Theor Appl Genet (2002) 105:50–57
DOI 10.1007/s00122-001-0846-4

M.A. Graham · L.F. Marek · R.C. Shoemaker

PCR Sampling of disease resistance-like sequences from a disease resistance gene cluster in soybean

Received: 5 May 2001 / Accepted: 23 August 2001 / Published online: 25 May 2002

Abstract Clusters of Resistance-like genes (RLGs) have been identified from a variety of plant species. In soybean, RLG-specific primers and BAC-fingerprinting were used to develop a contig of overlapping BACs for a cluster of RLGs on soybean linkage group J. The resistance genes *Rps2* (*Phytophthora* stem and root rot) and *Rmd-c* (powdery mildew) and the ineffective nodulation gene *Rj2* were previously mapped to this region of linkage group J. PCR hybridization was used to place two TIR/NBD/LRR cDNAs on overlapping BACs from this contig. Both of the cDNAs were present on BAC 34P7. Fingerprinting of this BAC suggested as many as twelve different RLGs were present. Given the high nucleotide identity shared between cDNAs LM6 and MG13 (>90%), direct sequencing of this region would be difficult. More sequence information was needed about the RLGs within this region before sequencing could be undertaken. By comparing the genomic sequences of cDNAs LM6 and MG13 we identified conserved regions from which oligonucleotide primers specific to BAC 34P7 RLGs could be designed. The nine primer pairs spanned the genomic sequence of LM6 and produced overlapping RLG products upon amplification of BAC 34P7. Amplification products from 12 different RLGs were identified. On average, nucleotide identity between RLG sequences was greater than 95%. Examination of RLG sequences also revealed evidence of additions, deletions and duplications within targeted regions of these genes. Using previously mapped cDNAs we were able to quickly and inexpensively access multiple RLGs within a single specific cluster.

Keywords Disease · Resistance · Cluster soybean

Communicated by M.A. Saghai Maroof

M.A. Graham
Interdepartmental Plant Physiology Major,
Iowa State University, Ames, IA 50011, USA

L.F. Marek
Department of Agronomy, Iowa State Unversity,
Ames, IA 50011, USA

R.C. Shoemaker (✉)
USDA-ARS, Corn Insect and Crop Genetics Research Unit,
Iowa State University, Ames, IA 50011, USA
e-mail: rcsshoe@iastate.edu
Tel.: +1-515-294-6233, Fax: +1-515-294-2299

Introduction

Genes with conserved sequence homology to disease resistance genes can be found throughout plant genomes. Common disease resistance signatures include Toll/Interleukin-1 receptor domains (TIR), nucleotide binding domains (NBD), leucine rich repeats (LRR), coiled-coil domains (CC) and protein kinase domains (PK; Hammond-Kosack and Jones 1997). Resistance-like genes (RLGs), or resistance gene analogs (RGAs), have been identified in a variety of plant species including soybean (Kanazin et al. 1996; Yu et al. 1996), potato (Leister et al. 1996), lettuce (Shen et al. 1998), maize (Collins et al. 1998), common bean (Rivkin et al. 1999), and *Arabidopsis* (Aarts et al. 1998; Speulman et al. 1998). Estimates from The Arabidopsis Genome Initiative (2000) suggest that RLGs with homology to the TIR/NBD/LRR class of R-genes account for 0.4% of the *Arabidopsis* genome. In many cases, RLGs are found clustered in known disease resistance loci. The tomato *Cf-5* locus contains six other RLGs (Dixon et al. 1998). The *Dm3* locus in lettuce is made up of 22 RLGs within a 3.5-Mb region (Meyers et al. 1998). The *Xa21* gene family is composed of seven homologs contained within a 230-kb region (Song et al. 1997).

Within a cluster of RLGs, there may be multiple genes conferring resistance to different isolates of a particular pathogen or to biologically diverse taxa of pathogens. For example, the *Arabidopsis HRT*/*RPP8* cluster confers resistance to a viral pathogen and an oomycete pathogen (Cooley et al. 2000). In potato, the *Gpa2* locus for potato cyst nematode resistance also contains *Rx1*, a gene for potato virus X resistance (Van der Vossen et al. 2000).

During the analysis of candidate R-gene clusters it becomes very important to determine the types of sequence variations that exist within the cluster. These differences reveal much about the evolution and divergence of the candidates. Minor sequence differences between homologs have proven to be very important in determining differences in pathogen specificity. For example, a single amino-acid difference in the leucine-rich domain of the rice blast resistance gene *Pi-ta* results in susceptibility to rice blast (Bryan et al. 2000). In addition, analysis of the *Cf-2/Cf-5* family using three tomato haplotypes suggests that variation in LRR copy number and recombination play a role in generating diversity among R-genes (Dixon et al. 1998). The predicted solvent-exposed regions of the LRR have been found to be hypervariable and correlate with differential pathogen recognition in several studies (Anderson et al. 1997; Parniske et al. 1997; Botella et al. 1998; Dixon et al. 1998; Meyers et al. 1998; Warren et al. 1998; Ellis et al. 1999).

In hard to transform species, identification of an R-gene from a cluster of candidate RLGs remains difficult. Extensive sequence information is required in the region surrounding the candidate genes. High nucleotide identity shared between members of an R-gene cluster and the number of RLGs within a cluster may make accurate sequence assembly difficult. In the *Xa21* R-gene family, the seven homologs share greater than 90% nucleotide identity. The number of similar RLGs in a cluster may dictate the conditions of a sequencing approach.

In soybean, several clusters of NBD RLGs were mapped to soybean linkage group J (LG J; Kanazin et al. 1996). One of the RLG clusters mapped to a region of LG J to which resistance to powdery mildew (*Rmd-c*) and *Phytophthora* stem and root rot (*Rps2*) and an ineffective nodulation gene (*Rj2*) had been mapped. Using RLG-specific primers and BAC fingerprinting, an assembly of overlapping BACs was developed for this region for the cultivar 'Williams 82' [*rps2*, *Rmd* (adult onset), *rj2*; Graham et al. 2000; Marek and Shoemaker 1997]. Fingerprint analyses of the BACs in this region suggested that more than twelve RLGs were present in a 700-kb interval. Two TIR/NBD/LRR cDNAs (LM6 and MG13) were associated with several overlapping BACs within this cluster (Graham et al. 2000).

We have developed a PCR-based approach that takes advantage of the high nucleotide identity among R-genes to evaluate a cluster of RLGs in soybean. By comparing the sequences of the TIR/NBD/LRR cDNAs LM6 and MG13, we identified conserved regions along the length of the cDNAs. Within the conserved regions from the LM6 sequence, a series of oligonucleotide primer pairs was designed to span the length of the cDNAs. PCR was used to amplify overlapping RLG sequences from a 200-kb BAC in the linkage group J cluster. This strategy was used to quickly and efficiently examine the TIR/NBD/LRR RLGs within this region. We estimated the nucleotide identities between genes in the cluster and examined amino-acid substitution rates within this cluster to determine which regions of these genes are under selective pressure. In addition, there are sequence differences between genes consistent with unequal recombination events. Sequencing of the RLG domains from BAC 34P7 was fast, efficient and yielded information on the structure of the RLGs within this cluster without sequencing the entire BAC. This technique can be applied to other R-gene clusters to quickly and inexpensively obtain sequence information.

Materials and methods

Placement of cDNA clones LM6 and MG13 onto BAC 34P7

Previously, two cDNA clones (LM6 and MG13; GenBank accession numbers AF175388 and AF175399) were identified by hybridization using probes for the conserved nucleotide binding domain of R-genes. To correlate cDNA clones with specific BACs, primers were designed from the LRRs of cDNAs LM6 and MG13 and were used to screen the USDA/ISU 'Williams 82' BAC library under high stringency PCR conditions. Positive BACs were used as templates for PCR using the cDNA 3′ LRR primers. The PCR products were sequenced and the sequences were compared to the cDNA sequence. The cDNAs were localized to a group of overlapping BACs on linkage group J (*rps2*, *Rmd*, *rj2*; Graham et al. 2000). BAC-end sequencing confirmed the positions of the cDNAs and was used to obtain genomic sequences of LM6 and MG13. The localization of cDNAs MG13 and LM6 is described in detail in Graham et al. (2000). BAC 34P7 (Gm_ISb001_034_P07; 200 kb) was chosen for the following experiments because it was the largest of the overlapping BACs corresponding to cDNAs LM6 and MG13. cDNAs LM6 and MG13 shared greater than 90% nucleotide identity. MG13, however, has several large deletions relative to LM6 and is missing almost all of the LRR (Graham et al. 2000).

Sequencing of genomic fragments

To determine the genomic sequences of RLGs within BAC 34P7, we designed primers from conserved regions of cDNAs LM6 and MG13. Sequence comparisons between the cDNAs were made us-

Table 1 Overlapping primer pairs designed from cDNA LM6 to amplify overlapping genomic segments (see Table 2) of R-genes from BAC 34P7

Primer name	Primer sequence	Primer size
GS 0L	5′ TCA ACC ATT ATC ATA CTG AAC 3′	21
GS 0R	5′ AGT TCT GAT GAA GTC AGC G 3′	19
GS 1L	5′ TGG CTT TGC ATC AAG TAG C 3′	19
GS 1R	5′ GAA CTG CCA GAG CAA GTG 3′	18
GS 2L	5′ TTG TCC ATA TCA TAG GGA TC 3′	20
GS 2R	5′ CAC GAT TCA AGA CAT CCT C 3′	19
GS 3L	5′ AAT CAG AGT GCT GCT CTT C 3′	19
GS 3R	5′ CTC ACT CAC CGT GTT GTC 3′	18
GS 3BL	5′ CCA GTG ATG AAA TCC AAG AG 3′	20
GS 3BR	5′ AGA TAT GAG GAA ATC CAG AC 3′	20
GS 4L	5′ GGA AGT GCA AGA GAT TAT TG 3′	20
GS 4R	5′ GGA ACT CAA ATG ACG TAA TG 3′	20
GS 4BL	5′ GTC TGG ATT TCT CCA TAT CT 3′	20
GS 4BR	5′ CCT GCA ACC ATA AGC ACT C 3′	19
GS 5L	5′ CCA TCT AAC TTT GAT CCT ATC 3′	21
GS 5R	5′ ACA ATT CCA CAA CTA TCC AG 3′	20
GS 6L	5′ GCT GGA TAG TTG TGG AAT TG 3′	20
GS 6R	5′ GGA AGT CAA GGA TGC ACA G 3′	19

ing AutoAssembler (Applied Biosystems, Foster City, Calif) and Lasergene (DNAstar, Inc., Madison, Wis.) software. Within the conserved regions from the LM6 sequence, a series of nine primer pairs was designed (Oligo 6.0, Molecular Biology Insights, Cascade Colo; Table 1). The primers were targeted across the length of LM6 so that overlapping fragments were produced upon amplification. LM6 sequence was used in primer design because of deletions present in cDNA MG13 (Graham et al. 2000).

The nine primer pairs were used for PCR amplification from BAC 34P7. For each primer pair a 50-µl reaction volume was used and 100 ng of BAC 34P7 DNA was added as template. PCR cycling conditions were 94 °C for 2 min, 35 cycles of 94 °C for 1 min, 52 °C for 30 s, 72 °C for 1 min, followed by 72 ° for 2 min. The amplification products were purified from a 1% low melt gel and cloned using the pGEMT Easy Vector System I (Promega, Madison, Wis.). Plasmid DNA was isolated according to Sambrook et al. (1989) and sequenced at the Iowa State DNA Synthesis and Sequencing Facility. Ten clones from each primer pair were chosen at random for sequencing and an additional 30 clones were sequenced from the control reactions. Automated dideoxy sequencing was carried out on both strands using an ABI 377 Automated Sequencer. Reactions were set up using the Applied Biosystems (Foster City, Calif.) Prism BigDye terminator cycle sequencing kit with Ampli*Taq* DNA polymerase, FS and electrophoresed on an Applied Biosystems Prism 377 DNA sequencer. Unique clones have been given GenBank accession numbers AF403250–AF403298.

Controls

To test for reproducibility, the PCR reaction using the GS 1 primer pair and BAC 34P7 template DNA was repeated. PCR conditions and cloning were performed as described previously and ten clones were chosen at random for sequencing. In repeating the experiment, we expected to find the original sequences identified. This would demonstrate that the PCR sequences accurately represented the genomic sequence and were not artifacts due to polymerase derived errors.

To demonstrate that sequence differences between the PCR generated clones were not due to misannealing or other PCR errors, the following controls were also included. Primer pairs GS 2 and GS 3B (Table 1) were used in control reactions to test for PCR-generated recombinants. From the clones initially amplified and sequenced from primers GS 2 and GS 3B, two unique clones were chosen from each primer pair for use in the control PCR reaction; 100 ng of each of the DNAs was mixed and used as template in a single PCR reaction with the corresponding primer pair. The PCR product was cloned and ten clones were chosen from each of the two control reactions for sequencing. The sequences of the control clones were compared to the sequences of the parent clones to determine if PCR-generated recombinants could form. Templates for the GS 2 reaction were each 463 basepairs (bp) in length and showed 93% nucleotide identity. Templates for the GS 3B reaction differed by a 199-bp direct repeat and seven additional single-bp differences. The clones were 550 and 749 bp in length and, excluding the repeat, were 99% identical. PCR conditions and cloning were performed as above and ten clones from each primer pair were chosen at random for sequencing.

Sequence analysis

Proofreading and vector sequence removal was done using the Sequencher (GeneCodes, Madison, Wis.) program. To eliminate redundant clones, sequences generated from a specific primer pair were aligned in Sequencher and compared. Clones that differed by at least three bases from all other clones generated by the same primer pair were considered unique. The following equation was used to determine the probability of finding a template with k errors (personal communication with Dr. E. C. Luschei, University of Wisconsin, Madison):

$$P(k|n,m,p_e) = \frac{\sum_{i=0}^{m} \binom{m}{i} B(k|n, i \cdot p_e)}{\sum_{k}\sum_{i} \binom{m}{i} B(k|n, i \cdot p_e)}.$$

In this equation $B(k|n, i{\cdot}p_e)$ is the binomial probability distribution function for k errors and n Bernoulli trials with a probability, p_e, of *Taq* DNA Polymerase errors by cycle i. Using an average template size of 600 bases, 30 template doublings or cycles (m) and an average *Taq* error rate of 1×10^{-5} errors/bp (published rates range from 8.9×10^{-5} to 1.1×10^{-4} errors/bp; Barnes 1992; Cariello et al. 1991), the probability of finding a clone with three or more errors introduced by *Taq* is ≤ 0.013%. Calculations were performed using Mathematica 4.0 software (Wolfram Research, Inc., Champaign, Ill.).

The location of introns was predicted based on the sequences of cDNAs LM6 and MG13 (Graham et al. 2000) and using the NetPlantGene intron prediction program (Hebsgaard et al. 1996). Sequence analyses of the 34P7 RLGs were performed using GCG software (Genetics Computer Group, Madison, Wis.). Exon sequences were combined using the Assemble program. Alignment of genomic sequences and open reading frames were generated using the Pileup program. Amino-acid substitution rates of exons were determined using the Diverge program.

Results

To verify that sequence differences between clones were not due to misannealing occurring during PCR, DNA from two distinct clones was mixed and used as template in a single PCR reaction. The PCR products were recloned and ten clones were randomly sequenced to determine if recombinants were formed during the PCR reaction. For primer pair GS 2, six of the clones were identical to one parent clone while the other four were identical to the other parent. For primer pair GS 3B, five clones were identical to one parent, while the other five were identical to the other parent. In each of these cases, no evidence of recombination occurring during PCR was found. Additionally, no sequencing errors were found in the sequences of these clones. This suggests that the few sequence differences seen throughout the 120 clones were due to *Taq* error.

Repeating the PCR reaction and cloning using the GS 1 primers and BAC 34P7 DNA identified four of the five clones identified in the first reaction. No novel clones were identified. If the generation of unique clones were due to misannealing, many of the clones identified in a PCR reaction would be unique and few duplicate copies of a clone would be present in either the same PCR reaction or a duplicate reaction. These results imply that sequence differences between clones are truly the result of differences found in genomic DNA.

Using the primers designed from cDNAs LM6 and MG13 we were able to generate genomic sequences from multiple RLGs within the cluster on BAC 34P7. Based on sequence differences within the overlapping amplified fragments of primer pairs GS3, GS3B and GS4, we determined that the amplification products corresponded to at least twelve different RLGs on BAC 34P7. On average, nucleotide identities from the RLGs within this region are greater than 95% (Table 2).

Table 2 Amino-acid and nucleotide sequence comparisons of BAC 34P7 clones

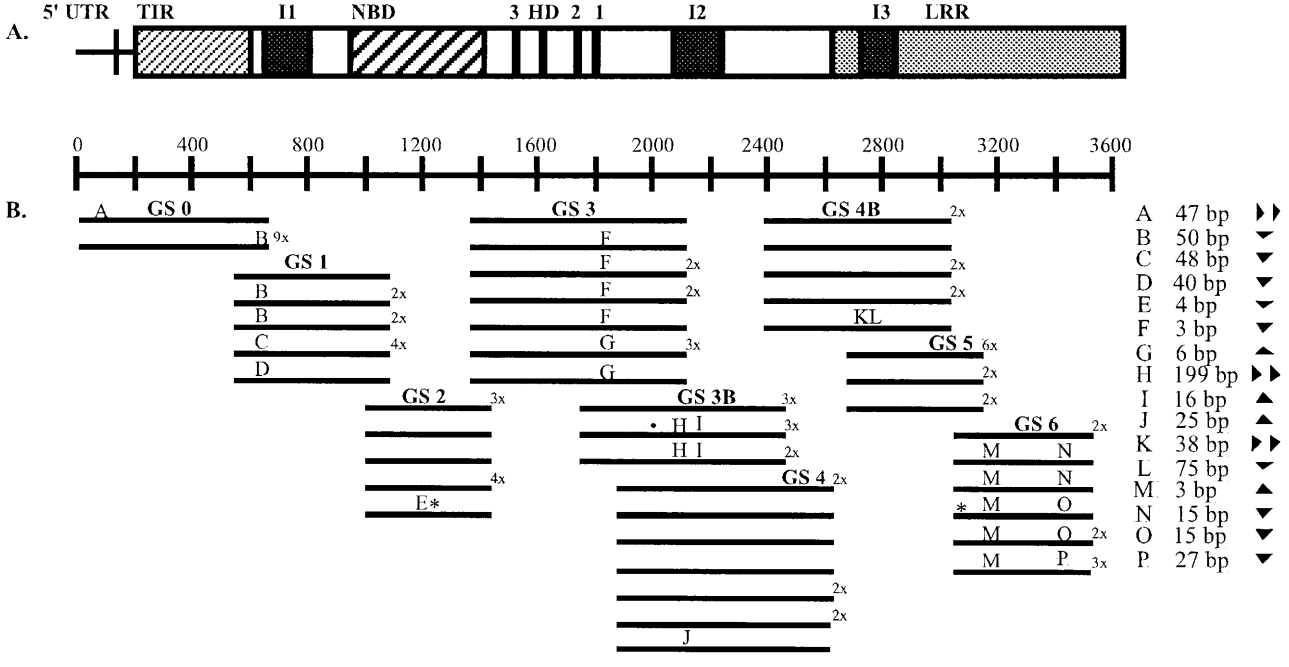
Comparison[a]	BP[b]	K_a/K_s[c]	N.I.[d]
GS 0 Range	1	–	1.00
GS 0 Average	–	–	–
G S1 Range	11–29	0.12–3.70	0.94–0.98
GS 1 Average	20.4	1.59	0.96
GS 2 Range	9–27	0.63–2.30	0.95–0.98
GS 2 Average	19.8	1.39	0.96
GS 3 Range	6–56	0.34–0.89	0.90–0.99
GS 3 Average	37.5	0.68	0.93
GS 3B Range	0–6	0.00–0.37	0.99–1.00
GS 3B Average	3.3	0.37	0.99
GS 4 Range	7–21	0.07–0.69	0.96–0.99
GS 4 Average	13.6	0.30	0.97
GS 4B Range	8–42	0.05–1.08	0.91–0.98
GS 4B Average	28.1	0.56	0.95
GS 5 Range	3–27	1.08–1.16	0.95–0.99
GS 5 Average	18.7	1.12	0.97
GS 6 Range	14–45	0.62–2.48	0.89–0.96
GS 6 Average	33.9	1.39	0.91

[a] All pairwise comparisons within the PCR generated genomic clones for a given primer pair
[b] Number of base pair differences found within a two by two comparison of clones
[c] Ratio of nonsynonymous to synonymous substitutions
[d] Nucleotide identity found in two by two comparisons

The positions of the PCR amplification products relative to the genomic sequence of cDNA clone LM6 are shown in Fig. 1. Each horizontal line represents a unique clone generated from a given primer pair. The top line from primer groups GS 2, GS 3, GS 5 and GS 6 corresponds to the genomic sequence of cDNA LM6. The remaining primer pairs did not amplify a genomic sequence corresponding to LM6. No genomic sequences were obtained for cDNA MG13. Since the end of BAC 34P7 falls within the gene corresponding to MG13 (data not shown) and the gene itself contains deletions within the TIR, NBD and LRR, only primer pair GS1 could have amplified a genomic sequence corresponding to MG13.

For some primer pairs, such as GS 0, only two or three unique clones were identified out of the ten that were sequenced. In contrast, primer pairs GS3 and GS4 amplified seven unique clones each. There are several possible explanations for the low number of unique clones identified by some of the primers. First, the genes could be so divergent in the primer binding site that only a subset could be amplified. Second, the primer binding site could have been deleted. Or third, only a few types

Fig. 1A, B Overlapping genomic segments from RLGs amplified from BAC 34P7. **A** General structure of cDNA LM6. The locations of the 5′ untranslated region (5′ UTR), several introns (*I*), the Toll/Interleukin-1 receptor homology (*TIR*), nucleotide binding domain (*NBD*), conserved domains of unknown function (*3, HD, 2 and 1*) and the leucine rich repeats (*LRR*) are *shaded*. The ruler below indicates the size of LM6 in base pairs. **B** Unique PCR amplified genomic regions corresponding to RLGs amplified from BAC 34P7. The grouped *horizontal lines* represent all unique genomic clones amplified by a specific primer pair. If more than one copy of the same clone was identified, the total number of copies is shown to the right of the line representing the specific clone. The *name* of the primer pair is positioned above each group. The position of the group is shown relative to LM6 above. Size range of PCR products for each primer pair: GS 0 (634–730 bp), GS 1 (413–463 bp), GS 2 (504–508 bp) GS 3 (546–555 bp), GS 3B (527–742 bp), GS 4 (491–507 bp), GS 4B (552–665 bp), GS 5 (585 bp) and GS 6 (397–412 bp). The *letters* above some of the lines indicate structural changes (insertions ▼, deletions ▲ and duplications ►►) found in the clones, relative to LM6. An *asterisk* (*) is used to indicate the location of a frameshift in potential pseudogenes. A *period* is used to indicate the location of an in frame stop codon in a potential pseudogene



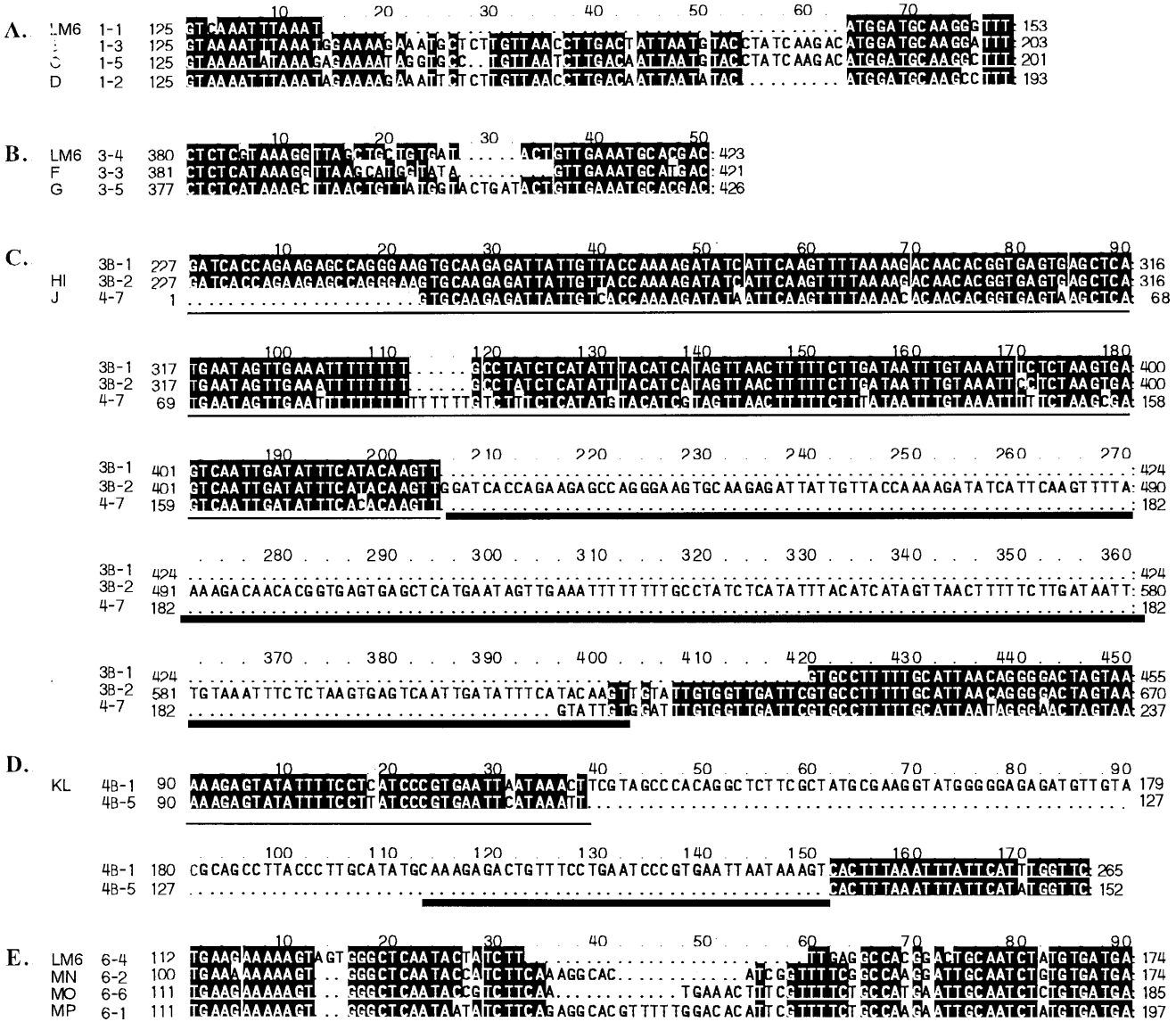


Fig. 2A–E Partial sequence alignments of structural differences in PCR-generated RLG segments from BAC 34P7. The five panels are labeled **A**–**E**. To the right of the panel name is the letter designating the structural change found in Fig. 1. Sequences corresponding to the genomic sequence of LM6 are labeled. Specific fragment names immediately precede the sequences. *Dotted regions* indicate gaps in the sequences introduced to maximize alignment. Panel **A** represents a region within the first intron. Panel **B** corresponds to a portion of the open reading frame following the conserved domains of unknown function. In panels **C and D**, the duplicated region is *underlined with a thin line* and the duplication is underlined with a *thick line*. In panel **C** the intron begins at consensus base 79 and ends at consensus base 441. In panel **D** the intron begins at consensus base 15. Panel **E** corresponds to a portion of the open reading frame in the LRR

of sequences comprise a specific region. All of these scenarios are possible.

While most of the amplified products contained intact open reading frames, primer pairs GS 2, GS 3B and GS 6 each amplified a single clone representing potential pseudogenes (Fig. 1). Frameshifts in the products from primer pairs GS 2 and GS 6 lead to premature stop codons. The pseudogene product from primer pair GS 3B contains an in frame stop codon.

The number of base differences, the ratio of nonsynonymous to synonymous amino acid substitutions (K_a/K_s) and the nucleotide identity of pairwise comparisons of amplification products generated by a given primer pair are shown in Table 2. By determining the K_a to K_s ratios (K_a/K_s), it is possible to determine if a gene has recently evolved under selective pressure. For the Toll/Interleukin-1 receptor homology region of the genes (primer pair GS 0), the K_a/K_s ratio was not determined since only two unique clones were identified. A ratio determined from just two clones would be uninformative. The GS-1 products, which also overlap portions of both the TIR and NBD, had an average K_a/K_s ratio of 1.59. The GS-2 products, which overlap the NBD, had an average K_a/K_s ratio of 1.39. While the GS 1 products overlap the NBD, the GS 2 products completely span it. The GS 3 products, which span several conserved domains

with unknown function (3, HD, 2, 1; Hammond-Kosack and Jones 1997), had an average K_a/K_s value of 0.68. Products from primer pair GS 4, which include an intron, had an average K_a/K_s value of 0.30. GS 4B, GS 5, and GS 6 products, which overlap with the LRR region of the genes, had average K_a/K_s values of 0.56, 1.12 and 1.39 respectively.

In addition to basepair differences, we found insertions, deletions and duplications within the predicted exons and introns of RLG PCR products from BAC 34P7. In Fig. 1, the letters above some of the amplified clones represent structural differences between the clones and the genomic sequence corresponding to cDNA LM6. The same letter indicates regions with the same overall structure. The vertical alignment of the letters demonstrates that the structural differences occur within the same region of multiple genes. These changes are examined in more detail in Fig. 2.

Panel A of Fig. 2 shows a portion of the alignment of some of the clones amplified from GS 1 (structural changes B, C and D located within the first intron). Clone 1-1 represents the genomic sequence corresponding to cDNA LM6. Relative to LM6, clone 1-3 has a 50-bp insertion. Clone 1-5 has an insertion of 48 of those 50 bases and clone 1-2 has 40 of the 50 inserted bases. The alignment of the sequences suggests that the three clones contained the same insertion, and that two of the clones have undergone additional independent deletions.

The second panel (B) in Fig. 2 examines a portion of the coding region following the conserved domains of unknown function. Relative to clone 3–4, representative of LM6, clone 3–3 has a three base pair deletion while clone 3–5 has an overlapping six base-pair addition.

Panel C compares the general structure of two of the clones amplified by primer pair GS 3B and one of the clones amplified by primer pair GS 4. Clone 3B-2 has a perfect 199 base pair repeat relative to 3B-1. The repeat includes parts of the exon and intron. Clone 3B-2 has 16 of the 25 bases inserted in clone 4–7. In panel D, clone 4B-5 has an imperfect 38-bp repeat, split by a 75-bp insertion. This portion of the sequence corresponds to the final intron in the genomic sequence for LM6. A BLASTn nucleotide homology search (Altschul et al. 1997) of the insertion against the GenBank nonredundant database shows highest homology with a soybean carboxyl transferase alpha subunit (GenBank accession number AF164511) with an expected value of 9×10^{-18} (GenBank, July 2001).

In the final panel of Fig. 2, a partial alignment of some of the clones amplified by GS 6 is shown. Clone 6-4 corresponds to the genomic sequence of cDNA LM6. All of the other clones amplified have a 3-bp deletion relative to LM6. In addition, clones 6-2 and 6-6 have overlapping 15-bp insertions and clone 6-1 has a 27-bp insertion. These insertions occur in the part of the open reading frame that codes for leucine-rich repeats.

Discussion

Using a PCR sampling approach we have examined the structures of RLGs within a single BAC from soybean Linkage Group J. We designed nine primer pairs that amplify overlapping segments from multiple putative resistance genes on BAC 34P7. The aim of this study was to determine the best possible approach for subsequent shotgun sequencing projects and to examine sequence differences between RLGs in a single cluster. Given the high nucleotide identity shared between BAC 34P7 RLGs, subclones with larger inserts would be needed to verify differences between potential genes. Dubcovsky et al. (2001) sequenced an approximately 65 kb BAC from rice at 5×, 10×, 15× and 20× redundancy. Even at 20× redundancy, two gaps remained in the sequence and 39 problem areas were identified. Problems included low coverage within certain regions and difficulties in sequence assembly caused by duplications and inversions. To achieve 20× redundancy 300,000 clones were sequenced. In contrast, we were able to identify and compare portions of twelve different R-genes by sequencing only 120 clones. PCR sampling of a gene cluster could also be applied to different disease loci and different classes of disease resistance genes.

An interesting feature of the BAC 34P7 RLGs is the high K_a/K_s ratio found within the TIR and NBD regions. By examining the K_a/K_s ratio it is possible to examine the evolutionary forces acting upon a group of genes. Within the protein coding region, a K_a/K_s ratio greater than one reflects diversifying selection. A ratio less than one suggests that purifying selection is occurring (Parniske et al. 1997). In tomato (Parniske et al. 1997), *Arabidopsis* (Botella et al. 1998) and lettuce (Meyers et al. 1998), K_a/K_s ratios greater than one have been limited to the β-strand/β-turn structural motif of the LRR. In other regions of the genes, including the NBD and TIR, the K_a/K_s ratios tended to be smaller than 1. The results of Botella et al. (1998), Meyers et al. (1998) and Parniske et al. (1997) supported the hypothesis that only the LRR regions of R-genes are involved in determining specificity to pathogens. The NBD and TIR are thought to be effector regions that would undergo purifying selection. However, recent studies in flax demonstrate the importance of the NBD and TIR in changing specificities (Ellis et al. 1999; Luck et al. 2000). Ellis et al. (1999) sequenced thirteen alleles of the *L* locus in flax and demonstrated that two alleles with different pathogen specificities differed only within the TIR. Luck et al. (2000) swapped TIR domains and a portion of the NBD between alleles of the *L* locus while maintaining the original LRR. These changes resulted in novel disease resistance specificities. These data suggest that changes in the TIR and portions of the NBD result in novel disease resistance specificities and therefore these regions may also diverge. The data presented in the current study support the divergence hypothesis.

Two other explanations can also account for the high K_a/K_s ratios detected within the TIR and NBD regions of

RLGs on BAC 34P7. First, exon/intron boundaries were predicted by comparison with cDNAs LM6 and MG13 and by computer algorithm. Including intron sequences in our analyses could effect the K_a/K_s ratios. While amplification products of primer pair GS1 do contain a predicted intron, products of primer pair GS2 do not. Second, including a pseudogene in our analyses could also effect the observed K_a/K_s ratios. Three different primer pairs amplified products containing stop codons or frame shifts in predicted exons suggesting the presence of a pseudogene. These may represent a single gene, or may come from three different genes. In addition cDNA MG13, which is also derived from this cluster, may be a pseudogene (Graham et al. 2000). MG13 contains deletions within the NBD, the conserved domains of unknown function and the LRR. However, no PCR derived genomic sequences with structural similarities to MG13 were detected, so if similar pseudogenes exist in this cluster, they were not included in our comparisons. Since pseudogenes do not encode functional products, no purifying selection takes place and the number of polymorphisms in these genes could be higher. Glusman et al. (2001) used principle component analysis to examine sequence differences in 906 human olfactory receptor (OR) genes. Like R-genes, OR genes are involved in signal recognition; they cluster in the genome and evolve by duplication and divergence. In addition, OR genes are prolific, accounting for as much as 1% of the human genome. Principle component analysis of OR genes revealed that OR pseudogenes are more divergent than functional OR genes. Similar results were obtained by Gilad et al. (2000). Including pseudogenes in the principle component analysis of functional OR genes increased the overall levels of divergence found relative to functional OR genes alone (Glusman et al. 2001). In the case of the BAC 34P7 RLGs, including a single pseudogene in a cluster of nine functional genes would mean that nine of the 45 possible pairwise comparisons would include a pseudogene. By combining RLG pseudogenes with functional genes, we could significantly skew amino-acid substitution ratios.

In addition to single nucleotide differences, we found evidence of insertions, deletions and duplications within this cluster. Within each of the three putative introns we found evidence of clones with insertions. Some of the clones corresponding to the second and third introns also contained relatively large duplications. In the second intron the duplication spanned part of the exon and part of the intron. Small insertions and deletions were also found in the coding regions of the genes. Sequence alignment of these regions revealed that multiple structural changes appear targeted to specific regions. Initial duplications have been followed by insertions and some insertions have been followed by deletions. While the resulting genes maintain a high nucleotide identity, comparison of their structural differences allows us to identify the different events that have occurred within a specific region. Structural changes of this nature are consistent with models of unequal exchange between R-genes. Decreases or increases in LRR number are associated with altered specificity in the *L* and *M* loci in flax (Anderson et al. 1997; Ellis et al. 1999; Luck et al. 2000), the *Cf-5* locus in tomato (Dixon et al. 1998) and the *RPP5* locus in *Arabidopsis* (Noël et al. 1999). Evaluation of the tomato *Cf-4/Cf-9* locus by Parniske et al. (1997) implicated intergenic regions of the clusters in the regulation of recombination rates.

Using a step-by-step amplification technique we have been able to quickly and efficiently generate sequences from many different resistance-like genes from within a cluster. Sequence comparisons of the RLGs on BAC 34P7 have revealed several features. First, the RLGs are closely related and have an average nucleotide identity >95%. Second, the K_a/K_s ratios of the NBD and LRR are greater than one. This suggests that both the NBD and LRR domains are diverging. Third, structural changes in these genes occur within defined regions and are consistent with a model of recombination-driven mutations. The results of this experiment demonstrate that selective amplification of an RLG cluster, rather than complete sequencing of a disease resistance locus, can be used to examine sequence differences in R-genes and the mechanisms which may lead to the development of novel genes.

Acknowledgements Names are necessary to report factually on the available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable. Contribution of the Field Crops Research Unit, USDA-ARS, Midwest Area and Project No. 3236 of the Iowa Agriculture and Home Economics Experiment Station, Ames, IA 50011. Journal paper No. 19212.

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. 11:251–258

Altschul SF, Madden TL, Schaffer AA, Zhang J, 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

Anderson PA, Lawrence GJ, Morrish BC, Ayliffe MA, Finnegan EJ, Ellis JG (1997) Inactivation of the flax rust resistance gene *M* associated with the loss of a repeated unit within the leucine-rich repeat coding region. Plant Cell 9:641–651

Barnes WM (1992) The fidelity of *Taq* polymerase catalyzing PCR is improved by an N-terminal deletion. Gene 112:29–35

Botella MA, Parker JE, Frost LN, Bittner-Eddy PD, Beynon JL, Daniels MJ, Holub EB, Jones JD (1998) Three genes of the Arabadopsis *RPP1* complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. Plant Cell 10:1847–1860

Bryan GT, Wu KS, Farrall L, Jia Y, Hershey HP, McAdams SA, Faulk KN, Donaldson GK, Rarchini R, Valent B (2000) tA single amino-acid difference distinguished resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. Plant Cell 12:2033–2045

Cariello NF, Swenberg JA, Skopek TR (1991) Fidelity of *Thermococcus litoralis* DNA polymerase (Vent) in PCR determined by denaturing gradient gel electrophoresis. Nucleic Acids Res 19:4193–4198

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

Cooley M, Pathirana S, Wu HJ, Kachroo P, Klessig D (2000) Members of the *Arabidopsis HRT/RPP8* family of resistance genes confer resistance to both viral and oomycete pathogens. Plant Cell 12:663–676

Dixon M, Hatzixanthis K, Jones D, Harrison K, Jones J (1998) The tomato *Cf-5* disease resistance gene and six homologs show pronounced allelic variation in leucine-rich repeat copy number. Plant Cell 10:1915–1925.

Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, Bennetzen JL (2001) Comparative sequence analysis of collinear barley and rice bacterial artificial chromosomes. Plant Physiol 125:1342–1353

Ellis JG, Lawrence GJ, Luck JE, Dodds P (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determines differences in gene-for-gene specificity. Plant Cell 11:495–506

Gilad Y, Segré D, Skorecki K, Nachman MW, Lancet D, Sharon D (2000) Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nature Genet 26:221–224

Glusman G, Yanai I, Rubin I, Lancet D (2001) The complete human olfactory subgenome. Genome Res 11:685–702

Graham M, Marek L, Lohnes D, Cregan P, Shoemaker R (2000) Expression and genome organization of resistance gene analogs in soybean. Genome 43:86–93

Hammond-Kosack K, Jones J (1997) Plant disease resistance genes. Annu Rev Plant Physiol Plant Mol Biol 48:575–608

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S (1996) Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. Nucleic Acid Res 24:3439–3452

Kanazin V, Marek LF, Shoemaker RC (1996) Resistance gene analogs are conserved and clustered in soybean. Proc Natl Acad Sci USA 93:11746–11750

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. Nature Genet 14:421–429

Luck J, Lawrence G, Dodds P, Shepher K, Ellis J (2000) Regions outside of the leucine-rich repeats of flax rust resistance proteins play a role in specificity determination. Plant Cell 12: 1367–1377

Marek LF, Shoemaker RC (1997) BAC contig development by fingerprint analysis in soybean. Genome 40:420–427

Meyers B, Shen K, Rohani P, Gaut B, Michelmore R (1998) Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell 10:1833–1846

Noël L, Moores TL, van der Biezen EA, Parniske M, Daniels M, Parker JE, Jones JDG (1999) Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. Plant Cell 11:2099–2111

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones D, Harrison K, Wulff B, Jones J (1997) Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. Cell 91:821–832

Rivkin MI, Vallejos CE, McClean PE (1999) Disease-resistance related sequences in common bean. Genome 42:41–47

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Press, Cold Spring Harbor, New York

Shen KA, Meyers BC, Islam-Faridi MN, Chin D, 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 11:815–823

Song W, Pi L, Wang G, Gardner J, Holsten T, Ronald P (1997) Evolution of the rice *Xa21* disease resistance gene family. Plant Cell 9:1279–1287

Speulman E, Bouchez D, Holub EB, Beynon JL (1998) Disease resistance gene homologs correlate with disease resistance loci of *Arabidopsis thaliana*. Plant J 14:467–474

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Van der Vossen E, van der Voort J, Kanyuka K, Bendahmane A, Sandbrink H, Baulcombe D, Baaker J, Stiekema W, Klein-Lankhorst R (2000) Homologues of a single resistance gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode. Plant J 23:567–576

Warren RF, Henk A, Mowery P, Holub E, Innes RW (1998) A mutation within the leucine-rich repeat domain of the *Arabidopsis* disease resistance gene *RPS5* partially suppresses multiple bacterial and downy mildew resistance genes. Plant Cell 10: 1439–1452

Yu YG, Buss GR, Sagaimaroof M (1996) Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding motif. Proc Natl Acad Sci USA 93:11751–11756