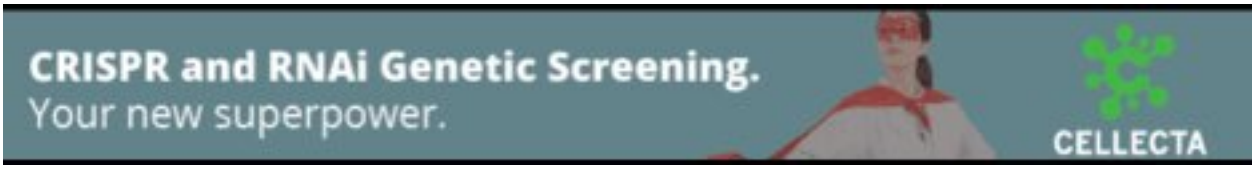
Patterns of Positive Selection in the Complete NBS-LRR Gene Family of *Arabidopsis thaliana*

Mariana Mondragón-Palomino, Blake C. Meyers, Richard W. Michelmore, et al.

Genome Res. 2002 12: 1305-1315
Access the most recent version at doi:10.1101/gr.159402



Letter

Patterns of Positive Selection in the Complete NBS-LRR Gene Family of *Arabidopsis thaliana*

Mariana Mondragón-Palomino,[1] Blake C. Meyers,[2,3] Richard W. Michelmore,[2] and Brandon S. Gaut[1,4]

[1]*Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, California 92612, USA;*
[2]*Department of Vegetable Crops, University of California, Davis, California 95616, USA*

Plant disease resistance genes have been shown to be subject to positive selection, particularly in the leucine rich repeat (LRR) region that may determine resistance specificity. We performed a genome-wide analysis of positive selection in members of the nucleotide binding site (NBS)-LRR gene family of *Arabidopsis thaliana*. Analyses were possible for 103 of 163 NBS-LRR nucleotide sequences in the genome, and the analyses uncovered substantial evidence of positive selection. Sites under positive selection were detected and identified for 10 sequence groups representing 53 NBS-LRR sequences. Functionally characterized *Arabidopsis* resistance genes were in these 10 groups, but several groups with extensive evidence of positive selection contained no previously characterized resistance genes. Amino acid residues under positive selection were identified, and these residues were mapped onto protein secondary structure. Positively selected positions were disproportionately located in the LRR domain ($P < 0.001$), particularly a nine–amino acid β-strand submotif that is likely to be solvent exposed. However, a substantial proportion (30%) of positively selected sites were located outside LRRs, suggesting that regions other than the LRR may function in determining resistance specificity. Because of the unusual sequence variability in the LRRs of this class of proteins, secondary-structure analysis identifies LRRs that are not identified by similarity analyses alone. LRRs also contain substantial indel variation, suggesting elasticity in LRR length could also influence resistance specificity.

Disease resistance genes (R genes) are crucial components of the hypersensitive response (HR), a plant defense mechanism that results in localized cell death. The HR is triggered when pathogen molecules, possibly virulence factors, are detected by plant receptors; genetic analysis of the HR has lead to the cloning of R genes, many of which encode receptor-like proteins. Based on their predicted domain structure, R proteins encoded by the R genes have been classified into four groups: intracellular kinases, extracellular receptors, extracellular receptors coupled to kinases, and intracellular receptors (Bent 1996).

Most characterized R genes encode putative intracellular receptors (Dangl and Jones 2001), which contain either a coiled-coil (CC) or a Toll/Interleukin-1 receptor (TIR) domain at their N-terminal end, followed by a nucleotide binding site (NBS). At the C-terminal end, these proteins consist of a series of leucine rich repeats (LRRs). The functions of the CC, TIR, and NBS domains are not known fully, but all similar proteins identified in animal systems play roles in protein–protein interactions and signal transduction (Srinivasula et al. 1998; Kopp and Medzhitov 1999; Inohara et al. 1999; Burkhard et al. 2001). The function of LRR domains is clearer because recent data suggest that LRRs in R proteins mediate direct or indirect interaction with pathogen molecules (Jia et al. 2000; Dangl and Jones 2001). The tertiary structure of LRRs has been experimentally determined for a diverse group of proteins (Price et al. 1998; Marino et al. 1999; Liker et al. 2000; Zhang et al. 2000), most notably porcine ribonuclease inhibitor (PRI; Kobe and Deisenhofer 1993, 1995b). Generally, individual LRRs form repeats of β-strand-loop and α-helix-loop units with nonleucine residues exposed and compose a binding surface predicted involved in protein recognition (Kobe and Kajava 2001). In R proteins, putatively solvent-exposed residues in β-sheets may interact with pathogen ligands and hence determine specificity for pathogen elicitors (Thomas et al. 1997; Ellis et al. 1999, 2000).

Comparative analyses of R genes from tomato, lettuce, rice, flax, and *Arabidopsis* have revealed that solvent-exposed positions of the LRRs are hypervariable and subject to positive natural selection (Parniske et al. 1997; Meyers et al. 1998; Wang et al. 1998; Noel et al. 1999; Ellis et al. 2000). Evidence for positive selection is consistent with host–pathogen coevolution (Endo et al. 1996) and selection for new resistance specificities. A corollary of this observation, and the underlying basis of our work, is that positive selection may be used as an evolutionary profile that identifies NBS-LRR-encoding genes that are likely to function in disease resistance.

The genome sequence of *Arabidopsis* provides an opportunity to investigate genomic patterns of positive selection in the NBS-LRR gene family. To detect positive selection, we estimate ratios of nonsynonymous to synonymous nucleotide substitutions, also known as ω, on NBS-LRR gene family members. ω is a molecular evolutionary measure of selection (Kimura and Ohta 1974). When ω is equal to 1, a gene is evolving without constraint on nonsynonymous substitutions relative to synonymous substitutions, a condition interpreted as neutral evolution. In contrast, an $\omega > 1$ is strong evidence of positive selection (Hughes and Nei 1988) and an $\omega < 1$ is consistent with purifying selection, although the possibility of positive selection cannot be excluded. To calculate ω, we have employed a maximum likelihood (ML) method that identifies the specific amino acid residues on which posi-

[3]Present address: Department of Plant and Soil Sciences, University of Delaware, Newark, Delaware 19711, USA.
[4]Corresponding author.
E-MAIL bgaut@uci.edu; FAX (949) 824-2181.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.159402.

tive selection has acted (Nielsen and Yang 1998; Yang and Bielawski 2000). The ML approach differs substantially from approaches used in previous studies of positive selection in NBS-LRR genes because previous studies partitioned nucleotide codons into predicted solvent-exposed regions and the remainder of the LRR (e.g., Parker et al. 1997; Botella et al. 1998; Warren et al. 1998; Bittner-Eddy et al. 2000). Such a priori partitioning does not permit identification of individual amino acids under positive selection, and it does not provide an accurate picture of the extent and genic location of positive selection.

The location of positively selected residues is important for inferring gene function. For example, previous studies have shown that solvent-exposed regions of the LRR are subject to positive selection; these results have been interpreted as evidence that solvent-exposed regions mediate pathogen recognition (Parniske et al. 1997). Here we map the position of all positively selected amino acid residues onto NBS-LRR genes and find that most but not all are found within the LRR region. In addition, we apply secondary-structure prediction methods to LRR regions to characterize structural motifs and also to determine whether positively selected sites fall predominantly in solvent-exposed residues. Altogether, this study of *Arabidopsis* NBS-LRR genes has two main goals. First, we use selection as an evolutionary profile, hypothesizing that positive selection may help identify the subset of NBS-LRR genes that are most likely to function in plant defense. Second, we elucidate the relationship among structure, function, and evolution by mapping positively selected sites onto NBS-LRR gene secondary structure.

RESULTS

Sequence Groups

We retrieved complete amino acid sequences of 163 genes from the *Arabidopsis* Resistance Genes database (At-RGenes), aligned the sequences, and reconstructed a neighbor-joining phylogeny. The phylogeny based on 163 NBS-LRR genes was similar to the At-RGenes database phylogeny in that there was a clear separation between the TIR-NBS-LRR and CC-NBS-LRR sequences, and there were also some similarities in the grouping of sequences within clades (data not shown). However, the At-RGenes phylogeny was based only on the NBS region, while our analyses were based on complete sequences of NBS-LRR proteins.

The aligned genes were too divergent for analysis of positive selection, and thus we partitioned sequences into individual groups. Based on the initial phylogeny and sequence characteristics (see Methods), we pared the data to 103 sequences and assigned them to 22 phylogenetically clustered groups (Table 1). During grouping, some sequences that could not be assigned to groups were discarded as orphans, including the characterized resistance genes *RPM1* and *RPS2*, which

Table 1. NBS-LRR Sequences Analyzed for Positive Selection

Group[1]	Type[2]	Sequences[3]	Length[4]	Identity[5]
1	TNL	At1g64070, At2g16870, At5g58120, At1g63750, At1g63870, At1g63880, At1g63730, At1g63740, At1g56510, At1g56520, At1g56540	1364	58.0
2	TNL	At5g40910, At5g40920, At5g41550, At5g41540, At5g41750, At5g41740	1011	65.4
3	TNL	At4g11170, At5g49140	1183	50.6
4	TNL	At5g17970, At5g18360	911	57.5
5	TNL	At2g14080, At1g65850, At5g38340, At5g44510, At1g69550, At5g11250, At3g44630, At3g44480, At3g44400, At3g44670 (*RPP1*)[6]	1436	59.0
6	TNL	At5g18350, At5g18370	1272	60.3
7	TNL	At5g46450, At5g46470, At4g08450, At5g40060, At1g31540, At5g46260, At5g46270, At5g46510, At5g46520	1249	60.9
8	TNL	At5g40100, At1g17600, At1g72840, At4g09430	1025	53.4
9	CNL	At1g63360, At1g62630, At5g63020, At5g05400	934	53.9
10	CNL	At4g14610, At1g12210, At1g12220, At1g12280, (*RPS5*), At1g12290, At4g10780	911	59.4
11	NL	At1g61180, At1g61310, At1g61300, At1g61190	978	86.3
12	NL	At5g43740, At1g51480, At1g15890	868	69.6
13	NL	At4g27190, At4g27220	995	50.3
14	CNL	At1g58400, At1g58410, At1g58390, At1g59620, At1g59780, At5g43470 (*RPP8*)[6], At5g48620, At5g35450, At1g53350, At1g10920	949	59.1
15	CNL	At3g46530 (*RPP13*)[6], At3g46730, At3g46710	872	66.3
16	NL	At5g04720, At5g47280, At1g33560, At4g33300	858	66.6
17	NL	At5g66900, At5g66910	822	73.2
18	TNL	At3g51570, At4g19530, At5g45060, At545250 (*RPS4*)[7], At5g17880	1202	54.1
19	TNL	At1g27170, At1g27180	1596	81.5
20	TNL	At4g16920, At4g16950 (*RPP5*)[6], At4g16860, At4g16890, At4g16940, At4g16960, At4g16900, At5g51630	1231	69.2
21	TNL	At5g45260, At5g45050	1440	70.9
22	TNL	At5g45200, At4g36150	1271	50.5

[1]Group designation is arbitrary.
[2]TNL, TIR-NBS-LRR; CNL, coiled-coil-NBS-LRR; NL, NBS-LRR.
[3]Names from the Munich Information Center for Protein Sequence *Arabidopsis* database (http://mips.gsf.de/proj/thal/db/index.html).
[4]Length of amino acid alignment, with gaps.
[5]average within group amino acid sequence identity.
[6]In Col-0, this sequence is the most similar to the R gene (in parentheses) characterized in other ecotypes.
[7]Characterized R gene in Col-0 (gene name in parentheses).

have been noted as orphans previously (Richly et al. 2002). After grouping, the average group size was 4.6 sequences, with the largest containing 11 sequences and the smallest containing 2. Eight of 22 groups contained only 2 sequences, and for these groups we could not apply the full range of ML analyses (Table 1). Group alignments ranged in length from 822 to 1596 amino acid positions, with an average length of 1108 positions (Table 1).

Detection of Positive Selection

We applied likelihood ratio (LR) tests of positive selection based on the ML methods and codon substitution models (M) of Yang, Nielsen, and colleagues (Yang 1997; Nielsen and Yang 1998; Yang et al. 2000). We applied two tests. The first LR test compared M1, the free-ratio model that assumes independent ω values for every branch of the phylogeny, versus M0, the null one-ratio model that constrains ω to be equal on all phylogenetic branches (Yang 1998; Yang and Nielsen 1998). This test was applied to all 22 sequence groups. After Bonferroni correction for 22 tests, six groups both fit M1 significantly better than M0 and also had at least one lineage with $\omega > 1$ (Table 2). Thus, the comparison of M1 and M0 detected positive selection in six sequence groups.

Comparison between M1 and M0 yields an average ω value among codons along a phylogenetic branch. If positive selection took place in only a few codons, the effect of positive selection on nucleotide substitution may not be detected (Anisimova et al. 2001). We therefore employed a second, more specific test that examines variation in ω among sites by comparing models M7 and M8 (Yang et al. 2000). This test could only be applied to the 14 sequence groups with more than two sequences (Table 2). For complete sequence alignments, LR tests with M7 and M8 identified 10 groups that fit the selective model better than the null model and also had an $\omega > 1$. Results remained significant after Bonferroni correction for 14 tests and an experiment-wide error of 5% (Table 2). When the LR test suggested positive selection action had occurred, positively selected sites were identified under M8 using a Bayesian method (Nielsen and Yang 1998; Yang et al. 2000). The number of inferred positively selected sites varied among the 10 groups in which positive selection was detected. For example, only one site was identified from group 18, but group 14 had 26 positively selected sites (Fig. 1; Table 2).

We also applied the two LR tests separately to the TIR, NBS, and LRR regions. CC domains were analyzed together with the NBS because the short length of CC domains made separate tests impractical, and groups with two sequences were not divided into domains because of low information content. Of the 14 groups divided into domains, 8 groups contained at least one domain that had (1) an estimate of $\omega > 1$ under M8, (2) sites identified to be under positive selec-

Table 2. Likelihood Ratio Test Results

Group	n[1]	*P*-value[2] M0:M1	*P*-value M7:M8	M8 Estimates[3]	Positively Selected Positions[4]
1	11	**6.0e-6**	1.6e-04	$\omega = 0.75$; $P_1 = 0.00$	
2	6	**<1.0e-6**	**<1.0e-6**	$\omega = 3.69$; $P_1 = 0.08$	263 C 489 W 573 S 599 S 621 F 667 G 693 I 730 R 754 F 756 Q 774 H 779 N 796 V 797 S 839 S 842 M 938 T 939 E 940 S 942 Y 945 V
3	2	0.009			
4	2	<1.0e-6			
5	10	**0.002**	<1.0e-6	$\omega = 0.23$; $P_1 = 0.55$	
6	2	0.023			
7	9	**3.1e-5**	1.0	$\omega = 0.50$; $P_1 = 0.0$	
8	4	0.040	**8.5e-5**	$\omega = 1.81$; $P_1 = 0.20$	97 Q 714 V 730 L 759 E
9	4	**<1.0e-6**	**<1.0e-6**	$\omega = 3.73$; $P_1 = 0.12$	111 R 305 H 372 H 474 R 562 Q 703 E 795 H
10	6	0.094	**<1.0e-6**	$\omega = 3.80$; $P_1 = 0.04$	553 K 577 S 722 G 784 F 843 E
11	4	0.006	**<1.0e-6**	$\omega = 8.89$; $P_1 = 0.10$	7 L 10 S 388 Q 427 K 431 Y 481 Y 534 D 604 N 626 T 629 D 676 V 717 S 719 R 738 E 758 E 779 V 781 L 817 N 840 S 841 N 862 E
12	3	0.002	**<1.0e-6**	$\omega = 7.18$; $P_1 = 0.06$	2 L 423 P 594 L 782 C 815 D 819 R 859 E 862 P
13	2	0.008			
14	10	0.004	**<1.0e-6**	$\omega = 3.80$; $P_1 = 0.07$	108 V 581 W 590 K 609 K 634 L 636 R 659 T 662 S 679 R 682 F 683 N 724 T 726 G 753 R 754 T 755 P 756 D 780 N 802 S 804 D 805 G 826 R 828 D 829 F 830 R 853 W
15	3	0.641	**6.0e-6**	$\omega = 4.04$; $P_1 = 0.08$	581 W 640 E 699 M 729 P 759 S 797 L 798 S 852 S 853 K
16	4	0.024	0.258	$\omega = 1.16$; $P_1 = 0.14$	
17	2	0.001			
18	5	0.076	**6.1e-5**	$\omega = 2.28$; $P_1 = 0.07$	805 L
19	2	0.003			
20	8	**0.002**	**<1.0e-6**	$\omega = 3.43$; $P_1 = 0.08$	602 E 649 W 650 Y 662 L 672 S 950 Y 967 E 1126 G 1195 F 1218 T 1224 S 1225 C 1227 P 1229 S
21	2	7.1e-4			
22	2	2.0e-6			

[1]Number of sequences in the group.
[2]*P*-values in bold represent significant tests in which ω is inferred to be >1.0. *P*-values are Bonferroni corrected for the number of tests within a column and an experimental Type I error of 0.05.
[3]ω is d_N:d_S estimated under M8; P_1 is the inferred proportion of positively selected sites.
[4]Position locations are based on alignments with gaps. Underlined positions are those common to the results of the analyses of the complete and domain alignments.

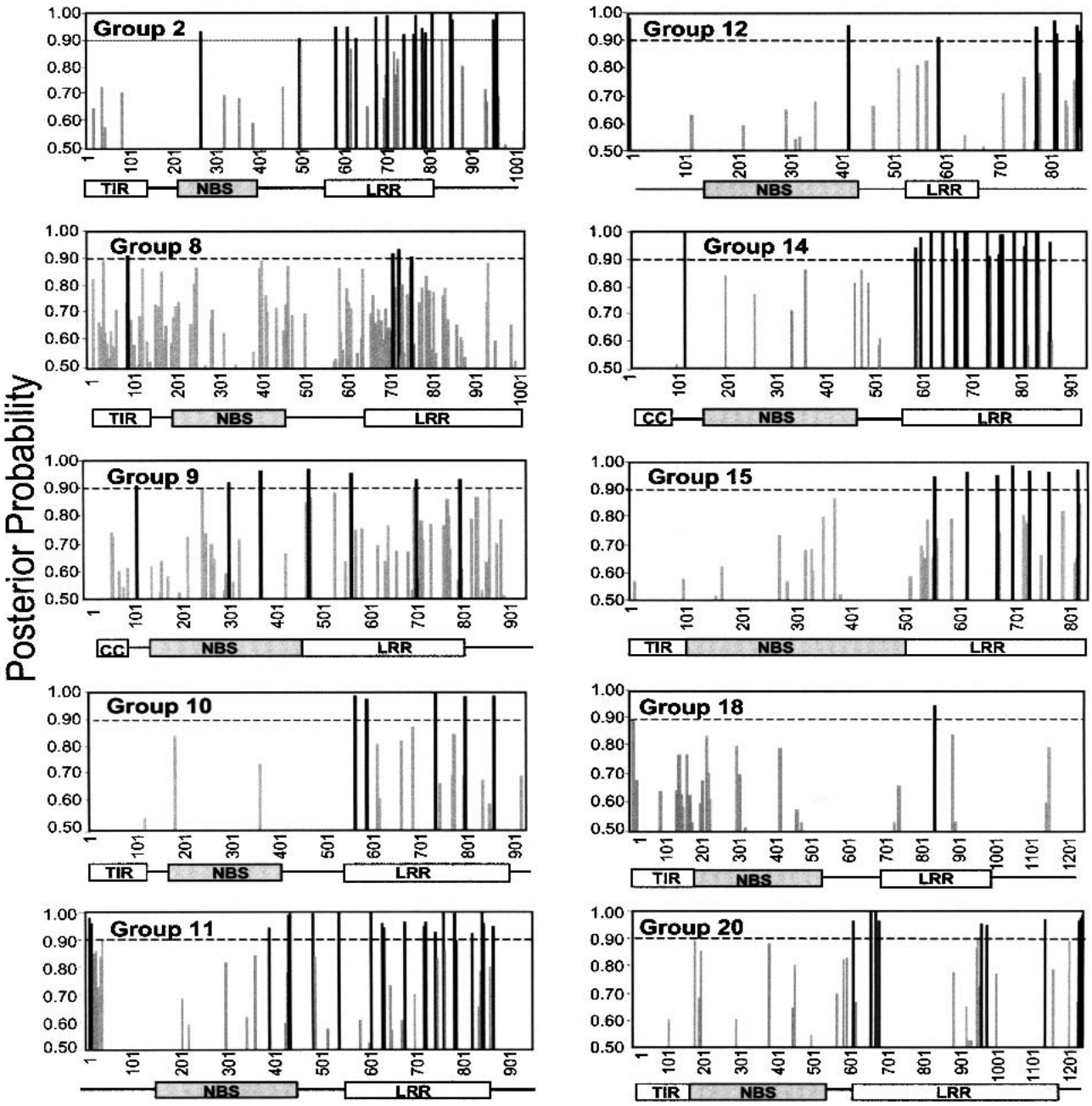


Figure 1 The posterior probability for sites in the positively selected class (ω > 1). Each graph represents 1 of the 10 sequence groups for which positive selection was detected by comparison of M7 and M8. The *X*-axis denotes position in the amino acid alignment. Sites with black bars had a posterior probability >0.9 under M8; sites with gray bars did not have posterior probabilities >0.9. Boxes under each graph denote domain structures of nucleotide sequences in the group, as identified either by Pfam (groups 2, 8, 9, 11, and 12) or by comparison to groups containing previously described R genes (groups 10, 14, 15, 18, and 20).

tion, and (3) a significant LR test (Table 3). For 7 of these 8 groups, positive selection was also detected with whole-sequence analysis (Table 2); the lone exception was group 7, which contained positively selected sites in the LRR region alone but no positively selected sites with complete data (Table 2). Six of the 8 domains that exhibited evidence of positive selection were LRRs, and they included 102 of the 105 positively selected sites detected in domain analyses (Table 3).

Location of Positively Selected Sites and Sequence Variation

We plotted the genic location of positively selected sites for the 10 groups that had sites detected from whole-sequence analysis (Fig. 1). Positively selected sites were not homogeneously distributed among regions; 69% (83 of 116) of sites were located in LRRs. The heterogeneous distribution of positively selected sites was clear from the comparison of the pro-

Table 3. Domains With Significant Tests With M3 and M8 After Bonferroni Correction

Group	Domain	n[1]	L[2]	*P*-Value[3] M0:M1	*P*-value[3] M7:M8	ω	Positively selected sites[4]
2	LRR	6	222	**4.5e-5**	**<1.0e-6**	5.73	599 S 607 L 621 F 664 T 667 G 669 E 686 M 688 Y 690 S 691 G 693 I 709 E 714 N 730 R 754 F 756 Q 774 H 776 D 779 N 796 V797 S 818 T
7	LRR	9	351	0.03	**<1.0e-6**	2.43	664 C 676 H 687 R 690 E 710 S 711 N 713 T 723 Q 734 E 735 R 737 E 753 C 757 N 774 E 776 Y 778 S 779 E 845 F 847 S 869 N 871 A 872 R 890 Q 894 S 911 S 916 G 938 I 958 T 969 H 974 T 977 S
8	NBS	4	252	0.64	**7.1e-3**	2.71	405 A 406 N
9	LRR	4	212	0.12	**5.5e-3**	3.93	474 R
10	LRR	6	246	0.42	**<1.0e-6**	4.02	553 K 577 S 599 V 672 F 722 G 760 G 761 Q 781 D 784 F 843 E
11	LRR	4	306	0.43	**<1.0e-6**	15.42	580 Y 586 K 599 F 604 N 606 S 626 T 629 D 645 L 647 R 667 Q 673 A 676 V 695 C 716 S 717 S 719 R 721 E 738 E 743 R 758 E 760 M 779 V 781 L 783 E 817 N 833 V 836 T 838 D 840 S 841 N 857 K
14	CC-NBS	10	435	0.03	**4.6e-3**	1.91	108 V
15	LRR	3	284	0.11	**7.5e-4**	4.66	581 W 640 E 699 M 729 P 759 S 797 L 798 S

[1]Number of sequences per group.
[2]Length without gaps.
[3]*P*-values in bold are these that remained significant after Bonferroni correction for the total number of tests.
[4]Positions are described by using the amino acid of the first sequence in the alignment and its corresponding position considering gaps; underlined positions are those common to the results of the analyses of the complete and domain alignments.

portion of sites under selection within the NBS and LRR regions, the two domains that occur in all proteins. Two-by-two contingency tests revealed that sites under positive selection occur significantly more frequently in the LRR domains ($P \ll 0.001$; $\chi^2 = 56.13$). Nonetheless, 33 positively selected sites were located in non-LRR regions.

We also studied the distribution of indels across regions in the groups in which positive selection was detected. In a two-by-two contingency test, LRRs had a significantly larger proportion of indels than non-LRR domains ($P \ll 0.001$; $\chi^2 = 145.14$). However, the high incidence of indels in the LRR was not unique to proteins under positive selection. Positively selected sites were not detected in groups 1, 5, 7, and 16, but indels were also more frequent in the LRR than the NBS for these sequence groups ($P \ll 0.001$, $\chi^2 = 91.36$). These observations are important for two reasons. First, the high incidence of gaps in the LRR regions provides additional evidence that LRRs are more labile than other domains. Second, gaps alone do not account for the high incidence of positively selected sites in LRRs.

LRR Secondary Structure and Residues Under Positive Selection

Non-leucine residues in the β-sheet of LRRs can be exposed to the solvent phase (Kobe and Deisenhofer 1994) and may interact with pathogen ligands (Jones and Jones 1997; Ellis et al. 2000), suggesting that the structural arrangement of variable sites in the LRR is important. To investigate this, we analyzed the predicted secondary structure of aligned LRR motifs and then mapped the distribution of sites under positive selection onto these structural predictions.

Prior analyses of plant NBS-LRR R proteins indicate that the LRR typically has a consensus sequence similar to LXXLXXLXLXX(N/C/T)X(X)LXXIPXX, where X represents any amino acid and the other letters denote specific amino acid residues (Hammond-Kosack and Jones 1997; Jones and Jones 1997). Our secondary-structure analyses of these repeats revealed that LRR structure is in general characterized by a coil (C) structure up to the third leucine (L) of this consensus, followed by 3 to 6 residues that have a β-strand (E) structure. The XXLXLXX motif within the LRR has been predicted to form a solvent-exposed β-sheet (Jones and Jones 1997). In our analyses, the first 3 to 6 of these residues consistently adopted a β-strand structure, followed by 3 to 6 residues in a coil (Fig. 2); we refer to this ~9-residue region as E_4C_5. The LRR consensus of the sequences we analyzed (Fig. 2) starts at the sixth residue of the consensus cited above. The β strand predicted for this consensus is centered on the second conserved L and the remaining E4C5 residues adopt a coil structure. In roughly one-third of the LRRs, the basic secondary structure of the remaining LRR is modified by 3 or 4 residues that adopt an α-helix configuration (H)—these residues tend to be aliphatic L, V, and I—and are located ~9–11 residues after the last residue in the β strand (Fig.2). The resulting secondary-structure pattern of CCEEEECCCCCCCCCCHHHHCC recurs throughout predicted LRR regions and, in some cases, within protein regions not predicted to contain LRR domains. The latter occurred in groups 11 and 18, in which six and eight LRR motifs were detected by Pfam analysis and an additional 6 and 3 regions, respectively, had a secondary structure consistent with an LRR (Fig. 2).

When the sites under positive selection were plotted onto the predicted secondary structure of each protein, we found that most sites fell into E_4C_5. A two-by-two contingency test comparing the E_4C_5 with the rest of the LRR showed that this motif contained a significantly higher proportion of positively selected sites ($\chi^2 = 48.60$), suggesting that these sites are evolutionarily, and perhaps functionally, distinct.

Contrasts Between Groups With and Without Positively Selected Sites

Some groups did not have identifiable positively selected sites, and it is useful to explore potential differences between

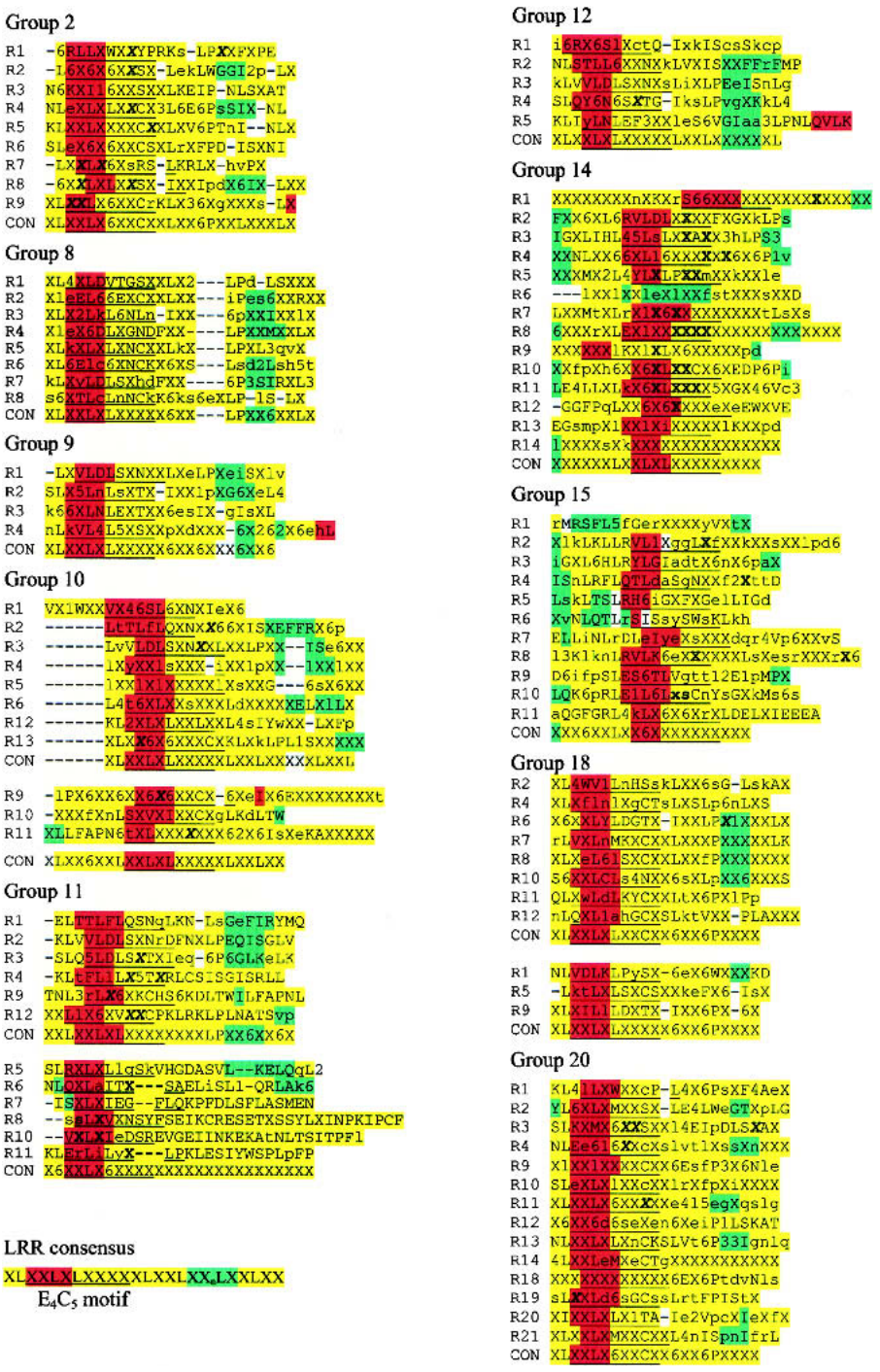


Figure 2 Secondary-structure predictions for LRRs from four groups with the highest number of positively selected sites. R# on left indicates the ordinal number for each LRR, and the sequence within each LRR represents the consensus from the alignment. Secondary-structure predictions are colored (yellow, coil; red, β-sheet; and blue, α-helix). The positions in which positive selection was detected are in bold and the sites that are part of the E_4C_5 motif are underlined. CON is the consensus sequence among LRRs for each group. Amino acids represented in uppercase are invariant among sequences. Notation for variable amino acid sites are as follows: X, any amino acid; 1, D,N; 2, E,Q; 3, S,T; 4, K,R; 5, F,Y,W; and 6, L,I,V,M. Dashes are gaps introduced to better align consensus LRRs and do not necessarily represent gaps in the original sequence alignments. In group 11, repeats 5, 6, 7, 8, 10, and 11 were defined by their predicted secondary structure. Similarly, repeats 1, 5, and 9 of group 18 were defined by predicted secondary structure. For other groups, LRRs were defined either by Pfam or by their similarity to characterized R genes.

groups that did and did not have positively selected sites. Because selected sites could only be identified in groups with more than two sequences with the M8 test, the ensuing section only considers groups with more than two sequences

Differences between groups with and without positively selected sites could be a consequence of sampling, so we contrasted four statistics between groups: (1) the number of sequences in the groups, (2) the length of the alignments, (3) the mean sequence identities in groups, and (4) tree lengths, which reflects the amount of sequence evolution among the sequences in a group. Each of these statistics contributes to the statistical power of the LR tests (Anismova et al. 2001), and significant differences in any of these characters between positively and nonpositively selected groups could indicate that sampling properties (like sequence length or sequence identity) are primarily responsible for our results. However, we detected no significant difference between positively selected and nonpositively selected groups in any of the four statistics (alignment length: $t = 1.43$, $P = 0.17$; number of sequences: $t = 1.61$, $P = 0.13$; average sequence identity: $t = 0.69$, $P = 0.5$; tree length: $t = 0.87$, $P = 0.4$). These results suggest that sampling characteristics alone do not play the primary role in discriminating between groups with and without positively selected sites.

Although there are no obvious sampling differences between groups with and without positively selected sites, there could be biological differences that do not relate to resistance function. We examined two potential biological differences. The first was gene expression, using EST hits as a proxy for expression. This proxy is particularly rough given the diverse origin and preparation methods of cDNA libraries used for EST sequencing, but we used ESTs to determine if there were substantial differences in expression between the two sequence classes. EST information was retrieved from the Munich Information Center for Protein Sequences (MIPS) *Arabidopsis thaliana* database (http://mips.gsf.de/proj/thal/) and the

RIKEN cDNA collection (Seki et al. 2002). We summed the number of EST hits for each sequence group (data not shown) and contrasted the total number of hits between positively selected and nonpositively selected groups. Although the groups with positively selected sites had a slightly higher average number of EST hits (12 hits vs. 8.1 hits), the difference in hits was not significant ($t = 0.91$; $P = 0.37$). Thus, by this method there is no detectable difference in gene expression between the two classes of sequences.

Ectopic recombination and gene conversion among sequences is a second biological parameter that could conceivably affect tests for positive selection. Recombination and gene conversion among sequences could affect test statistics because the ML test assumes a single phylogeny adequately represents the evolution of a group of sequences. If ectopic exchange (or gene conversion) occurs in only one genic region (for example, the LRR), it is possible that different genic regions have different evolutionary histories, so that the sequence did not evolve with a single phylogenetic pattern. Although the ML approach is known to be reasonably robust to incorrect phylogenetic assumptions (Yang et al. 2000), the effects of gene conversion and ectopic exchange on test statistics is not known. Nonetheless, we tested for gene conversion and ectopic recombination with Sawyer's (1989) test. Of the 14 groups with more than two sequences, only two groups (groups 20 and 7) showed significant evidence of ectopic exchange at the 5% significance level, and there is evidence for only one exchange event in each group (data not shown). Of these two, group 20 showed evidence for positive selection. In contrast, group 7 contains no evidence of positively selected sites based on whole-sequence data but some evidence when the LRR is examined separately (Tables 2 and 3; also see Discussion). Overall, however, Sawyer's test provided little evidence of ectopic exchange within groups, and there is no indication that ectopic events contribute substantially to differences between groups with and without positively selected sites.

DISCUSSION

Positive selection has been documented in genes that encode pathogen surface proteins (Bush 2001; Peek et al. 2001), reproductive proteins (Swanson et al. 2001a,b), and host defense systems like the human major histocompatibility complex (Hughes and Nei 1988), plant chitinases (Bishop et al. 2000), and NBS-LRR R genes (Meyers et al. 1998; Wang et al. 1998; Bergelson et al. 2001). The correlation between positive selection and host–pathogen interactions is particularly strong. For example, a GenBank survey uncovered remarkably few sequences (0.45%) evolving under positive selection, but more than half of these sequences were involved in host–pathogen interaction (Endo et al. 1996).

The close relationship between host–pathogen interactions and positive selection suggests that positive selection can form the basis for an evolutionary profile to identify NBS-LRR genes that are likely to be involved in *Arabidopsis* disease resistance. However, there are at least two caveats to evolutionary profiling in this gene family. The first caveat is that it is difficult to determine whether the detection of positive selection is a predictor of function. If profiling is accurate, characterized resistance genes should fall into groups for which positive selection is consistently inferred. Our *Arabidopsis* data include sequences for R genes *RPS4* and *RPS5*, as well as the Col-0 sequences that are most similar to the defense genes *RPP1*, *RPP5*, *RPP8/HRT*, and *RPP13* (Table 1), and these R genes have been inferred to be subject to positive selection (Parker et al. 1997; Botella et al. 1998; McDowell et al. 1998; Warren et al. 1998; Bittner-Eddy et al. 2000). Five of these six characterized R genes, or their likely orthologs, are in groups that have positively selected amino acid sites under M8 (Tables 1 and 2); hence, profiling correctly identifies these groups as containing functional resistance genes. Positive selection in these groups was also detected after known defense genes were removed from analysis (data not shown), raising the possibility that the groups contain multiple functional R genes.

The sixth characterized R-gene, *RPP1*, has a putative Col-0 ortholog in group 5 (Table 1), in which positively selected sites were not detected (Table 2). There is, however, evidence for positive selection in this group based on the test of M0 versus M1 (Table 2). It is more interesting that there is no evidence for positive selection in this group, either by the test of M7 versus M8 or the test of M0 versus M1, when the *RPP1* ortholog is removed from analysis (data not shown). Thus, group 5 appears to consist primarily of sequences that lack a signature of positive selection. One cannot ascribe function to sequences based solely on tests for positive selection, but it is tempting to speculate that most genes in group 5 either do not play a role in defense or do not directly mediate pathogen interaction.

The most similar Col-0 homolog to the characterized resistance gene *RRS1* is in group 21. To date, there has been no evidence of positive selection in *RRS1* (Deslandes et al. 2002), and we detect no evidence for positive selection in group 21 (Table 2). We should note, however, that group 21 consists of only two sequences, and the power to detect positive selection in groups with two sequences appears to be low (see below). We should also note that the Col-0 homolog of *RRS1* contains a WRKY domain but unlike *RRS1* is not predicted to contain a nuclear localization signal downstream of the LRR region (data not shown). It thus is unclear to what extent *RRS1* and its putative Col-0 homolog share functions.

Altogether, there is a strong correspondence between characterized resistance genes and positive selection. Six characterized resistance genes or their putative Col-0 orthologs fall into groups in which positive selection was detected. The putatative Col-0 ortholog of a seventh gene, *RRS1*, does not belong to a group in which we detected positive selection, but the Col-0 ortholog appears to differ in domain structure relative to *RRS1*, suggesting it may not be functionally equivalent. More importantly, we have also identified sequence groups with extensive evidence of positive selection (e.g., groups 2, 12, and 13) that do not contain known R genes. These groups may contain uncharacterized, functionally active R genes.

The second caveat to evolutionary profiling is that it is subject to analytical limitations. For example, positive selection was not detected in groups with two sequences (Table 2), probably reflecting low statistical power in tests with few sequences and also in tests that average ω among codons (Anisimova et al. 2001). The statistical power of the test is also sensitive to factors such as sequence length and identity (Anisimova et al. 2001). To determine whether these factors underlie our results, we contrasted four characteristics (the number of sequences, sequence length, sequence identity, and tree length) among groups. None of these four factors differed significantly between groups with and without positively selected sites, suggesting that sampling biases do not underlie detection of positively selected sites. A final analyti-

cal consideration is that LR tests with M8 tend to be conservative when the LR statistic is assumed to be χ^2 distributed (Anisimova et al. 2001). The limitations of the ML method, as applied here, tend to make it conservative, and it therefore is likely that some positively selected sites were not detected in our analyses (a Type II error). On the other hand, this conservative bias suggests Type I errors may be rare.

Groups with and without positively selected sites could vary in biological factors other than their function (or lack thereof) in disease resistance. One such difference is ectopic recombination, or gene conversion, which could affect test statistics. We analyzed sequence groups for evidence of ectopic recombination. Only two groups (groups 7 and 20) contained evidence of gene conversion, suggesting both that gene conversion is not widespread within groups and that gene conversion does not contribute substantially to differences between groups with and without positively selected sites. We should note, however, that gene conversion may help explain some of the inconsistent results based on group 7 (see below). A second potential biological difference is that there could be differences in gene activity between the positively selected groups and the groups without positively selected sites. To investigate this possibility, we measured EST hits as a proxy for gene expression. There was no overall difference between groups with and without positively selected sites.

Genic Location of Positively Selected Sites

The ML method is thought to be more effective when applied to entire genes, as opposed to separate sequence domains (Swanson and Yang 2002). It therefore is not surprising that analyses on separate domains identified fewer positively selected sites (105 vs.116) in fewer groups (8 vs. 10) than whole-sequence analysis. Nonetheless, the results of domain and whole-sequence analyses were fairly consistent except for group 7, in which 31 positively selected sites were identified in the LRR domain analysis, but no positively selected sites were detected with whole-sequence analysis (Tables 2 and 3). At present, the reasons for this discrepancy are unclear, but it is possible that gene conversion in this group (see Results) contributes to differences between whole-sequence and domain analyses. Positive selection for group 7 was detected with the test of M1 versus M0. With the exception of group 7, results were consistent between domain and whole-sequence analyses in two ways. First, the same amino acid sites were identified as positively selected. For example, without group 7, 75% of the 74 sites identified in domain analyses were also identified in whole-sequence analyses. Second, both analyses detected positive selection primarily in LRR domains (Table 3). The proportion of LRR sites identified with domain analyses (97%) was greater than that detected with whole-sequence analyses (70%), but this difference may reflect relatively low statistical power in relatively short CC, TIR, and NBS domains.

One important characteristic of the ML approach is that it identifies positively selected sites without a priori delimitation of regions. Thus, our approach differs substantially from previous analyses of NBS-LRR sequences, because most previous analyses first targeted subsequences within LRRs before applying tests for selection (Parniske et al. 1997; Wang et al. 1998; Bergelson et al. 2001). Nonetheless, prior studies have documented that positive selection is present in particular subdomains of the LRR (Parniske et al. 1997; Meyers et al. 1998; Wang et al. 1998). Our results corroborate these earlier findings and indicate that positive selection is predominantly targeted on the LRR region, particularly the E_4C_5 submotif. Unfortunately, we cannot make direct comparisons between our results and previous results because previous papers calculated ω by averaging among codons and thus did not identify individual codons in which positive selection has occurred.

We mapped the location of positively selected sites from whole-sequence analysis onto LRR secondary structure. One intriguing result from this exercise is that secondary-structure prediction identified LRR regions that were not detected by Pfam. Although this discrepancy occurred in only two sequence groups (groups 11 and 18; Fig. 2), these results suggest that secondary-structure analyses could be employed for prediction and delineation of LRR domains. In PRI, for which the tertiary structure has been solved, each LRR forms a short region of β strand, followed by a loop, a region of α-helix, and another loop that leads to a second LRR (Kobe and Deisenhofer 1994). The LRRs are arranged so that the molecule resembles a horseshoe, with the β-sheets lining the inner face and the α-helices lining the outer face. It is hypothesized that the interactions of PRI with its ligands are mediated primarily by β-sheets (Kobe and Deisenhofer 1995b). In *Arabidopsis* NBS-LRR proteins, the E_4C_5 regions adopt a β-strand-loop structure and are frequent targets of positive selection. Amino acid residues outside the E_4C_5 are often conserved among the repeats of a single-sequence group, consistent with the hypothesis that non-E_4C_5 residues are involved in interactions between consecutive motifs (Fig. 2; Kobe and Deisenhofer 1995a; Jones and Jones 1997).

One surprising aspect of our study is that 30% (34 of 116) of positively selected sites identified in whole-sequence analyses were not in LRRs. Based on whole-sequence analysis, 4 sites were in either CC or TIR domains, 7 sites were in the NBS domain, and 23 sites were in domain regions not identified by Pfam. Some of the latter may actually be located in poorly defined LRR domains, but there remains an appreciable number of positively selected sites outside the LRR. Positive selection in non-LRR regions has been documented previously. For example, a study of the flax *L* locus showed that the TIR region contributes to resistance specificity and may be under positive selection (Luck et al. 2000). We cannot assess directly the functional importance of the positively selected sites in non-LRR regions, but it is possible that these sites also play a role in intra- or intermolecular interactions in protein complexes during recognition and signaling.

We found a high incidence of alignment gaps (or indels) in LRR regions. This LRR elasticity was found in groups both with and without evidence of positive selection, but the consensus LRRs contained few gaps in the E_4C_5 region and predicted β-sheets (data not shown). These observations may have structural implications; indels in the predicted coil-helix-coil region may confer additional conformational variability that could lead to altered recognition specificities. Furthermore, predicted α-helices do not appear in a regular pattern through all groups (Fig. 2); α-Helices occur in alternate repeats, consecutive repeats, or not at all. The distribution of α-helices in the secondary structure may also have an affect on the tertiary structure of the LRR domain. Taken together, we believe that this study suggests that several factors may both individually and collectively influence the evolution of new resistance specificities (Fig. 3). These factors include variation in the position of indels in the backbone of the LRR

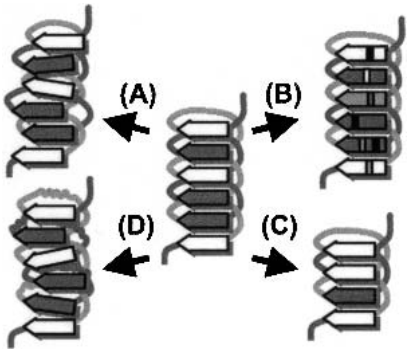


Figure 3 A model of the evolutionary processes in LRRs that generate novel resistance specificities. (*A*) Indels in the backbone of the LRR domain. (*B*) Hypervariability in the E_4C_5 region. (*C*) Expansion/contraction in the overall number of LRR units. (*D*) Changes in secondary structure resulting from amino acid changes outside of the E_4C_5 region.

domain, hypervariability in the E_4C_5 region, changes in secondary structure resulting from amino acid substitutions in the backbone, and expansion/contraction in the overall number of LRR units.

METHODS

Sequences and Alignment

Complete amino acid sequences of 163 genes were retrieved from At-RGenes (http://www.niblrrs.ucdavis.edu/At_RGenes/) in January 2001. The 163 sequences were aligned with CLUSTALW (Thompson et al. 1994), using default settings. Because the genes were too divergent for analysis of positive selection, we partitioned sequences into individual groups in two steps. First we obtained a neighbor-joining phylogeny of the 163 genes using PAUP* version 4.0b6 (Swofford 2000). Second, based on this phylogeny and an identity matrix, we partitioned 22 phylogenetic clades into sequence groups by three criteria: (1) the average identity in a group was >50% at the amino acid level (Bergelson et al. 2001), (2) the number of conservative amino acid substitutions was >50%, and (3) the percentage of gapped residues was <25%. The average identity within a group was calculated with PAUP*; other statistics were calculated with GeneDoc version 2.5 (Nicholas et al. 1997).

After grouping, we iteratively realigned sequences within each group using CLUSTALW and reestimated sequence identities. During the grouping process we eliminated 60 sequences that either did not fall into groups with the criteria outlined above or were lacking NBS-LRR protein domains. In some cases, we also trimmed the C-terminal ends of sequences that could not be aligned reliably (details available from authors). The remaining sequences contained NBS and LRR regions, and most contained either a TIR or a CC region (Table 1). Amino acid alignments were converted back into nucleotide sequence alignments, which were used in analyses. Alignments are available at http://bgbox.bio.uci.edu; gapped regions of alignments were not considered in subsequent positive selection analysis.

For some analyses, we divided amino acid and nucleotide alignments from groups with more than two sequences into putative TIR, NBS, CC-NBS, and LRR domains. For the groups that have sequences similar to known R genes, domain boundaries were determined from characterization of *Arabidopsis* R genes *RPP5* (Parker et al. 1997), *RPS5* (Warren et al. 1998), *RPP1* (Botella et al. 1998), *RPP8* (McDowell et al. 1998), *RPS4* (Gassmann et al. 1999), and *RPP13* (Bittner-Eddy et al. 2000) genes (Table 1). For groups that had no published information, domains were determined for each sequence with Pfam (Sonnhammer et al. 1998) and consensus domain regions were identified within each group.

Sequence Analyses

The ω ratio was calculated with the computer program Codeml from PAML (Yang 1997; Yang et al. 2000). The relative fit of codon substitution models was evaluated with likelihood ratio (LR) statistics, which are assumed to be χ^2 distributed with degrees of freedom equal to the difference in the number of parameters between models. LR tests for positive selection compare a model in which there is a class of sites with $\omega > 1$ against a model that does not allow for this class. We employed two LR tests to compare codon substitution models (M). Yang, Nielsen, and colleagues described the substitution models in detail (Nielsen and Yang 1998; Yang et al. 2000), and here we use their notation. The first compared M1 and M0; comparison between the two models identified phylogenetic branches with $\omega > 1$ in which positive selection had acted.

A second, more specific approach to detect positive selection is to study variation in ω among sites. This variation is tested with an additional LR test between M7 and M8. This test has been applied widely (Yang et al. 2000; Swanson et al. 2001b), but for this study it is important to note three test characteristics. First, detection of positive selection requires significant differences between M7 and M8 and estimates of ω that exceed 1. Second, under M8 it is possible to estimate the proportion of sites that are under positive selection, and this proportion is denoted P_1. Third, the application of these models requires a topological, or phylogenetic, assumption. For each sequence group, PAML analyses were applied assuming the maximum parsimony (MP) tree obtained from PAUP* branch-and-bound searches. For groups in which there was no single MP tree, the neighbor-joining (NJ) tree was assumed. It should be noted, however, that the ML approach is relatively insensitive to topological assumptions (Yang et al. 2000).

Positively selected sites were identified under M8 with the Bayesian approach implemented in PAML (Nielsen and Yang 1998; Yang et al. 2000). From groups with evidence of positive selection, based on the LR test, we further examined sites that had a >90% posterior probability of being in the $\omega > 1$ class. It is also important to note that for groups of two sequences, the only appropriate LR test is that between M1 and M0. In these cases, ω was fixed at 1 for M0, whereas ω was estimated for M1.

We mapped positively selected sites onto the secondary structure of each protein. The protein secondary structure was predicted on the complete amino acid sequences of each group with SSPro (Baldi et al. 1999), using default settings. This program assigns the highest probability secondary structure—either α-helix (H), β-strand (E), or coil (C)—to each amino acid residue. Results were mapped onto the amino acid alignment with GeneDoc (Nicholas et al. 1997).

Gene conversion was assessed by the method of Sawyer (1989), as implemented in the program Geneconv (http://www.math.wustl.edu/~sawyer/geneconv/). The test was applied to the nucleotide alignments of the 14 groups that contained three or more sequences and considered only synonymous sites. Amino acid differences among sequences were not appropriate for this test, for two reasons. First, the amino acid differences may be driven by positive selection, but the test assumes sequence differences are selectively neutral. Second, amino acid differences among sequences are clustered in LRR regions, thus potentially causing spurious results. Geneconv reports global *P*-values based on an entire alignment; significance was based on these *P*-values after correction for multiple tests.

ACKNOWLEDGMENTS

We thank Z. Yang and anonymous reviewers for useful advice and suggestions, and P. Tiffin and R. Michelmore for discussion. This work was supported by a UC-MEXUS Scholarship and a fellowship from the School of Biological Sciences, University of California Irvine to M.M.P., by NSF grants 98-15855 and 01-13498 to B.S.G. B.C.M. is supported by NSF grant 99-75971.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Anisimova, M., Bielawski, J.P., and Yang, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Bio. Evol.* **18:** 1585–1592.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. 1999. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* **15:** 937–946.

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

Bergelson, J., Kreitman, M., Stahl, E.A., and Tian, D. 2001. Evolutionary dynamics of plant R-genes. *Science* **292:** 2281–2285.

Bishop, J.G., Dean, A.M., and Mitchell-Olds, T. 2000. Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. *Proc. Natl. Acad. Sci.* **97:** 5322–5327.

Bittner-Eddy, P.D., Crute, E.B., Holub, J.L., and Beynon, J.L. 2000. *RPP13* is a simple locus in *Arabidopsis thaliana* for alleles that specifiy downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. *Plant J.* **21:** 177–188.

Botella, M.A., Parker, J.E., Frost, L.N., Bittner-Eddy, P.D., Beynon, J.L., Daniels, M.J., Holub, E.B., and Jones, J.D.G. 1998. Three genes of the *Arabidopsis RPP1* complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *Plant Cell* **10:** 1847–1860.

Burkhard, P., Stetefeld, J., and Strelkov, S.V. 2001. Coiled coils: A highly versatile protein folding motif. *Trends Cell Biol.* **11:** 82–88.

Bush, R.M. 2001. Predicting adaptive evolution. *Nat. Rev. Genet.* **2:** 387–392.

Dangl, J.L. and Jones, J.D.G. 2001. Plant pathogens and integrated defence responses to infection. *Nature* **411:** 826–833.

Deslandes, L., Olivier, J., Theulières, F., Hirsch, J., Feng, D.-X., Bittner-Eddy, P., Beynon, J., and Marco, Y 2002. Resistance to *Ralstonia solaneacearum* in *Arabidopsis thaliana* is conferred by the recessive *RRS1-R* gene, a member of a novel family of resistance genes. *Proc. Natl. Acad. Sci.* **99:** 2404–2409

Ellis, J.G., Lawrence, G.J., Luck, J.E., and Dodds, P.N. 1999. Identification of regions in alleles of the flax rust resistance gene L that determine differences in gene-for-gene specificity. *Plant Cell* **11:** 495–506.

Ellis, J., Dodds, P., and Pryor, T. 2000. The generation of plant disease resistance gene specificities. *Trends Plant Sci.* **5:** 373–379.

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

Gassmann, W., Hinsch, M.E., and Staskawicz, B.J. 1999. The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant J.* **20:** 265–277.

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

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

Inohara, N., Koseki, T., del Peso, L., Hu, Y.M., Yee, C., Chen, S., Carrio, R., Merino, J., Liu, D., Ni, J., et al. 1999. Nod1, an Apaf-1-like activator of caspase-9 and nuclear factor-κ B. *J. Biol. Chem.* **274:** 14560–14567.

Jia, Y., McAdams, S.A., Bryan, G.T., Hershey, H.P., and Valent, B. 2000. Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *EMBO J.* **19:** 4004–4014.

Jones, D.A. and Jones, J.D.G. 1997. The role of leucine-rich repeat proteins in plant defences. *Adv. Bot. Res.* **24:** 90–167.

Kimura, M. and Ohta, T. 1974. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci.* **71:** 2848–2852.

Kobe, B. and Deisenhofer, J. 1993. Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature* **366:** 751–756.

———. 1994. The leucine-rich repeat—a versatile binding motif. *TIBS* **19:** 415–421.

———. 1995a. Proteins with leucine-rich repeats. *Curr. Oppin. Struct. Biol.* **5:** 409–416.

———. 1995b. A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature* **374:** 183–186.

Kobe, B. and Kajava, A.V. 2001. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **11:** 725–732.

Kopp, E.B. and Medzhitov, R. 1999. The Toll-receptor family and central of innate immunity. *Curr. Opin. Immun.* **11:** 13–18.

Liker, E., Fernandez, E., Izurralde, E., and Conti, E. 2000. The structure of the mRNA export factor TAP reveals a cis arrangement of a non/canonical RNP domain and an LRR domain. *EMBO J.* **19:** 5587–5598.

Luck, J.E., Lawrence, G.L., Dodds, P.N., Sheperd, K.W., and Ellis, J.G. 2000. Regions outside of the leucine-rich repeats of flax rust resistance proteins play a role in specificity determination. *Plant Cell* **12:** 1367–1377.

Marino, M., Braun, L., Cossart, P., and Ghosh, P. 1999. Structure of the InlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen *L. monocytogenes*. *Mol. Cell* **4:** 1063–1072.

McDowell, J.M., Dhandaydham, M., Long, T.A., Aarts, M.G.M., Goff, S., Holub, E.B., and Dangl, J.L. 1998. Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* locus of *Arabidopsis*. *Plant Cell* **10:** 1861–1874.

Meyers, B.C., Shen, K.A., Rohani, P., Gaut, B.S., and Michelmore, R.W. 1998. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. *Plant Cell* **11:** 1833–1846.

Nicholas, K.B., Nicholas, H.B., and Deerfield, D.W. 1997. GeneDoc: Analysis and Visualization of Genetic Variation. *EMBNEW.NEWS* **4:** 14.

Nielsen, R. and Yang, Z.H. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936.

Noel, L., Moores, T.L., van der Biezen, E.A., Parniske, M., Daniels, M.J., Parker, J.E., and Jones, J.D.G. 1999. Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. *Plant Cell* **11:** 2099–2111.

Parker, J.E., Coleman, M.J., Szabo, V., Frost, L.N., Schmidt, R., van der Biezen, E.A., Moores, T.L., 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.

Peek, A.S., Souza, V., Eguiarte, L.E., and Gaut, B.S. 2001. The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (*fimA*) from *Escherichia coli*. *J. Mol. Evol.* **52:** 193–204.

Price, S.R., Evans, P.R., and Nagai, K. 1998. Crystal structure of the spliceosomal U2B″-U2A′ protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **394:** 645–650.

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

Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6:** 526–538.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M. Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., et al. 2002 Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296:** 141–145

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26:** 320–322.

Srinivasula, S.M., Ahmad, M., Fernandes-Alnemri, T., and Alnemri, E.S. 1998. Autoactivation of procaspase-9 by Apaf-1-mediated oligomerization. *Mol. Cell* **1:** 949–957.

Swanson, W.J. and Yang, Z. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* **19:** 49–57.

Swanson, W.J., Aquadro, C.F., and Vacquier, V.D. 2001a.

Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive Darwinian selection of sperm lysin. *Mol. Biol. Evol.* **18:** 376–383.

Swanson, W.J., Yang, Z., Wolfner, M.F., and Aquadro, C.F. 2001b. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci.* **98:** 2509–2514.

Swofford, D.L. 2000. PAUP* phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, MA..

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.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

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., et al. 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.

Warren, R., Henk, A., Mowery, P., Holub, E.B., and Innes, R.W. 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.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13:** 555–556.

———. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Bio. Evol.* **15:** 568–573.

Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *TREE* **15:** 496–503.

Yang, Z. and Nielsen, R. 1998. Synonymous and non-synonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46:** 409–418.

Yang, Z.H., Nielsen, R., Goldman, N., and Pedersen, A.M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.

Zhang, H., Seabra, M.C., and Deisenhofer, J. 2000. Crystal structure of Rab geranylgeranyltransferase at 2.0 Å resolution. *Structure* **8:** 241–251.

WEB SITE REFERENCES

http://bgbox.bio.uci.edu; The Web site from which aligned LRR sequences from this study can be downloaded.

http://mips.gsf.de/proj/thal/; The Munich Information Center for Protein Sequences contains *Arabidopsis thaliana* EST information.

http://www.math.wustl.edu/~sawyer/geneconv/; The location of Genecov, a program that tests for gene conversion.

http://www.niblrrs.ucdavis.edu/At_RGenes/; The database of *Arabidopsis* NBS-LRR encoding disease resistance gene homologs.

Received February 5, 2002; accepted in revised form June 12, 2002.