The Plant Cell, Vol. 11, 1833–1846, November 1998, www.plantcell.org © 1998 American Society of Plant Physiologists

Receptor-like Genes in the Major Resistance Locus of Lettuce Are Subject to Divergent Selection

Blake C. Meyers,[a,1] Katherine A. Shen,[a] Pejman Rohani,[a] Brandon S. Gaut,[b] and Richard W. Michelmore[a,2]

[a]Department of Vegetable Crops, University of California, Davis, California 95616
[b]Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525

Disease resistance genes in plants are often found in complex multigene families. The largest known cluster of disease resistance specificities in lettuce contains the *RGC2* family of genes. We compared the sequences of nine full-length genomic copies of *RGC2* representing the diversity in the cluster to determine the structure of genes within this family and to examine the evolution of its members. The transcribed regions range from at least 7.0 to 13.1 kb, and the cDNAs contain deduced open reading frames of ~5.5 kb. The predicted RGC2 proteins contain a nucleotide binding site and irregular leucine-rich repeats (LRRs) that are characteristic of resistance genes cloned from other species. Unique features of the *RGC2* gene products include a bipartite LRR region with >40 repeats. At least eight members of this family are transcribed. The level of sequence diversity between family members varied in different regions of the gene. The ratio of nonsynonymous (K_a) to synonymous (K_s) nucleotide substitutions was lowest in the region encoding the nucleotide binding site, which is the presumed effector domain of the protein. The LRR-encoding region showed an alternating pattern of conservation and hypervariability. This alternating pattern of variation was also found in all comparisons within families of resistance genes cloned from other species. The K_a/K_s ratios indicate that diversifying selection has resulted in increased variation at these codons. The patterns of variation support the predicted structure of LRR regions with solvent-exposed hypervariable residues that are potentially involved in binding pathogen-derived ligands.

INTRODUCTION

Plant disease resistance is often inherited as single Mendelian resistance genes that determine the reaction to specific pathogen avirulence genes. These genes fall into several mechanistic and structural classes (Michelmore, 1995; Baker et al., 1997). Genes encoding similar amino acid motifs are found in diverse plant species and are effective against a wide range of pathogens, including viruses, bacteria, nematodes, and fungi. The most common class of cloned genes encodes proteins containing a nucleotide binding site (NBS) and leucine-rich repeats (LRRs; Staskawicz et al., 1995; Bent, 1996). These domains are often components of signal transduction proteins (Kobe and Deisenhofer, 1994; Traut, 1994), which supports the hypothesis that these genes encode receptors and may act early in a signal transduction pathway (reviewed in Baker et al., 1997).

Genetic studies have determined that resistance genes are often members of complex loci comprised of linked resistance specificities (Pryor, 1987; Crute and Pink, 1996). Molecular data from at least 10 families of resistance genes, including loci from tomato, lettuce, rice, flax, and Arabidopsis, indicate that these loci frequently contain arrays of related genes. Sequencing of the tomato *Cf-4* and *Cf-9* haplotypes, which confer resistance to *Cladosporium fulvum*, demonstrated the presence of five closely related members in each genotype spanning ~35 kb (Parniske et al., 1997). The *Xa21* gene from rice, which confers resistance to the bacterial pathogen *Xanthomonas oryzae* pv *oryzae*, belongs to a multigene family containing at least eight members distributed over ~230 kb (Williams et al., 1996). DNA gel blot analysis indicates that the *M* locus may contain ⩾15 members of a multigene family contained within <1 Mb (Anderson et al., 1997). In Arabidopsis, eight *RPP5* homologs are clustered over 90 kb (Bevan et al., 1998).

Plants are challenged by rapidly evolving pathogen populations and must be able to evolve new resistance specificities to detect virulent variants. However, little is known about the mechanisms that have influenced the evolution of both individual plant resistance genes and the multigene families that contain such genes. High levels of meiotic instability have been detected in some resistance gene clusters, particularly in the *Rp1* complex of maize (Sudupak et al., 1993; Hulbert, 1997). At the *Cf-4*/*9* locus of tomato, pairing between dissimilar haplotypes may increase variation by stimulating unequal intragenic recombination (Parniske et al., 1997). Intragenic recombination has probably resulted in

[1]Current address: DuPont Agricultural Biotechnology, Delaware Technology Park, Newark, DE 19714.
[2]To whom correspondence should be addressed. E-mail rwmichelmore@ucdavis.edu; fax 530-752-9659.

variation in the LRR-encoding region of the *L6* and *M* genes in flax (Ellis et al., 1995; Anderson et al., 1997). The multigene nature of resistance loci may facilitate meiotic instability in a heterozygous state. Published models for the generation of novel resistance gene specificities propose recombination, gene conversion, and unequal crossing over as the primary mechanisms in generating haplotype diversity (Shepherd and Mayo, 1972; Pryor, 1987; Richter et al., 1995; Hammond-Kosack and Jones, 1997).

The major cluster of resistance genes (the *Dm3* locus) of lettuce is the most complex and largest family of plant resistance genes characterized to date. Genetic analysis of different lettuce genotypes has demonstrated >10 resistance specificities at this locus, most of which are *Dm* genes, encoding resistance to lettuce downy mildew (*Bremia lactucae*; Farrara et al., 1987; T. Nakahara and R.W. Michelmore, unpublished data). The *Dm3* haplotype in cultivar Diana contains at least 24 diverse resistance gene candidate (RGC) sequences distributed over ~3.5 Mb (Meyers et al., 1998). Genomic bacterial artificial chromosome (BAC) clones containing 22 members of the *RGC2* gene family (*RGC2A* to *RGC2W*) have been identified and mapped in the region encompassing *Dm3* (Meyers et al., 1998). Limited genomic sequencing of two *RGC2* sequences detected the presence of both NBS and LRR motifs (Shen et al., 1998). The family exhibits a high level of sequence divergence between members in the NBS region (Meyers et al., 1998). Deletion mutant mapping data and the molecular analysis of two additional mutants have identified one member of the *RGC2* family as *Dm3* (Okubara et al., 1997; Meyers et al., 1998; D.B. Chin, R. Arroyo-Garcia, B.C. Meyers, K.A. Shen, and R.W. Michelmore, unpublished data).

In this study, we analyzed the complete sequences of nine of the 24 genes from the *Dm3* cluster, including all members of the subfamily most closely related to *Dm3* and several of the more divergent members of the family. This analysis demonstrated that the genes clustered at the *Dm3* locus are among the largest thus far reported for plants. They have multiple introns, one of which varies greatly in size. The LRR-encoding region seems to be bipartite and contains a polymorphic, compound trinucleotide simple sequence repeat in the open reading frame. Most of the genes studied were transcribed and contained intact open reading frames. Regions of hypervariability were identified in regions encoding amino acids in the LRR that may comprise a solvent-exposed surface. Sequence diversity within these regions may affect ligand binding and therefore contribute to the evolution of novel specificities.

RESULTS

Choice of *RGC2* Copies for Sequencing

A total of nine *RGC2* family members were selected for sequencing; these included the candidate *Dm3* gene, members of a subfamily closely related to *Dm3*, and additional members to sample diversity in the family. Both mutant analysis and mapping data indicate that copy *RGC2B* is *Dm3* (Okubara et al., 1997; Meyers et al., 1998). Genetic complementation with *RGC2B* is currently under way. Three *RGC2* copies, *RGC2C*, *RGC2D*, and *RGC2S*, were selected because they comprise the subfamily most closely related to *RGC2B.* This subfamily shares a set of markers that reside in intron 3 and are lacking in other family members, including the low-copy markers AM14 (Anderson et al., 1996), $IPCR_{800}$, and the microsatellite MSAT15-34 (Okubara et al., 1997; Meyers et al., 1998). In addition, we sequenced five divergent *RGC2* copies to determine the degree and nature of evolutionary changes that have occurred within the *RGC2* family. Prior sequence analysis had revealed that the regions encoding the NBS of *RGC2A*, *RGC2N*, *RGC2J*, *RGC2K*, and *RGC2O* were only 61 to 74% identical to each other and *RGC2B.* Mapping and sequence analysis indicate that these sequences include the range of sequence diversity and physical positions observed within the family (Meyers et al., 1998).

Analysis and Structure of the *RGC2* Genes

The size and genomic structure of *RGC2* genes were determined by analysis of genomic and cDNA sequences. Rapid amplification of cDNA ends (RACE; Frohman et al., 1988) products was obtained using primers in exon 2, 5′ of the region encoding the NBS in *RGC2A* and *RGC2B* (Table 1). The initiation of the cDNA occurred ~700 bp upstream of the first coding exon (Figure 1). However, there is an intron of 59 and 85 bp in the 5′ untranslated region (5′ UTR) in the mature mRNA of *RCG2A* and *RGC2B*, respectively. A 3.4-kb 3′ RACE product was isolated using a primer located at the end of exon 2 (Table 1). This cDNA was identical to the exon sequences of *RGC2C* and included a 283-bp 3′ UTR and a poly(A) tail. RACE using primers designed 5′ to or within the 3′ UTR did not amplify any larger products, indicating that we had identified the 3′ end of the gene.

The genomic structure of the nine *RGC2* genes was similar. All genes, except for *RGC2D* (described below), had

Table 1. Oligonucleotide Primers Used in This Study

Purpose	Primer Name	Sequence (5′ to 3′)
5′ RACE	5RACE3B	CCACATTTGCTCTGATCCCTTC
	5RACE3A	CACACAAGGCTACCATGTGGA
3′ RACE	3RACE2A	CAGACCCCATTGCTATTCAATC
	3RACE3C	GCAAACACTTTGTCAAGACTTGAG
	3RACECDNAF4	CGTCACAAACAACTACCACTAC
	3RACECDNAF5	GTAGAAGAAGACAAACAGAAAGAA
MSATE6	5MSATE6-1	CCCAAGAAGAATCCTACCA
	3EXON4C	AGTGATTGTGAAGAAGGAAGAA

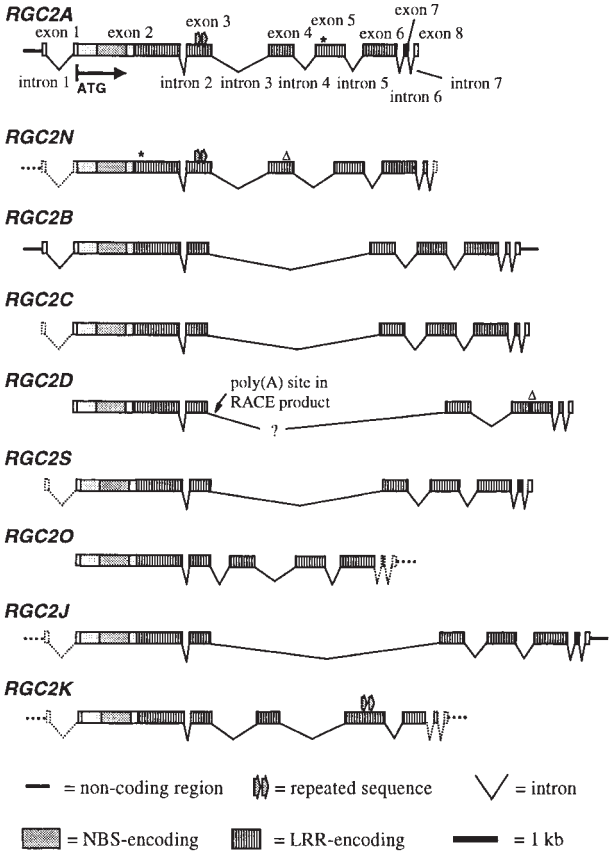


Figure 1. Structure of Nine *RGC2* Genes.

The coding region starts from the ATG, as marked. The 5′ untranslated leader sequence was identified in a subset of the genes by analysis of RACE products (see text). Dotted lines in the 5′ and 3′ regions of some copies indicate that genomic sequence was obtained but the intron–exon boundaries were not determined. *RGC2A* and *RGC2K* contain repeats as indicated (arrowheads). *RGC2A* and *RGC2N* contain stop codons at the positions indicated (asterisks); *RGC2N* and *RGC2D* contain deletions resulting in frameshift mutations at the positions indicated by the open triangle. The size of the intron 3 in *RGC2D* could not be determined, as indicated by the question mark (see text). The position of the poly(A) site identified in a 3′ RACE product of *RGC2D* is shown.

eight exons. Intron–exon splice boundaries were identified by comparison to the 5′ and 3′ RACE products described above and by computer analysis of the genomic sequence to predict putative splice sites (Hebsgaard et al., 1996). Computer analysis predicted splice sites in all genes at the same locations as those identified by comparisons of cDNA and genomic sequences. The predicted mRNA is ~5.9 kb. The open reading frame accounted for 5274 to 5757 bp of the genomic sequence, encoding a predicted protein of 1758 to 1919 amino acids. The length of the complete genomic sequence varied from 7 kb (*RGC2O*) to 13.6 kb (*RGC2J*) (Figure 1 and Table 2). Therefore, the *RGC2* genes are among the largest genes thus far reported in plants. Most of the size difference between copies was due to differences in intron 3 (Figure 1 and Table 2). Comparisons between the sequences of the *RGC2* genes and the insertion site of a T-DNA that destroyed *Dm3* activity (Okubara et al., 1997) demonstrated that the insertion had occurred in intron 3 of *RGC2B*.

The predicted amino acid sequence of the RGC2B protein contains several distinct motifs as well as regions with no obvious homologies. The N-terminal and C-terminal regions of the protein have no significant similarities to sequences in the databases, including other disease resistance genes (Figures 2A and 2F). Unlike the class of NBS-LRR resistance genes, which includes the tobacco gene *N*, the flax gene *L6*, and the Arabidopsis gene *RPP5*, we found no homology to the N terminus of the Toll–interleukin-1 homology domain (TIR domain; Baker et al., 1997; Hammond-Kosack and Jones, 1997; Parker et al., 1997). No leucine zipper motifs were identified; this motif has been predicted from the nucleotide sequence of the resistance genes *RPS2*, *RPM1*, and *Prf* (Bent et al., 1994; Mindrinos et al., 1994; Grant et al., 1995; Salmeron et al., 1996). Adjacent to the N-terminal region, the RGC2B protein contains an NBS that is identifiable by the presence of the conserved P loop domain with the sequence GMGGVGKT, which is followed by four other characteristic motifs (Figure 2B). The sequence and spacing of these motifs and the position of the NBS in the protein are consistent with other known plant resistance genes (Hammond-Kosack and Jones, 1997). A short region with no significant similarity to other known genes separates the NBS from a C-terminal LRR region (Figure 2C).

The C-terminal two-thirds of the predicted protein is rich in leucine and other aliphatic residues and comprises a series of irregular repeats (Figures 2D and 2E). The LxxLxxaxaxxCxxaxxa (where x is any amino acid and a is a conserved aliphatic amino acid) consensus of RGC2 LRRs is more closely related to the predicted cytoplasmic LRR consensus LxxLxxLxLxx(N/C/T)x(x)LxxIPxxaxx than to the extracytoplasmic consensus LxxLxxLxLxxNxLxGxIPxxLx (Jones and Jones, 1997). However, the RGC2 LRRs are degenerate in comparison with this consensus and vary in length (Figures 2D and 2E). *RGC2* genes encode ~20 LRRs 5′ of intron 3 and ~21 LRRs 3′ of intron 3. The highly variable intron bisects the LRR region and defines a bipartite configuration in which the C-terminal region exhibits a more evident alternating pattern of hypervariable and conserved amino acids (see below). The total of ~41 LRRs is larger than any previously reported LRR region (Kobe and Deisenhofer, 1994). Several regions with few aliphatic residues and a poor match to known LRR consensus sequences interrupt the LRR region. It is possible that these are “loop-out” regions, providing some sort of a molecular hinge between LRR regions, as proposed for the *Cf-4* and *Cf-9* resistance genes of tomato (Jones and Jones, 1997).

Table 2. Sizes of Coding Regions and Introns in Nine *RGC2* Genes Sequenced[a]

RGC2 Copy	Exon 2	Intron 2	Exon 3	Intron 3	Exon 4	Intron 4	Exon 5	Intron 5	Exon 6	Intron 6	Exon 7	Intron 7
A	2676	131	717	1566	669	674	717	683	834	129	43	139
B	2667	131	507	5058	672	663	732	685	831	129	43	139
C	2655	131	507	~5200[b]	669	663	717	644	834	129	43	139
D	2661	131	510	>5930[c]	672	1009	375/Δ[d]	Δ	Δ/550	126	43	139
J	2682	134	525	6097	672	665	765	642	867	125	53	156
K	2676	112	546	1149	612	1815	1182	319	759	ND[e]	ND	ND
N	2679	130	717	1562	665	1022	729	685	834	126	43	ND
O	2685	140	537	363	603	993	687	219	762	ND	ND	ND
S	2655	131	507	~5300[c]	681	664	732	689	834	129	43	139

[a] Intron and exon lengths are given in base pairs, based on RACE cDNA analysis and a computer splice site prediction program from the website www.cbs.dtu.dk/services/NetPgene/ (Hebsgaard et al., 1996). Data for exon 1 and intron 1 in the 5′ UTR are not given because cDNA sequences and splice sites were not available for many of the copies in this region.

[b] Sequences are incomplete for an ~200-bp region containing three adjacent microsatellite sequences.

[c] This is the amount sequenced; the actual intron is much larger or discontinuous (see text).

[d] Δ, deleted region.

[e] ND, the intron–exon splice boundaries for the 3′ ends of *RGC2K* and *RGC2O* were not determined because 3′ RACE products were not identified.

One region in exon 5, encoding residues that do not match the LRR consensus, contains a small, compound, in-frame trinucleotide repeat designated MSATE6. This sequence is a derivative of the consensus $(ACA)_x$ACGAAGGGG$(TCT)_y$ and encodes polythreonine, a three–amino acid intervening sequence, and an adjacent stretch of polyserine (Figure 2E). The microsatellite MSATE6 is hypervariable among *RGC2* copies and was quite useful for mapping and detecting transcripts of particular members of the *RGC2* gene family (Meyers et al., 1998). In *RGC2K*, the microsatellite is $(ACA)_2$AAGGCA-$(TCT)_2$, representing the minimal repeat size observed in the *RGC2* family. The largest array, $(ACA)_5$ACGAAGGGG-$(TCT)_{21}$, is in *RGC2J.* The function of this region in the protein is not known. However, this microsatellite sequence is the site of differences in half of the nine *RGC2* copies sequenced: a 1.2-kb deletion in *RGC2D*, two large direct repeats in *RGC2K*, a 45-bp deletion in *RGC2O*, and a stop codon that occurs just 5′ of the microsatellite in *RGC2A* (Figure 1).

Transcript Analysis of the *RGC2* Family

Six of the nine genes contain complete open reading frames; however, a variety of mutations indicated that the remaining three are pseudogenes. *RGC2A* contains a nonsense mutation in exon 5, 3 bp 5′ of the microsatellite MSATE6. The *RGC2A* microsatellite allele is missing from the cDNA, indicating that the 3′ end of this gene is not expressed. *RGC2N* contains a nonsense mutation in exon 2, ~2.2 kb downstream of the start codon, as well as a 1-bp deletion in exon 4 (Figure 1). Comparisons of the genomic sequence of *RGC2D* to other copies revealed an ~1.3-kb deletion that fused exon 5 to exon 6 and introduced a frameshift. This fusion in exon 5 occurred 51 bp 5′ of the microsatellite MSATE6 and eliminated parts of both exons, all of intron 5, and the microsatellite. A 3′ RACE product was obtained that was identical to the 5′ coding sequence of *RGC2D* but contained a poly(A) tail 176 bp 3′ of exon 3 (Figure 1). Intron donor and acceptor splice sites are present at both ends of intron 3 in *RGC2D.* However, we sequenced almost 6 kb from both ends of intron 3, although polymerase chain reaction (PCR) failed to amplify across the predicted gap in the sequence. Therefore, either intron 3 of *RGC2D* is too large to be amplified by PCR (at least ~10 kb total) or the two ends of the gene are rearranged with respect to each other.

Members of the *RGC2* family that are transcribed were identified by analyzing RACE products and assaying the MSATE6 microsatellite in cDNA. Sequences of four *RGC2* family members, *RGC2B*, *RGC2C*, *RGC2D*, and *RGC2N*, were identified as RACE products. Microsatellite MSATE6 (as described above) was amplified from at least seven copies by using a cDNA template, including an additional four, *RGC2J*, *RGC2I*, *RGC2E*, and *RGC2S*, which had not been identified by RACE analysis (Figure 3). Interestingly, the largest allele of MSATE6 from *RGC2J* is transcribed. The MSAT data, together with the sequenced RACE products, indicate that at least eight *RGC2* copies are transcribed.

Genomic Comparisons between *RGC2* Family Members

Large variations in exon size were observed only in the LRR-encoding regions. Throughout the coding regions, there

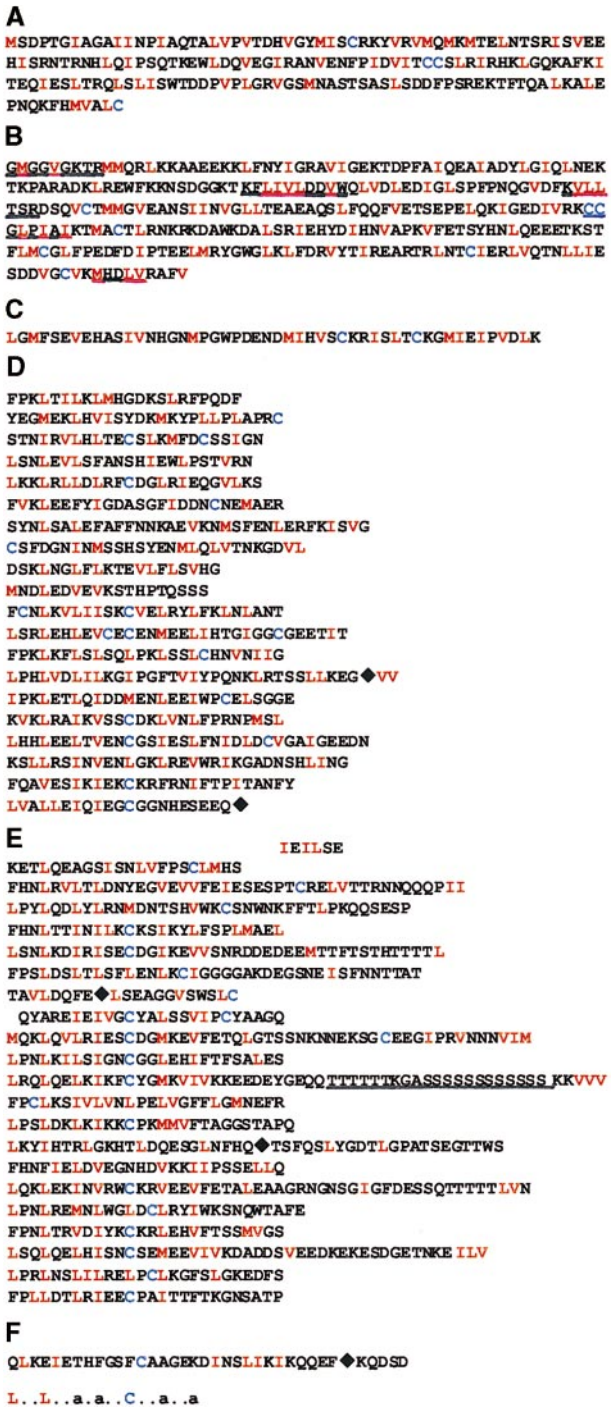


Figure 2. Amino Acid Sequence for the Predicted Full-Length Transcript of *RGC2B.*

The amino acid sequence is shown in single-letter code and is divided into six regions.
(A) The N terminus.
(B) The nucleotide binding site.
(C) A connecting region.
(D) The N-terminal LRR region.
(E) The C-terminal LRR region.
(F) The C terminus.
The conserved motifs of the NBS and the microsatellite MSATE6 are underlined in **(B)** and **(E)**, respectively. LRRs have been aligned according to the consensus sequence given at bottom, which approximates the consensus for cytoplasmic LRRs (Jones and Jones, 1997); "a" indicates the positions of aliphatic amino acids. The positions of introns are shown by diamonds. Aliphatic and cysteine residues are in red and blue, respectively.

were small indels of between one and seven codons that maintained intact open reading frames. Greater variation resulted in changes in the number of encoded LRRs. Relative to other *RGC2* genes, *RGC2A* and *RGC2N* contain a direct repeat in exon 3 (80% nucleotide identity) that encodes approximately two LRRs (Figure 1). *RGC2K* contains a direct repeat of 480 bp in exon 5 (78% nucleotide identity) that encodes approximately six LRRs (Figure 1).

Intron positions but not sizes were found to be conserved between copies. *RGC2* genes have five introns in the coding region, one in the 5′ UTR and one in the 3′ UTR. In the coding region, all introns are 3′ to the NBS-encoding region. Most introns range in length from 59 to 1815 bp. Extreme size variation was found to occur in intron 3, whereas the other introns show less variation (Table 2). The smallest intron 3 was 363 bp in *RGC2O.* This intron was four to 16 times longer in other family members; intron 3 was 6097 bp in *RGC2J.*

The similarities between intron 3 sequences are limited to regions adjacent to the splice sites (Figure 4). Sequence comparisons between the copies indicated that the ends of the intron have a high degree of similarity (>70% identical); however, this is lost within 400 to 500 bp. The middle of the introns contain DNA with no homology to distantly related *RGC2* copies or to any sequences in the databases. Intron 3 was found to be closely related and to exceed 5 kb in four *RGC2* genes: *RGC2B*, *RGC2C*, *RGC2D*, and *RGC2S.* Exon sequences indicate that *RGC2O* is also closely related to these four copies; however, intron 3 in *RGC2O* is <8% of intron 3 in *RGC2B*, the smallest of the above four genes. In the more diverse genes (*RGC2K*, *RGC2A*, and *RGC2J*), intron 3 varies from 1 to 6.1 kb, yet these introns have little sequence similarity to each other and to the *RGC2B* subfamily. Although genome expansion in intergenic regions of many plant species has been attributed to insertions of transposable elements (SanMiguel et al., 1996), no sequences homologous to known transposable elements were found in intron 3. We also were unable to identify terminal repeats characteristically associated with long terminal repeat elements or miniature inverted-repeat transposable elements that are often present in plant genes (Wessler et al., 1995).

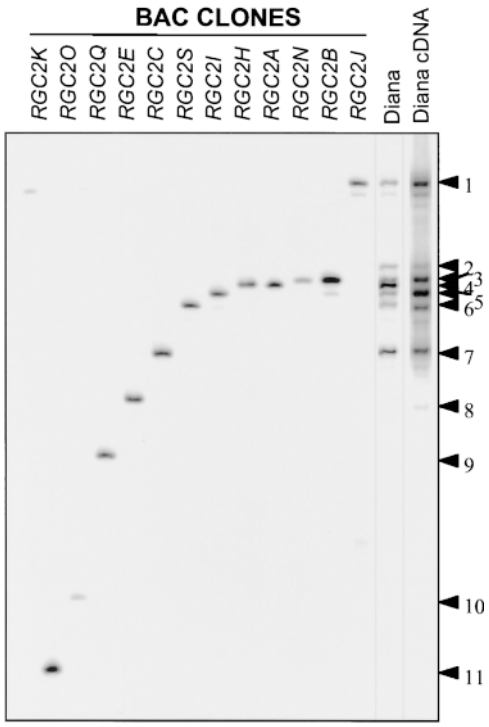


Figure 3. MSATE6 from BAC, Genomic, and cDNA Templates.

Primers were designed from *RGC2A* to amplify a microsatellite marker from exon 5. MSATE6 was amplified from genomic DNA of cultivar Diana, cDNA of cultivar Diana, and 12 BAC clones each containing a single copy of *RGC2.* Numbers to the right designate individual bands. The cDNA lacked band 4, which represents the *RGC2A* sequence from which the primers were designed, indicating no contamination of the cDNA with genomic DNA. No BAC clones containing band 2 were identified (Meyers et al., 1998). Amplification of the microsatellite was not expected from the more divergent members of the *RGC2* family in lettuce genomic DNA because of mismatches at the priming sites; however, microsatellites were amplified from BAC templates containing the divergent members.

Sequence conservation outside of the coding region was detected only for closely related genes. We obtained from 0.8 to 4.1 kb of genomic sequence 5′ to the ATG for *RGC2A*, *RGC2B*, *RGC2D*, *RGC2K, RGC2O*, and *RGC2N.* Sequence similarity was >96% between two sets of closely related copies: *RGC2A* and *RGC2N*, and *RGC2B* and *RGC2S.* Little sequence similarity was found between other sequences, indicating a high degree of divergence in both intron 3 and upstream sequences. Beyond the 3′ end of the open reading frame, 0.3 to 2.1 kb of sequence information was obtained for *RGC2B*, *RGC2K*, and *RGC2J*; again, the sequences outside of these divergent coding regions were unrelated. However, a probe from 4 kb 3′ of *RGC2B* hybridized with at least 11 members of the *RGC2* family (Meyers et al., 1998); therefore, there may be regions of sequence conservation beyond the 3′ end of the genes.

Comparisons of Nucleotide Substitution Patterns in Different Regions of Resistance Genes

A comparison between the aligned deduced amino acid sequences revealed an alternating pattern of variable and conserved amino acids in the LRR region. This pattern was more pronounced in the C-terminal half of the LRR region, which is encoded 3′ of intron 3. The hypervariable amino acids in each repeat are positioned around two conserved aliphatic amino acid sites in the consensus xx(a)x(a)xx. In the porcine ribonuclease inhibitor, these amino acids form parallel β sheets flanked by β turns (Kobe and Deisenhofer, 1994; Jones and Jones, 1997); these comprise a solvent-exposed surface that interacts with the ligand (Kobe and Deisenhofer, 1995).

Frequencies of nonsynonymous (K_a) and synonymous (K_s) nucleotide substitutions and K_a/K_s ratios were calculated for five different regions of the open reading frame: the 5′ end, the NBS-encoding region, the spacer between the NBS- and LRR-encoding regions, the 5′-encoded LRR region, and the 3′-encoded LRR region (Figure 5). Similar analyses in mammalian genes involved in pathogen recognition have detected higher rates of nonsynonymous than synonymous substitution in ligand binding regions (Hughes and Nei, 1988; Tanaka and Nei, 1989). K_a/K_s ratios <1 may result from the elimination of most nonsynonymous substitutions through purifying selection. K_a/K_s ratios >1 indicate diversifying selection (Li, 1997). When the complete open reading

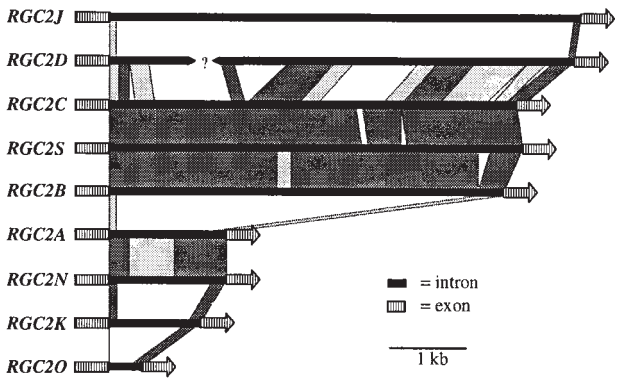
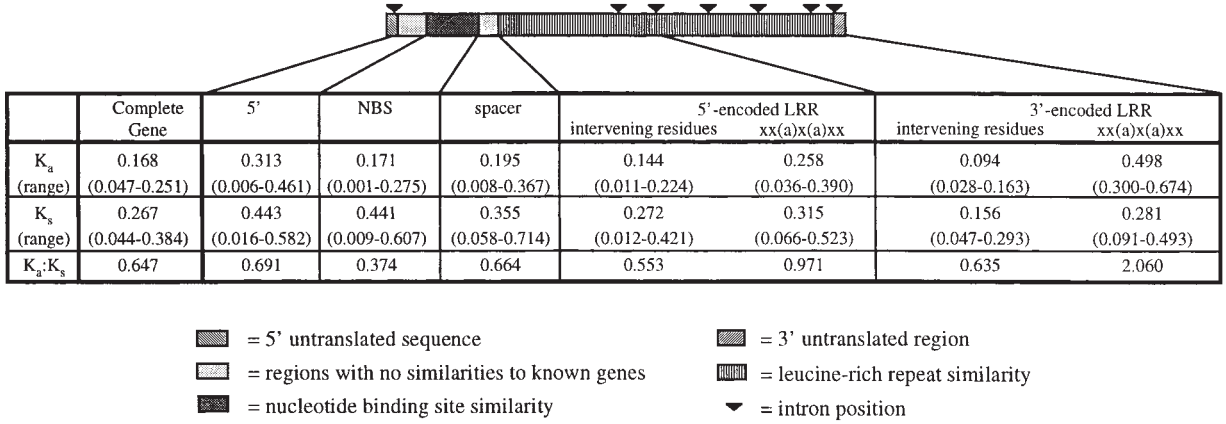


Figure 4. Sequence Similarity between Intron 3 of Nine *RGC2* Copies.

Pairwise comparisons were performed using intron 3 sequences from nine *RGC2* copies. Sequences demonstrating the greatest similarity within intron 3 were placed together. Regions with 65 to 90% similarity between copies are shown in light gray; regions with >90% similarity are dark gray. Unshaded regions are <65% similar. *RGC2D* contains a gap in the sequence (?; see text).



	Complete Gene	5'	NBS	spacer	5'-encoded LRR intervening residues	5'-encoded LRR xx(a)x(a)xx	3'-encoded LRR intervening residues	3'-encoded LRR xx(a)x(a)xx
K_a	0.168	0.313	0.171	0.195	0.144	0.258	0.094	0.498
(range)	(0.047-0.251)	(0.006-0.461)	(0.001-0.275)	(0.008-0.367)	(0.011-0.224)	(0.036-0.390)	(0.028-0.163)	(0.300-0.674)
K_s	0.267	0.443	0.441	0.355	0.272	0.315	0.156	0.281
(range)	(0.044-0.384)	(0.016-0.582)	(0.009-0.607)	(0.058-0.714)	(0.012-0.421)	(0.066-0.523)	(0.047-0.293)	(0.091-0.493)
K_a:K_s	0.647	0.691	0.374	0.664	0.553	0.971	0.635	2.060

Figure 5. K_a and K_s Values among *RGC2* Genes.

Values were calculated for nonsynonymous (K_a) and synonymous (K_s) substitutions in the protein coding regions of the gene. Values were calculated for 36 pairwise comparisons and averaged. The range of the values is given below the K_a and K_s averages to indicate the diversity of the sequences compared. The K_a/K_s ratio was calculated by averaging the ratio for each comparison. Individual K_a and K_s values are plotted in Figure 6.

frames of *RGC2* genes are compared, K_s exceeds K_a, indicating conservation of the gene as a whole (Figure 5). The NBS-encoding region was the most highly conserved portion of the gene, with an average K_a/K_s ratio of 0.374 (Figures 5 and 6). Statistical analysis using a *G* test strongly rejected the null hypothesis for neutral evolution (33 of 36 pairwise comparisons were significantly <1 at $P < 0.000001$). The 5′ and NBS-LRR spacer regions also had K_a/K_s ratios <1 (Figures 5 and 6A to 6F).

The LRR-encoding region of *RGC2* genes had unusual substitution patterns. The two halves of the LRR-encoding region separated by intron 3 were considered independently because of the more pronounced pattern of variability in the 3′ portion. In the porcine ribonuclease inhibitor, the aliphatic residues in the xx(a)x(a)xx consensus are buried in the hydrophobic core of the protein and do not interact with the ligand (Kobe and Deisenhofer, 1994, 1995). Therefore, the K_a/K_s ratio was calculated for the nucleotides encoding the xx(a)x(a)xx region of the LRR repeats in RGC2, omitting the codons for the conserved aliphatic positions. The K_a/K_s ratio of the codons corresponding to the xx(a)x(a)xx amino acids of the C-terminal LRR region was significantly >1 in 11 of 36 comparisons, indicating that these residues are under divergent selection (Figure 6 and Table 3). The same positions in the 5′ end of the encoded LRR showed elevated K_a/K_s ratios, but the average was not >1. The K_a/K_s ratio was calculated separately for the LRR-encoding sequence between encoded xx(a)x(a)xx motifs, designated the "intervening residues." The K_a/K_s ratios for the intervening residues were 0.553 in the 5′ and 0.635 in the 3′ regions (Figure 5). The ratios for many pairwise comparisons were significantly <1, indicating purifying rather than divergent selection. In a *G* test, 33 of 36 pairwise comparisons were significant at the $P < 0.01$ level for the 5′-encoded LRR, and 19 of 36 pairwise comparisons were significant at the $P < 0.05$ level for the 3′-encoded LRR; the lower number of significant comparisons in the 3′-encoded LRR reflects a low proportion of variable sites in this region. The statistically highly significant pattern of K_a/K_s ratios is evidence for a conserved backbone alternating with arrays of solvent-exposed β-sheet surfaces that are under diversifying selection in the LRR region of RGC2 proteins.

To determine whether alternating patterns of variability are a common feature of LRR regions in other plant disease resistance genes, we calculated the K_a/K_s ratios for the LRR- and putative effector-encoding regions of three types of LRR-containing resistance genes from other plant species. Alignments were made within the *I2C* and *Mi* families of tomato (Ori et al., 1997; Milligan et al., 1998), the *Xa21* family of rice (Song et al., 1997), and between the *L6* and *M* genes of flax (Lawrence et al., 1995; Anderson et al., 1997). The *I2C*, *Mi*, and *Xa21* families represent paralogs within localized clusters of genes. *L* and *M* are homoeologous loci derived from an ancient polyploidization event (Ellis et al., 1995). The K_a/K_s ratios were calculated for the same regions as for *RGC2*, except that the LRR-encoding regions were not split in two because no bipartite structure was apparent. Comparisons were also made to data for the *Cf-4/9* cluster from tomato (Parniske et al., 1997).

In each comparison, the K_a/K_s ratios indicated that all regions are under purifying selection except for the xx(a)x(a)xx residues of the LRRs (Figure 6 and Table 4). In every comparison, the codons for the xx(a)x(a)xx residues had a K_a/K_s ratio >1, indicating diversifying selection, whereas the different putative effector-encoding regions for each type of resistance gene had K_a/K_s ratios of <1, indicating selection for conservation. Many of these comparisons were significant only at the 5 to 10% level (Table 4). However, they are consistent with the

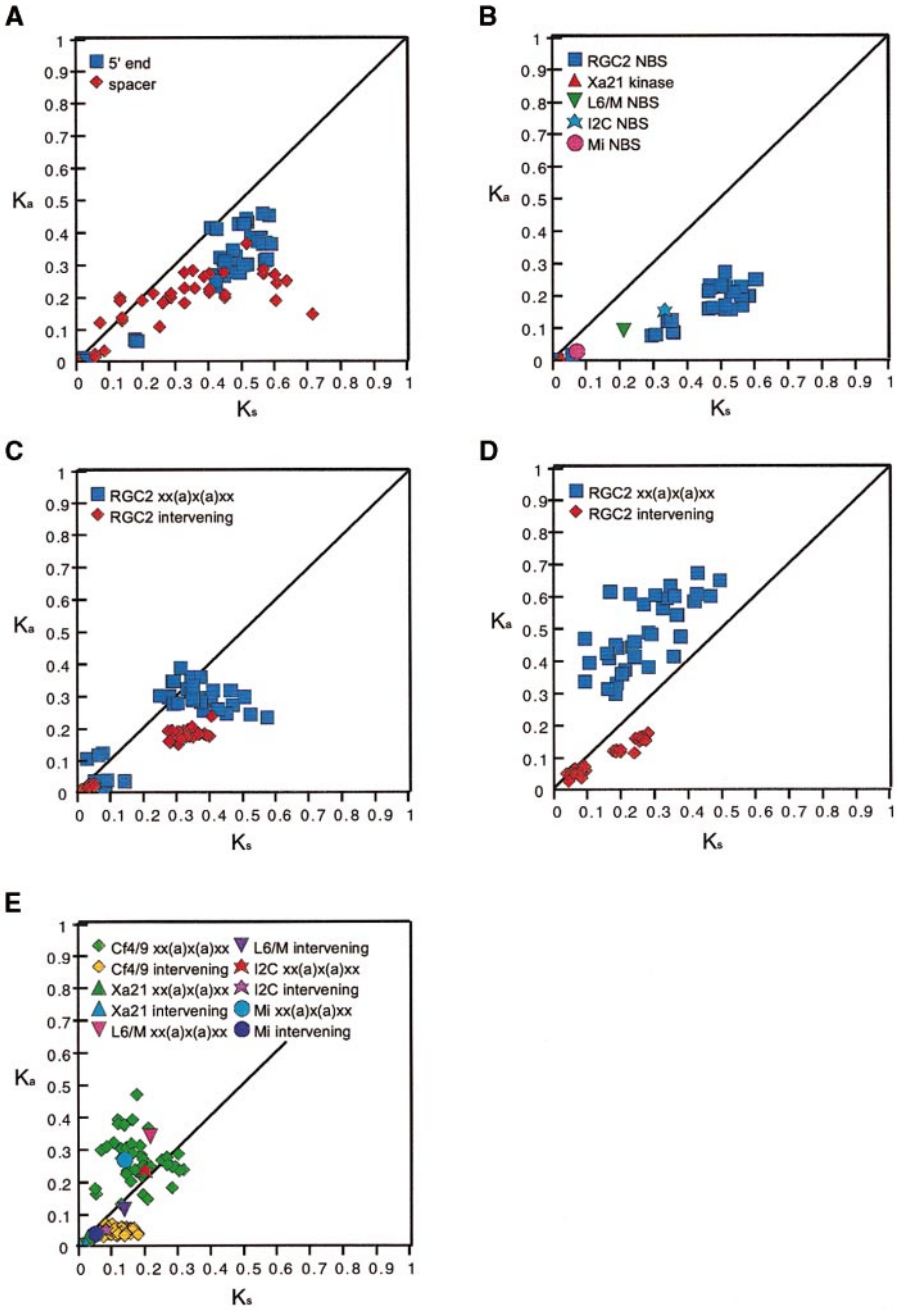


Figure 6. Synonymous and Nonsynonymous Substitution Frequencies in the *RGC2* Family and Other Plant Disease Resistance Genes.

(A) 5′ and spacer regions of *RGC2* genes.
(B) Regions encoding putative effector domains.
(C) *RGC2* 5′-encoded LRR.
(D) *RGC2* 3′-encoded LRR.
(E) The LRR-encoding region of other resistance gene families.
K_s and K_a substitutions were calculated for pairwise comparisons within resistance gene families. These values were plotted in the form (K_s, K_a). The diagonal corresponds to $K_s = K_a$, representing neutral evolution; points above this line provide evidence for diversifying selection, without implying statistical significance. Points below the diagonal suggest selection for conservation.

highly significant data for *RGC2. Xa21* paralogs have been grouped into two distinct classes based on sequence similarity (Song et al., 1997). Only comparisons within the *Xa21* class of paralogs (i.e., B [the functional *Xa21* gene], D, and F) showed elevated K_a/K_s ratios, and only one of these comparisons is significant at the 5% level (Table 4). Comparisons including other *Xa21* family members (the A2 class) had K_a/K_s ratios <1 (data not shown), possibly indicating that these genes are not functional in disease resistance.

Wang et al. (1998) found evidence for diversifying selection only in the *Xa21B*/*Xa21D* comparison by using comparisons of the entire LRR-encoding region. Elevated K_a/K_s ratios previously have been reported for the *Cf-4*/*9* genes in tomato (Parniske et al., 1997); however, in our reanalysis of these sequences, most pairwise comparisons are not significant (Table 4), suggesting that only some members of the *Cf-4*/*9* family are currently under diversifying selection. In summary, evidence for diversifying selection of varying levels of significance was found in all comparisons within LRR-containing families of plant disease resistance genes. In each case, this selection was acting on the region that may comprise a solvent-exposed surface in other LRR-containing proteins.

DISCUSSION

The *Dm3* downy mildew resistance locus of lettuce is composed of at least 24 diverse copies that span an estimated 3.5 Mb (Meyers et al., 1998). This locus is larger and more complex than are clusters of resistance genes described to date. Sequencing of nine full-length genomic copies demonstrated that the *RGC2* NBS-LRR genes are among the largest and most diverse plant resistance genes. The distribution of variable amino acids and patterns of nucleotide substitution support a model for divergent selection acting on amino acid residues that comprise the putative ligand binding surfaces.

RGC2 Genes Are Similar to but Distinct from Other NBS-LRR Resistance Genes

The *RGC2* genes are similar to many of the other known plant disease resistance genes but have several distinct features. The *RGC2* genes encode polypeptides of 1758 to 1919 amino acids; these are among the largest of any disease resistance genes known in plants. Xa1 in rice, conferring resistance to *X. o. oryzae*, is 1802 amino acids (Yoshimura et al., 1998); Prf in tomato, required for resistance to *Pseudomonas syringae* pv *tomato*, is 1824 amino acids (Salmeron et al., 1996). The remainder of known NBS-LRR–type resistance products in plants vary from 909 (RPS2, Bent et al., 1994; Mindrinos et al., 1994) to 1361 (RPP5, Parker et al., 1997) amino acids. The N terminus of the RGC2 proteins comprises a short region with no similarity to proteins in databases, including other resistance gene products. The NBS region is similar in size and motifs to other resistance gene products. A short region that again has no similarity to sequences in the databases separates the NBS from the LRR domain.

The C-terminal region of RGC2 proteins contains many LRRs. The RGC2 consensus sequence is related to cytoplasmic LRR proteins (Kobe and Deisenhofer, 1994; Jones and Jones, 1997), although it is more degenerate and more variable in length. The total of >40 LRRs spanning 1249 to 1380 amino acids is larger than any previously reported LRR region (Kobe and Deisenhofer, 1994; Jones and Jones, 1997). It is considerably larger than LRR regions found in similarly sized resistance gene products; those of Xa1, RPP5, and Prf include ~558 (the number of LRRs cannot be determined because the structure of the *Xa1* gene is atypical), 575 (21 LRRs), and 417 (18 LRRs) amino acids, respectively. Therefore, each LRR domain of the RGC2 proteins encoded either side of intron 3 (~20 and ~21 LRRs) is of similar size to the entire LRR region encoded by other NBS-LRR–type resistance genes.

We identified a hypervariable compound microsatellite within the coding region of *RGC2* genes. Although members of this family containing different sizes of this repeat are transcribed, it is not known whether this repeat influences gene function. Some animal genes tolerate hypervariable

Table 3. Subset of Comparisons within the *RGC2* Family That Show an Alternating Pattern of Positive and Purifying Selection[a]

	LRR: Intervening[b]		LRR: xx(a)x(a)xx[b]	
Gene Pair	*G* Value	P (χ^2)[b]	*G* Value	P (χ^2)[b]
RGC2N vs *RGC2O*	4.010	0.0452[c]	6.668	0.0098[d]
RGC2N vs *RGC2J*	0.003	0.9558	11.673	0.0006[e]
RGC2N vs *RGC2B*	0.207	0.6491	7.698	0.0055[d]
RGC2N vs *RGC2S*	0.184	0.6679	7.195	0.0073[d]
RGC2N vs *RGC2D*	5.865	0.0154[c]	9.199	0.0024[d]
RGC2N vs *RGC2K*	12.781	0.0004[e]	9.504	0.0021[d]
RGC2O vs *RGC2J*	5.918	0.0150[c]	14.203	0.0002[e]
RGC2O vs *RGC2D*	14.124	0.0002[e]	7.567	0.0059[d]
RGC2J vs *RGC2D*	1.450	0.2285	12.459	0.0004[e]
RGC2B vs *RGC2K*	17.426	0.00003[e]	7.730	0.0054[d]
RGC2S vs *RGC2C*	1.984	0.1590	7.211	0.0072[d]

[a] All comparisons that show a highly significant deviation ($P < 0.01$) from $K_a/K_s = 1$ for the region encoding the xx(a)x(a)xx motif are shown. Another 14 comparisons that were significant at the $P < 0.05$ level are not shown. The intervening and xx(a)x(a)xx-encoding regions deviated from the null hypothesis of $K_a/K_s = 1$ in opposite directions (Figure 6). Low *G* values for the codons of the intervening sequence reflect the low proportion of polymorphic sites in the region.
[b] Calculated for the LRR region encoded 3′ to intron 3.
[c] Significant at the 5% level.
[d] Significant at the 1% level.
[e] Significant at the 0.1% level.

Table 4. K_a/K_s Ratios in Different Regions of Plant Resistance Genes

Species	Genes	Putative Effector Regions[a]	LRR: Intervening	LRR: xx(a)x(a)xx	No. of Significant Comparisons[b] <0.1%	<1%	<5%	<10%	Σ	GenBank Accession Numbers
Lettuce	*RGC2* family	0.374	0.635[c]	2.060[c]	3	11	25	28	36	AF072268 to AF072275
Tomato	*I2C-1* vs *I2C-2*	0.470	0.580	1.167			1	1	1	AF004878, AF004879
Tomato	*Mi* copy 1 vs copy 2	0.425	0.761	1.927				1	1	AF039681, AF039682
Flax	*L6* vs *M*	0.444	0.827	1.571			1	1	1	U27081, U73916
Tomato	*Cf-4/9* homologs	0.859	0.593[d]	1.323[d]		3	9	17	55	AJ002235 to AJ002237
Rice	*Xa21* copies *B*, *D*, *F*[e]	0.423	0.518	2.106			1	2	1	U37133, U72726, U72728

[a] These regions include the NBS for *RGC2*, *I2C*, *L6*, *M*, and *Mi*; the 3′ LRR for *Cf-4/9*; and the kinase region for *Xa21*.

[b] The level of significance using the *G* test for deviation from $K_a/K_s = 1$ is shown for the xx(a)x(a)xx-encoding region. The number of significant pairwise comparisons at each level is given; therefore, the values at lower levels of significance are cumulative. Σ, total number of pairwise comparisons made.

[c] For the 3′ LRR region (abstracted from Figure 5).

[d] For the 5′ LRR region (Parniske et al., 1997); includes both orthologous and paralogous comparisons.

[e] The *Xa21* class of paralogs only (see text).

trinucleotide repeats in coding sequences; repeats encoding polyglutamine have been identified within animal receptors involved in growth and development, but the function of these repeats also is not known (Edwards et al., 1991). Trimeric repeats within transcribed sequences also have been identified as the cause of dysfunctional alleles in several human genes, including the fragile-X syndrome and myotonic dystrophy in humans (Fu et al., 1991; Brook et al., 1992). In both cases, phenotypically normal individuals may have many repeats (<46 for fragile-X or <27 for myotonic dystrophy; Fu et al., 1991; Brook et al., 1992); expansion beyond a threshold results in a dysfunctional genotype. It remains to be determined whether the size of the repeat affects the activity or the recognition specificity of *RGC2* genes. Expansion or contraction of the microsatellite sequence could alter the spacing of binding surfaces determined by LRRs flanking the repeat.

Intron position and number are conserved between *RGC2* family members but differ from other resistance genes. *RGC2* genes have seven introns, with five in the coding region. *I2C*, *RPS2*, and *RPM1* lack introns (reviewed in Hammond-Kossack and Jones, 1997). *Xa1* has three introns, of which two are in the coding region 5′ to the region encoding the NBS (Yoshimura et al., 1998). Three intron positions are shared among the TIR-NBS-LRR class of resistance genes *N*, *L6*, and *RPP5* (Parker et al., 1997). Therefore, the size of the LRR region and the position of the introns indicate that *RGC2* genes are members of a family of resistance genes that is distinct from those characterized to date.

Diversity in Intron 3 Suggests Distinct Lineages of *RGC2* Genes

Most of the variation in the size of *RGC2* genes is due to intron 3, which ranged from <400 bp to >6 kb. The disparate sequences and sizes of this intron suggest a complex evolutionary history. The diversity of sequences indicates that intron 3 has evolved independently in different lineages of the *RGC2* gene family. Without knowing the progenitor or ancestral gene sequence, we cannot determine whether the variation in size was due to insertions or deletions. The numerous indels in intron 3 had no homologs in the databases, terminal inverted repeats characteristic of transposable elements, obvious secondary structure, or duplications of sequences flanking the indels (Wessler et al., 1995; Bennetzen, 1996). Therefore, although the *Dm3* region contains transposable elements and retrotransposable elements in the intergenic regions (Meyers et al., 1998), we found no evidence for such elements within the *RGC2* gene sequences. Consequently, the mechanisms generating the variation in intron size are not apparent.

The degree of sequence divergence in the intron might influence meiotic pairing and hence the frequency of unequal crossing over between paralogs. Pairing between more diverse sequences would tend to be repressed and result in decreased levels of recombination and gene conversion. Consequently, a high level of sequence diversity would be maintained, and individual members would tend to evolve independently. Genetic analysis of the *Rp1* disease resistance cluster of maize indicates that recombination is lower between more distantly related haplotypes (Sudupak et al., 1993). In the *Cf-4*/*9* resistance gene locus of tomato, it has been proposed that dissimilar intergenic regions suppress mispairing in homozygotes (Parniske et al., 1997). The coding region of *RGC2* is more than twice as large as that of the genes in the *Cf-4*/*9* cluster, and divergent intron sequences in the middle of the gene would be expected to affect pairing behavior. The consistent sequence diversity observed among members of the *RGC2* family supports the hypothesis that there is little sequence exchange to homogenize these genes.

Selective Influences Differ across the *RGC2* Gene

Nucleotide substitution patterns in the *RGC2* family vary across the gene, particularly within the region encoding the LRRs. Synonymous substitutions have occurred at a higher frequency in the 5′ end of *RGC2.* Nonsynonymous substitutions were found at a significantly lower frequency than that of the synonymous substitutions, particularly within the NBS-encoding region; therefore, the low K_a/K_s ratio of the NBS-encoding region, which is the putative effector region, indicates that it has undergone the highest level of purifying selection within the gene. Within the region encoding the LRR domain, the putative ligand binding domain, there was an alternating pattern of conserved and variable amino acids. This was particularly evident in the 3′ end of the *RGC2* LRR-encoding region. The conserved regions correspond to amino acids that may form a structural backbone of the LRR; the hypervariable amino acids are predicted to form β sheets that are involved in ligand binding (Kajava et al., 1995; Kobe and Deisenhofer, 1995; Jones and Jones, 1997). Furthermore, the K_a/K_s ratios in the putative ligand binding surfaces of the 3′-encoded LRRs were >1, implying that divergent selection occurs at these positions. This pattern was found in the LRR-encoding region of genes from diverse plant species that confer resistance to pathogens that include fungi, bacteria, and nematodes. These genes represent three of four described structural classes for plant resistance genes (reviewed in Baker et al., 1997), indicating that the LRR sequence of these disparate resistance gene products is influenced by natural selection in a similar manner. This pattern is further evidence that the LRR region of resistance gene products is composed of a conserved backbone with variation localized in solvent-exposed surfaces.

Because patterns of amino acid variability and calculations of nucleotide substitutions evaluate groups of amino acids or nucleotides, they are inherently limited and cannot identify particular amino acids or LRRs critical to ligand binding. Mutations that alter receptor activity and ligand binding could occur anywhere in the gene. Also, some LRRs or portions of the LRR region may have a greater functional role in ligand binding than others. LRRs in the C-terminal half of Cf proteins are highly conserved, and the hypervariable residues are localized to the N-terminal half of the protein (Parniske et al., 1997; Thomas et al., 1997). In *RGC2* genes, the 3′ half of the LRR-encoding region was the more variable and showed higher K_a/K_s ratios than did the 5′-encoded LRR region. The detection of a hypervariable region within the LRR in all types of resistance genes analyzed suggests some biological significance to this pattern. However, further experimentation is necessary to confirm and refine conclusions derived from K_a and K_s calculations. In flax, domain swaps between alleles of the *L* gene indicate that the LRR is an important determinant of specificity (Ellis et al., 1997). Domain swaps between homologs followed by site-directed mutagenesis will focus on delineating regions of *RGC2* genes critical for specificity determination.

All classes of LRR-containing resistance gene products contain hypervariable surfaces within the putative receptor domain. The statistically significant evidence for divergent selection acting on many *RGC2* genes indicates that they must have been active in recognition of pathogens. This is consistent with our expression data. It is in contrast to the *Cf-4*/*9* data in which many comparisons result in K_a/K_s ratios ≤ 1 (Figure 6), indicating that numerous copies have not been under recent divergent selection. A large number of variable LRRs, particularly as found in RGC2, could increase binding opportunities or form multiple LRR subdomains. Such diversity in LRRs could increase the breadth of protein–ligand interactions and provide the flexibility for plants to coevolve with diverse pathogens.

Few proteins seem to be subject to diversifying selection, as indicated by a K_a/K_s ratio >1. Analyses of nucleotide substitution frequencies have detected evidence for diversifying selection in only 17 of 3595 families of sequences, and more than half of the 17 are antigenic surface proteins of parasites and viruses (Endo et al., 1996). The selective advantage of elevated occurrences of nonsynonymous substitutions has been most studied in the antigen recognition site (ARS) of class I mammalian major histocompatability complex (MHC) genes and the complementarity-determining region (CDR) of Ig genes (Hughes and Nei, 1988; Tanaka and Nei, 1989; Nei et al., 1997). Both the ARS and the CDR are responsible for recognizing and binding a wide range of potential ligands; evidence for diversifying selection in these genes suggests that variation is evolutionarily advantageous. The proposed ligand binding region of the *RGC2* gene products and other plant resistance gene products also appears to be under diversifying selection. The evolution of new recognition specificities in MHC and Ig genes involves variation in the individual amino acids as well as recombination and gene conversion. The relative importance of each of these sources of variation in the evolution of plant resistance genes remains to be determined.

Mechanisms Involved in the Evolution of Resistance Genes

Models describing the evolution of plant resistance genes have proposed recombination and gene conversion as the primary forces that generate diversity within these multigene families (Shepherd and Mayo, 1972; Pryor, 1987; Richter et al., 1995; Hammond-Kosack and Jones, 1997). The models assume that these events result in the rapid evolution of novel resistance specificities to counteract variable pathogen populations. Unequal crossing over is clearly involved in the generation of duplicated arrays of resistance genes. The size of multigene families at resistance loci varies between haplotypes (Parniske et al., 1997; D.T. Lavelle and R.W. Michelmore, unpublished data). The *Cf-2* locus of tomato includes two functional genes that are >99.9% identical, suggesting recent sequence duplication (Dixon et al., 1996). Novel specificities

may also result from recombination or gene conversion. Recombination and unequal crossing over at the *Rp1* resistance gene complex of maize are involved in meiotic instability and the generation of new specificities (Sudupak et al., 1993; Richter et al., 1995; Hulbert, 1997). Sequence analysis of the *Cf-4*/*9* locus of tomato indicates that sequence exchange has occurred between gene family members, resulting from either recombination or gene conversion (Dixon et al., 1996; Parniske et al., 1997). Also, mutants identified at the *M* locus of flax may have resulted from intragenic recombination (Anderson et al., 1997). Recombination is therefore involved in alterations in the copy number of resistance genes and in the generation of novel resistance specificities.

The importance of single base changes in the evolution of plant disease resistance genes may have been overlooked. There is now evidence for diversifying selection in the predicted β-sheet portion of the LRR consensus in all types of LRR-encoding resistance genes (Parniske et al., 1997; Wang et al., 1998; this study). The nonsynonymous substitutions that have accumulated in the region encoding the putative ligand binding domain may substantially affect recognition specificity. Recombination between alleles and paralogs could result in enhanced or novel binding properties by shuffling LRRs between related genes. The relative importance of recombination versus single base changes depends on the rate at which each of these events occurs. Although unequal crossing over has been measured at some resistance gene clusters (Richter et al., 1995), there are few data on the point mutation rates in these genes. The high level of sequence diversity in the *RGC2* family indicates that recombination and gene conversion are not homogenizing these sequences and are infrequent events (this study; Meyers et al., 1998). Therefore, novel amino acid substitutions in the solvent-exposed surfaces of the LRR region may be more important than intergenic recombination and gene conversion in the rapid evolution of novel specificities.

METHODS

Sequencing of *RGC2* Copies

Genomic sequences of the *RGC2* genes and flanking regions were obtained using a combination of several methods. All *RGC2* copies were sequenced from the bacterial artificial chromosome (BAC) clones on which they were originally identified (Meyers et al., 1998). *RGC2A* and *RGC2B* copies were primarily sequenced using a primer walking strategy. Primers were synthesized by Gibco Life Technologies (Grand Island, NY). When possible, primers generated during the sequencing of *RGC2A* and *RGC2B* were used to sequence further *RGC2* copies; additional primers for the more divergent members were synthesized as required. Some sequencing was performed using a modified version of the polymerase chain reaction (PCR)–based long-distance sequencer method (Hagiwara and Harris, 1996). To simplify the Hagiwara and Harris (1996) method, we performed a standard PCR, with 35 cycles, annealing for 30 sec at 58°C, and extension for 2 min at 72°C. PCR fragments were purified before sequencing by exonuclease I/shrimp alkaline phosphatase treatment (U.S. Biochemical Corp.) to remove unincorporated deoxynucleotide triphosphates and excess primers. Restriction digests, ligations, and agarose gel analysis of DNA fragments were performed according to standard protocols (Sambrook et al., 1989).

DNA Sequencing and Analysis

DNA sequencing was performed using an ABI 377 automated sequencer (Applied Biosystems Inc., Foster City, CA) and the PRISM Ready Reaction DyeDeoxy Terminator cycle sequencing kit (Applied Biosystems Inc.) with custom primers or standard Sp6, T7, M13 (-21), or M13 reverse primers. Sequence data were evaluated using Sequencher (GeneCodes, Ann Arbor, MI) for contig assembly and sequence editing. Splice site analysis and intron–exon boundaries were determined by comparison with cDNA sequences and by using the software program NetPlantGene available at www.cbs.dtu.dk/services/NetPgene/ (Hebsgaard et al., 1996). GeneDoc (www.cris.com/~ketchup/genedoc.shtml), DNAstar (Lasergene, Madison, WI), and Genetics Computer Group (Madison, WI) software packages were used for multiple sequence alignments and sequence comparisons. Nucleotide substitution rates were calculated by the method of Li (1993) in the Diverge program in the Genetics Computer Group software package. A 2×2 contingency table *G* test was used to test for the significance of differences in synonymous and nonsynonymous substitution rates (Zhang et al., 1997). The values for the 2×2 contingency table were estimated by using the model of Nei and Gojobori (1986). K_a and K_s values were plotted with the Freelance Graphics program (Lotus, Cambridge, MA). Phylogenetic studies were performed with PAUP* version 4.0 (Sinauer Associates, Sunderland, MA).

cDNA Analysis

RNA was isolated from lettuce (*Lactuca sativa*) cultivar Diana via the procedure of Jones et al. (1995). First-strand cDNA was synthesized by use of Superscript reverse transcriptase from 1 μg of total RNA, as specified by the manufacturer (Gibco Life Technologies). 5′ and 3′ rapid amplification of cDNA ends (RACE; Frohman et al., 1988) was performed using the Marathon kit (Clontech, Palo Alto, CA), according to the manufacturer's instructions, with primers designed from the predicted open reading frames observed in the genomic sequence.

GenBank accession numbers are as follows: *RGC2A*, AF072268; *RGC2B*, AF072267; *RGC2C*, AF072269; *RGC2D*, AF072270; *RGC2J*, AF072271; *RGC2K*, AF072272; *RGC2N*, AF072273; *RGC2O*, AF072274; and *RGC2S*, AF072275.

ACKNOWLEDGMENTS

We thank Dean O. Lavelle for technical assistance. This work was supported by the U.S. Department of Agriculture National Research Initiative Competitive Grant Program (Grant No. 95-37300-1571). Partial support for B.C.M. was provided by a National Science Foundation graduate research fellowship.

Received June 12, 1998; accepted September 14, 1998.

REFERENCES

Anderson, P.A., Okubara, P.A., Arroyo-Garcia, R., Meyers, B.C., and Michelmore, R.W. (1996). Molecular analysis of irradiation-induced and spontaneous deletion mutants at a disease resistance locus in *Lactuca sativa.* Mol. Gen. Genet. **251,** 316–325.

Anderson, P.A., Lawrence, G.J., Morrish, B.C., Ayliffe, M.A., Finnegan, E.J., and Ellis, J.G. (1997). Inactivation of the flax rust resistance gene *M* associated with loss of a repeated unit within the leucine-rich repeat coding region. Plant Cell **9,** 641–651.

Baker, B., Zambryski, P., Staskawicz, B., and Dinesh-Kumar, S.P. (1997). Signaling in plant–microbe interactions. Science **276,** 726–733.

Bennetzen, J.L. (1996). The contributions of retroelements to plant genome organization, function and evolution. Trends Microbiol. **4,** 347–353.

Bent, A.F. (1996). Plant disease resistance genes: Function meets structure. Plant Cell **8,** 1757–1771.

Bent, A.F., Kunkel, B.N., Dahlbeck, D., Brown, K.L., Schmidt, R., Giraudat, J., Leung, J., and Staskawicz, B.J. (1994). *RPS2* of *Arabidopsis thaliana*: A leucine-rich repeat class of plant disease resistance genes. Science **265,** 1856–1860.

Bevan, M., et al. (1998). Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana.* Nature **391,** 485–488.

Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, J.P., Hudson, T., Sohn, R., Zemelman, B., Snell, R.G., Rundle, S.A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P.S., Shaw, D.J., and Housman, D.E. (1992). Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell **68,** 799–808.

Crute, I.R., and Pink, D.A.C. (1996). Genetics and utilization of pathogen resistance in plants. Plant Cell **8,** 1747–1755.

Dixon, M.S., Jones, D.A., Keddie, J.S., Thomas, C.M., Harrison, K., and Jones, J.D.G. (1996). The tomato *Cf-2* disease resistance locus comprises two functional genes encoding leucine-rich repeat proteins. Cell **84,** 451–459.

Edwards, A., Civitello, A., Hammond, H.A., and Caskey, C.T. (1991). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am. J. Hum. Genet. **49,** 746–756.

Ellis, J.G., Lawrence, G.J., Finnegan, E.J., and Anderson, P.A. (1995). Contrasting complexity of two rust resistance loci in flax. Proc. Natl. Acad. Sci. USA **92,** 4185–4188.

Ellis, J.G., Lawrence, G.J., Ayliffe, M., Anderson, P., Collins, N., Finnegan, J., Frost, D., Luck, J., and Pryor, T. (1997). Advances in the molecular genetic analysis of the flax–flax rust interaction. Annu. Rev. Phytopathol. **35,** 271–291.

Endo, T., Ikeo, K., and Gojobori, T. (1996). Large-scale search for genes upon which positive selection may operate. Mol. Biol. Evol. **13,** 685–690.

Farrara, B., Illott, T.W., and Michelmore, R.W. (1987). Genetic analysis of factors for resistance to downy mildew (*Bremia lactucae*) in lettuce (*Lactuca sativa*). Plant Pathol. **36,** 499–514.

Frohman, M.A., Dush, M.K., and Martin, G.R. (1988). Rapid production of full-length cDNAs from rare transcripts by amplification using a single gene-specific primer. Proc. Natl. Acad. Sci. USA **85,** 8998–9002.

Fu, Y.H., Kuhl., D.P.A., Pizzuti, A., Pieretti, M., Sutcliffe, J.S., Richards, S., Verkerk, A.J.M.H., Holden, J.J.A., Fenwick, R.G., Warren, S.T., Oostra, B.A., Nelson, D.L., and Caskey, C.T. (1991). Variation of the CGG repeat at the fragile-X site results in genetic instability: Resolution of the Sherman paradox. Cell **67,** 1047–1058.

Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W., and Dangl, J.L. (1995). Structure of the Arabidopsis *RPM1* gene enabling dual specificity disease resistance. Science **269,** 843–846.

Hagiwara, K., and Harris, C.C. (1996). 'Long distance sequencer' method: A novel strategy for large DNA sequencing projects. Nucleic Acids Res. **24,** 2460–2461.

Hammond-Kosack, K.E., and Jones, J.D.G. (1997). Plant disease resistance genes. Annu. Rev. Plant Physiol. Plant Mol. Biol. **48,** 575–607.

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

Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatability complex class I loci reveals overdominant selection. Nature **335,** 167–170.

Hulbert, S.H. (1997). Structure and evolution of the *Rp1* complex conferring rust resistance in maize. Annu. Rev. Phytopathol. **35,** 293–310.

Jones, A., Davies, H.M., and Voelker, T.A. (1995). Palmitoyl–acyl carrier protein (ACP) thioesterase and the evolutionary origin of plant acyl-ACP thioesterases. Plant Cell **7,** 359–371.

Jones, D., and Jones, J.D.G. (1997). The role of leucine-rich repeat proteins in plant defenses. Adv. Bot. Res. Adv. Plant Pathol. **24,** 89–167.

Kajava, A.V., Vassart, G., and Wodak, S.J. (1995). Modeling of the three-dimensional structure of proteins with the typical leucine-rich repeats. Structure **3,** 867–877.

Kobe, B., and Deisenhofer, J. (1994). The leucine-rich repeat: A versatile binding motif. Trends Bio. Sci. **19,** 415–421.

Kobe, B., and Deisenhofer, J. (1995). A structural basis of the interactions between leucine-rich repeats and protein ligands. Nature **374,** 183–186.

Lawrence, G.J., Finnegan, E.J., Ayliffe, M.A., and Ellis, J.G. (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.

Li, W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36,** 96–99.

Li, W.H. (1997). Molecular Evolution. (Sunderland, MA: Sinauer Associates).

Meyers, B.C., Chin, D.B., Shen, K.A., Sivaramakrishnan, S., Lavelle, D.O., Zhang, Z., and Michelmore, R.W. (1998). The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. Plant Cell **10,** 1817–1832.

Michelmore, R.W. (1995). Molecular approaches to manipulation of disease resistance genes. Annu. Rev. Phytopathol. **15,** 393–427.

Milligan, S.B., Bodeau, J., Yaghoobi, J., Kaloshian, I., Zabel, P., and Williamson, V.M. (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.

Mindrinos, M., Katagiri, F., Yu, G.L., and Ausubel, F.M. (1994). The A. thaliana disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. Cell **78,** 1089–1099.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3,** 418–426.

Nei, M., Gu, X., and Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA **94,** 7799–7806.

Okubara, P.A., Arroyo-Garcia, R., Shen, K.A., Mazier, M., Kim, S.J., Yang, C.-H., and Michelmore, R.W. (1997). A transgenic mutant of *Lactuca sativa* (lettuce) with a T-DNA tightly linked to loss of downy mildew resistance. Mol. Plant-Microbe Interact. **10,** 970–977.

Ori, N., Eshed, Y., Paran, I., Presting, G., Aviv, D., Tanksley, S., Zamir, D., and Fluhr, R. (1997). The *I2C* family from the wilt disease resistance locus *I2* belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. Plant Cell **9,** 521–532.

Parker, J.E., Coleman, M.J., Szabò, V., Frost, L.N., Schmidt, R., Van der Biezen, E.A., Moores, T., Dean, C., Daniels, M.J., and Jones, J.D.G. (1997). The Arabidopsis downy mildew resistance gene *RPP5* shares similarity to the Toll and interleukin-1 receptors with *N* and *L6.* Plant Cell **9,** 879–894.

Parniske, M., Hammond-Kosack, K.E., Golstein, C., Thomas, C.M., Jones, D.A., Harrison, K., Wulff, B.B.H., and Jones, J.D.G. (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.

Pryor, T. (1987). The origin and structure of fungal disease resistance in plants. Trends Genet. **3,** 157–161.

Richter, T.E., Pryor, T.J., Bennetzen, J.L., and Hulbert, S.H. (1995). New rust resistance specificities associated with recombination in the *Rp1* complex in maize. Genetics **141,** 373–381.

Salmeron, J.M., Oldroyd, G.E.D., Rommens, C.M.T., Scofield, S.R., Kim, H.-S., Lavelle, D.T., Dahlbeck, D., and Staskawicz, B.J. (1996). Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. Cell **86,** 123–133.

Sambrook, J., Fritisch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, 2nd ed. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science **274,** 765–768.

Shen, K.A., Meyers, B.C., Islam-Faridi, N., Stelly, D.M., and Michelmore, R.W. (1998). Resistance gene candidates identified using PCR with degenerate oligonucleotide primers map to resistance gene clusters in lettuce. Mol. Plant-Microbe Interact. **11,** 815–823.

Shepherd, K.W., and Mayo, G.M.E. (1972). Genes conferring specific plant disease resistance. Science **175,** 375–380.

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

Staskawicz, B.J., Ausubel, F.M., Baker, B.J., Ellis, J.G., and Jones, J.D.G. (1995). Molecular genetics of plant disease resistance. Science **268,** 661–667.

Sudupak, M.A., Bennetzen, J.L., and Hulbert, S.H. (1993). Unequal exchange and meiotic instability of disease-resistance genes in the *Rp1* region of maize. Genetics **133,** 119–125.

Tanaka, T., and Nei, M. (1989). Positive Darwinian selection observed at the variable region genes of immunoglobulins. Mol. Biol. Evol. **6,** 447–459.

Thomas, C.M., Jones, D.A., Parniske, M., Harrison, K., Balint-Kurti, P.J., Hatzixanthis, K., and Jones, J.D.G. (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.

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

Wang, G.-L., Ruan, D.-L., Song, W.-Y., Sideris, S., Chen, L., Pi, L.-Y., Zhang, S., Zhang, Z., Fauquet, C., Gaut, B.S., Whalen, M.C., and Ronald, P.C. (1998). *Xa21D* encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. Plant Cell **10,** 765–779.

Wessler, S.R., Bureau, T.E., and White, S.E. (1995). LTR-retrotransposons and MITEs: Important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. **5,** 814–821.

Williams, C.E, Wang, B., Holsten, T.E., Scambray, J., Dasilva, F.D.G., and Ronald, P.C. (1996). Markers for selection of the rice *Xa21* disease resistance gene. Theor. Appl. Genet. **93,** 1119–1122.

Yoshimura, S., Yamanouchi, U., Katayose, Y., Toki, S., Wang, Z.X., Kono, I., Kurata, N., Yano, M., Iwata, N., and Sasaki, T. (1998). Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. Proc. Natl. Acad. Sci. USA **95,** 1663–1668.

Zhang, Z., Khumar, S., and Nei, M. (1997). Small-sample tests of episodic adaptive evolution: A case study of primate lysozymes. Mol. Biol. Evol. **14,** 1335–1338.