12) identified in Fig. 2 than representatives from *M. truncatula* (6) *G. max* (7) *P. vulgaris* (6) or *L. japonicus* (2). It is to be expected that continued sampling in other legume taxa will also fill out these clades, but the broad representation of *Arachis* sequences throughout the gene family does provide a useful early indication of legume R-gene diversity. This, sequence diversity, together with the more basal placement of *Arachis* spp. sequences as important references for this gene family. As would be expected, in the more sparse clades, fewer genera tend to be represented. This might be due to insufficient sampling and/or indicate that these rarer sequence types were lost from one or more taxa. If any of these smaller clades are found to contain sequence types that are responsible for unusual disease specificities or functions, it might be worthwhile to try to isolate the orthologous sequences from the other agronomically important crop species. For example, it will be interesting to find whether clade D contains *M. truncatula* or *P. vulgaris* orthologs of the sequences from the *Arachis* spp., *G. max*, or *L. japonicus*.

In plants, the only function attributed to NBS sequences so far is in pest and disease resistance. In numerous studies, NBS markers have been shown to be linked to disease resistance. A notable recent example of this is the work done in tomato by Zhang et al. (2002), in which 29 RGHS were mapped near 25 R genes or QTL loci.

Our group has developed a mapping population of F2 plants from a cross between the wild diploids *A. stenosperma* and *A. duranensis*. The parents of this population show contrasting responses to two root-knot nematode species: *Meloidogyne arenaria* and a population of *M. javanica* isolated from the forage peanut *A. pintoi*; and two leaf-spot agents, *Cercosporidium personatum* and *Cercospora arachidicola*. We intend to

place the RGHS isolated in this work, together with disease resistances, on a diploid map of *Arachis* spp.

Acknowledgments The authors would like to thank Dr. José F. M. Valls for providing *Arachis* spp. seeds and information about the species, and Juliana G.O. Dias and Divino L. Miguel for assistance in the greenhouse. Thanks are also due to Wellington Martins and Felipe R. Silva for bioinformatics support, and to Nevin Young, Nicholas Collins and anonymous referees for helpful comments.

References

- Aarts MG, te Lintel Hekkert B, Holub EB, Beynon JL, Stiekema WJ, Pereira A (1998) Identification of R-gene homologous DNA fragments genetically linked to disease resistance loci in *Arabidopsis thaliana*. *Mol Plant-Microbe Interact* 11:251–8
- Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol* 42:459–468
- Agrios GN (1997) *Plant pathology*. Academic Press, San Diego, pp 93–114
- Altschul SF, Madden TL, Schaffer AA, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs *Nucleic Acids Res* 25:3389–3402
- Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL (2002) The Pfam protein families database. *Nucleic Acids Res* 30:276–280
- Bennet MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. *Phil Trans Royal Soc London B* 274:227–274
- Bertioli DJ, Schlichter UHA, Adams MJ, Burrows PR, Steinbiss H-H, Antoniw JF (1995) An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Res* 23:4520–4523
- Bowles DJ (1990) Defense-related proteins in higher plants. *Annu Rev Biochem* 59:873–907
- Cannon SB, Zhu H, Baumgarten AM, Spangler R, May G, Cook DR, Young ND (2002) Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. *J Mol Evol* 54:548–562
- Collins NC, Webb CA, Seah S, Ellis JG, Hulbert SH, Pryor A (1998) The isolation and mapping of disease resistance gene analogs in maize. *Mol Plant-Microbe Interact* 11:968–978
- Collins N, Drake J, Ayliffe M, Sun Q, Ellis J, Hulbert S, Pryor T (1999) Molecular characterization of the maize *Rp1-D* rust resistance haplotype and its mutants. *Plant Cell* 11:1365–76
- Collins N, Park R, Spielmeier W, Ellis J, Pryor AJ (2001) Resistance gene analogs in barley and their relationship to rust resistance genes. *Genome* 44:375–381
- Dangl JL, Jones JDG (2001) Plant pathogens and integrated defence responses to infection. *Nature* 411:826–833
- Deslandes L, Olivier J, Theulieres F, Hirsch J, Feng DX, Bittner-Eddy P, Beynon J, Marco Y (2002) Resistance to *Ralstonia solanacearum* in *Arabidopsis thaliana* is conferred by the recessive *RRS1-R* gene, a member of a novel family of resistance genes. *Proc Natl Acad Sci* 99:2404–2409
- De Wit PLGM (1995) Fungal avirulence genes and plant resistance genes: unraveling the molecular basis of gene-for-gene interactions. *Adv Bot Res* 21:147–185
- Donald TM, Pellerone F, Adam-Blondon AF, Bouquet A (2002) Identification of resistance gene analogs linked to a powdery mildew resistance locus in grapevine. *Theor Appl Genet* 104:610–618
- Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763
- Galgaro L, Lopes CR, Gimenes M, Valls JFM, Kochert G (1997) Genetic variation between species of sections *Extranervosae*, *Caulorrhizae*, *Heteranthae*, and *Triseminatae* (genus *Arachis*) estimated by DNA polymorphism. *Genome* 41:445–454
- Gassmann W, Hinsch ME, Staskawicz BJ (1999) The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant J* 20:265–277
- Halward T, Stalker T, LaRue E, Kochert G (1992) Use of single-primer DNA amplifications in genetic studies of peanut (*Arachis hypogaea* L.). *Plant Mol Biol* 18:315–325
- Hayes AJ, Saghai-Marooof MA (2000) Targeted resistance gene mapping in soybean using modified AFLPs. *Theor Appl Genet* 100:1279–1283
- Heath MC (2000) Hypersensitive response-related death. *Plant Mol Biol* 44:321–34
- Jones DA, Jones JDG (1997) The roles of leucine-rich repeat proteins in plant defences. *Adv Bot Res* 24:89–167
- Kanazin V, Marek LF, Shoemaker RC (1996) Resistance gene analogs are conserved and clustered in soybean. *Proc Natl Acad Sci USA* 93:11746–11750
- Kobe B, Deisenhofer J (1994) The leucine-rich repeat: a versatile binding motif. *Trends Biochem Sci* 19:415–420
- Kochert G, Halward T, Branch WD, Simpson CE (1991) RFLP variability in peanut (*Arachis hypogaea* L.) cultivars and wild species. *Theor Appl Genet* 81:563–570
- Kochert G, Halward T, Stalker HT (1996) Genetic variation in peanut and its implications in plant breeding. In: Pickersgill B, Lock JM (eds) *Legumes of economic importance (Advances in legume systematics, part 8)*. Royal Botanic Gardens Kew, London
- Lawrence GJ, Finnegan EJ, Ayliffe MA, Ellis JG (1995) The L6 gene for flax rust resistance is related to the *Arabidopsis* bacterial resistance gene *RPS2* and the tobacco viral resistance gene *N*. *Plant Cell* 7:1195–1206
- Leister D, Ballvora A, Salamini F, Gebhardt C (1996) A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants. *Nat Genet* 14:421–429
- Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J* 20:317–332
- Meyers BC, Morgante M, 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
- Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res* 11:1113–30
- Milligan SB, Bodeau J, Yaghoobi J, Kaloshian I, Zabel P, Williamson VM (1998) The root knot nematode resistance gene *Mi* from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. *Plant Cell* 10:1307–1319
- Nelson SC, Simpson CE, Starr JL (1989) Resistance to *Meloidogyne arenaria* in *Arachis* spp. germplasm. *J Nematol* 21:654–660 (Suppl S)
- Nicholson RL, Hammerschmidt R (1992) Phenolic compounds and their role in disease resistance. *Annu Rev Phytopathol* 30:369–389
- Noel L, Moores TL, van der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JDG (1999) Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. *Plant Cell* 11:2099–2111
- Notredame C, Higgins D, Heringa J (2000) T-Coffee: a novel method for multiple sequence alignments. *J Mol Biol* 30:205–217
- Pan Q, Wendel J, Fluhr R (2000) Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J Mol Evol* 50:203–213
- Peñuela S, Danesh D, Young ND (2002) Targeted isolation, sequence analysis, and physical mapping of non-TIR NBS-LRR genes in soybean. *Theor Appl Genet* 104:261–272

- Peso L, Gonzalez VM, Inoharat N, Ellis RE, Nunez G (2000) Disruption of the CED-9.CED-4 complex by EGL-1 is a critical step for programmed cell death in *Caenorhabditis elegans*. *J Biol Chem* 275:27205–27211
- Richly E, Kurth J, Leister D (2002) Mode of amplification and reorganisation of resistance genes during recent *Arabidopsis thaliana* evolution. *Mol Biol Evol* 19:76–84
- Rogers SO, Bendich AJ (1988) Extraction of DNA from plant tissues. In: Gelvin S, Schilperoot RA (eds). *Plant molecular biology manual*. Kluwer Academic Press, Boston, pp A6:1-10
- Ryals J, Ukness S, Ward E (1994) Systemic acquired resistance. *Plant Physiol* 104:1109–1112
- Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: a laboratory manual* (2nd edn). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504
- Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18:6097–6100
- Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KF (2002) MIPS *Arabidopsis thaliana* Database (MAtdB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res* 30: 91–93
- Shen KA, Meyers BC, Islam-Faridi MN, Chin DB, Stelly DM, Michelmore RW (1998) Resistance gene candidates identified by PCR with degenerate oligonucleotide primers map to clusters of resistance genes in lettuce. *Mol Plant-Microbe Interact* 11:815–23
- Simonich MT, Innes RW (1995) A disease resistance gene in *Arabidopsis* with specificity for the *avrPph3* gene of *Pseudomonas syringae* pv. *phaseolicola*. *Mol Plant-Microbe Interact* 8:637–640
- Simpson CE (2001) Use of wild *Arachis* species/introgression of genes into *A. hypogaea* L. *Peanut Sci* 28:114–116
- Staden R (1996) The Staden sequence analysis package. *Mol Biotechnol* 5:233–241
- Tameling WI, Elzinga SD, Darmin PS, Vossen JH, Takken FL, Haring MA, Cornelissen BJ (2002) The tomato R gene products I-2 and MI-1 are functional ATP binding proteins with ATPase activity. *Plant Cell* 14:2929–2939
- The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Thomas CM, Jones DA, Parniske M, Harrison K, Balint-Kurti PJ, Hatzixanthis K, Jones JD (1997) Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognitional specificity in *Cf-4* and *Cf-9*. *Plant Cell* 9:2209–2224
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 24:4876–4882
- Van der Biezen EA, Jones JDG (1998) The NB-ARC domain: a novel signaling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr Biol* 8:R226–R227
- Wang ZX, Yano M, Yamanouchi U, Iwamoto M, Monna L, Hayasaka H, Katayose Y, Sasaki T (1999) The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. *Plant J* 19:55–64
- Young ND (2000) The genetic architecture of resistance. *Curr Opin Plant Biol* 3:285–290
- Yu YG, Buss GR, Maroof MA (1996) Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site. *Proc Natl Acad Sci USA* 93:11751–11756
- Zhang LP, Khan A, Niño-Liu D, Foolad MR (2002) A molecular linkage map of tomato displaying chromosomal locations of resistance gene analogs based on a *Lycopersicon esculentum* x *Lycopersicon hirsutum* cross. *Genome* 45:133–146