Proyecto Programa de Formación

"Actualización de conocimientos, metodologías y tendencias en biología molecular de plantas"

Seminario de Difusión

"Evolución de los genes de resistencia a patógenos en plantas"

Instituto de Biología Vegetal y Biotecnología
Universidad de Talca
Talca, 28 de Julio de 2003

Resistance gene evolution

Pamela C Ronald

Plant resistance genes are highly polymorphic and have diverse recognition specificities. These genes often occur as members of clustered gene families that have evolved through duplication and diversification. Regions of nucleotides conserved between family members and flanking sequences facilitate equal or unequal recombination events. Transposition contributes to allelic diversity. Resistance gene clusters appear to evolve more rapidly than other regions of the genome, and domains responsible for recognitional specificity, such as the leucine-rich repeat domain, are subject to adaptive selection.

Addresses
Department of Plant Pathology, 1 Shields Avenue, University of California at Davis, Davis, CA 95616, USA; e-mail: pcronald@ucdavis.edu

Current Opinion in Plant Biology 1998, 1:294–298

http://biomednet.com/elecref/1369526600100294

 ISSN 1369-5266

Abbreviations

dn	nonsynonymous substitutions
ds	synonymous substitutions
kb	kilobase
LRR	leucine-rich repeat
MHC	major histocompatibility complex
pb	base pair
R	resistance
RLA	resistance-like analog
TE	transposable element

Introduction

The continued survival of most organisms depends on the presence of specific genetic systems that maintain diversity in the face of a changing environment. Classic examples include antigenic variation in trypanosomes and immunoglobulin gene formation in mammals. Similarly, most plant species contain a large number of highly polymorphic disease resistance (R) genes, most of which share common structural domains [1]. It has long been speculated that DNA rearrangements play a key role in the evolution of these genes, thus allowing plants to generate new resistances to match the changing pattern of pathogen virulence [2,3]. In support of this hypothesis, studies of the maize disease resistance locus *rp1* revealed that recombination of flanking markers was associated with the creation of novel resistance phenotypes [4]. This review focuses on the evolution of R genes using recent information gained from molecular genetic analysis of R genes.

Genomic organization of resistance genes

R genes of different structural classes conferring resistance to diverse pathogens are present in the plant genome in nonoverlapping discrete clusters (groups of genes of related structure and/or function) [5,6••]. Within a discrete cluster, members of an R gene family are often arranged as tandem direct repeats, which is consistent with an origin through gene duplication and their continued evolution through unequal exchange. There are also R loci which consist of a single gene with multiple distinct alleles. For example, the L rust resistance locus in flax has 13 different specificities [7].

In addition to R genes with known specificities, resistance-like analogs (RGAs sequences, whose function is unknown, map as clusters in rice, *Arabidopsis*, potato, tomato and soybean [8,9•,10,11•]. On the basis of comparative mapping studies of monocot RGAs, Leister *et al.* [9•] suggest that R genes diverge more rapidly than the rest of the genome through sequence divergence or ectopic recombination. For example, using rice and barley RGAs, for mapping on the foxtail millet map, 17 loci were identified but only five were found at syntenic map locations. Similarly, the barley *mlo* and *Rpg1* genes, conferring resistance to the powdery mildew and stem rust fungi respectively, are not found in the syntenic region in the rice genome although the order of flanking markers is conserved between barley and rice [12,13]. These results contrast to the synteny observed in most other cereal genes [14].

Although the clustering and rapid evolution of R genes suggests that a gene conferring resistance to one pathogen species could evolve to recognize a different pathogen species, there is no direct evidence yet to support this hypothesis. Future cloning and sequencing of linked genes conferring resistance to different pathogens may eventually demonstrate such a common evolutionary origin.

In plants, leucine-rich repeat (LRR) domains of R gene products show similarity to domains in diverse proteins controlling cell–cell communication in development and signaling, suggesting that both classes of genes may have evolved through duplication and divergence of common ancestors [15–17]. To date, genes controlling development have not been found within R gene clusters.

Duplication and recombination

Duplication plays a central role in creating complex genetic systems [18]. Duplication can create new loci, alter gene family number through recombination, or generate repeated sequences within a gene. For example, studies of the major histocompatibility complex (MHC) showed that human and mouse genomes contain regions that apparently emerged as a result of chromosomal duplication [19]. Similarly, at least two additional clusters of the

resistance gene *Cf-9* homologues on the short arm of chromosome 1 of tomato have been found [20] and chromosomal duplications created entirely new clusters of R genes in lettuce [21].

Recombination can lead to amplification or reduction of the number of R gene family members. For example, the presence of two nearly identical functional *Cf-2* genes suggests that they arose through a recent gene duplication event [22]. Analysis of the *Cf-2/Cf-5* locus, where only a few sequences homologous to *Cf* genes reside, has revealed a rare disease-susceptible recombinant that arose via an unequal crossover event leading to a reduction of the *Cf* homologue numbers [22]. Molecular analysis of five *Cf-4/Cf-9* disease-sensitive recombinants demonstrated that each was generated by chromosomal mispairing of intergenic sequences and unequal crossing over [23•]. The *Xa21* multigene family encoding resistance to bacterial blight in rice contains a large duplication of at least 17 kb; one of the duplicated genes confers the same race-specific resistance as *Xa21* [24•,25•]. The presumed duplication and diversification of the tomato *Pto* gene family led to the generation of alternative recognition capabilities of the encoded proteins [26,27]. Finally, it has been proposed that the flax *M* rust resistance locus, which carries tandemly arrayed specificities, evolved from a rare duplication of an ancestral *M* gene [28]. Repeated DNA flanking the locus may have enhanced subsequent duplication through unequal crossing over events. These results indicate that gene duplication is a major force in R gene evolution.

In some cases, recombination between diverged family members occurs at highly conserved stretches of nucleotides. For instance, a large proportion of recombination events at the *Xa21* locus were localized to a highly conserved domain in the 5′ coding domain, resulting in new promoter/gene combinations [24•]. Similarly, the recombination exchange site in the *M* mutants can be localized to a 45 bp region that is invariant between LRR repeats [29•]. Such recombination events can lead to gross structural changes.

In addition to swapping of large gene regions, recombination can lead to fine structural changes within a gene. The repetitive structure of LRR coding regions could facilitate intragenic (and intergenic) genic recombination leading to expansion and contraction of the LRR number, as demonstrated in mutants of *M* and *Rpp5*. Whereas the wild-type *M* gene contains two DNA repeats encoding LRRs, spontaneous mutants contain a single repeat [29•]. The mutant alleles with a single LRR repeat may have been generated by an unequal exchange between the first repeat in one *M* gene and the second repeat in its homolog [29•]. A fast-neutron generated susceptible Rpp5 mutant contains an intragenic duplication of four complete LRRs. This duplication may have arisen from an unequal crossing over event between two sequences of identity in the LRRs [30•].

Rapid sequence exchange among tandemly repeated gene families generally leads to sequence homogenization between members [6••]. How can variability, therefore be maintained in R gene family members? To address this question, Parniske *et al.* [6••] sequenced three haplotypes at the *Cf-4/Cf-9* locus. Comparison of intergenic regions revealed a high degree of sequence rearrangements, whereas in the coding regions a patchwork of sequence similarities was observed [6••]. The observed variable sequence patches could result either from successive rounds of reciprocal recombination or from gene conversion events. In a homozygous background, the *Cf-9* gene was found to be very stable. In contrast, the meiotic stability of *Cf-9* was dramatically reduced in a *Cf-4/Cf-9* transheterozygous background. Parniske *et al.* [6••] propose that the polymorphism of the intergenic regions suppresses unequal recombination in homozygotes and sister chromatids, thereby preventing sequence homogenization of the gene family. In this situation, recombination between regions of high homology within a coding region may actually contribute to the maintenance of a useful combination of R gene specificities. In a *Cf-4/Cf-9* transheterozygous background, homologous sequences aligned unequally are used as recombination templates. Such unequal recombination alters the number of gene family members as well as the composition of the clusters, resulting in increased variation within the population.

Lesion-mimic mutants

Recombination at R loci can also lead to the generation of lesion-mimic mutants which display a phenotype similar to the hypersensitive response controlled by R genes, but in the absence of pathogen. This observation led to the hypothesis that similar types of genes are involved in both phenotypes and that the lesion-mimic mutant genes may be derived from R gene loci [2]. The recovery of four rust resistance *Rp1* alleles with lesion-mimic phenotypes is the most direct evidence to date that at least some of the lesion-mimic mutants are variants of race-specific R genes [31]. Flanking marker analysis indicated that at least two of the four mutants were derived from crossover events.

The barley powdery mildew resistance gene *mlo* and the *lsd* (lesion stimulating disease) and *acd2* (accelerated cell death) genes from *Arabidopsis* provide other examples of genes displaying a lesion-mimic mutant phenotype together with defense responses associated with disease resistance [32–34]. In these cases, however, no genes conferring race-specific resistance have yet been mapped to these loci. Moreover, these lesion-mimic genes encode proteins with structures distinct from other cloned R genes, indicating that not all lesion mimic-mutants have a direct evolutionary link to R genes [32,35].

Adaptive selection of pathogen recognition domains

Characterization of nucleotide substitution patterns in R gene families has provided insight into the function and evolution of particular coding domains. For the investigation of function, the ratio of nucleotide substitutions that lead to amino acid replacements (nonsynonymous substitutions, dn) and nucleotide substitutions that do not alter amino acids (synonymous substitutions, ds) is particularly informative. In most protein-coding genes, the dn/ds ratio is less than one; this observation is consistent with functional constraint against amino acid replacements [36]. Conversely, a dn/ds ratio significantly greater than one indicates that adaptive selection events have fueled divergence between genes [37,38]. Evidence of adaptive selection is rare but appears to be most common in gene regions encoding surface antigens of parasites or viruses [39]. It is expected that regions that bind ligand will be subject to stronger adaptive selection than regions that play a structural role. For example, the antigen recognition site of alleles at the class I MHC loci in human and mouse displays a dn/ds ratio greater than one, indicating that the antigen recognition site is subject to strong adaptive selection events, whereas structural regions of the protein are not [37].

Analysis of 11 *Cf* gene family members revealed that the predicted solvent-exposed residues of the β-strand/β-turn region of the LRR domain exhibit increased dn/ds ratios relative to other residues in the LRR domain, suggesting that solvent-exposed residues play a role in ligand binding [6••]. Similarly, a comparison of nucleotide substitutions in the LRR coding regions of *Xa21* and gene family member *Xa21D* revealed that, although *Xa21* and *Xa21D* share 99.1% sequence identity, nonsynonymous substitutions occur significantly more frequently than do synonymous substitutions in the LRR; this result is consistent with the LRR's putative role in ligand binding [25•]. These results indicate that the LRR domain, which governs race-specific pathogen recognition, is subject to adaptive evolution. Diversity at the LRR domain would provide an evolutionary advantage for recognizing, binding, and defending against a broad array of pathogens.

Diversification of *R* gene family members by transposon-like elements

The human MHC class 2 region is among the most polymorphic part of the human genome. Multiple repetitive sequences representing more than 20 different families have been characterized in the MHC region [40]. Part of the interspecific and intraspecific variation observed in the MHC is caused by different integration patterns of retroelements. Comparative studies of different human haplotypes and primate species revealed that retroelement insertions have contributed to genome plasticity of the MHC during primate evolution. Retroelements also contribute to recombination and genomic instability by serving as sites for recombination and translocation events [40].

In plants, it has long been hypothesized that transposable elements (TEs, or transposons) play a role in the reconstruction of genomes in response to environmental stresses such as tissue culture, irradiation or pathogen infection [41,42]. In partial support of this hypothesis, Pouteau *et al.* [43] demonstrated that the transcription of the tobacco retrotransposon *Tnt1* is induced by a broad spectrum of microbial and fungal elicitors. TE insertion into and excision from regulatory and coding regions can change the coding capacity and expression patterns of the gene [44–46].

There is no evidence yet for the generation of new specificity at R gene loci as a result of the insertion and excision of a TE. It has been shown, however, that TE-induced gene alterations can cause R gene inactivation. For example, in the case of the maize fungal resistance gene *Hm1*, which confers resistance to *Cochliobolus carbonum* race 1, a 315 bp insertion (designated *dHBr*) was found in a mutant allele of this gene [46]. Moreover, the insertion of a transposon (a 256 bp element named *Drone*) disrupted the *Hm1*-conferred resistance in an inbred line of maize and, as a result, led to the genesis of the leaf spot and ear rot disease of maize in 1938 [47]. In flax, two mutants of the L6 gene for rust resistance carry small (300 bp) insertion elements which inactivate the gene [48].

In rice, transposon-like elements appear to be a major source of variability of the *Xa21*-gene family members. Seventeen transposon-like elements grouped into 11 families, including three families of miniature inverted repeat TEs (MITEs), five novel elements, *Ds*-like elements, a *CACTA*-like element and a retrotransposable element are present at the *Xa21* locus [24•,49,50]. Integration of two of these elements into coding sequences creates open reading frames (ORFs) that encode truncated proteins. At least one of these truncated proteins can confer an attenuated resistance with *Xa21* specificity [25•]. TE insertion into 5′ and 3′ flanking regulatory regions was also observed. Many of the elements seem to have been active over the entire evolutionary period of the *Xa21* gene family members [24•].

It is tempting to speculate that TEs contribute to the evolution of R gene diversity. Movement of these transposons in response to pathogen-induced stress would provide genetic plasticity with a possible selective advantage. As the insertion of TEs at the *Hm1*, *I-6* and *Xa21* loci results in loss of function or impaired function, however, the question of whether TE movement can confer a selective advantage remains open. In addition, more sequence information is needed to determine if TEs are more abundant or more active at R gene loci than in other regions of the genome.

Conclusions

Common themes in the evolution of R genes and gene families are emerging, on the basis of sequence analysis of cloned R genes. First, duplication and subsequent divergence of a progenitor R gene can amplify or create additional clusters of a gene family. Second, unequal recombination at intergenic regions between family members creates additional copy number variability within the population. Third, recombination at highly conserved regions in intragenic regions allows for the formation of novel gene combinations. Fourth, adaptive evolution of LRR domains allows for rapid generation of altered recognition specificities. Finally, movement of TEs may result in further allelic diversity.

Acknowledgements

The author is grateful to PA Anderson, JL Dangl, SH Hulbert, G Johal, DA Jones, JDG Jones, J McDowell, RW Michelmore, B Meyers, T Richter and W-Y Song for helpful discussions and critical reading of the manuscript.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Baker B, Zambryski P, Staskawicz B, Dinesh-Kumar SP: **Signaling in plant–microbe interactions.** *Science* 1997, **276**:726-733.

2. Pryor AJ: **The origin and structure of fungal disease resistance genes in plants.** *Trends Genet* 1987, **3**:157-161.

3. Pryor AJ, Ellis J: **The genetic complexity of fungal disease resistance genes in plants.** *Adv Plant Pathol* 1993, **10**:281-305.

4. Richter TE, Pryor AJ, Bennetzen JL, Hulbert SH: **New rust resistance specificities associated with recombinations at the Rp1 complex in maize.** *Genetics* 1995, **141**:373-381.

5. Witsenboer H, Vogel J, Michelmore RW: **Identification, genetic localization, and allelic diversity of selectively amplified microsatellite polymorphic loci in lettuce and wild relatives (*Lactuca spp.*).** *Genome* 1997, **40**:923-936.

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

This paper represents the first comprehensive sequence comparisons of three R loci haplotypes. Their analysis indicates that solvent-exposed residues play a role in ligand binding.

7. Islam MR, Shepherd KW: **Present status of genetics of rust resistance in flax.** *Euph* 1991, **55**:255-267.

8. Leister D, Ballvora A, Salamini F, Gebhardt C: **A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants.** *Nat Genet* 1996, **14**:421-429.

9. • Leister D, Kurth J, Laurie DA, Yano M, Sasaki T, Devos K, Graner A, Schulze-Lefert P: **Rapid reorganization of resistance gene homologues in cereal genomes.** *Proc Natl Acad Sci USA* 1998, **95**:370-375.

This paper provides additional evidence that resistance-like gene sequences (RLGs) map as clusters. On the basis of comparative mapping studies of monocot RLGs, the authors suggest that R genes diverge more rapidly than the rest of the genome.

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

11. • Botella MA, Coleman MJ, Hughes DE, Nishimura MT, Jones JDG, Somerville SC: **Map positions of 47 *Arabidopsis* sequences with sequence similarity to disease resistance genes.** *Plant J* 1997, **12**:1197-1211.

This impressive mapping of 47 RLGs in *Arabidopsis* provides a starting point for a more in depth characterization of each RLG.

12. Panstruga R, Buschges R, Freialdenhoven A, Ropenack E, Schulze-Lefert P: **Insights into non-race-specific resistance: the *mlo*-controlled resistance in barley to powdery mildew.** *Plant disease resistance gene function. An EMBO workshop: 1997, May 18–20:* Maratea, Italy.

13. Kilian A, Chen J, Han F, Steffenson B, Kleinhofs A: **Towards map-based cloning of the barley stem rust resistance genes *Rpg1* and *rpg4* using rice as an intergenomic cloning vehicle.** *Plant Mol Biol* 1997, **35**:187-195.

14. Devos KM, Gale MD: **Comparative genetics in the grasses.** *Plant Mol Biol* 1997, **35**:3-15.

15. Li J, Chory J: **A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction.** *Cell* 1997, **90**:929-938.

16. Torii KU, Mitsukawa N, Oosumi T, Matsuura Y, Yokoyama R, Whittier RF, Komeda Y: **The *Arabidopsis ERECTA* gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats.** *Plant Cell* 1996, **8**:735-746.

17. Clark SE, Williams RW, Meyerowitz EM: **The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in *Arabidopsis*.** *Cell* 1997, **89**:575-585.

18. Ohno S: *Evolution by Gene Duplication.* Berlin: Springer Verlag; 1970.

19. Kasahara M, Nakaya J, Satta Y, Takahata N: **Chromosomal duplication and the emergence of the adaptive immune system.** *Trends Genet* 1997, **13**:90-92.

20. Jones JDG, Parniske M, Thomas T, Hammond-Kosack K, Romeis T, Piedras P, Tai T, Torres MA, Hatzixanthis K, Brading P, Wulff BH: **Evolution and function of tomato Cf-disease resistance genes.** *Plant disease resistance gene function. An EMBO workshop: 1997, May 18–20:* Maratea, Italy.

21. Paran I, Yesseli RV, Michelmore RW: **Recent amplification of triose phosphate isomerase related sequences in lettuce** 1992, *Genome* **35**:627-635.

22. Dixon MS, Jones DA, Keddie JS, Thomas CM, Harrison K, Jones JDG: **The tomato *Cf-2* disease resistance locus comprises two functional genes encoding leucine-rich repeat proteins.** *Cell* 1996, **84**:451-459.

23. • Thomas CM, Jones DA, Parniske M, Harrison K, Balint-Kurti PJ, Hatzixanthis K, Jones JDG: **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* 1997, **9**:2209-2224.

This work carefully analyzes five Cf-4/Cf-9 disease sensitive recombinants and demonstrates that each was generated by chromosomal mispairing of intergenic sequences and unequal crossing over.

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

This work demonstrates that recombination, duplication and transposition contribute to the sequence diversity of the *Xa21* gene family. A large proportion of recombination events at the *Xa21* locus were localized to a highly conserved domain in the 5′coding domain resulting in new promoter/gene combinations. Seventeen transposable elements were identified, making this R gene locus the most transposon-rich of those so far characterized.

25. • Wang G-L, Ruan D-L, Song WY, Sideris S, Chen L-L, Pi L-Y, Zhang S, Zhang Z, Fauquet C, Gaut B, Ronald P: ***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* 1998, **10**:765-779.

This paper indicates that a partial resistance phenotype is conferred by an *Xa21* gene family member encoding a new class of R genes closely related to PGIPs: a presumed secreted receptor-like molecule. Its novel structure illustrates that there are likely many more R gene classes yet to be identified.

26. Martin GB, Frary A, Wu T, Brommonschenkel S, Chunwongse J, Earle ED, Tanksley SD: **A member of the tomato *Pto* gene family confers sensitivity to fenthion resulting in rapid cell death.** *Plant Cell* 1994, **6**:1543-1552.

27. Zhou J, Loh YT, Bressan RA, Martin GB: **The tomato gene *Pti1* encodes a serine/threonine kinase that is phosphorylated by *Pto* and is involved in the hypersensitive response.** *Cell* 1995, **83**:925-935.

28. Ellis J, Lawrence GJ, Finnegan EJ, Anderson PA: **Contrasting complexity of two rust resistance loci in flax.** *Proc Natl Acad Sci USA* 1995, **92**:4185-4188.

29. • Anderson PA, Lawrence GJ, Morrish BC, Ayliffe MA, Finnegan EJ, Ellis JG: **Inactivation of the flax rust resistance gene M associated with loss of a repeated unit within the leucine-rich repeat coding region.** *Plant Cell* 1997, **9**:641-651.
This work indicates that the LRRs can serve as a recombination template leading to variation in the presumed ligand recognition domain. The wild-type *M* gene contains two DNA repeats encoding LRRs, whereas spontaneous mutants contain a single repeat.

30. • Parker JE, Coleman MJ, Szabo V, Frost LN, Schmidt R, van der Biezen EA, Moores T, Dean C, Daniels MJ, Jones JDG: **The *Arabidopsis* Downy mildew resistance gene *Rpp5* shares similarity to the Toll and INterleukin-1 receptors with *N* and *L6*.** *Plant Cell* 1997, **9**:879-894.
A fast-neutron generated *Rpp5* mutant contains an intragenic duplication of four complete LRRs. Another example that the LRRs can serve as a recombination template leading to variation in the presumed ligand recognition domain.

31. Hu G, Richter T, Hulbert S, Pryor T: **Disease lesion mimicry caused by mutations in the rust resistance gene *rp1*.** *Plant Cell* 1996, **8**:1367-1376.

32. Buschges R, Hollricher K, Panstruga R, Simons G, Wolter M, Frijters A, Van Daelen R, van der Lee T, Groenendijk J, Topsch S *et al.*: **The barley *Mlo* gene: a novel control element of plant pathogen resistance.** *Cell* 1997, **88**:695-705.

33. Greenberg JT, Guo A, Klessig DF, Ausubel FM: **Programmed cell death in plants: a pathogen-triggered response activated coordinately with multiple defense functions.** *Cell* 1994, **77**:551-563.

34. Dietrich RA, Delaney TP, Uknes SJ, Ward ER, Ryals JA, Dangl JL: ***Arabidopsis* mutants simulating disease resistance response.** *Cell* 1994, **77**:565-577.

35. Dietrich RA, Richberg MH, Schmidt R, Dean C, Dangl JL: **A novel zinc finger protein is encoded by the *Arabidopsis LSD1* gene and functions as a negative regulator of plant cell death.** *Cell* 1997, **88**:685-694.

36. Kimura M: *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press; 1983.

37. Hughes AL, Nei M: **Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection.** *Nature* 1988, **335**:167-170.

38. Messier W, Stewart CB: **Episodic adaptive evolution of primate lysozymes.** *Nature* 1997, **385**:151-154.

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

40. Andersson G, Svensson A, Setterblad N, Rask L: **Retroelements in the human MHC class II region.** *Trends Genet* 1998, **14**:109-114.

41. McClintock B: **The significance of responses of the genome to challenge.** *Science* 1984, **226**:792-801.

42. Wessler S, Bureau TE, White SE: **LTR-retrotransposons and MITEs – important players in the evolution of plant genomes.** *Curr Opin Genet Dev* 1995, **5**:814-821.

43. Pouteau S, Boccara M, Grandbastion MA: **Microbial elicitors of plant defence responses activate transcription of a retrotransposon.** *Plant J* 1994, **5**:535-542.

44. McDonald JF: **Transposable elements - possible catalysts of organismic evolution.** *Trends Ecol Evol* 1995, **10**:123-126.

45. Marionette S, Wessler SR: **Retrotransposon insertion into the maize waxy gene results in tissue-specific RNA processing.** *Plant Cell* 1997, **9**:967-978.

46. Johal GS, Briggs SP: **Reductase activity encoded by the HM1 disease resistance gene in maize.** *Science* 1992, **258**:985-987.

47. Multani DS, Meeley RB, Paterson AH, Gray J, Briggs SP, Johal GS: **Plant-pathogen microevolution: molecular basis for the origin of a fungal disease in maize.** *Proc Natl Acad Sci USA* 1998, **95**:1686-1691.

48. Lawrence GJ, Finnegan EJ, Ayliffe MA, Ellis JG: **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* 1995, **7**:1195-1206.

49. Song WY, Pi LY, Bureau T, Ronald PR: **Identification and characterization of 14 transposable-like elements in the noncoding regions of the rice *Xa21* disease resistance gene family members.** *Mol Gen Genet* 1998, in press.

50. Bureau T, Ronald P, Wessler S: **A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes.** *Proc Natl Acad Sci USA* 1996, **93**:8524-8529.

Genes controlling expression of defense responses in *Arabidopsis*

Jane Glazebrook

In the past year, two regulatory defense-related genes, *EDS1l* and *COI1*, have been cloned. Several other genes with regulatory functions have been identified by mutation, including *DND1*, *PAD4*, *CPR6*, and *SSI1*. It has become clear that jasmonate signaling plays an important role in defense response signaling, and that the jasmonate and salicylic acid signaling pathways are interconnected.

Addresses
Novartis Agricultural Discovery, Institute, Inc., 3050 Science Park Rd, Suite 102, San Diego CA 92121, USA
e-mail: jane.glazebrook@nadii.novartis.com

Current Opinion in Plant Biology 1999, 2:280–286

http://biomednet.com/elecref/1369526600200280

 ISSN 1369-5266

Abbreviations

avr	avirulence
HR	hypersensitive response
ISR	induced systemic resistance
JA	jasmonic acid
LRR	leucine rich repeat
LZ	leucine zipper
NBS	nucleotide binding site
PR	pathogenesis related
R	resistance
SA	salicylic acid
SAR	systemic acquired resistance

Introduction

Plants are capable of activating a large array of defense mechanisms in response to pathogen attack. A crucial factor determining the success of these mechanisms is the speed of their activation. Consequently, there is considerable interest in understanding how plants recognize pathogen attack and control expression of defense mechanisms.

Some potential pathogens trigger a very rapid resistance response called gene-for-gene resistance. This occurs when the pathogen carries an avirulence (*avr*) gene that triggers specific recognition by a corresponding host resistance (*R*) gene. *R* gene specificity is generally quite narrow, in most cases only pathogens carrying a particular *avr* gene are recognized. Recognition is thought to be mediated by ligand–receptor binding. *R* genes have been studied extensively in recent years and several excellent reviews are available [1–3].

One of the defense mechanisms triggered by gene-for-gene resistance is programmed cell death at the infection site. This is called the hypersensitive response, or HR. Pathogens that induce the HR, or cause cell death by other means, activate a systemic resistance response called systemic acquired resistance (SAR). During SAR, levels of salicylic acid (SA) rise throughout the plant, defense genes such as pathogenesis related (PR) genes are expressed, and the plant becomes more resistant to pathogen attack. SA is a crucial component of this response. Plants that cannot accumulate SA due to the presence of a transgene that encodes an SA-degrading enzyme (*nahG*), develop an HR in response to challenge by avirulent pathogens, but do not exhibit systemic expression of defense genes and do not develop resistance to subsequent pathogen attack [4]. The nature of the systemic signal that triggers SAR is a subject of debate [5,6]. SA clearly moves from the site of the HR to other parts of the plant, but if this is the signal, it must be effective at extremely low concentration [7].

SAR is quite similar to some reactions that occur locally in response to attack by virulent (those that cause disease) or avirulent (those that trigger gene-for-gene resistance) pathogens. In general, activation of defense gene expression occurs more slowly in response to virulent pathogens than in response to avirulent pathogens. Some pathogens trigger expression of defense genes through a different signaling pathway that requires components of the jasmonic acid (JA) and ethylene signaling pathways [8]. The SA and JA pathways interact in a complicated manner that is poorly understood.

One approach to understanding the signal transduction networks that control defense mechanisms is to use genetic methods to identify signaling components and determine their roles within the network. Considerable progress has been made using this approach in *Arabidopsis*–pathogen model systems. This review will focus on recent (published in 1998 and early 1999) progress in identifying *Arabidopsis* genes that affect regulation of defense gene expression, and on what is known about their roles and relative positions in the signal transduction network. Figure 1 shows a model of how the network might be arranged (see [9], for a discussion on earlier work). Due to space limitations, *R* genes, genes studied in other plant species, and insights gained from other types of analysis will not be discussed in detail.

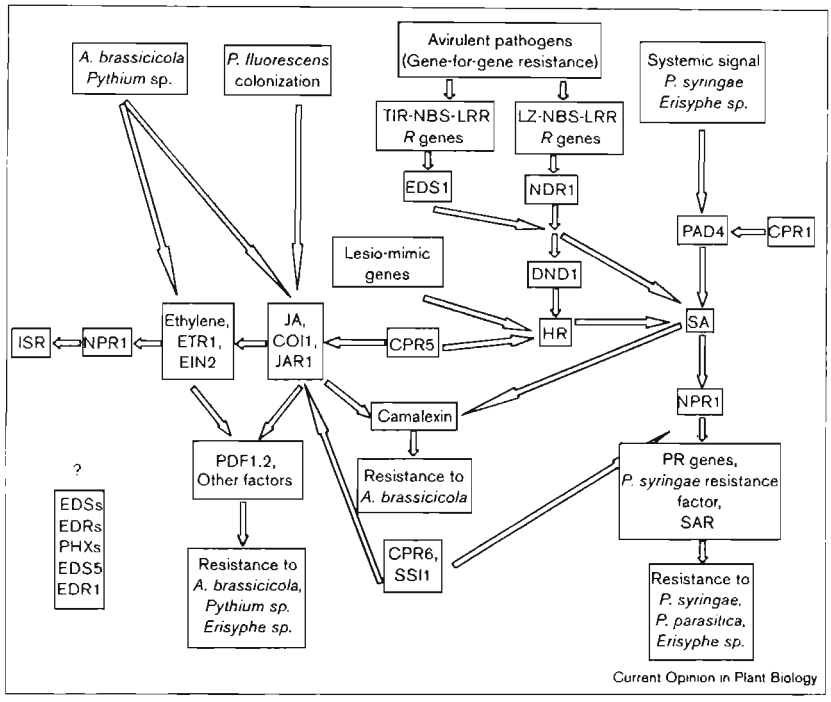
R gene signal transduction

Genes such as *NDR1*, *EDS1*, *DND1*, and the lesion-mimic genes probably act in signal transduction pathways downstream from *R*–*avr* recognition.

NDR1 and *EDS1* are required for gene-for-gene mediated resistance to avirulent strains of the bacterial pathogen *Pseudomonas syringae* and the oomycete pathogen *Peronospora parasitica*. Curiously, *ndr1* mutants are susceptible to one set of avirulent pathogens, whereas *eds1* mutants are susceptible to a non-overlapping set [10••]. The five cloned *R* genes that require *EDS1* all belong to the subset of the nucleotide binding site–leucine rich repeat (NBS–LRR) class of *R* genes

Figure 1

A model of the defense response signaling network showing the relative sites of action of genes discussed in this review. This model is almost certain to be found incorrect before this article is published, and is intended only as a means to stimulate discussion. The SA amplification loop is not shown, as it is not clear which genes might be involved in this. The mutual inhibition between the JA and SA pathways is not shown for the same reason. The rationale for the arrangement of genes in the network is presented in the text. This figure is adapted from Figure 1 of last year's review of this topic [9], with alterations to incorporate results reported in the last year.



that contain sequences similar to the cytoplasmic domains of *Drosophila* Toll and mammalian interleukin 1 transmembrane receptors. The two genes that require *NDR1* belong to the leucine-zipper (LZ) subclass of NBS–LRR genes. There is another LZ–NBS–LRR gene that does not require *EDS1* or *NDR1*, so the correlation between R gene structure and requirement for *EDS1* or *NDR1* is not perfect. Nevertheless, these results show that *R* genes differ in their requirements for downstream factors and that these differences are correlated with *R* gene structural type.

NDR1 encodes a protein with two predicted transmembrane domains [11]. RPM1, which requires NDR1 to mediate resistance, is membrane-associated, despite the fact that its primary sequence does not include any likely membrane-integral stretches [12]. It is possible that part of the function of NDR1 is to hold *R* proteins close to the membrane. *EDS1* encodes a protein with blocks of homology to triacyl glycerol lipases [13••]. The significance of this homology is not known, but it is tempting to speculate that EDS1 is involved in synthesis or degradation of a signal molecule. EDS1 expression is inducible by SA and pathogen infection, suggesting that EDS1 may be involved in signal amplification [13••].

It has been extremely difficult to isolate mutations in genes other than the *R* genes that are required for gene-for-gene resistance. McNellis *et al.* have devised a selection procedure on the basis of precisely controlled inducible expression of the *avr* gene *avrRpt2* in plants carrying the corresponding resistance gene *RPS2* [14•]. Expression of *avrRpt2* in this background is lethal, as it triggers a systemic HR. It is now possible to select for mutants with subtle defects in gene-for-gene signaling by requiring growth on a concentration of inducer slightly higher than the lethal dose. This is a very promising approach for identifying loci involved in gene-for-gene resistance and/or the HR.

Characterization of *dnd1* mutants has provided genetic evidence that the HR is separable from gene-for-gene resistance [15••]. When *dnd1* plants are infected with avirulent pathogens, no HR occurs, but the level of resistance is comparable to that in wild-type plants. One possibility is that DND1 is a regulator of cell death. However, *dnd1* mutants also have elevated SA levels and constitutively express the defense gene *PR1*, raising the possibility that SAR activation leads indirectly to suppression of cell death. This idea could be tested by constructing a *dnd1 nahG* line.

Lesion-mimic mutants develop HR-like lesions, have high levels of SA, and express defense genes, all in the absence of pathogen attack. It is likely that some of the lesion-mimic gene products have important roles in regulation of the HR. These mutants have been studied quite extensively, but few results have been reported in the last year. The reader may refer to recent reviews describing this interesting class of mutants [16,17].

SA-dependent signaling

SA levels increase locally in response to pathogen attack, and systemically in response to the SAR-inducing signal. SA is necessary and sufficient for activation of PR gene expression and enhanced disease resistance. Physiological analyses and characterization of certain lesion-mimic mutants strongly suggest that there is a positive autoregulatory loop affecting SA concentrations [18–20]. Several mutants with defects in SA signaling have been characterized. These include *npr1*, in which expression of *PR* genes in response to SA is blocked; *cpr1*, *cpr5*, and *cpr6*, which constitutively express *PR* genes; the *npr1* suppressor *ssi1*; *pad4*, which has a defect in SA accumulation; and *eds5*, which has a defect in *PR1* expression.

Expression of the defense genes *PR1*, *BG2*, and *PR5* in response to SA treatment requires a gene called *NPR1* or *NIM1*. Mutations in *npr1* abolish SAR, and cause enhanced susceptibility to infection by various pathogens [21–24]. NPR1 appears to be a positive regulator of *PR* gene expression that acts downstream from SA. *NPR1* encodes a novel protein that contains ankyrin repeats (which are often involved in protein–protein interactions [25,26]), and that is localized to the nucleus in the presence of SA [9]. Consequently, it is unlikely that NPR1 acts as a transcription factor to directly control *PR* gene expression, but its nuclear localization suggests that it may interact with such transcription factors.

The *cpr1*, *cpr5*, and *cpr6* mutations cause elevated SA levels, constitutive expression of *PR1*, *BG2*, and *PR5*, and resistance to *P. syringae* and *P. parasitica* [27,28,29••]. In all cases, *cpr nahG* plants do not exhibit elevated gene expression or resistance to *P. syringae*, suggesting that the *CPR* genes act upstream from SA. In *cpr5 npr1* double mutants, defense gene expression and resistance to *P. syringae* are abolished, confirming that CPR5 is acting upstream from NPR1 [28]. The case of *cpr6* mutants is more complicated. The *cpr6* mutation is dominant, so it is likely that the mutant phenotype represents a gain of function rather than a loss of function [29••]. In *cpr6 npr1* plants, constitutive expression of *PR1*, *BG2*, and *PR5* is retained, but resistance to *P. syringae* is lost [29••]. This result leads to two interesting conclusions. First, there must be an SA-dependent, NPR1/NIM1-independent mechanism for activation of *PR1*, *BG2*, and *PR5* [29••]. This could explain the observation that in *npr1* plants infected with *P. syringae*, expression of *PR1* is reduced but not abolished, and expression of *BG2* and *PR5* is wild-type [23]. Second, the factor responsible for *P. syringae* resistance in *cpr6* plants is not *PR1*, *BG2*, or *PR5*, implying that the relationship between expression of these genes and *P. syringae* resistance is merely correlative, not causal [29••]. The challenge now is to find a defense mechanism that is constitutively expressed in *cpr6* in an NPR1-dependent manner, and to determine if this mechanism confers resistance to *P. syringae*.

The phenotypes caused by the dominant *ssi1* mutation superficially resemble those of *cpr* mutants, with the important difference that *ssi1* suppresses *npr1* mutations [30••]. In *ssi1* plants, *PR1*, *BG2*, and *PR5* are constitutively expressed [30••]. In *ssi1 npr1* plants, this expression remains, and unlike *cpr6 npr1* plants, the enhanced sensitivity of *npr1* to *P. syringae* infection is suppressed [30••]. All of the *ssi1* phenotypes are abolished by *nahG*, demonstrating that they are SA-dependent [30••].

PAD4 seems to act upstream from SA. In *pad4* plants infected with a virulent *P. syringae* strain, SA levels, synthesis of the antimicrobial compound camalexin, and *PR1* expression are all reduced [31•]. SA is necessary, but not sufficient, for activation of camalexin synthesis [31•,32]. The camalexin defect in *pad4* plants is reversible by exogenous SA [31•]. Mutations in *pad4* do not affect SA levels, camalexin synthesis, or *PR1* when plants are infected with an avirulent *P. syringae* strain [31•]. Taken together, these results suggest that PAD4 is required for signal amplification to activate the SA pathway in response to pathogens that do not elicit a strong defense response [31•]. The phenotypes of *cpr1 pad4* plants are indistinguishable from those of *pad4* plants, indicating that CPR1 acts upstream from PAD4 to activate *PR* gene expression (N Zhou and J Glazebrook, unpublished data).

Expression of *PR1* is also reduced in *eds5* mutants infected with a virulent *P. syringae* strain [33]. It is likely that EDS5 acts somewhere in the SA pathway. The phenotypes of the various mutants suggest that CPR1 and CPR5 act upstream from SA as negative regulators of SA signaling. CPR6 may also be a positive regulator acting upstream from SA. NPR1 appears to be a positive regulator that functions downstream from SA to activate a subset of SA-dependent responses. SSI1 and EDS5 also affect SA signaling, but their positions in the signal transduction network are not yet clear.

JA-dependent signaling

JA signaling affects diverse processes including fruit ripening, pollen development, root growth, and response to wounding [8]. The *jar1* and *coi1* mutants fail to respond to JA [34,35]. *COI1* has been cloned, and found to encode a protein containing leucine-rich repeats and a degenerate F-box motif [36••]. These features are characteristic of proteins that function in complexes that ubiquitinate proteins targeted for degradation. It follows that COI1 may act by promoting degradation of a factor that exerts a negative regulatory effect in the JA signal transduction pathway.

In the past few years it has become apparent that JA plays an important role in regulation of pathogen defenses. Inoculation of *Arabidopsis* with the avirulent fungal pathogen *Alternaria brassicicola* induces expression of the defensin gene *PDF1.2* [37]. This induction does not require SA or NPR1, but it does require ethylene and JA signaling [37]. Studies of the effect of mutations in *ETR1* (the ethylene receptor), *EIN2* (required for responses to ethylene) or *COI1* on *PDF1.2* expression in response to

A. brassicicola, ethylene, JA, or combinations of JA and ethylene suggest a model in which ethylene and JA are required simultaneously for *PDF1.2* expression [38••].

Like SA signaling, JA signaling has systemic effects. Plants in which only a few leaves were infected with *A. brassicicola* express *PDF1.2* throughout the plant [37]. Although *A. brassicicola* fails to infect wild-type plants, it is able to infect *coi1* mutants, suggesting that JA signaling is required for resistance to *A. brassicicola*. JA-dependent responses are also sufficient to confer resistance to *A. brassicicola*. This was demonstrated using *pad3* mutants, which are unable to synthesize camalexin and are susceptible to *A. brassicicola* [39,40•]. Treatment of *pad3* plants with JA prior to infection greatly reduced *A. brassicicola* growth [40•].

SA signaling and JA signaling pathways are interconnected in complicated ways. Studies in other systems have shown that SA signaling and JA signaling are mutually inhibitory [8,41]. However, synthesis of camalexin in response to *P. syringae* infection is blocked in *nahG* [31•,32] and *coi1* (J Glazebrook, unpublished data) plants, strongly suggesting that camalexin synthesis requires both SA and JA signaling. The *cpr5*, *cpr6*, and *acd2* mutations cause constitutive expression of both *PR1* and *PDF1.2*, suggesting that there may be a common control point for activation of both pathways. [28,29••,37]. *PDF1.2* is also constitutively expressed in *ssi1* plants. Curiously, this expression is SA-dependent, in contrast with wild-type plants, in which activation of *PDF1.2* expression is completely SA-independent [30••]. The proposed explanation for this effect is that *ssi1* acts as a switch between the two pathways [30••]. An alternative possibility is that *ssi1* perturbs the balance of SA-dependent and JA-dependent signaling in a way that shifts *PDF1.2* expression toward SA-dependence.

Induced systemic resistance (ISR)

Some rhizosphere-associated bacteria promote disease resistance [42]. This phenomenon, called ISR, has been studied using *Pseudomonas fluorescens* strain WCS417r to colonize *Arabidopsis* roots [43]. Colonized plants are more resistant to infection by the fungal pathogen *Fusarium oxysporum* f sp *raphani* and *P. syringae* [43]. ISR occurs in *nahG* plants, indicating that it is not an SA-dependent phenomenon [43]. Rather, ISR appears to be JA- and ethylene-dependent. The observation that ethylene can induce ISR in *jar1* mutants led to the hypothesis that ISR requires a JA signal followed by an ethylene signal [44••]. No changes in gene expression associated with ISR have been detected [44••], suggesting that it is different from activation of *PDF1.2* expression by *A. brassicicola*.

Curiously, ISR requires *NPR1* [44••]. This was unexpected in light of the facts that NPR1 was previously known to be involved only in SA-dependent processes, and ISR is SA-independent. This result implies that NPR1 can respond to signals from at least two different sources, one that is SA-dependent and one that is derived from ISR signaling. If the SA-dependent signal is received, NPR1 mediates a resistance response characterized by *PR1* expression, whereas if the ISR signal is received, NPR1 mediates a different resistance response. It is difficult to imagine how this could occur, unless NPR1 is interacting with different 'adapter' molecules to mediate the different signals. The ankyrin repeats found in NPR1 could function in protein–protein interactions between NPR1 and adapter proteins. Identification of proteins that interact with NPR1, and characterization of plants with loss-of-function mutations affecting those proteins, would be very helpful for understanding how NPR1 acts in each pathway. It would also be worthwhile to determine if the *ssi1 or cpr6* mutations suppress the ISR defect of *npr1* mutants.

Relevance to disease resistance

Characterization of the effects of various mutations on resistance to different pathogens has revealed that there is considerable variation in the extent to which pathogens are affected by defense mechanisms. SAR is known to confer resistance to a wide array of pathogens, including bacteria, fungi, oomycetes, and viruses. In *Arabidopsis*, the SA pathway mutants *npr1* and *pad4* show enhanced susceptibility to *P. syringae* and *P. parasitica* [21,22,24,31•,45]. The fungus *Erisyphe orontii* also seems to be sensitive to SA-dependent responses. Among a collection of mutants that display enhanced susceptibility to *P. syringae*, only mutants that had defects in expression of *PR1* were also more susceptible to *E. orontii* [46•]. *P. parasitica* may be inhibited by JA-dependent mechanisms as well as by SA-dependent ones. In *cpr5 npr1* double mutants, the *PR1* expression and resistance to *P. syringae* caused by *cpr5* is abolished, but *PDF1.2* expression and *P. parasitica* resistance are retained, suggesting that activation of the JA pathway is causing *P. parasitica* resistance [28].

JA signaling is important for limiting the growth of certain fungal pathogens. The *fad3-2 fad7-2 fad8* triple mutant is unable to synthesize JA due to an inability to produce linoleic acid, a precursor of JA. These plants and *jar1* plants are much more susceptible to infection by *Pythium* species than wild-type plants are [47•,48•]. JA treatment enhances resistance to *A. brassicicola*, and *coi1* mutants show enhanced susceptibility, whereas the *nahG* transgene and an *npr1* mutation have no effect [40]. These observations suggest that JA signaling is important for resistance to fungi such as *Pythium* species and *A. brassicicola*, while SA signaling has little effect on resistance to *A. brassiciciola*.

Overexpression of rate-limiting defense response regulators may cause the signaling network to respond faster or more strongly to pathogen attack, thereby improving resistance. Overexpression of *NPR1* caused increased resistance to *P. syringae* and *P. parasitica* in a dosage dependent manner [49••]. Importantly, *NPR1*-overexpression had no obvious deleterious effects on plant growth, in contrast to mutations that lead to constitutive overexpression of defense responses, which generally cause dwarfism. In

the future, the effect of overexpression of other cloned regulatory genes, such as *NDR1*, *EDS1*, and *COI1*, should be tested.

Other mutations that may affect signaling

There are several mutants that affect disease resistance that may prove to be involved in control of defense responses, but have not yet been characterized in detail. These include *eds* mutants, that show enhanced disease susceptibility to virulent *P. syringae* strains [23,33,50], *phx* mutants, isolated as suppressors of the lesion-mimic mutant *lsd5* [51], and *edr* mutants, which display enhanced resistance to *P. syringae* and/or *Erisyphe cichoracearum* infection [52•]. *EDR1* almost certainly affects SA signaling, since expression of *PR1* in response to *E. cichoracearum* infection occurs more rapidly in *edr1* mutants than in wild-type plants [52•].

Conclusions

Many genes that function in regulation of defense responses have been identified. Progress has been made in determining the positions of various genes in the signal transduction network. However, current models seem to have little predictive value, in that characterization of new mutants often requires wholesale rearrangements of the existing models in order to explain observed phenotypes. Obviously, the signal transduction network is not well understood.

The field is now in a position to develop second-generation approaches to identify additional components of the signaling networks. These include screening for suppressors and enhancers of known mutations, and using two-hybrid screens to identify proteins that may interact with the products of cloned genes. The biological significance of two-hybrid interactions can be tested using a reverse-genetic approach to obtain loss-of-function mutations in the relevant genes.

For determining the roles of each gene in the signal transduction network, it would be very helpful if all mutants were tested for all phenotypes. It is also important to construct double mutants for epistasis testing. Both of these approaches require free exchange of mutants among various laboratories. The sequencing of the *Arabidopsis* genome, which should be complete in late 2000, will make it possible to apply powerful new techniques to the study of signaling. For example, 'gene chips' could be used to monitor expression levels of every gene simultaneously, so that the effects of mutations on gene expression patterns can be determined completely and efficiently. This will be useful for discovery of pathogen-inducible genes that are not yet known, as well as for elucidation of signal transduction networks.

Acknowledgements

I apologize to scientists whose work I overlooked, or was not able to include due to space limitations.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ellis J, Jones D: **Structure and function of proteins controlling strain-specific pathogen resistance in plants.** *Curr Opin Plant Biol* 1998, 1:288-293.

2. Jones DA, Jones JDG: **The role of leucine-rich repeat proteins in plant defenses.** *Adv Bot Res* 1997, 24:89-167.

3. Ronald P: **Resistance gene evolution.** *Curr Opin Plant Biol* 1998, 1:294-298.

4. Ryals JA, Neuenschwander UH, Willits MG, Molina A, Steiner H-Y, Hunt MD: **Systemic acquired resistance.** *Plant Cell* 1996, 8:1809-1819.

5. Shulaev V, Leon J, Raskin I: **Is salicylic acid a translocated signal of systemic acquired resistance in tobacco?** *Plant Cell* 1995, 7:1691-1701.

6. Vernooij B, Friedrich L, Morse A, Reist R, Kolditz-Jawhar R, Ward E, Uknes S, Kessmann H, Ryals J: **Salicylic acid is not the translocated signal responsible for inducing systemic acquired resistance but is required in signal transduction.** *Plant Cell* 1994, 6:959-965.

7. Willits MG, Ryals JA: **Determining the relationship between salicylic acid levels and systemic acquired resistance induction in tobacco.** *Mol Plant–Microbe Interact* 1998, 11:795-800.

8. Creelman RA, Mullet JE: **Biosynthesis and action of jasmonates in plants.** *Annu Rev Plant Physiol Plant Mol Biol* 1997, 48:355-381.

9. Dong X: **SA, JA, ethylene, and disease resistance in plants.** *Curr Opin Plant Biol* 1998, 1:316-323.

10. •• Aarts N, Metz M, Holub E, Staskawicz BJ, Daniels MJ, Parker JE: **Different requirements for *EDS1* and *NDR1* by disease resistance genes define at least two *R* gene-mediated signaling pathways in *Arabidopsis*.** *Proc Natl Acad Sci* 1998, 95:10306-10311.
An elegant series of crosses were used to bring *ndr1* or *eds1* together with relevant *R* genes. The results make a strong genetic argument that *EDS1* and *NDR1* operate in separate signaling pathways.

11. Century KS, Shapiro AD, Repetti PP, Dahlbeck D, Holub E, Staskawicz BJ: ***NDR1*, a pathogen-induced component required for *Arabidopsis* disease resistance.** *Science* 1997, 278:1963-1965.

12. Boyes DC, Nam J, Dangl JL: **The *Arabidopsis thaliana RPM1* disease resistance gene product is a peripheral plasma membrane protein that is degraded coincident with the hypersensitive response.** *Proc Natl Acad Sci USA* 1998, 95:15849-15854.

13. •• Falk A, Feys BJ, Frost LN, Jones JDG, Daniels MJ, Parker JE: ***EDS1*, an essential component of *R* gene-mediated disease resistance in *Arabidopsis* has homology to eukaryotic lipases.** *Proc Natl Acad Sci USA* 1999, 96:3292-3297.
This paper describes the isolation of *EDS1*. The homology to lipases is intriguing; it may indicate a role for an undiscovered small signal molecule in gene-for-gene resistance. The pathogen-inducibility of *EDS1* suggests that there is either another factor acting between *R* gene products and EDS1, or that *EDS1* is regulated by a feedback loop.

14. • McNellis TW, Mudgett MB, Li K, Aoyama T, Horvath D, Chua N-H, Staskawicz BJ: **Glucocorticoid-inducible expression of a bacterial avirulence gene in transgenic *Arabidopsis* induces hypersensitive cell death.** *Plant J* 1998, 14:247-257.
This report describes a selection that should be a powerful tool for identifying factors acting downstream from *R* genes in gene-for-gene resistance.

15. •• Yu I-C, Parker J, Bent AF: **Gene-for-gene disease resistance without the hypersensitive response in *Arabidopsis dnd1* mutant.** *Proc Natl Acad Sci USA* 1998, 95:7819-7824.
This interesting paper makes a genetic argument that the HR is not required for gene-for-gene resistance. Is the high SA level in *dnd1* causing inhibition of cell death?

16. Dangl JL, Dietrich RA, Richberg MA: **Death don't have no mercy: cell death programs in plant–microbe interactions.** *Plant Cell* 1996, 8:1793-1807.

17. Richberg MH, Aviv DH, Dangl JL: **Dead cells do tell tales.** *Curr Opin Plant Biol* 1998, 1:480-485.

18. Shirasu K, Nakajima H, Rajasekhar VK, Dixon RA, Lamb C: **Salicylic acid potentiates an agonist-dependent gain control that amplifies pathogen signals in the activation of defense mechanisms.** *Plant Cell* 1997, **9**:261-270.

19. Hunt MD, Delaney TP, Dietrich RA, Weymann KB, Dangl JL, Ryals JA: **Salicylate-independent lesion formation in *Arabidopsis lsd* mutants.** *Mol Plant-Microbe Int* 1997, **10**:531-536.

20. Weymann K, Hunt M, Uknes S, Neuenschwander U, Lawton K, Steiner HY, Ryals J: **Suppression and restoration of lesion formation in *Arabidopsis lsd* mutants.** *Plant Cell* 1995, **7**:2013-2022.

21. Cao H, Bowling SA, Gordon S, Dong X: **Characterization of an *Arabidopsis* mutant that is nonresponsive to inducers of systemic acquired resistance.** *Plant Cell* 1994, **6**:1583-1592.

22. Delaney TP, Friedrich L, Ryals JA: ***Arabidopsis* signal transduction mutant defective in chemically and biologically induced disease resistance.** *Proc Natl Acad Sci USA* 1995, **92**:6602-6606.

23. Glazebrook J, Rogers EE, Ausubel FM: **Isolation of *Arabidopsis* mutants with enhanced disease susceptibility by direct screening.** *Genetics* 1996, **143**:973-982.

24. Shah J, Tsui F, Klessig DF: **Characterization of a salicylic acid insensitive mutant (*sai1*) of *Arabidopsis thaliana*, identified in a selective screen utilizing the SA-inducible expression of the *tms2* gene.** *Mol Plant-Microbe Interact* 1997, **10**:69-78.

25. Cao H, Glazebrook J, Clarke JD, Volko S, Dong X: **The *Arabidopsis NPR1* gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats.** *Cell* 1997, **88**:57-63.

26. Ryals J, Weymann K, Lawton K, Friedrich L, Ellis D, Steiner H-Y, Johnson J, Delaney TP, Jesse T, Vos P, Uknes S: **The *Arabidopsis NIM1* protein shows homology to the mammalian transcription factor IκB.** *Plant Cell* 1997, **9**:425-439.

27. Bowling SA, Guo A, Cao H, Gordon AS, Klessig DF, Dong X: **A mutation in *Arabidopsis* that leads to constitutive expression of systemic acquired resistance.** *Plant Cell* 1994, **6**:1845-1857.

28. Bowling SA, Clarke JD, Liu Y, Klessig DF, Dong X: **The *cpr5* mutant of *Arabidopsis* expresses both *NPR1*-dependent and *NPR1*-independent resistance.** *Plant Cell* 1997, **9**:1573-1584.

29. •• Clarke JD, Liu Y, Klessig DF, Dong X: **Uncoupling PR gene expression from NPR1 and bacterial resistance: characterization of the dominant *Arabidopsis cpr6-1* mutant.** *Plant Cell* 1998, **10**:557-569.

Thorough characterization of *cpr6* plants leads to two important conclusions. First, there is an SA-dependent, NPR1-independent way to activate *PR* gene expression. Second, there is an NPR1-dependent resistance mechanism that is distinct from expression of known *PR* genes. Possible models to explain the role of *CPR6* in defense response signaling are presented.

30. •• Shah J, Kachroo P, Klessig DF: **The *Arabidopsis ssi1* mutation restores pathogenesis-related gene expression in *npr1* plants and renders defensin gene expression salicylic acid dependent.** *Plant Cell* 1999, **11**:191-206.

This study is a good example of the usefulness of supressor genetics for identifying new genes. The surprising discovery that *PDF1 2* expression can be SA-dependent is described. The introduction and discussion present an excellent description of the current understanding of SA and JA signaling pathways, and the implications of *ssi1* phenotypes for models of the signaling network.

31. • Zhou N, Tootle TL, Tsui F, Klessig DF, Glazebrook J: ***PAD4* functions upstream from salicylic acid to control defense responses in *Arabidopsis*.** *Plant Cell* 1998, **10**:1021-1030.

This paper describes a series of experiments leading to the conclusion that *PAD4* acts upstream of SA in defense response signaling. The observation that *PAD4* is not required for responses to avirulent *P syringae* strains indicates that there is a *PAD4*-independent mechanism for activation of SA synthesis.

32. Zhao J, Last RL: **Coordinate regulation of the tryptophan biosynthetic pathway and indolic phytoalexin accumulation in *Arabidopsis*.** *Plant Cell* 1996, **8**:2235-2244.

33. Rogers EE, Ausubel FM: ***Arabidopsis* enhanced disease susceptibility mutants exhibit enhanced susceptibility to several bacterial pathogens and alterations in *PR-1* gene expression.** *Plant Cell* 1997, **9**:305-316.

34. Feys BJF, Benedetti CE, Penfold CN, Turner JG: ***Arabidopsis* mutants selected for resistance to the phytotoxin coronatine are male sterile, insensitive to methyl jasmonate, and resistant to a bacterial pathogen.** *Plant Cell* 1994, **6**:751-759.

35. Staswick PE, Su W, Howell SH: **Methyl jasmonate inhibition of root growth and induction of a leaf protein are decreased in an *Arabidopsis thaliana* mutant.** *Proc Natl Acad Sci USA* 1992, **89**:6837-6840.

36. •• Xie D-X, Feys BF, James S, Nieto-Rostro M, Turner JG: ***COI1*: An *Arabidopsis* gene required for jasmonate-regulated defense and fertility.** *Science* 1998, **280**:1091-1094.

This paper describes isolation of *COI1*. Sequence similarities lead to the hypothesis that protein degradation is involved in JA signaling.

37. Penninckx IAMA, Eggermont K, Terras FFG, Thomma BPHJ, De Samblancx GW, Buchala A, Metraux J-P, Manners JM, Broekaert WF: **Pathogen-induced systemic activation of a plant defensin gene in *Arabidopsis* follows a salicylic acid-independent pathway.** *Plant Cell* 1996, **8**:2309-2323.

38. •• Penninckx IAMA, Thomma BPHJ, Buchala A, Metraux J-P, Broekaert WF: **Concomitant activation of jasmonate and ethylene response pathways is required for induction of a plant defensin gene in *Arabidopsis*.** *Plant Cell* 1998, **10**:2103-2113.

A nice series of experiments show that jasmonate and ethylene are required concomitantly, rather than sequentially, for activation of *PDF1 2* expression. Confusion about the requirement of ETR1 for *PDF1 2* expression is resolved by using a stronger allele.

39. Glazebrook J, Ausubel FM: **Isolation of phytoalexin-deficient mutants of *Arabidopsis thaliana* and characterization of their interactions with bacterial pathogens.** *Proc Natl Acad Sci USA* 1994, **91**:8955-8959.

40. • Thomma BPHJ, Eggermont K, Penninckx IAMA, Mauch-Mani B, Vogelsang R, Cammue BPA, Broekaert WF: **Separate jasmonate-dependent and salicylate-dependent defense-response pathways in *Arabidopsis* are essential for resistance to distinct microbial pathogens.** *Proc Natl Acad Sci USA* 1998, **95**:15107-15111.

This work demonstrates that resistance to *A brassicicola* requires JA signaling, and not SA signaling. These results show that JA signaling has an important functional role in disease resistance. To get a complete understanding of disease resistance pathways, it is necessary to work with a range of pathogens, as the significance of various defense mechanisms varies greatly between different pathogens.

41. Harms K, Ramirez I, Pena-Cortes H: **Inhibition of wound-induced accumulation of allene oxide synthase transcripts in flax leaves by aspirin and salicylic acid.** *Plant Physiol* 1998, **118**:1057-1065.

42. van Loon LC, Bakker PAHM, Pieterse CMJ: **Systemic resistance induced by rhizosphere bacteria.** *Annu Rev Phytopathol* 1998, **36**:453-483.

43. Pieterse CMJ, van Wees S, Hoffland E, van Pelt JA, van Loon LC: **Systemic resistance in *Arabidopsis* induced by biocontrol bacteria is independent of salicylic acid accumulation and pathogenesis-related gene expression.** *Plant Cell* 1996, **8**:1225-1237.

44. •• Pieterse CMJ, van Wees SCM, van Pelt JA, Knoester M, Laan R, Gerrits H, Weisbeek PJ, van Loon LC: **A novel signaling pathway controlling induced systemic resistance in *Arabidopsis*.** *Plant Cell* 1998, **10**:1571-1580.

This is the latest installment of the ISR in *Arabidopsis* story. Mutations affecting the ethylene, JA and SA pathways are used to show that ISR is JA and ethylene dependent, and SA-independent. The surprising finding that ISR requires NPR1 forces revision of the role of NPR1 in signaling pathways.

45. Glazebrook J, Zook M, Mert F, Kagan I, Rogers EE, Crute IR, Holub EB, Hammerschmidt R, Ausubel FM: **Phytoalexin-deficient mutants of *Arabidopsis* reveal that *PAD4* encodes a regulatory factor and that four *PAD* genes contribute to downy mildew resistance.** *Genetics* 1997, **146**:381-392.

46. • Reuber TL, Plotnikova JM, Dewdney J, Rogers EE, Wood W, Ausubel FM: **Correlation of defense gene induction defects with powdery mildew susceptibility in *Arabidopsis* enhanced disease susceptibility mutants.** *Plant J* 1998, **16**:473-485.

A collection of mutants with enhanced susceptibility to *P syringae* were tested for enhanced susceptibility to *Erisyphe orontii*. Only mutants with SA signaling pathway defects were more susceptible, suggesting that SA-dependent responses contribute to *E. orontii* resistance. The results also indicate that many factors that are important for *P. syringae* resistance do not have a significant effect on *E. orontii*

47. • Staswick PE, Yuen GY, Lehman CC: **Jasmonate signaling mutants of *Arabidopsis* are susceptible to the soil fungus *Pythium irregulare*.** *Plant J* 1998, **15**:747-754.

This paper shows that *jar1* mutants are highly susceptible to *Phythium* infection, demonstrating a role of SA signaling in resistance to *Pythium*.

48. Vijayan P, Shockey J, Levesque CA, Cook RJ, Browse J: **A role for jasmonate in pathogen defense of *Arabidopsis*** *Proc Natl Acad Sci USA* 1998, **95**:7209-7214.
This paper shows that the ability to synthesize JA is important for resistance to *Phythium*, demonstrating a role for *JA* in resistance to *Pythium*.

49. Cao H, Li X, Dong X: **Generation of broad-spectrum disease resistance by overexpression of an essential regulatory gene in systemic acquired resistance.** *Proc Natl Acad Sci* 1998, **95**:6531-6536.
This paper shows that *NPR1* overexpression causes increased resistance to *P. syringae* and *P. parasitica*. This is an important result that is relevant to genetic engineering strategies for improving disease resistance. It suggests that sensitization of signaling pathways by increasing the expression level of key regulatory factors may be an effective method for improving disease resistance in crops.

50. Volko SM, Boller T, Ausubel FM: **Isolation of new *Arabidopsis* mutants with enhanced disease susceptibility by direct screening.** *Genetics* 1998, **149**:537-548.

51. Morel J-B, Dangl JL: **Suppressors of the *Arabidopsis lsd5* cell death mutation identify genes involved in regulating disease resistance responses.** *Genetics* 1999, **151**:305-319.

52. Frye CA, Innes RW: **An *Arabidopsis* mutant with enhanced resistance to powdery mildew.** *Plant Cell* 1998, **10**:947-956.
This paper describes isolation of several enhanced resistance mutants, and characterization of the *edr1* mutant. *EDR1* almost certainly plays a role in SA signaling, as expression of *PR1* occurs more rapidly in *edr1* mutants. This is the first mutant with this characteristic to be described. Other mutants with altered *PR1* expression either fail to express *PR1*, or express it constitutively.

The Plant Cell, Vol. 15, 809–834, April 2003, www.plantcell.org © 2003 American Society of Plant Biologists

GENOMICS ARTICLE

Genome-Wide Analysis of NBS-LRR–Encoding Genes in Arabidopsis[W]

Blake C. Meyers,[a,b] Alexander Kozik,[a] Alyssa Griego,[a] Hanhui Kuang,[a] and Richard W. Michelmore[a,1]

[a] Department of Vegetable Crops, University of California, Davis, California 95616
[b] Department of Plant and Soil Sciences, University of Delaware, Newark, Delaware 19711

The Arabidopsis genome contains ~200 genes that encode proteins with similarity to the nucleotide binding site and other domains characteristic of plant resistance proteins. Through a reiterative process of sequence analysis and reannotation, we identified 149 NBS-LRR–encoding genes in the Arabidopsis (ecotype Columbia) genomic sequence. Fifty-six of these genes were corrected from earlier annotations. At least 12 are predicted to be pseudogenes. As described previously, two distinct groups of sequences were identified: those that encoded an N-terminal domain with Toll/Interleukin-1 Receptor homology (TIR-NBS-LRR, or TNL), and those that encoded an N-terminal coiled-coil motif (CC-NBS-LRR, or CNL). The encoded proteins are distinct from the 58 predicted adapter proteins in the previously described TIR-X, TIR-NBS, and CC-NBS groups. Classification based on protein domains, intron positions, sequence conservation, and genome distribution defined four subgroups of CNL proteins, eight subgroups of TNL proteins, and a pair of divergent NL proteins that lack a defined N-terminal motif. CNL proteins generally were encoded in single exons, although two subclasses were identified that contained introns in unique positions. TNL proteins were encoded in modular exons, with conserved intron positions separating distinct protein domains. Conserved motifs were identified in the LRRs of both CNL and TNL proteins. In contrast to CNL proteins, TNL proteins contained large and variable C-terminal domains. The extant distribution and diversity of the NBS-LRR sequences has been generated by extensive duplication and ectopic rearrangements that involved segmental duplications as well as microscale events. The observed diversity of these NBS-LRR proteins indicates the variety of recognition molecules available in an individual genotype to detect diverse biotic challenges.

INTRODUCTION

Preliminary sequence analysis suggested that a significant proportion of the Arabidopsis ecotype Columbia (Col-0) genome is devoted to encoding various components of a defense system (Arabidopsis Genome Initiative, 2000). We can now evaluate in detail the repertoire of genes available in a single genotype to defend against diverse biotic challenges. Resistance (*R*) genes have been shown frequently by classic genetics to be single loci that confer resistance against pathogens that express matching avirulence genes in a "gene-for-gene" manner (Flor, 1956, 1971). This type of specific resistance often is associated with a localized hypersensitive response, a form of programmed cell death, in the plant cells proximal to the site of infection triggered by recognition of a pathogen product (Dangl et al., 1996; Heath, 2000). The plant resistance response triggered by *R* gene recognition also includes increased expression of defense genes, generation of reactive oxygen species, production or release of salicylic acid, ion fluxes, and other factors (Heath, 2000).

During the last 8 years, numerous *R* genes have been cloned from many plant species (Dangl and Jones, 2001; Hulbert et al., 2001). *R* genes encode at least five diverse classes of proteins (R proteins) (Dangl and Jones, 2001). The largest class of known R proteins includes those that contain a nucleotide binding site and leucine-rich repeat domains (NBS-LRR proteins). NBS-LRR proteins may recognize the presence of the pathogen directly or indirectly. In theory, specific recognition of multiple pathogens could necessitate the activity of numerous *R* genes. The guard hypothesis proposes that NBS-LRR proteins guard plant targets against pathogen effector proteins; in this scenario, these pathogen products act as virulence factors to enhance the susceptibility of the host plant in the absence of recognition (van der Biezen and Jones, 1998a; Dangl and Jones, 2001). A small number of *R* genes can provide defense against diverse pathogens if a limited number of effector targets are present. The definition of a complete set of NBS-LRR proteins in a plant genome will provide insights into the diversity of defense genes available in a single plant.

The NBS-LRR R proteins contain distinct domains, several of which are composed of characteristic motifs. Nucleotide binding sites are found in diverse proteins and are required for ATP and GTP binding (Walker et al., 1982; Saraste et al., 1990). The ability of plant NBS-LRR proteins to bind nucleotides has been demonstrated for the tomato I2 and Mi R proteins (Tameling et

[1] To whom correspondence should be addressed. E-mail rwmichelmore@ucdavis.edu; fax 530-752-9659.
[W] Online version contains Web-only data.
Article, publication date, and citation information can be found at www.plantcell.org/cgi/doi/10.1105/tpc.009308.

al., 2002). The NBS contains conserved motifs that can be used to classify the sequences into subgroups with discrete functions (Saraste et al., 1990; Bourne et al., 1991; Traut, 1994). The NBS-LRR plant R proteins are members of a specific and distinct subgroup of NBS proteins that contain additional protein domains, such as a C-terminal LRR region of variable length (Bent, 1996; Hammond-Kosack and Jones, 1996; Baker et al., 1997; van der Biezen and Jones, 1998b; Meyers et al., 1999; Cannon et al., 2002). The NBS-LRR family of proteins has been subdivided further based on the presence or absence of an N-terminal Toll/Interleukin-1 Receptor (TIR) homology region (Meyers et al., 1999; Pan et al., 2000; Cannon et al., 2002; Richly et al., 2002). Most of those proteins lacking a TIR have a coiled-coil (CC) motif in the N-terminal region (Pan et al., 2000). Detailed comparative analyses of the complete set of Arabidopsis R proteins have not been made.

Genetic and genomic studies have provided insights into the evolution of *R* genes and the mechanisms that generate variation in these genes. Classic genetic studies demonstrated that many but not all *R* genes are clustered in plant genomes (reviewed by Hulbert et al., 2001). Consistent with this finding, genome sequencing demonstrated that the majority of NBS-LRR–encoding genes are clustered in the genomes of both Arabidopsis and rice (Meyers et al., 1999; Bai et al., 2002; Richly et al., 2002). The clustered arrangement of these genes may be a critical attribute allowing the generation of novel resistance specificities via recombination or gene conversion (Hulbert et al., 2001). In addition, analyses of individual clusters provided evidence of diversifying selection in the majority of plant *R* genes studied, suggesting that variation may be concentrated within predicted binding surfaces (Parniske et al., 1997; Botella et al., 1998; Meyers et al., 1998b; Wang et al., 1998; Cooley et al., 2000; Luck et al., 2000; Mondragon-Palomino et al., 2002). The combined data from classic and molecular studies have led to models describing the predicted evolutionary constraints on these proteins and the ways in which variation is produced and maintained (Michelmore and Meyers, 1998; Mondragon-Palomino et al., 2002). Additional NBS-LRR proteins identified through ongoing genomics projects are contributing to our understanding of the mechanisms that generate sequence diversity in these proteins.

Here, we characterize the complete set of plant *R* gene–related NBS-encoding genes in the Col-0 Arabidopsis genome. Bioinformatics analysis combined with experimental validation demonstrated the presence of 149 NBS-LRR–encoding genes and an additional 58 related genes lacking LRRs (Meyers et al., 2002). As demonstrated previously, the NBS-LRR–encoding genes can be subdivided into two distinct classes: those with or without a TIR region. Numerous subgroups existed in both classes, as defined by intron numbers and positions, phylogenetic analyses, and encoded protein motifs. Their distribution within the Arabidopsis Col-0 genome is the consequence of numerous duplication events and ectopic rearrangements as well as conservation and preferential amplification of particular gene pairs. This bioinformatics analysis of the *R* gene homologs provides a definitive resource for ongoing functional and evolutionary studies of this large family of plant genes.

RESULTS

Identification and Classification of NBS-LRR–Encoding Genes

The complete set of NBS-encoding sequences was identified from the Arabidopsis genome of ecotype Col-0 in a reiterative process (Table 1, Figure 1). Four analytical steps were used to compile the final set of sequences. First, a set of 159 genes with the NBS motif was selected from the complete set of predicted Arabidopsis proteins (http://mips.gsf.de) using a hidden Markov model (HMM) (Eddy, 1998) for the NBS domain from the Pfam database (PF0931; http://pfam.wustl.edu).

In the second analytical step, selected protein sequences were aligned based only on the NBS domain using CLUSTAL W. This alignment then was used to develop an Arabidopsis-specific HMM model to identify related sequences. The refined HMM was compared again against the complete set of predicted Arabidopsis proteins. All sequences that matched the model with a score of 0.05 or greater were incorporated into the HMM. The refined HMM was compared again with the entire set of Arabidopsis open reading frames (ORFs) with the threshold for acceptance decreased to 0.001. The 10 sequences with scores just above this threshold and the 15 sequences with scores just below this threshold were analyzed for the presence of the TIR, NBS, or LRR motifs using Pfam and

Table 1. Numbers of Arabidopsis Genes That Encode Domains Similar to Plant R Proteins

Predicted Protein Domains[a]	Letter Code	Previous No.[b]	Full Manual[c]
CC-NBS-LRR	CNL	48	51
NBS_{CC}-LRR	NL	2	4
TIR-NBS-LRR	TNL	82	83
NBS_{TIR}-LRR	NL	2	2
TIR-NBS-LRR-X	TNLX	5	5
TIR-NBS-TIR-NBS-LRR	TNTNL	2	2
TIR-TIR-NBS-LRR	TTNL	0	2
Total with LRRs		141	149
TIR-NBS	TN	14	21
TIR-X	TX	23	30
X-TIR-NBS-X	XTNX	0	2
CC-NBS	CN	4	4
CC-NBS-X	CNX	1	1
CC (related to CNL)	C	0	1
NBS_{CC}	N	1	1
Total without LRRs		43	58

Table updated from Meyers et al. (2002).

[a] Protein domains present in the predicted protein. NBS domains from CNL or TNL proteins are distinct (Meyers et al., 1999); the CC or TIR subscript indicates NBS motifs predictive of a CC or TIR domain N-terminal to the NBS. Sequences can be accessed at http://niblrrs.ucdavis.edu.

[b] Number of genes identified by automated analysis before this analysis and in the public databases.

[c] Number of genes identified in this study by manual assessment of the genomic DNA sequence, automated annotations, and predicted protein domains.

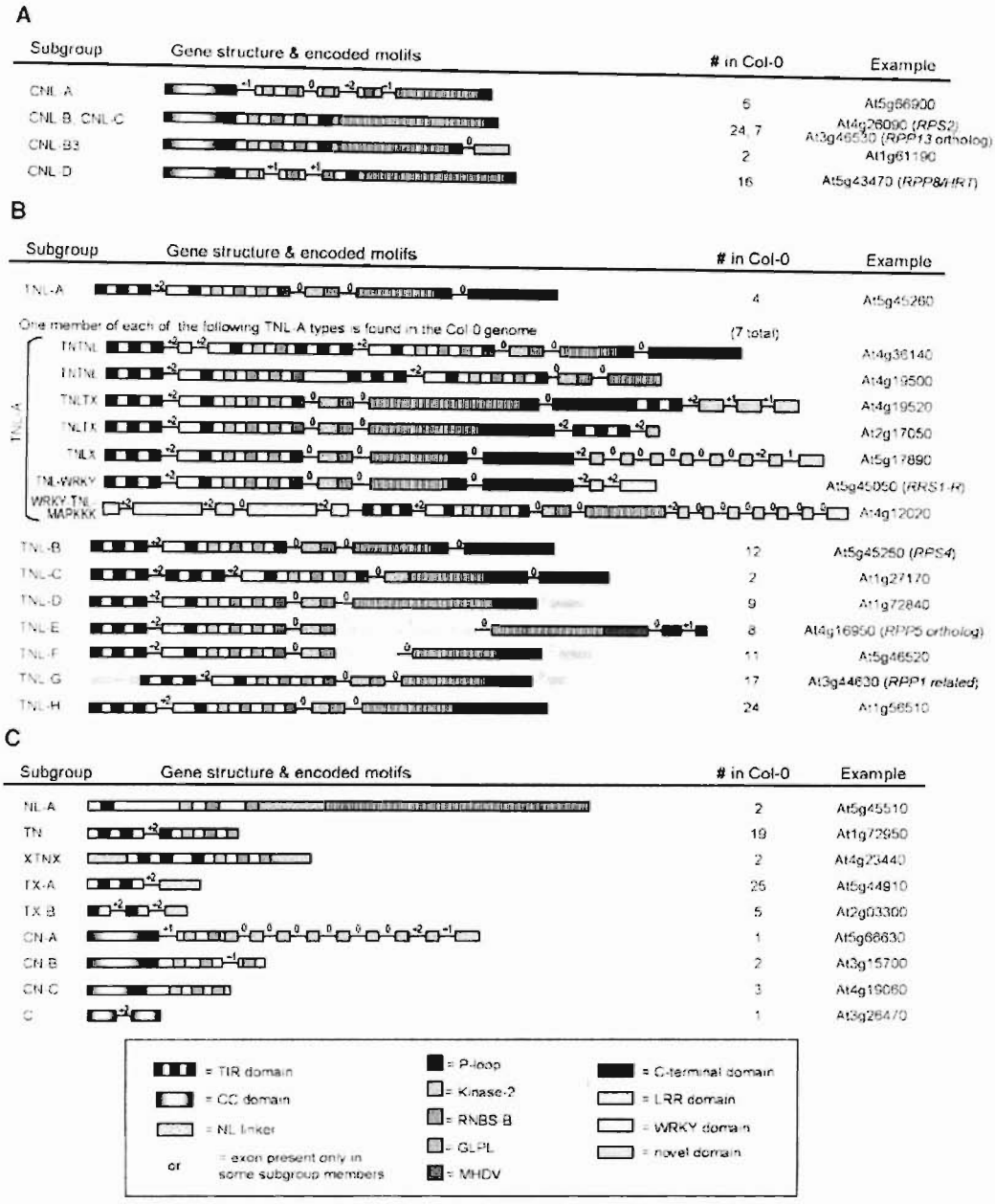


Figure 1. Intron/Exon Configurations and Protein Motifs of NBS-LRR–Encoding Genes in Arabidopsis.

(A) *CNL* genes.

(B) *TNL* genes. All members of the variable *TNL-A* subgroup are shown; only one member of the more homogeneous subgroups is diagrammed.

(C) Additional genes that encode CC, TIR, or NBS domains similar to the CNL or TNL proteins. *TN* and *TX* genes are described in more detail by Meyers et al. (2002).

Encoded protein domains are indicated with shading and colors. Exons are drawn approximately to scale as boxes; connecting thin lines indicate the positions of introns, which are not drawn to scale. Numbers above introns indicate the phase of the intron (see text). Numbers under "# in Col-0" indicate the total number found in the Col-0 genomic sequence; the "representative" columns list the diagrammed gene for each type. Genes of known function are shown where available.

visual inspection. Four of the 10 sequences just above the 0.001 threshold value did not contain TIR, NBS, or LRR motifs and were discarded; all sequences above these 10 contained NBS motifs. Below this threshold, only 2 of the next 15 proteins contained the NBS motif by Pfam analysis and therefore were retained in the analysis. The remaining 13 low-scoring proteins were either predominantly LRRs or were receptor-like kinases; all lacked any recognizable NBS motifs. This analysis identified 194 annotated genes that encoded homologs of NBS-LRR R proteins.

In the third step, we performed TBLASTN analyses using eight sequences selected to represent the diversity of NBS-LRR proteins to search the entire Arabidopsis genomic sequence to ensure that there were no additional related genes that had not been identified as ORFs by the automated annotation. All resulting sequences in the BLAST (Basic Local Alignment Search Tool) output (up to E = 1.0) were assessed manually for the presence of TIR, NBS, LRR, or R protein–like CC domains. This procedure identified four additional sequences. Finally, manual reannotation, intron/exon analysis, and protein motif comparisons were performed on all of the selected sequences to correct misannotation (as described below). Combined, these analyses identified 207 distinct genes encoding R protein–like TIR, CC, and NBS-LRR domains.

The predicted proteins encoded by these genes were classified initially based on Pfam protein motif analyses (Table 1). We restricted our current analyses to the 149 genes that encode both NBS and LRR domains because the LRR motif is present in diverse proteins unrelated to plant *R* genes. These 149 NBS sequences included 11 cloned *R* genes or the closest Col-0 homologs to *R* genes cloned from other Arabidopsis ecotypes. The additional 58 Arabidopsis genes identified during our search, most of which encode TIR motifs but not LRRs, have been described elsewhere (Meyers et al., 2002).

Detailed information about these NBS-encoding sequences is presented in our online database (http://www.niblrrs.ucdavis.edu). This database of NBS sequences includes links to the MIPS and TIGR Arabidopsis databases, gene locations, Pfam analyses of motifs, EST matches, and FASTA results for these sequences compared with either the complete Arabidopsis genome or the GenBank nonredundant set.

Predicted Pseudogenes and Annotation Errors Identified by Manual Reannotation

The initial sequence comparisons indicated that numerous NBS-LRR sequences had been partially misannotated during the automated annotation process. The automated annotations available in GenBank, MIPS, and TIGR represent powerful and useful initial attempts at annotation but generally have not been verified and corrected for individual genes and gene families (Haas et al., 2002). Therefore, we undertook the complete manual reannotation and analysis of the NBS-LRR gene family to rectify incorrect start codon predictions, splicing errors, missed or extra exons, fused genes, split genes, and incorrectly predicted pseudogenes. Nonfunctional genes, or "pseudogenes," were predicted on the basis of frameshift mutations or premature stop codons (Table 2); such reading frame disruptions were not identified by automated annotation programs, which instead inserted introns around the frameshift or nonsense mutations (data not shown). Mutations were identified by comparing DNA and protein sequences and by comparing intron positions and numbers of closely related gene homologs.

For each gene, the number of introns and their positions relative to encoded protein motifs and domains were determined. Intron positions and numbers generally were consistent with phylogenetic data, allowing the identification of anomalous exons and introns. Introns occurring in nonconserved locations

Table 2. Pseudogenes and Annotation Errors in Arabidopsis *CNL* and *TNL* Genes

Annotation Error	Identifiers, *CNL* Genes	Identifiers, *TNL* Genes
Incorrect intron/exon splice boundaries or numbers of exons	At1g51485, At1g58400, At1g59124, At5g45510, At1g58807, At1g61180, At1g61300, At1g61310	At1g72860, At5g22690, At4g16890, At1g31540, At4g11170, At4g16860, At4g16920, At4g16950, At4g16960, At4g19510, At4g19520, At4g19530, At5g17880, At5g44510, At5g45230, At5g46470, At5g51630
Misidentified frameshift (extra introns)[a]	At1g10920,[b] At1g59620[b]	At5g40060,[b] At2g17060,[b] At4g09360,[b] At3g25515,[b] At4g09430,[b] At4g16900,[b] At5g45240, At5g41740
Wrong start codon	At1g59780	At4g16940, At1g65850,[b] At1g63740, At5g46520
Gene fusion	At4g19050	At1g64070, At3g25510, At4g14370
Split gene	None	At1g57630, At2g17050, At5g46490
Truncated gene (from BAC terminus)	At1g58842, At1g63350	At5g38350
Wrong terminal exon	None	At1g56520
Premature stop codon (extra introns)[a]	At1g50180	At5g40920, At1g63860[b]
Error in genomic sequence	At4g14610[c]	At4g19500[c]
Annotation correct; motif analysis indicates deletion in protein	At5g47280, At4g27220, At1g61300	At5g45210, At4g09430, At4g16900, At5g40060, At3g04220, At3g25515, At5g17970, At5g40920, At1g56520

[a] Frameshifts or premature stop codons not identified by automated annotation programs resulting in erroneous splice predictions; some of these genes contained additional predicted annotation errors.

[b] Frameshifts or premature stop codons resequenced and verified, confirming the predicted pseudogene.

[c] Frameshifts resequenced and not confirmed. Genome sequence corrected, resulting in uninterrupted ORFs.

were reanalyzed by BLASTX comparisons using the intron sequence plus ~100 bp of 5′ and 3′ exon sequences. In 37 genes, either (1) translation and BLAST comparison of a small predicted intron matched the predicted protein sequence of another NBS-LRR protein (indicating that the intron prediction probably was incorrect), or (2) small additional nonconserved exons (<50 bp) were identified for which no similar exons could be found in comparisons with closely related genes (Table 2). In total, our reannotation of the *CNL* and *TNL* genes (genes that encode an N-terminal CC motif [CNL] or an N-terminal domain with TIR homology [TNL]) differed from the automated annotation in 56 of 149 genes. Combined with the reannotated *TX* (TIR-X) and *TN* (TIR-NBS) genes (Meyers et al., 2002), we calculated that ~36% of automated annotations contained errors. This value is consistent with that found in previous large-scale analyses of other Arabidopsis genes (Haas et al., 2002).

We amplified by PCR and resequenced genomic DNA from Col-0 to verify experimentally the predicted frameshift and nonsense mutations in the Arabidopsis Col-0 *CNL* and *TNL* genes. Our reannotation identified 13 genes for which the translation of a predicted intron sequence encoded protein sequence that matched other NBS-LRR proteins but included either a frameshift or a nonsense mutation (Table 2). We were able to amplify the regions encoding these mutations in 11 of the 13 genes; these 11 predicted pseudogenes contained 14 predicted mutations (Table 2; two sites each in At4g14610, At1g59620, and At4g09360). In 9 of the 11 genes, containing 11 of the 14 putative mutations, the sequences matched perfectly the published genomic sequence, indicating that these genes did contain disrupted reading frames and are likely pseudogenes. Neither of two frameshift mutations predicted in At4g14610 was found in the Col-0 accession that we analyzed, indicating a single complete ORF for this gene and errors in the published sequence. In addition, an error was identified in the sequence and annotation of the *TNL* gene At4g19500 (Meyers et al., 2002).

Additional pseudogenes were predicted as those that lacked specific motifs or contained large deletions even though they had apparently intact ORFs (Table 2). For example, At5g47280 lacks a CC motif in the predicted protein as a result of a deletion at the 5′ end of the gene. At5g45210 lacks most of the encoded LRR and C terminus present in other homologs. In the absence of functional data for these genes, it cannot be inferred with certainty whether these are pseudogenes. However, we identified 12 potential pseudogenes with uninterrupted ORFs that had deletions, in addition to the nine predicted pseudogenes with disrupted reading frames (Table 2).

In a few groups of closely related sequences, variable numbers of exons were observed, and these differences could not be attributed to disrupted reading frames or incorrect annotation (Figure 1). Among the *CNL* genes, At1g61180 and At1g61190 have an additional 3′ exon. Greater diversity in exon numbers was observed among the *TNL* genes than among the *CNL* genes, with most *TNL* genes containing four exons and most *CNL* genes containing only one exon (Figure 1). The Col-0 homologs of the *RPP1* genes (Botella et al., 1998), including genes At3g44480, At3g44510, At3g44630, At3g44670, and At3g44400, show an unusual exon configuration; some of these genes contain an additional 5′ exon and/or 3′ exon. Database searches with these genes identified two ESTs, providing evidence of alternative splicing of the exons at the 3′ end of the gene. This finding indicates that there may be additional variation in the exon number that cannot be determined without full-length cDNA clones. In addition, we have not considered noncoding exons in the 5′ and 3′ untranslated regions in this analysis, although among known *R* genes in Arabidopsis, noncoding exons have been reported only for *RPP1* (Botella et al., 1998). Analysis of cDNA sequences from the 5′ and 3′ ends of the NBS-LRR–encoding genes demonstrates that 10 of 80 analyzed genes contain noncoding exons (X. Tan, B. Meyers, and R.W. Michelmore, unpublished data).

Intron Positions and Phases Distinguish Subgroups and Indicate the Modular Nature of TNL Proteins

We analyzed the intron positions and phases in the different subgroups of the 149 *CNL* and *TNL* genes and in the closely related genes to assess the diversity within and between each group. Intron phases in spliceosomal introns can be classified based on the position of the intron with respect to the reading frame of the gene: phase-0 introns lie between two codons; phase-1 introns interrupt a codon between the first and second bases; and phase-2 introns interrupt a codon between the second and third bases (Sharp, 1981). Intron phases usually are conserved, because a modification of the phase on one side of the intron requires a concordant change at the distal location to maintain the reading frame (Long and Deutsch, 1999). Three distinct patterns of intron phases and positions were identified in *CN* and *CNL* genes (Figure 1A). These probably reflect the independent acquisition or loss of introns; a fourth pattern exhibited by two genes reflects the addition of a 3′ exon separated by a phase-0 intron. A greater degree of variation in the number of introns was observed among *TN*, *TX*, and *TNL* genes, but the positions and phases of individual introns were highly conserved with respect to the protein motifs encoded by flanking exons (Figures 1B and 1C). Much of the variation in intron numbers in the *TNL* genes was caused by the addition of 3′ exons that encode LRR motifs separated by phase-0 introns (Figure 1B). The greater diversity of intron positions and phases in the *CN*/*CNL* genes (as opposed to intron and exon numbers) may indicate that this group is more ancient than the *TN*/*TNL* gene family. Recent studies also have found shorter branch lengths for phylogenetic trees of *TNL* genes (Cannon et al., 2002), also suggesting that this group may have evolved more recently than the *CNL* genes.

Conserved Domains and Motifs in CNL and TNL Proteins

The 149 reannotated *CNL* and *TNL* genes were translated and subjected to protein domain and motif analyses. The protein analysis programs hmmpfam and hmmsearch (Eddy, 1998) were used initially to identify the major domains encoded in these genes. These programs were suitable for defining the presence or absence of the TIR, NBS, and LRR domains, but they could not recognize smaller individual motifs or more dispersed patterns, such as those present in the CC domain. Based on preliminary Pfam analyses of the entire predicted

proteins as well as homology with previously described motifs within the NBS (Meyers et al., 1999, 2002; Cannon et al., 2002), we initially divided the 149 genes into two major classes that encode either 55 CC-NBS-LRR or 94 TIR-NBS-LRR proteins. The NBS domain was defined by Pfam analysis; the NBS, N-terminal, and LRR plus C-terminal regions then were analyzed individually using the program MEME (Multiple Expectation Maximization for Motif Elicitation) (Bailey and Elkan, 1995). These analyses are described below in the order in which the domains are positioned in the proteins, starting at the N terminus (Figure 1).

The N-Terminal Domain

Immediately adjacent to the translation initiation codon of the majority of TNL proteins, we identified N-terminal amino acid residues similar to those that may enhance gene expression and protein stability. Analysis with MEME identified the motif SSSSSRNWRY N-terminal to the first TIR motif with a score of $<e^{-04}$ in 67 of 93 proteins classified as TNLs (MEME output 1; see supplemental data online). Similar Ala-polyserine sequences immediately after the N-terminal Met [$MA(S)_n$] have been found in a variety of highly expressed genes, and mutations in these sequences have been shown to reduce reporter protein stability in plants (Sawant et al., 2001). Twenty-nine of the 67 TNL proteins with the Ser-rich motif at the N terminus had sequences close to the consensus $MA(S)_n$; an additional 23 more TNL proteins had variants of $MA(S)_n$ with several nonconserved substitutions (see supplemental data online). The Ser-rich motif was present in 12 of the closest homologs of RPP28 (At2g14080) (N. Sepahvand, P.D. Bittner-Eddy, and J.L. Beynon, unpublished data); however, it was preceded by an ~40–amino acid N-terminal region containing a unique conserved motif (motif 13 in MEME output 1; see supplemental data online). The three closest homologs to the *R* gene *RPP1* in the ecotype Wassilewskija also encoded motif 13 as well as an additional N-terminal novel motif encoded by a separate 5′ exon that was described previously by Botella et al. (1998). No sequences related to $MA(S)_n$ were present at the N terminus of CNL proteins.

Several conserved motifs were confirmed that had been identified previously in the TIR domain of plant NBS-LRRs and related proteins (motifs TIR-1, TIR-2, TIR-3, and TIR-4) (Meyers et al., 1999, 2002). The order of these motifs was well conserved. Previous findings had noted duplications of the TIR motifs in some Arabidopsis proteins (Meyers et al., 1999); these unusual proteins in the TNL-A subgroup (Figure 1) are considered in more detail below and by Meyers et al. (2002). Within the group of TNL proteins, only the TNL-A subgroup contained a slight variation on the TIR-A motif (MEME output 1; see supplemental data online). Overall, the TIR motifs of the TNL proteins were essentially as described previously (Meyers et al., 2002) and included ~175 amino acids.

The presence of an N-terminal CC domain has been identified as a characteristic motif in the N terminus of the CNL R proteins (Pan et al., 2000), and the presence or absence of a CC motif can be anticipated on the basis of characteristic motifs present in the NBS (Meyers et al., 1999, 2002). We had initially defined the group of 55 CNL proteins based on motifs in the NBS and a lack of TIR motifs (Table 1). Because CC motifs are not defined in the Pfam database, motifs within the N-terminal region of CN and CNL proteins were analyzed using the program COILS (Lupas et al., 1991) to assess the positions and prevalence of CC motifs. In total, the CC domain of the CNL proteins spanned ~175 amino acids N terminal to the NBS. The predicted CC motif was positioned from 25 to 50 amino acids from the N terminus in most CNL proteins. There was strong evidence of an N-terminal CC motif in 50 of 55 CNL proteins; evidence for a CC motif was weak in At3g14460. Four proteins (NL proteins [Table 1]) had NBS motifs similar to CNLs but lacked a CC motif. At5g47280 and At1g61310 contained apparent N-terminal deletions that removed the region of the protein in which the CC motif was found in closely related homologs of these proteins. At4g19050 and At5g45510 were divergent NBS-LRR proteins that showed no evidence of a CC motif and contained few amino acids N terminal to the NBS (Figure 1C). Four of five CN proteins had a clear CC motif; At5g45440 did not. Using COILS, CC motifs were not identified in the N terminus of TN or TNL proteins, demonstrating the specificity of this motif to the CNL group.

We identified 20 distinct motifs in the N-terminal domain from the 60 CNL proteins using MEME (Figure 2; MEME output 4; see supplemental data online). Fourteen motifs were common and found in more than six CNL proteins. Up to seven motifs were present in individual proteins. In 49 proteins, one of two distinct MEME motifs, 1 or 7, was coincident with the CC pattern identified by COILS. We identified three patterns of CC domains based on shared MEME motifs (see supplemental data online). These three CC motif patterns (CNL-A, CNL-B, and CNL-C/D) matched the subgroups defined by intron position (Figure 1) and the clades identified in phylogenetic analyses using the NBS domain (see below). Pair-wise comparisons of motifs demonstrated little sequence similarity or overlap between distinct motifs located in similar positions in the CC domains of these three subgroups. One subgroup was divided further; the CNL-C motif pattern was closely related to but distinct from the CNL-D pattern. Among the five CN proteins, the CC domain of the CN-B class was closely related to that of the CNL-B class, whereas the CN-C class was more divergent (see supplemental data online). Although At5g45440 did not contain a predicted CC motif, it did have conserved N-terminal motifs (MEME output 4; see supplemental data online). The BLAST search of the Arabidopsis genomic sequence described above also revealed a gene, At3g26470, that encodes only a CC domain related to the CNL-A subgroup (score of $5e^{-48}$); this is the C protein listed in Table 1.

The NBS Domain

Previous work had identified eight major motifs in the NBS region, and several of these motifs demonstrated different patterns depending on whether they were present in the TNL or CNL groups (van der Biezen and Jones, 1998b; Meyers et al., 1999). We analyzed the 149 TNL and CNL predicted proteins using MEME. MEME identified motifs that matched the eight major motifs identified previously. However, MEME identified more than eight motifs. The configuration of the motifs identi-

fied by MEME reflected conservation within subgroups and diversity between different subgroups of TNL and CNL sequences (Figure 2; see supplemental data online). The eight major motifs differed in their divergence within and between the CNL and TNL groups (Table 3). In the current study, the pre-P-loop sequence (described previously as part of the TIR [Meyers et al., 1999]) and the P-loop were considered as a single motif. The P-loop, kinase-2, RNBS-B, and GLPL motifs demonstrated a high level of similarity between CNL and TNL proteins (Table 3). The RNBS-A and RNBS-D motifs were dissimilar, and the RNBS-C motif had low similarity between the Arabidopsis CNL and TNL proteins (Table 3), as was observed for plant R protein homologs in general (Meyers et al., 1999).

Although not immediately apparent from the consensus sequence shown in Table 3, the second and third amino acids of the GLPL motif in the NBS of many TNL proteins did not match the commonly identified consensus core GLPL (see NBS alignment in the supplemental data online). Rather, the most common variations contained the consensus GNLPL or SGNPL and lacked contiguous GL residues within the core of the motif. This is critical to the design of degenerate oligonucleotide primers for the amplification of *R* genes that often have used this motif (see Discussion).

Finally, the eighth conserved major motif in the NBS has been called MHDV, based on clearly conserved amino acids in the CNL proteins (Collins et al., 1998). This motif was beyond the most C-terminal RNBS-D motif identified in our previous work (Meyers et al., 1999) and was highly conserved in CNL proteins, with a minor variation (QHDV) present in the CNL-A subgroup (Table 3; see supplemental data online). The MHDV motif is slightly different in the TNL proteins, but it is clearly present and also starts with a conserved Met followed by a His (Table 3). The MHDV motif was not identified in any of the proteins that lacked an LRR (CN or TN), nor was it present in the divergent NL proteins At5g45510 and At4g19050. We considered this motif to represent the C-terminal end of the NBS, at least when LRRs are present. Mutations in the conserved Asp of the CNL variant of the MHDV motif resulted in a gain-of-function phenotype in the potato Rx protein (Bendahmane et al., 2002). In total, the eight NBS motifs from P-loop to MHDV spanned ~300 amino acids in the CNL and TNL proteins.

The LRR Region

The LRR region is characterized by leucine-rich repeats C-terminal to the NBS in many *R* genes (Jones and Jones, 1997). However, the precise start and number of LRRs had not been well defined in many NBS-LRR proteins. Therefore, we analyzed all predicted protein sequences encoded 3′ to the NBS to define the boundaries, numbers, and diversity of repeats in this domain. Initially, MEME was used as described previously except that the length and number of sequences required two rounds of analysis. First, samples of the CNL and TNL groups were analyzed together; then, all sequences within each group were analyzed separately. Parallel to the MEME analysis, we used the method described by Mondragon-Palomino et al. (2002) to estimate the number of LRR units in each protein. We manually combined secondary structure analyses derived from the program SSPro (Pollastri et al., 2002) with LRR consensus sequences to identify the individual repeats.

As a first step in defining the full LRR, we sought to determine if the LRR domain began immediately C terminal to the MHDV motif (the last conserved NBS motif) or if a spacer region separated the two domains. We analyzed all amino acids encoded immediately 3′ to the encoded MHDV motif. In *TNL* genes, a short exon averaging ~300 bp was found between the encoded NBS described above and longer exons more 3′ that clearly encoded LRR motifs. This exon is conserved in diverse *TNL* genes from other plant species (see above). In the latter half of this exon, previous studies identified hypervariable amino acids and predicted up to two LRR motifs encoded for some Arabidopsis *TNL* genes (Noel et al., 1999). Our MEME analysis identified motifs matching the canonical LRR patterns (Jones and Jones, 1997) encoded at the 3′ end of this exon (identified as 5 or 14 in the NBS MEME analysis; see supplemental data online). The manual analysis confirmed two LRRs encoded in this exon. In addition, two conserved motifs that were not identified as LRRs were found between the NBS and LRR domains in TNL proteins. MEME motif 8 was bisected by the intron, and motif 11 was in the middle of the short exon N-terminal to the first LRR (MEME analysis 2; see supplemental data online). Therefore, there were ~65 amino acids between the NBS and LRR domains in TNL; we designated this non-LRR region the NL linker (NBS-LRR linker).

CNL genes predominantly lacked an intron between the NBS and the LRR. Only the CNL-A class had an intron in this position (Figure 1). Manual analysis of LRR motifs in the CNL proteins identified LRR repeats starting ~40 amino acids C terminal to the NBS MHDV motif, consistent with previous analyses of individual CNL proteins (Bent et al., 1994; Grant et al., 1995; Warren et al., 1998; Cooley et al., 2000). MEME motif analysis in this region of the CNL sequences identified a short conserved NL linker of ~40 amino acids. The motif for this linker was conserved within the different CNL classes but varied among classes (Table 3; motifs 9 [latter half], 14, and 28 in MEME analysis 5; see supplemental data online). In TN and CN proteins that lack the LRR (Meyers et al., 2002), we found no evidence of the NL linker protein sequences.

The C-terminal boundary of the LRR region was defined as the point at which LRRs no longer could be recognized. Based on the manual and MEME analyses, LRRs constituted approximately half of the C-terminal region in the TNL proteins and nearly the entire C-terminal region in CNL proteins. The average TNL LRR domain contained a mean of 14 LRRs (standard deviation of 4.2, range of 8 to 25; see supplemental data online). MEME analysis of the TNL LRR domains identified ~10 distinct MEME motifs that spanned ~350 amino acids. The CNL proteins also had a mean of 14 LRRs (standard deviation of 3.5, range of 9 to 25; see supplemental data online), including ~10 distinct MEME motifs with >350 amino acids. Although MEME motifs did not correspond precisely to individual LRR units, duplication patterns were observed clearly as repeated motifs in >18 CNL LRRs and 46 TNLs (MEME analyses 3 and 6; see supplemental data online). These data suggest that CNL and TNL LRR domains are similar in length and that duplications of LRRs accounted for much of the variation in length.

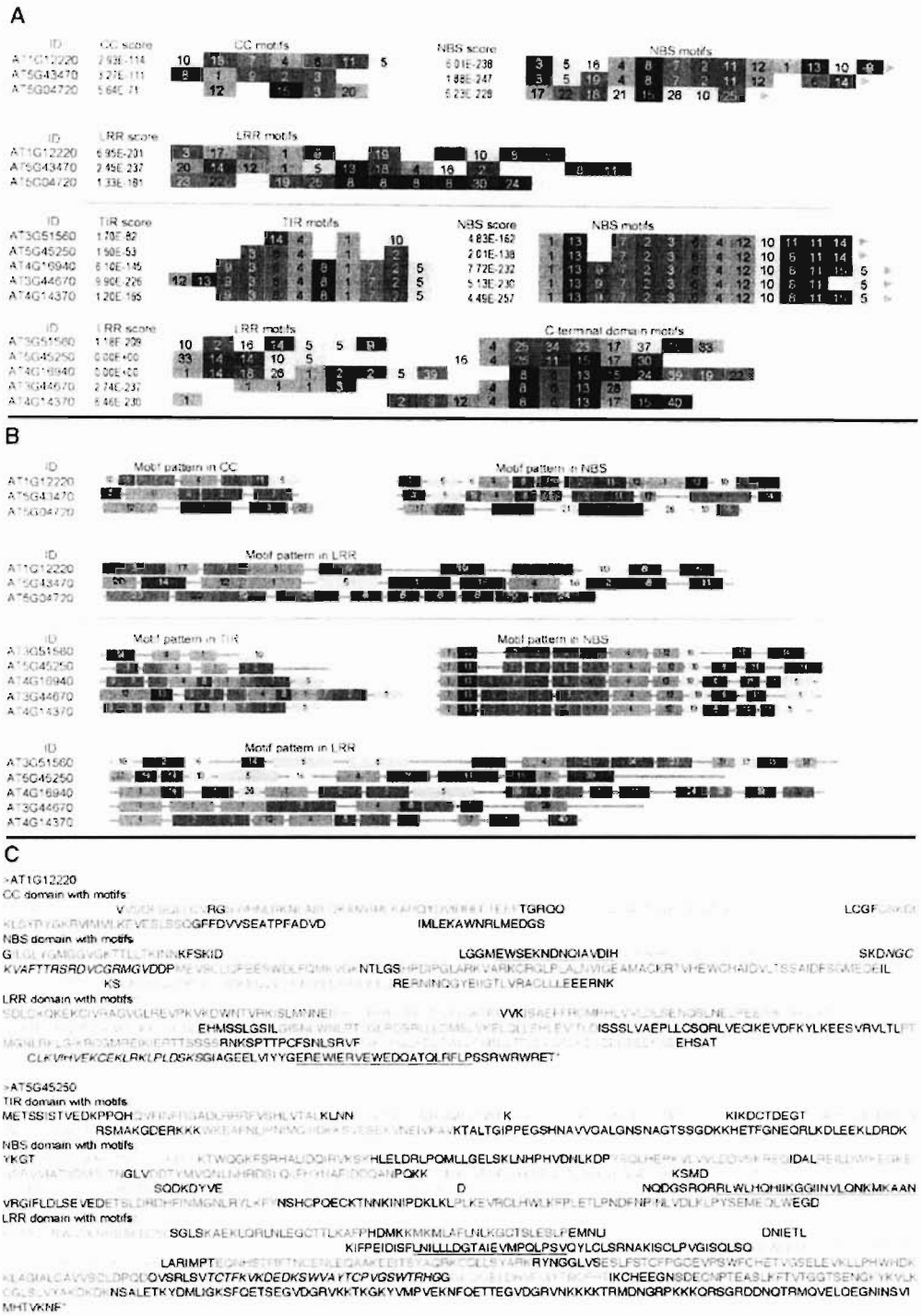


Figure 2. Motif Patterns in CNL and TNL Proteins.

Different colored boxes and numbers indicate separate and distinct motifs identified using MEME (Bailey and Elkan, 1995) and displayed by MAST (Bailey and Gribskov, 1998). Motifs are colored the same in **(A)**, **(B)**, and **(C)**. ID, identifier number.

(A) Examples of summarized and aligned MEME motifs for different domains of CNL and TNL proteins. All proteins are displayed in the supplemental data online. Thin dotted lines indicate their linear order. Motifs from the MEME analyses in supplemental data online (MEME outputs 1 to 6) were con-

Finally, the MEME motifs and patterns of repeats in the manually defined LRRs were examined to determine the conservation of LRRs within and among CNL and TNL proteins. MEME identified a variety of LRR-related motifs. These MEME motifs were less consistent in order, spacing, and number than MEME motifs identified in the other domains (see supplemental data online). Most proteins did not have a regular pattern; however, several predicted proteins had highly regular patterns of repeats, including At1g69550, At5g44510, and At2g14080 and to a lesser extent At1g27170 and At1g27180. Few motifs were similar between TNL and CNL proteins (MEME analysis 7; see supplemental data online). Motif 1 in the LRR domain of both TNL and CNL proteins was related (Table 3). This MEME-identified motif corresponds to the previously described, conserved third LRR, in which a mutation in the Arabidopsis CNL RPS5 had epistatic effects on disease resistance (Warren et al., 1998) and a mutation produced a gain-of-function phenotype in the potato Rx protein (Bendahmane et al., 2002).

In the TNL proteins, C terminal to the location of the motif-1 complex, duplicated patterns of LRR motifs were observed. In some subgroups, predominantly TNL-E, separate exons encoding duplications within the LRR region were common (Figure 1). These duplicated exons were recognizable by the repetition of LRR motif 1; this motif was encoded at the 5′ end of these exons. The 24 proteins in subgroup TNL-H were homogeneous in the composition and arrangement of their LRR motifs, probably reflecting the recent expansion of the subgroup (see supplemental data online). Motif 4 included the most C-terminal recognizable LRR motif in most TNL subgroups (Table 3; see supplemental data online).

In the CNL proteins, the LRR motif patterns were conserved within subgroups, but each subgroup was characterized by distinct sets of motifs. Motif 1 was conserved in all CNL subgroups except for CNL-A, which lacked this motif. Several motifs were unique to individual subgroups (see supplemental data online). The final LRR motif detectable in most CNL proteins was motif 8 (Table 3; see supplemental data online). The last occurrence of this motif typically ended 40 to 80 amino acids before the C terminus of the protein.

The C-Terminal Domain

The CNL and TNL groups differed markedly in the size and composition of sequences C-terminal to the LRR domain. The difference in the C-terminal domain accounted for much of the increased total length of TNL versus CNL proteins. The CNL proteins had conserved motifs present in the 40– to 80–amino acid C-terminal domain; like the NL linker, these motifs were specific to the CNL-A, CNL-B, and CNL-C/D subgroups (Table 3; see supplemental data online). By contrast, the C termini of the TNL proteins had a large number of non-LRR conserved motifs spanning ~200 to 300 amino acids. As reported previously for TNL proteins of known function (Gassmann et al., 1999; Dodds et al., 2001), the C-terminal non-LRR domain is approximately as large as the LRR domain. The two motifs, 8 and 25 (MEME analysis 3; see supplemental data online), began subsequent to the last LRR (motif 4) in most proteins of all TNL subgroups. C-terminal motifs were conserved within each subgroup but were less conserved among subgroups than were motifs within the TIR or NBS domains (see supplemental data online). In several members of the TNL-F subgroup, duplications of entire exons resulted in duplicated C-terminal motifs. Although the functional roles of these C-terminal motifs are unclear, their conservation and wide distribution throughout the TNL subgroup suggests that these domains are important for protein function.

A putative nuclear localization signal (NLS) was described by Deslandes et al. (2002) in the C-terminal domain of the Arabidopsis TNL:WRKY resistance protein RRS1 and cited as evidence for the nuclear localization of R genes (Lahaye, 2002). The motif patterns in the C-terminal domain of RRS1 and its putative Col-0 ortholog At5g45050 were similar to those of other TNL-A subgroup members. MEME motif 17 included the putative NLS identified by Deslandes et al. (2002) and was found in the C-terminal domain of most TNL proteins (MEME analysis 3; see supplemental data online). However, the particular amino acids representing the putative NLS sequence were not conserved among TNL proteins, suggesting that the proposed NLS in RRS1 is either spurious or a unique variant of the conserved C-terminal domain found in most TNL proteins.

Nonconserved Domains

Nine TNL proteins had unusual configurations or additions other than the TIR-NBS-LRR C-terminal domain structure described above (Figure 1). Most of these proteins were in either the TNL-A or the TNL-C subgroup. Several of these predicted anomalous domain configurations have been confirmed in pre-

Figure 2. (continued)

solidated and aligned manually in a spreadsheet. To allow alignment, the size of the colored and numbered box does not correspond to the size of the motif. Because motif analyses had to be performed for each domain separately for each of the CNL and TNL groups of proteins, numbers and colors are specific only to that domain. The MEME "score" for the overall match of the protein to the motif models is given as a P value. Missing motifs may indicate either a poor match ($>e^{-04}$) or a deleted domain.

(B) Examples of MEME output of the same proteins summarized in **(A)**. Data for all proteins are available in the supplemental data online (MEME outputs 1 to 6). The sizes of the boxes and the gaps between motifs are drawn according to scale to illustrate the relative sizes and positions of each domain and motif that is not displayed in **(A)**.

(C) Two examples of the motifs found in individual CNL and TNL protein sequences that are displayed in **(A)** and **(B)**. Colors were added manually to illustrate the motifs identified by MEME and displayed by MAST. MEME motif alignments with the sequences are available in the output of the MAST program in the supplemental data online (MAST outputs 1 to 6).

Table 3. Major Motifs in Predicted Arabidopsis CNL and TNL Proteins

Domain	(Sub)Group	Motif[a]	Sequence[b]
TIR	TNL	TIR-1	DVFPSFRGEDVRKTFLSHLLKEF
	TNL	TIR-2	IGPELIQAIRESRIAIVVLSKNYASSSWCLDELVEIMKC
	TNL	TIR-3	ELGQIVMPIFYGVDPSDVRKQ
	TNL	TIR-4	WRKALTDVANIAGEHS
TN linker	TNL		NxTPSRDFDDLVGIEAHLEKMKSLLCLES
CC	CNL-A to -D		See MEME outputs in supplemental data online
NBS	TNL	P-loop	VGIWGPPGIGKTTIARALF
	CNL	P-loop	VGIYGMGGVGKTTLARQIF
	TNL	RNBS-A	DYGMKLHLQEQFLSEILNQKDIKIxHLGV
	CNL	RNBS-A	VKxGFDIVIWVVVSQEFTLKKIQQDILEK
	TNL	Kinase 2	RLKDKKVLIVLDDVD
	CNL	Kinase 2	KRFLLVLDDIW
	TNL	RNBS-B	QLDALAGETxWFGPGSRIIVTTEDK
	CNL	RNBS-B	NGCKVLFTTRSEEVC
	TNL	RNBS-C	NHIYEVxFPSxEEALQIFCQYAFGQNSPP
	CNL	RNBS-C	KVECLTPEEAWELFQRKV
	TNL	GLPL	EVAxLAGGLPLGLKVL
	CNL	GLPL	EVAKKCGGLPLALKVI
	TNL	RNBS-D	EDKDLFLHIACFFNG
	CNL	RNBS-D	CFLYCALFPEDYEIxKEKLIDYWIAEGFI
	TNL	MHDV	MHNLLQQLGREIV
	CNL	MHDV	VKMHDVVREMALWIA
NL linker	TNL	NL	QFLVDAEDICDVLTDDTGTEK(x)$_{\sim 13}$ELxISEKAFKGMRNLRFLKIY(x)$_{\sim 18}$PPKLRLLHWDAYPLKSLPxxF NPENLVELNMPYSKLEKLWE
	CNL-B	NL	SDFGKQKENCIVQAGVGLREIPKVKNWGAVRRMSLMNNQIEHITCSPECPELTTLFLQYNQ
	CNL-C/D	NL	KEENFLQITSDPTSTANIQSQxxxTSRRFVYHYPTTLHVEGDINNPKLRSLVV
LRR	TNL	Motif 1 (LDL)	MDLSYSRNLKELPDLSNATNLERLDLSYCSSLVELPSSI
	CNL	Motif 1 (LDL)	IGNLVHLRYLDLSYTGITHLPYGLGNLKKLIYLNL
	TNL	Motif 4 (end)	LHWLDLKGCRKLVSLPQLPDSLQYLDAHGCESLETVACP
	CNL	Motif 8 (end)	LHTITIWNCPKLKKLPDGICF
C terminus	TNL		See MEME outputs in supplemental data online
	CNL-B	CT	EPEWIERVEWEDEATKNRFLP
	CNL-C/D	CT	WKERLSEGGEDYYKVQHIPSV

[a] Domains and motifs are listed in the order that they occurred in CNL and TNL proteins, starting with motifs most N terminal in the protein. Some of the motifs have been described previously (Meyers et al., 1999, 2002). Numbers for LRR motifs refer to MEME motifs described in the supplemental data online.

[b] Consensus amino acid sequence derived from MEME. Related motifs in the NBS and LRR domains of CNL and TNL proteins are aligned. The complete output is available in the supplemental data online. Underlined residues indicate possible LRR consensus matches (Jones and Jones, 1997). x indicates a nonconserved residue.

vious experimental analyses (Deslandes et al., 2002; Meyers et al., 2002). At1g27170 and At1g27180 encode duplications of the TIR domain; At4g36140 and At4g19500 encode TN:TNL fusions; and At2g17050 and At4g19520 encode TNL:TX fusions. TN or TX proteins have been suggested to play a role as adapter proteins (Meyers et al., 2002). In addition, the *R* gene *RRS-1* and its Col-0 homolog At5g45050 encode a WRKY motif fused at the C terminus (Deslandes et al., 2002). At4g12020 is the most unusual TNL protein; it contains a WRKY-related protein domain at the N terminus and a sequence similar to mitogen-activated protein kinase kinase kinases in place of the C-terminal domain. Based on the varied similarities of its 16 exons, At4g12020 appears to be a chimera composed of parts of five other genes, and it shares a predicted promoter region of only 273 bp with At4g12010 (see below) (Figure 3A). At5g17890 encodes a TNL protein with a C-terminal fusion to a neutral zinc metallopeptidase; a similar domain also is present in one unusual CNX protein, At5g66630. The chimeric At5g66630 apparently resulted from a small translocation of the 5′ end of At5g66890 and resides within a small cluster of homologs, At5g66610 to At5g66640 (Figure 3B). The neutral zinc metallopeptidase family is encoded by only seven paralogs in the Col-0 genome, and two of these seven are part of either CNX or TNLX proteins (Figure 1). The functional significance of these unusual domain configurations and additions is unknown.

Phylogenetic Analysis of Predicted Proteins Containing NBS Sequences Related to *R* Genes

We assessed sequence diversity and relationships by generating two phylogenetic trees, one for the CNL proteins and one for the TNL proteins (Figures 4A and 4B). NBS sequences were

used because the NBS domain is present in both CNL and TNL proteins and contains numerous conserved motifs that assist proper alignment. The availability of full-length sequences allowed the use of the entire NBS domain (from ~10 amino acids N terminal to the first Gly in the P-loop motif to ~30 amino acids beyond the MHDV motif), in contrast to the earlier analysis of Meyers et al. (1999), which used only the region between the P-loop and GLPL motifs. Both CNL and TNL trees showed long branch lengths and closely clustered nodes, reflecting a high level of sequence divergence (Figures 4A and 4B). The nodes closest to the branch tips were supported most highly, although increased support would have been found for more of the internal nodes if the number of sequences had been reduced. The trees are robust, however, because phylogenetic analysis using both distance and parsimony algorithms produced similar trees (data not shown).

The phylogenetic relationships based on the NBS predominantly recapitulated patterns of protein and gene structure (Figures 4A and 4B). The motif patterns defined by MEME for each of the domains identified monophyletic clades within each of the CNL and TNL groups. In addition, genes that encode sequences in these clades shared intron positions and to a lesser extent numbers (Figures 1, 4A, and 4B). Together, intron numbers and positions, protein motifs, and phylogenetic analyses defined four subgroups of CNL proteins, eight subgroups of TNL proteins, and a pair of divergent NL proteins (Figures 1, 4A, and 4B). Among the CNL and TNL subgroups, only CNL-C was not monophyletic; phylogenetic analysis suggested that the CNL-D subgroup was derived from the CNL-C subgroup (Figure 4A). TNL subgroups were consistent with our previous phylogenetic analysis using the TIR domain (Meyers et al., 2002). The consistency among these three distinct sources of data—protein motifs, intron positions, and sequence diversity for the NBS and TIR regions—suggests that shuffling of protein domains has been rare among distantly related CNL or TNL sequences.

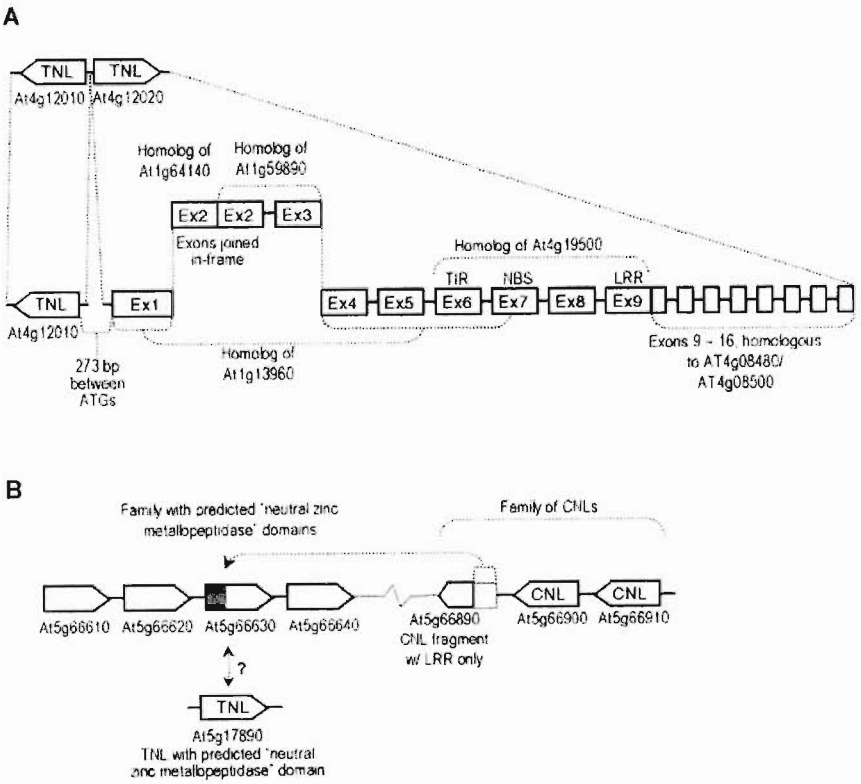


Figure 3. Modifications of Two TNL Proteins Caused by Genic Rearrangements.

(A) Gene At4g12020 encodes protein domains similar to five different genes. Exons (Ex) 2 and 9 encode in-frame fusions of distinct protein domains. Based on sequence homologies, exons 2 and 3 apparently were inserted into exons 1, 4, and 5. Exons 6 to 9 encode TNL domains fused at the 3′ end to a mitogen-activated protein kinase kinase kinase homolog. The complete gene was found in a head-to-head orientation with TNL At4g12010; 273 bp separates the predicted translational start codons of these genes.

(B) Gene At5g66630 encodes an NBS fused to neutral zinc metallopeptidase motifs; the NBS of this gene is related most closely to a nearby family of *CNL* genes, one of which is lacking the NBS region, suggesting a translocation of this domain. At5g17890 is a TNL fused to neutral zinc metallopeptidase motifs homologous with At5g66630 (BLAST E value = $3e^{-82}$).

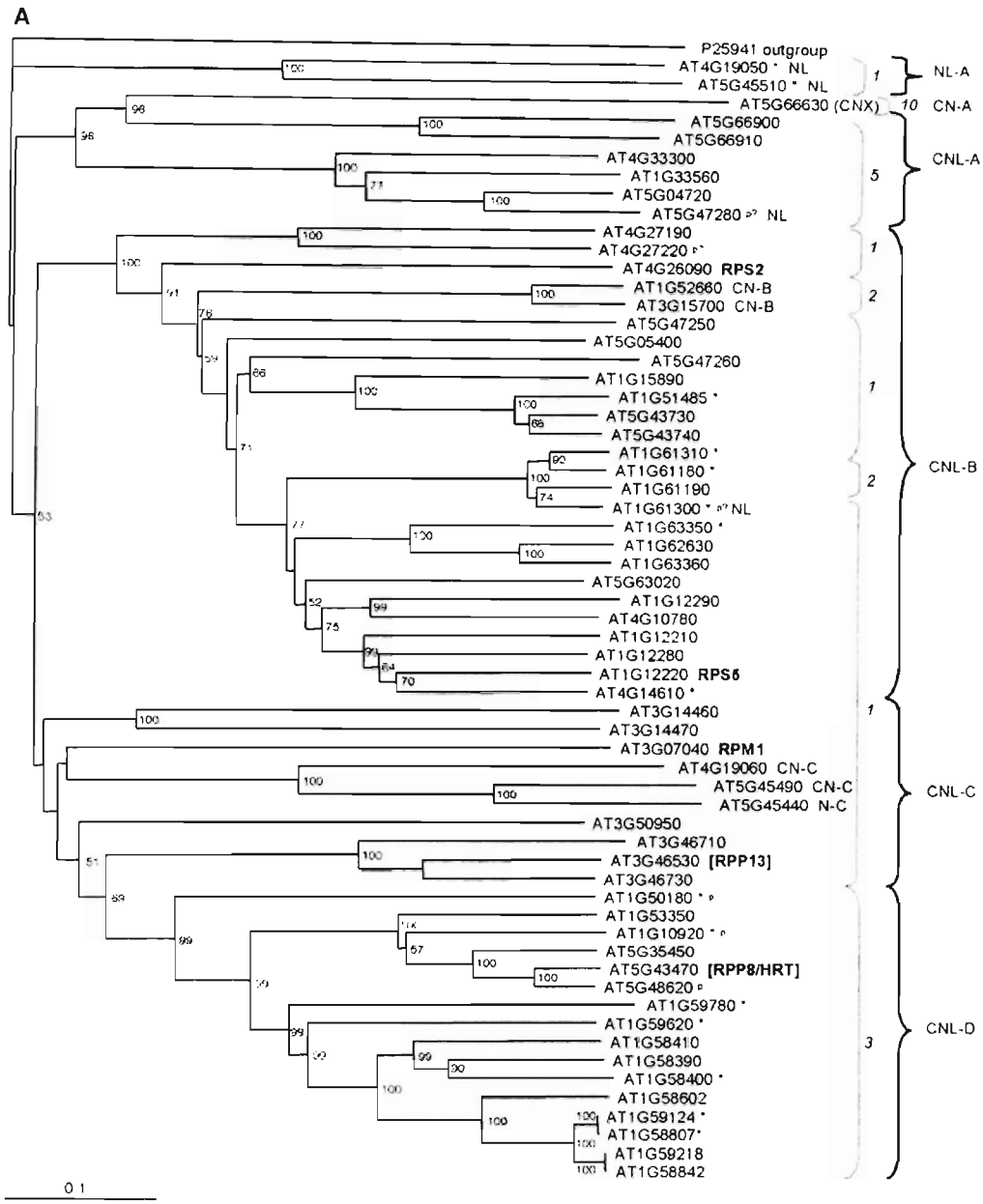


Figure 4. Phylogenetic Relationship of NBS-Containing Predicted Proteins from the Complete Arabidopsis Genome.

(A) Tree of CN and CNL proteins.

(B) Tree of TN and TNL proteins.

Neighbor-joining trees from distance matrices constructed according to the two-parameter method of Kimura (1980) using the aligned NBS protein sequences. Branch lengths are proportional to genetic distance. Sequence identifiers are given for each sequence as designated by the Arabidopsis Genome Initiative (2000). Names of known resistance gene products are indicated in boldface. The number of exons for each gene is indicated at right by gray brackets. Asterisks indicate that our gene prediction differed from that in MIPS and TIGR; superscript "p" indicates a predicted or potential pseudogene (see text). The *Streptomyces* sequence rooted both trees as the outgroup. Numbers on branches Indicate the percentage of 1000 bootstrap replicates that support the adjacent node; bootstrap results were not reported if the support was <50%. Black braces at right in each tree indicate the subgroup names; subgroups were defined based on phylogeny and intron position/number (see text). Proteins that contained either more or less than the CC-NBS-LRR domains (in **[A]**) or the TIR-NBS-LRR domains (in **[B]**) are indicated with a code after the identifier that refers to protein configurations in Table 1. Two sequences each had two NBS domains; these domains were included in the analysis with the primary subgroup (TNL-A) indicated in parentheses by the position of the second NBS. The trees are available at http://niblrrs.ucdavis.edu with links to data for each gene.

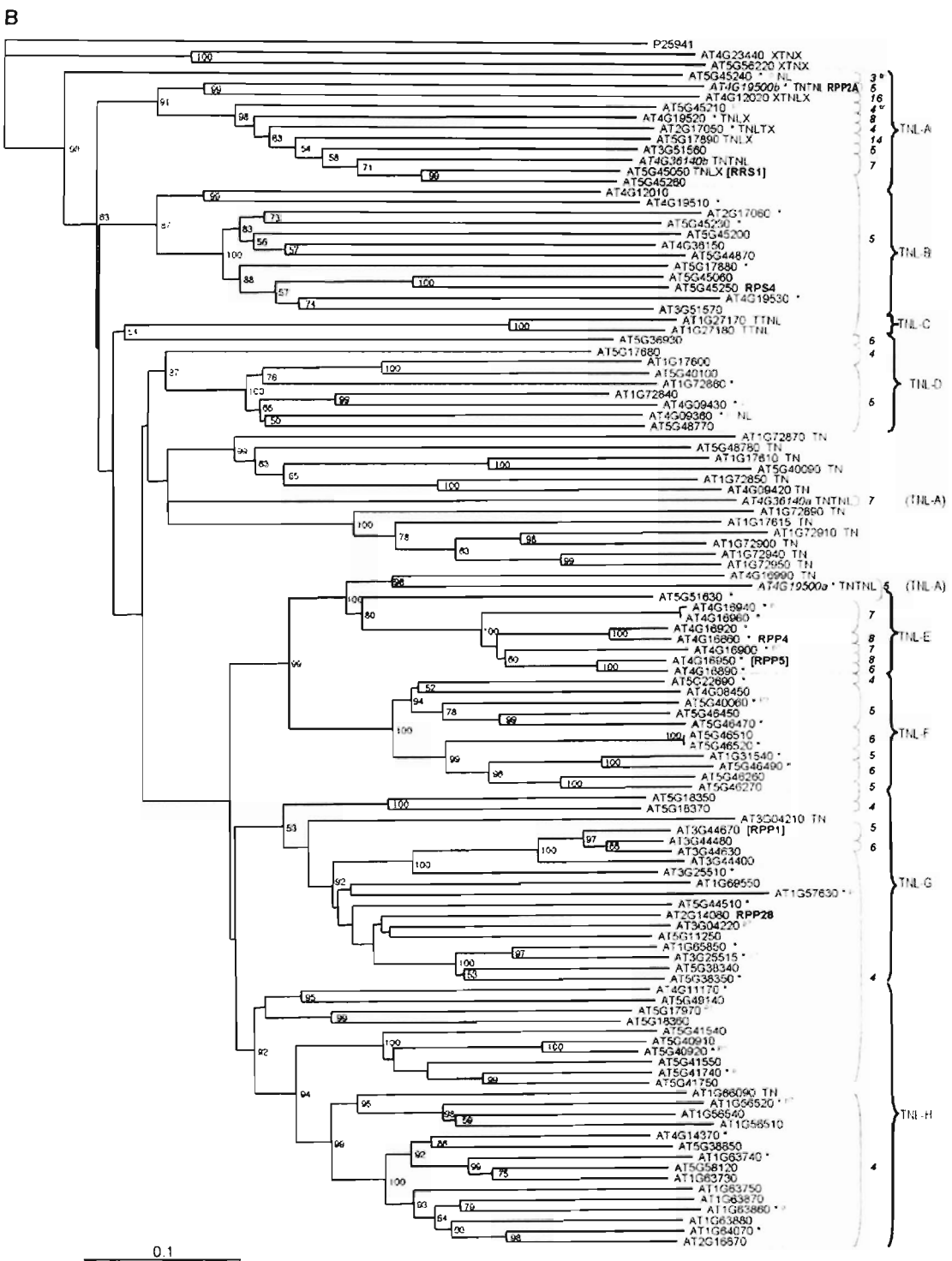


Figure 4. (continued).

Although TX, TN, and TNL sequences all contain TIR domains and presumably share an ancient ancestor, previous phylogenetic analyses of only the TIR-encoding domain demonstrated the diversification of two monophyletic clades of TN sequences and one clade of TX sequences (Meyers et al., 2002). Therefore, TIR domain relationships indicate that *TNL* genes evolved independently of most *TX* and *TN* genes. Phylogenetic analysis of the NBS region confirmed the existence of two major TN clades distinct from the TNL clades (Figure 4B). The NBS analysis also was consistent with several TN sequences being most closely related to TNL sequences rather than to other TN sequences (Meyers et al., 2002).

The known Col-0 R proteins and the closest homologs of the known Arabidopsis R proteins identified in ecotypes other than Col-0 were mapped onto the phylogenetic trees. Known R proteins were found in clades distributed throughout both trees. The TNL tree included RPS4, RPP4, RPP2A, and RPP28 from Col-0 as well as the closest Col-0 homologs of RPP1, RPP5, and RRS1. The CNL tree included RPM1, RPS2, and RPS5 from Col-0 and the closest Col-0 homologs of RPP8 and RPP13. Only five subgroups, NL-A, CNL-A, TNL-C, TNL-D, and TNL-H, did not include a known R protein. Therefore, more than two-thirds of all Arabidopsis Col-0 NBS-LRR proteins were within the same subgroup as at least one protein with a demonstrated role in disease resistance.

Genetic Events Resulting in the Expansion of the NBS-LRR Gene Family in Col-0

The physical distribution of NBS-LRR–encoding genes across the Col-0 genome was investigated to illustrate the genetic events that shaped the complexity and diversity of these genes. Both *CNL* and *TNL* genes showed obvious clustering in the genome (Figure 5). We also examined the distribution of *TX*, *TN*, and *CN* genes because these related genes are linked closely to some *TNL* genes (Meyers et al., 2002). We used the same parameters to define a cluster as Richly et al. (2002); two or more *CNL*, *TNL*, *TX*, *TN*, or *CN* genes that occurred within a maximum of eight ORFs were considered to be clustered. This is a useful operational definition because the numbers or sizes of clusters changed little when the maximum number of intervening ORFs was increased to 25 or even 50. In most cases, the function is not known for the other genes in the clusters that do not encode NBS-LRR proteins. Approximately two-thirds of *CNL* and *TNL* genes (109 of 149) were distributed in 43 clusters; the remaining 40 *CNL* and *TNL* genes were single-

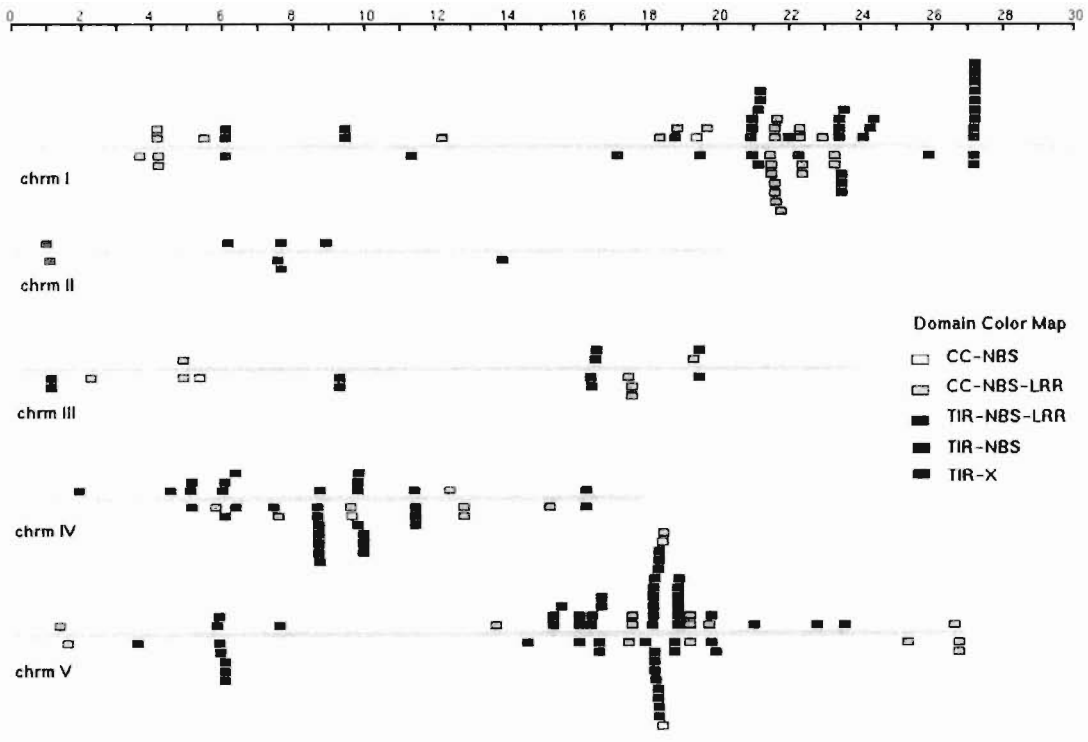


Figure 5. Physical Locations of Arabidopsis Sequences That Encode NBS Proteins Similar to Plant *R* Genes.

Boxes above and below each Arabidopsis chromosome (chrm; gray bars) designate the approximate locations of each gene. Chromosome lengths are shown in megabase pairs on the scale at top. A list of the clusters is given in the supplemental data online. Similar figures are available at http://niblrrs.ucdavis.edu with links to data for each gene.

tons (Table 4, Figure 5; see supplemental data online). The largest cluster consisting of only NBS-LRR–encoding genes was the *RPP4/RPP5* cluster, which constituted seven TNL sequences on chromosome IV (see supplemental data online). Sixteen clusters contained combinations of *TNL* or *CNL* genes with *TX*-, *TN*-, or *CN*-encoding genes (Table 4; see supplemental data online); the largest of these clusters contain *TNL* and *TN* genes or *TNL* and *TX* genes and have been described previously (Meyers et al., 2002). Of these 16 clusters, 12 contained *TNL* genes paired with *TX* or *TN* genes, one contained four *CNL* genes with a *TX* gene, and one contained three *TNL* genes with a *CN* gene (see supplemental data online). The two diverse *NL* genes, At4g19050 and At5g45510, were adjacent to one and two CN genes, respectively.

We compared the phylogenetic analysis and the physical clustering data to determine if clusters were composed solely of monophyletic clades (Figures 4A and 4B; see supplemental data online). Four clusters contained *CNL* and *TNL* genes from diverse subgroups, excluding the TNL-A/B pairs (see above). The clusters were At5g17880 to At5g17970 (representing subgroups TNL-A, -B, and -H), At5g18350 to At5g18370 (TNL-G and -H), At5g40060 to At5g40100 (TNL-F and -D), and At5g47250 to At5g47280 (CNL-A and -B). These clusters of mixed subgroups could have arisen as a result of either selective pressures (Richly et al., 2002) or chance events that colocalized the genes. Richly et al. (2002) estimated the number of heterogeneous clusters expected if the genes were arranged randomly in the genome, based on the total number of genes within the boundaries of the cluster. Using the same formula with the current estimated total of 29,028 genes in Arabidopsis (http://www.tigr.org), the number of mixed clusters predicted to occur at random was greater than the four that we identified. Therefore, in contrast to Richly et al. (2002), we conclude that these four mixed clusters are likely the result of random associations among the 149 NBS-LRR–encoding genes in the Col-0 genome and do not provide evidence for selection for mixed clusters.

The genes that encode the TNL-A and TNL-B proteins showed an unusual pattern of clustering. Seven clusters were identified that contained 11 paired sets of genes encoding members of the TNL-A and TNL-B subgroups (Figure 6A). Five clusters encoded one representative of each subgroup, and one cluster encoded 17 *TNL* and *TX* genes. Because the *TNL-A* and *TNL-B* genes each form a monophyletic group, the duplication of these genes took place after an ancestral pairing event and preserved their orientation. Ten of the 11 pairs of *TNL-A* and *TNL-B* genes maintained a head-to-head configuration (At4g19500 was inverted; Figure 6A). The most complex cluster included 17 *TNL* and *TX* genes (Meyers et al., 2002) and spanned a 246-kb region on chromosome V that included 39 predicted genes (Figure 6A). This cluster includes the known *R* genes *RPS4* (Gassmann et al., 1999) and *RRS1* (Deslandes et al., 2002). It is not known if the complexity of this cluster or the pairing of the TNL-A and TNL-B genes reflects selective pressure to maintain functional pairs of genes. It also is interesting that 9 of the 11 genes in the *TNL-A* subgroup encode proteins with very different and unusual additional domains (see above; Figures 1 and 6A). The additional domains do not share high sequence similarity and therefore apparently were acquired independently. The importance of these additional domains to the functions of most of these proteins is unknown; however, At5g45050 confers recessive resistance to *Ralstonia solanacearum* (Deslandes et al., 2002), and At4g19500 was identified recently as the *Peronospora parasitica* resistance gene *RPP2A* (E. Sinapidou, K. Williams, and J.L. Beynon, unpublished data).

Some of the *CNL* and *TNL* genes that were not in clusters (singletons) were related closely to clustered genes (Figures 4A and 4B; see supplemental data online). Small translocations apparently have separated these members of monophyletic clades and may have occurred quite frequently in the evolution of the Arabidopsis genome. These rearrangements have been local, to positions elsewhere on the same chromosome, or to other chromosomes. For example, two singletons, At1g59620 and At1g59780, are separated by ~17 and ~33 genes from the large cluster shown in Figure 6B on chromosome I. In the *TNL-H* subgroup, closely related sequences At1g63730 to At1g63750 are found as a cluster; however, the most closely related *TNL-H* homologs of these genes are found on chromosomes II, IV, and V (Figure 4B).

A comparison of the physical positions and the phylogenetic analysis revealed both local and distant duplications of *CNL* and *TNL* genes. The majority of the clusters contained closely

Table 4. Clusters of CNL- and TNL-Encoding Genes in Arabidopsis Col-0

Category[a]	No. of Clusters	No. of Genes
Monophyletic[b] duplicated *TNL* or *CNL*s	25	73
Mixed (*TN*, *TX*, and *CN* with *NL*, *TNL*, and *CNL*)	12	43
TNL-A/B pair	7	21
Mixed clusters of subgroups (not *TNL-A/B*)	4	11
Total in clusters with *NL*, *CNL*, and *TNL*	43	109 (+35 *TX*, *TN*, and *CN*)
Total in clusters with *TX* or *TN* only	4	11
CNL/TNL not clustered		40
Total genes[c] (*NL*, *CNL*, *TNL*, *TX*, *TN*, and *CN*)		207

[a] A complete listing and description of clusters is available in the supplemental data online. Categories are not mutually exclusive.
[b] Some clusters do not include all members of the monophyletic clade.
[c] See Meyers et al. (2002) for descriptions of the *TX*, *TN*, and *CN* genes included in this analysis.

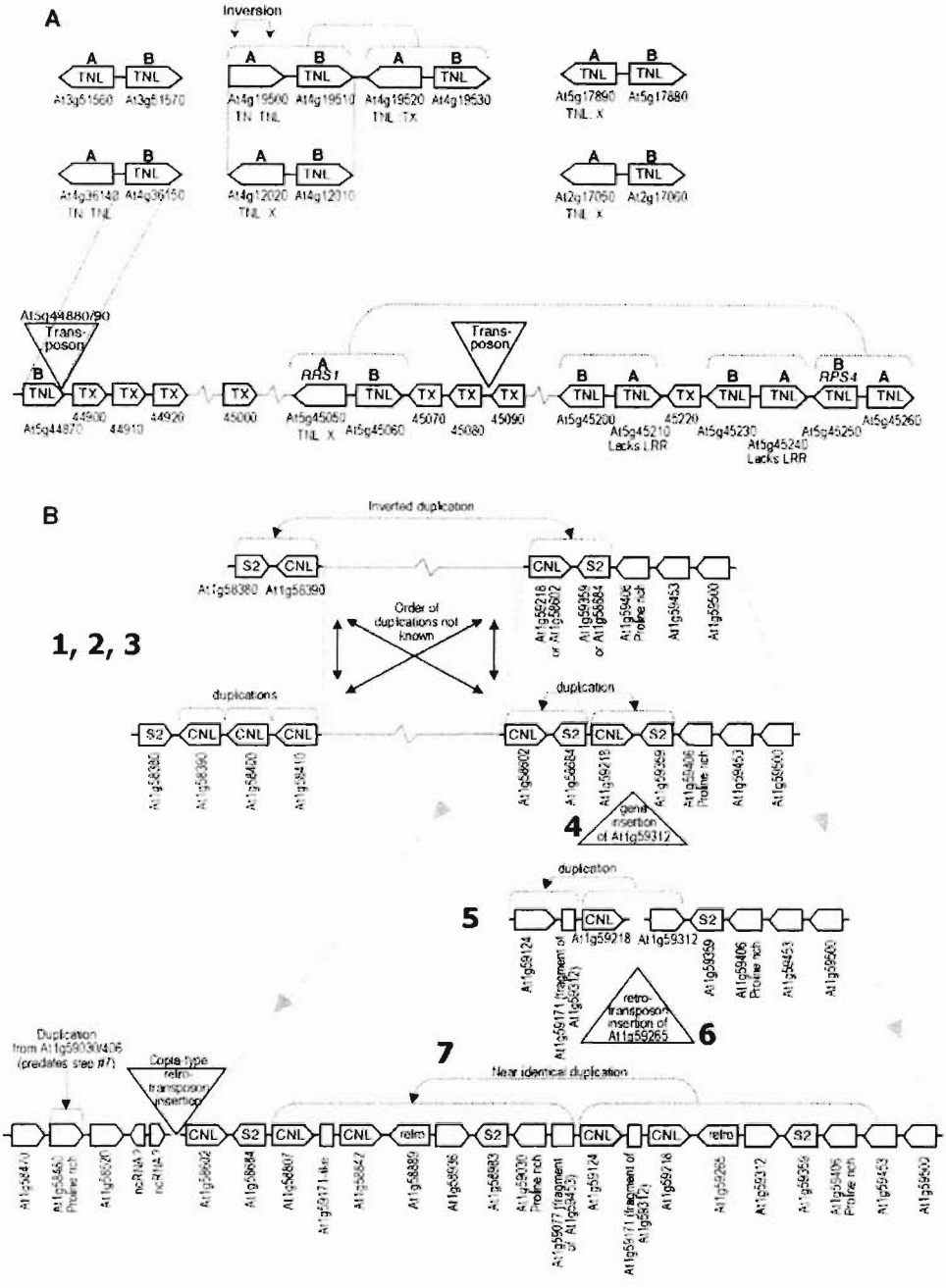


Figure 6. Multiple Localized Duplication Events That Resulted in Clusters of NBS-LRR–Encoding Genes.

Dotted lines designate the boundaries of duplication events inferred from closely related sequences. Triangles indicate the insertion site of a gene, transposon, or retrotransposon.

(A) An ancient pairing of genes that is present in ~11 occurrences in the Col-0 genomic sequence. Genes labeled A belong to the monophyletic subgroup *TNL-A*, and genes labeled B belong to the monophyletic subgroup *TNL-B*. See Figure 4 for more detailed phylogenetic relationships. B genes encode predicted TNLs, whereas A genes encode modified TNLs with additional protein motifs, as indicated below the gene identifier.

(B) A complex family of CNLs and unrelated genes on chromosome I. The evolutionary history of the cluster was inferred based on observed sequence homologies in the Col-0 genomic sequence. Boldface numerals indicate the order of events predicted in this region, as inferred from relationships of pairs of genes and gene fragments. Dashed lines that connect the ends of the clusters indicate the boundaries of a single region shown at different inferred evolutionary time points. The scheme at bottom represents the extant Col-0 sequence. The black arrows indicate that evidence of multiple duplication events was identified, but the order of these events could not be distinguished. ncRNA, noncoding RNA identified in the gene annotation.

related sequences from within the same CNL or TNL subgroup, indicating localized duplication events, most likely tandem duplications resulting from unequal crossing over. Several of these clusters have been noted previously and correspond to clusters of *R* genes defined by classic genetics (Holub, 2001). Expansion of a *TNL* cluster by tandem duplications and insertions of retrotransposons has been described for the *RPP4/RPP5* family (Noel et al., 1999). We examined the patterns of sequence similarity to infer the complex pattern of localized duplications and insertions that resulted in the expansion of two related *CNL* clusters on chromosome I (Figure 6B). The locations of gene fragments allowed us to infer the direction and boundaries of some of the duplication events. One of these clusters is a tightly clustered array of three *CNL* genes, whereas the other includes five *CNL* genes and numerous unrelated genes (Figure 6B). Early events in the expansion of these clusters included a distal duplication of single *CNL* genes and localized duplications of single genes, pairs of genes, and/or gene fragments. Later events included insertions of single genes and retrotransposons and finally a recent duplication of approximately eight genes, including two *CNL* genes (Figure 6B).

To investigate the role of large segmental duplications in the expansion of NBS-encoding genes, we analyzed the positions of *CNL*, *TNL*, and related genes relative to segmental duplications detected in the Col-0 genome. Boundaries of 81 previously described duplicated regions were derived as gene identifier numbers from http://www.psb.rug.ac.be/bioinformatics/simillion_pnas02/ (Simillion et al., 2002). These 81 duplications were all from those that contained at least 10 genes in common. We confirmed these genome duplications by BLAST comparison of all predicted Arabidopsis proteins against each other and displayed sequence similarities as a diagonal plot along each chromosome (see supplemental data online). Chromosomal positions using coordinates corresponding to the current annotation for each boundary gene as well as all of the *CNL*- and *TNL*-related genes also were displayed linearly using GenomePixelizer (see supplemental data online) (Kozik et al., 2002). The boundaries of the duplicated segments were joined by lines, as were *CNL*, *TNL*, and related genes with >60% amino acid identity.

The locations of *CNL*- and *TNL*-related genes relative to duplicated segments and their persistence in the duplicated regions then were assessed by visual inspection of the diagonal plot and the linear GenomePixelizer display. A total of 124 CNL- and TNL-encoding genes were located in duplicated regions (Table 5; see supplemental data online). These were distributed in 43 of the 162 segments involved in the 81 duplications. Twenty-five *CNL*- and *TNL*-related genes were not located in any of the 162 duplicated regions; however, some of these genes had paralogs with >60% identity that did reside in one segment of a pair of duplicated regions (e.g., At4g04110 and At5g58120). In 25 cases, the *CNL*- and *TNL*-related genes were present in only one of the two segments involved in the duplication: duplications 1.1.4 and 3.4.13 (Table 6; see supplemental data online). In only nine cases were the *CNL*- and *TNL*-related genes present in both segments involved in the duplication: duplications 1.1.2 and 3.5.1 (Table 6; see supplemental data online). However, close inspection of the diagonal plot revealed a more complex situation than simple duplication of a chromosomal region. Even when the genes resided in both members of a segmental duplication, only rarely were the NBS-LRR genes flanked by syntenic genes and therefore located along the diagonal line of the diagonal plot (see supplemental data online). Therefore, although some of the amplification of CNL- and TNL-encoding genes occurred as a result of segmental duplications that involved 10 or more genes, much of the amplification occurred independently of such duplications. The frequent presence of CNL- and TNL-encoding genes in only one segment of a duplication and at nonduplicated positions and their variable positions within duplicated segments suggest that microscale events involving translocations of NBS-LRR–encoding genes around the genome as well as deletions occurred after the segmental duplications by as yet undefined genetic mechanism(s).

We also analyzed sequence data from the Arabidopsis ecotype Landsberg *erecta* (L*er*) to examine the types of genetic events that shaped NBS-LRR gene clusters observed through intergenomic comparisons. In Col-0, the absence of clustering of the two *CNL* singletons (At5g43470 and At5g48620) belies the complexity of events that led to the Col-0 haplotype. In L*er*,

Table 5. Distribution of Three Multigene Families That Encode NBS-LRR, Cytochrome P450, and LRR Kinase Proteins in the Arabidopsis Col-0 Genome Relative to Segmental Duplications

	Gene Family		
Class	NBS-LRR	Cytochrome P450	LRR Kinase
No. of pairs of segmental duplications	81	81	81
No. of pairs with gene(s) in either or both segments	34	47	52
No. of pairs with gene(s) in only one segment	25	19	24
No. of pairs with gene(s) in both segments	9	28	28
No. of pairs with simple duplication of a gene[a]	4	15	21
Total genes in family	149	245	206
No. (%) of genes residing in segmental duplications	124 (83%)	199 (81%)	163 (79%)
No. (%) of genes in simple segmental duplications[a]	14 (9%)	81 (33%)	66 (32%)

[a] See text. Each pair of genes had to have at least 40% identity, and their element on the diagonal plot is located along the duplication diagonal (see supplemental data online).

Table 6. Relationships between Segmental Duplications and NBS-Encoding Genes

Duplication[a]	Boundary Gene Identifiers	*CNL* and *TNL* Gene Identifiers
Examples of persistence of *CNL* and *TNL* genes in duplicated segments		
1.1.2	At1g17230 to At1g22340	At1g17610
	At1g72180 to At1g78270	At1g72840, At1g72920, At1g72930
1.5.5	At1g65630 to At1g67270	At1g65850
	At5g36950 to At5g38690	At5g38340, At5g38350
3.5.1	At3g01015 to At3g04350	At3g04220
	At5g14060 to At5g18490	At5g18350 to At5g17890
Examples of *CNL* and *TNL* genes present in only one segment of the duplication		
1.1.4	At1g08970 to At1g10570	No *CNL*, *TNL*, and related genes
	At1g56170 to At1g60220	Contains 13 *CNL* and *TNL* genes
3.4.13	At3g21465 to At3g23870	No *CNL*, *TNL*, and related genes
	At4g13800 to At4g15640	At4g14370, At4g14610

[a] Segmental duplications as designated by Simillion et al. (2002).

there are four syntenic *CNL* genes that include *RPP8* (McDowell et al., 1998). Based on flanking genes and gene fragments, we were able to infer the history of rearrangements involving these CNL sequences (Figure 7). The initial event generating the locus that includes At5g43470 likely involved a small duplication from the locus that includes At5g48620 to a position ~2.3 Mb away on the same chromosome. A subsequent duplication event produced the functional *RPP8* gene and the homolog *RPH8* to generate the extant L*er* haplotype. This haplotype then underwent an unequal crossing-over event to produce the extant Col-0 haplotype (McDowell et al., 1998; Cooley et al., 2000). We sequenced 12.8 kb around the locus in L*er* syntenic with At5g48620 and found evidence of a duplication event that produced the pair of *CNL* genes in L*er* (Figure 7). These inferred complex histories demonstrate that gene duplications, translocations, and insertions of genes and mobile elements all have contributed to the configuration of several *CNL* and *TNL* clusters and singletons (Figures 6 and 7). As additional genomic sequence from other Arabidopsis ecotypes becomes available, it will become possible to infer the evolutionary history of many *CNL* and *TNL* genes and to determine the relative frequencies with which rearrangements, duplications, and deletions occurred.

DISCUSSION

The Col-0 Arabidopsis Genome Contains ~150 CNL and TNL Sequences in Distinct Subgroups

We have characterized the complete set of 149 CNL- and TNL-encoding genes in the current version of the Arabidopsis Col-0 genome. These represent ~0.5% of all predicted ORFs. Based on gene structure, protein motifs, and sequence divergence, we defined eight *TNL* subgroups and four *CNL* subgroups and identified one *NL* subgroup. Nearly two-thirds of all NBS-LRR–encoding genes were found in subgroups containing at least one known *R* gene or a Col-0 ortholog of a known *R* gene. In total, only four of eight *TNL* subgroups and one of four *CNL* subgroups did not include a known *R* gene or *R* gene ortholog. These genes could encode R proteins of as yet unknown specificities. The large number of NBS-LRR–encoding genes involved in defense that have been cloned from other plant species suggests that the frequency of NBS-LRR–encoding genes observed in Arabidopsis is not exceptional and that hundreds of NBS-LRR–encoding genes will be identified in each genome sequenced. The rice genome encodes >500 CNL proteins (Bai et al., 2002; Meyers et al., 2002). Several other types of proteins are encoded in plant genomes that also may be involved in early events leading to disease resistance, including kinases such as Pto in tomato (Martin et al., 1993), receptor-like kinases such as Xa21 in rice (Song et al., 1995), LRR proteins such as Cf-9 in tomato (Jones et al., 1994), and the CC-type protein RPW8 in Arabidopsis (Xiao et al., 2001). In the Arabidopsis Col-0 genome, an additional 58 genes encode proteins that lack LRRs and are related closely to the CNL and TNL proteins (Meyers et al., 2002). Therefore, including components of the signal transduction cascade and disease responses, a significant proportion of the plant genome encodes proteins potentially involved in defense against disease.

An essential component of our analysis was the manual reannotation of individual NBS-LRR–encoding genes. One-third of the genes contained errors resulting from automated annotation. Many of these minor errors resulted from the misannotation of genuine premature stop codons, frameshift errors, or retrotransposon insertions. We confirmed 10 pseudogenes by resequencing the predicted mutations; three predicted mutations in two genes reflected errors in the genomic sequence. Several genes had been annotated incorrectly with either additional or deleted protein motifs or domains. However, unusual domain structure was not an absolute predictor of misannotation; some of the most unusual protein configurations in the TNL-A subgroup were genuine (Meyers et al., 2002). When ~5000 full-length ESTs were compared with the Arabidopsis genomic sequence, again approximately one-third of auto-

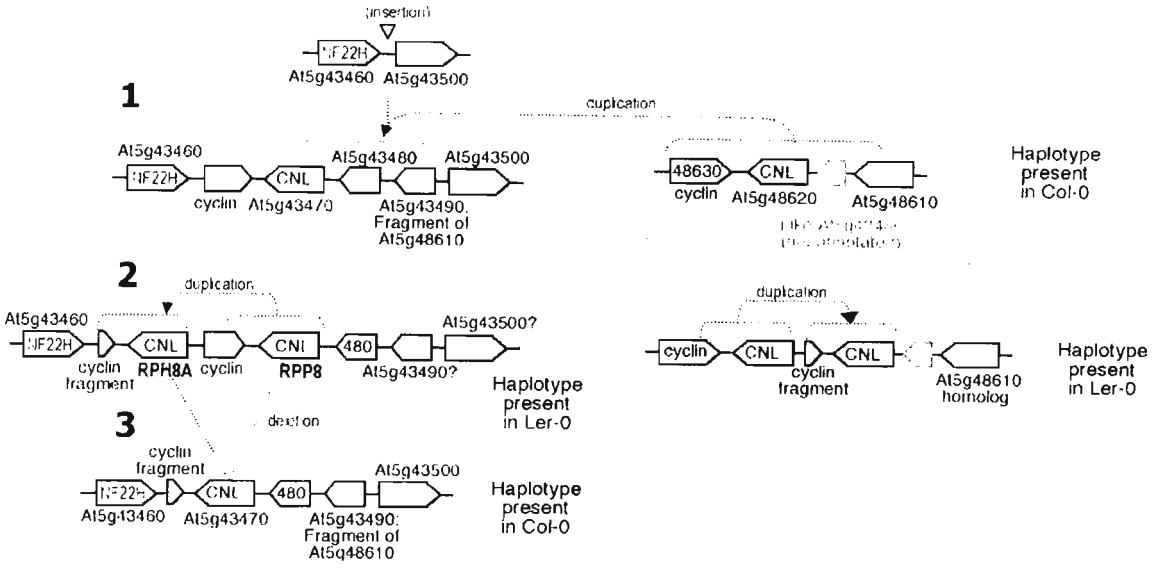


Figure 7. Rearrangements among RPP8 Homologs in Arabidopsis Ecotypes.

Two clusters were analyzed in Col-0 and L*er* to determine the genetic rearrangements in their evolutionary history. The inferred ancient arrangement of the cluster and the earliest events are indicated at top. Below, later events and the extant genomic arrangement in Col-0 and L*er* are shown. Dotted lines designate the boundaries of duplication events inferred from closely related sequences. Dashed lines that connect the ends of the clusters indicate the boundaries of a single region shown at different inferred evolutionary time points. Sequences for the L*er RPP8* cluster were obtained from GenBank (McDowell et al., 1998).

mated annotations contained errors (Haas et al., 2002). Therefore, analyses using only automated annotations without manual reassessment risk misinterpretation, particularly when large gene families are considered. Continual refinements to gene prediction programs may reduce the rate of errors in annotation.

Although *TNL* genes outnumber *CNL* genes by nearly two to one in the Arabidopsis genome, several lines of evidence suggested that the *CNL* genes may be the more ancient group. In the NBS-based phylogeny, longer branch lengths were found in the CNL tree compared with the TNL tree. Also, intron positions, which are expected to change infrequently over evolutionary time, were less conserved in *CNL* than in *TNL* genes. Comparisons across plant species also have demonstrated a greater degree of diversity among CNL proteins than TNL proteins (Cannon et al., 2002). Therefore, the *TNL* genes apparently have undergone a recent amplification relative to the *CNL* genes in the Arabidopsis lineage.

There have been different patterns of amplification of *CNL* and *TNL* genes during the evolution of other plant species. In contrast to Arabidopsis and other dicotyledonous plants, CNL sequences are more numerous and diverse in the rice genome than in Arabidopsis (Bai et al., 2002). Comparisons of NBS sequences characteristic of CNL proteins also showed that some CNL subgroups may have preferentially amplified and diversified in specific plant lineages (Cannon et al., 2002). Although a few TX- and TN-like sequences have been found in cereals, no *TNL* genes have been identified in cereal genomes (Bai et al., 2002; Meyers et al., 2002). However, the presence of *TNL* genes in coniferous genomes (Meyers et al., 2002) complicates attempts to deduce the evolution of *TNL* and *CNL* genes using data available at present. Analysis of the *TNL* and *CNL* genes in additional plant families is required to infer the evolutionary events leading to the differences in *R* gene composition.

TNL and CNL Gene and Protein Configurations Are Conserved in Arabidopsis

Few biochemical data exist to describe the functions of these proteins in plants, although the role of the various domains has been inferred based on homology with better characterized proteins in other organisms. Proteins that have homology with the plant NBS-LRR proteins function in mammalian defense responses. However, it is not known if the sequence similarity reflects conserved mechanisms and protein functions. In the innate immune responses of animal systems, small TIR-containing proteins such as the Arabidopsis TX and TN proteins play an important role in signaling (Medzhitov et al., 1998; Fitzgerald et al., 2001; Meyers et al., 2002). CC and TIR domains of mammalian defense proteins are involved in protein–protein interactions (Kopp and Medzhitov, 1999; Burkhard et al., 2001). The mammalian apoptotic response protein Apaf-1 includes a NBS domain similar to that of the plant R protein (van der Biezen and Jones, 1998b). Both NBS and LRR domains are present in the mammalian CARD/Nod family (Inohara et al., 2002) and in a family of >14 PYRIN-containing Apaf-1–like proteins (Wang et al., 2002). In these mammalian proteins, the

N-terminal domain is involved in protein–protein interactions with downstream signaling partners (adapter proteins), the NBS hydrolyzes ATP and functions as a regulatory domain, and the LRR binds upstream regulators (Hu et al., 1999; Wang et al., 2002). As predicted, the NBS of I2, a tomato CNL protein, has been shown to bind ATP (Tameling et al., 2002). Recent experiments using the CC, NBS, and LRR domains encoded by the potato *Rx*, the tomato *Mi*, and the flax *L* genes indicated that the CC or TIR and LRR domains may regulate downstream signaling events by intramolecular interactions (Hwang et al., 2000; Luck et al., 2000; Moffett et al., 2002).

Our study defined numerous motifs within each of the major domains. Some motifs were conserved in both CNL and TNL proteins, whereas others were characteristic of either the CNL or the TNL group. Furthermore, some motifs were specific to individual subgroups. In addition to the previously defined motifs in the NBS domain, we identified conserved motifs in the CC, TIR, and LRR domains of the CNL and TNL proteins. There were two major patterns of motifs in the CC domain of CNL proteins compared with the more homogeneous TIR domain of TNL proteins. Whether this finding reflects the more ancient origin of the CNL group or diversity in function is unknown. We also characterized the large C-terminal domain in TNL proteins that had distinct motifs from the LRR; this domain was much smaller in CNL proteins. Biochemical structure-function analyses, including mutation studies, now are necessary to determine the precise roles of the conserved and variable motifs. In other studies, mutations in a few of these motifs have resulted in either loss-of-function or gain-of-function phenotypes (Warren et al., 1998; Tao et al., 2000; Bendahmane et al., 2002; Shen et al., 2002; Tornero et al., 2002). Our studies have defined candidate sites for large-scale site-directed mutagenesis and for the interpretation of random mutagenesis experiments.

Intron positions in Arabidopsis *TNL* genes were similar to those in *TNL* genes from other plant species. The first *TNL* intron, separating the encoded TIR and NBS domains, also was present in three flax *TNL* genes, *L6*, *M*, and *P* (Lawrence et al., 1995; Anderson et al., 1997; Dodds et al., 2001), and in the tobacco *N* gene (Whitham et al., 1994). The second *TNL* exon, after the NBS, was conserved in the tobacco *N* gene and in flax *L6* and *M* genes but not in the flax *P* gene (Dodds et al., 2001). The third *TNL* exon, at the 5′ end of the encoded LRR domains (see below), was present in all of the flax and tobacco genes and was important for alternative splicing (Anderson et al., 1997; Dinesh-Kumar and Baker, 2000); this intron was not present in two Arabidopsis *TNL-C* genes (Figure 1B). Additional introns also occurred at the 3′ ends of the *TNL* genes within both the encoded LRR and the encoded non-LRR C-terminal domains (described below). Of *TNL* genes cloned from other plant species, only the *P* gene from flax contained an intron in a similar position (Dodds et al., 2001), although the tobacco *N* gene contained an intron close to the stop codon (Whitham et al., 1994). Introns in *CNL* genes were fewer and more variable in position than those in *TNL* genes in Arabidopsis and across different plant species (Meyers et al., 1998a; Milligan et al., 1998; Tai et al., 1999; Halterman et al., 2001; Bai et al., 2002; this study).

The intron positions of the *TNL* genes corresponded to the predicted boundaries of the encoded TIR, NBS, and LRR protein domains. This fact is indicative of the evolution of a modular protein composed of separate structural units, each with distinct functions. The extant gene configuration may reflect the ancient fusion of independent genes that encoded interacting proteins. *CNL* genes appear to be more ancient and have lost the modular gene structure but may have retained modular activity at the protein level. Distinct functions of the different domains are supported by the demonstration that the domains of the potato CNL protein Rx can act in trans to produce the hypersensitive response phenotype when either the CC or the LRR is expressed from separate genes (Moffett et al., 2002). The TIR, CC, NBS, and LRR domains initially may have evolved independently but were more selectively advantageous when fused into multidomain proteins. The exact order of the fusion events is unclear because of the variable representation of the *TX*, *TN*, *CN*, *CNL*, and *TNL* genes in different plant families (Bai et al., 2002; Meyers et al., 2002). The extra domains present at the N or C termini in members of the TNL-A subgroup are indicative of proteins with which TNL proteins interact.

Exon-defined protein modules would be conducive to the shuffling of domains by genetic rearrangements to generate chimeric proteins. However, in both comparisons of patterns of protein motifs and phylogenetic analyses, there was little evidence of shuffling between members of different subgroups. This subgroup-specific conservation may reflect selection acting on the protein as a unit rather than on the domains independently. The lack of the conserved intron positions separating the domains in the more ancient CNL group is consistent with a lack of selective advantage for domain shuffling between subgroups. Furthermore, domain swaps within the *Mi* gene of tomato and the *L* gene of flax indicated that intramolecular interactions occur between the N- and C-terminal domains of R proteins and demonstrated that specific combinations of the N terminus and the LRR are required for normal function (Hwang et al., 2000; Luck et al., 2000). The requirement for compatibility between different domains would drive coevolution of the interacting domains and confer selective advantage for genes that encode multidomain proteins over genes that encode the domains independently.

The definition of conserved and variable motifs has technical consequences for the use of PCR with degenerate primers as a strategy to isolate *R* gene homologs. Most studies to date have used primers designed to amplify sequences that encode the NBS from as many diverse genes as possible; however, a great diversity of sequences have not been amplified, and *CNL* genes have tended to be amplified preferentially (Yu et al., 1996; Aarts et al., 1998; Shen et al., 1998; Speulman et al., 1998; Deng et al., 2000; Noir et al., 2001; Donald et al., 2002), except in leguminous species, in which *TNL* genes predominate (Kanazin et al., 1996; Yu et al., 1996; Zhu et al., 2002). This bias and lack of diversity may be attributable to sequence polymorphisms in the conserved motifs. A particularly germane finding from our study was that there are two predominant versions of the GLPL motif of TNL proteins and that neither of these versions (GNLPL or SGNPL) included both the Gly and the Leu that were present in the core GGLPL sequence of CNL proteins. Most degenerate oligomers used previously to isolate *R* gene homologs have used one primer designed to amplify se-

quences that encode the consensus GLPL. This consensus was based on the first *R* genes to be cloned, which encoded either CNL or TNL proteins that fortuitously matched the GLPL consensus. Very few of the entire set of *TNL* genes in the Arabidopsis genome would be amplified by the primers used previously. Amplification of the complete set of *R* gene homologs may require the use of numerous pairs of degenerate primers. Primers now can be designed that should amplify either major groups of sequences, such as the *TNL* and *CNL* genes, or specific subgroups of sequences that may be underrepresented in initial analyses. These primers can be designed to any of the conserved motifs that we have identified in the CNL or TNL proteins and need not rely on the NBS domain.

Genetic Events Shaped the Composition of Specific Defense Responses in Arabidopsis

Various levels of duplication and rearrangement have occurred in the Arabidopsis genome, suggesting great genome plasticity over evolutionary time. Up to 80% of the Arabidopsis genome has been involved in segmental duplications (Arabidopsis Genome Initiative, 2000; Vision et al., 2000; Simillion et al., 2002). Segmental duplication apparently is responsible for some amplification of *CNL* and *TNL* genes. However, much of the expansion of these groups seems to have occurred independently of large duplications. Larger genomes, especially those with greater proportions of retrotransposons and (archeo)polyploidy, may have even more complex patterns and distributions of *CNL* and *TNL* genes than those observed in Arabidopsis. Segmental deletions as well as duplications will contribute to the extant distributions in the genome and obscure syntenic relationships (Leister et al., 1998; Simillion et al., 2002). However, complex distributions and variation between distantly related species is not evidence of rapid evolution (Michelmore and Meyers, 1998). Studies using intragenomic and intergenomic sequence comparisons between other Arabidopsis ecotypes are required to determine the relative stability of different clusters of *CNL* and *TNL* genes relative to other gene families and to reveal the genetic mechanisms responsible for the microscale rearrangements.

We found clear evidence of many microscale chromosomal duplications and deletions that involved NBS-LRR–encoding genes as well as unrelated neighboring genes or fragments of genes. These duplications were the result of translocations to both local and distant positions in the Arabidopsis Col-0 genome. Other large multigene families, such as those that encode cytochrome P450 proteins or receptor-like kinases, also are clustered in the genome (http://niblrrs.ucdavis.edu). Comparison of the distributions of NBS-LRR, cytochrome P450, and receptor-like kinases that encode genes within and between the segmental duplications revealed that the distribution of NBS-LRR–encoding genes was not dramatically different from that of these two other multigene families (Table 5; see supplemental data online). Although the lower frequency of NBS-LRR–encoding genes in simple duplications may indicate that they are more prone to deletions, comparisons between genotypes are required to investigate this possibility further. This fact indicates that the movement of individual genes or small sets of genes via ectopic rearrangement is a common phenomenon and that there is no evidence for genetic mechanisms that specifically amplify NBS-LRR–encoding genes. The small duplications and rearrangements described for *CNL* and *TNL* genes seem to exemplify a common type of microscale event that contributes to the dynamic nature of the Arabidopsis genome and that may be similar to events reported for grass species (Song et al., 2002).

Although small translocation events may be common, recombination among NBS-LRR–encoding genes in different subgroups seems to be rare. The patterns of motifs throughout the length of CNL and TNL proteins demonstrated consistent relationships within the subgroups; similarly, phylogenetic trees generated from NBS (this study) and TIR (Meyers et al., 2002) sequences were consistent and correlated with the patterns of motifs. Recombination between diverse NBS-LRR–encoding genes has been proposed to drive the evolution of resistance specificities (Richly et al., 2002); however, our data indicate that this occurs rarely, if at all.

Recombination is not uncommon within clusters of closely related paralogs that encode NBS-LRR and other types of plant R proteins; both intergenic and intragenic recombination have been observed in several species (Ellis et al., 1999; Chin et al., 2001; Hulbert et al., 2001). Evidence of duplications within the LRR region, found in this study and others (Noel et al., 1999), suggests that this region of the gene is either the most susceptible or the most permissive region for unequal crossing over. Nearly 10% of the genes were clearly pseudogenes. Such pseudogenes could be nonfunctional genes that have yet to be lost from the genome or reservoirs of genetic diversity that could be accessed by recombination or gene conversion.

Overall, the extant repertoire of diverse *CNL* and *TNL* genes has resulted from the accumulated consequences of numerous macroduplication and microduplication, translocation, and deletion events that have shaped the Arabidopsis genome.

Functional Roles for CNL and TNL Proteins

The observed number and diversity of CNL and TNL proteins in Arabidopsis represent a major part of the spectrum of recognition molecules available in an individual plant genotype to detect diverse pathogens. Although other types of proteins may play important roles in pathogen recognition, the majority of the *R* genes cloned to date encode CNL and TNL proteins (Dangl and Jones, 2001). The proportion of the ~150 NBS-LRR proteins in Arabidopsis that actively function in disease resistance remains to be demonstrated. At least 127 CNL and TNL genes in the Col-0 genome have uninterrupted full-length ORFs. Eleven of these or their orthologs have been shown to encode functional R proteins and are found in 5 of 13 subgroups. Therefore, the majority of NBS-LRR–encoding genes are at least similar in sequence to functional *R* genes. Furthermore, 53 *CNL* and *TNL* genes are found in subgroups that exhibit evidence of diversifying selection, consistent with the recognition of variable pathogen populations (Mondragon-Palomino et al., 2002). Even members of the most atypical TNL proteins (subgroup TNL-A) have been shown to function as R proteins, including the TNL:WRKY protein encoded by *RRS1* (Deslandes

et al., 2002) and the TN:TNL protein encoded by *RPP2a* (E. Sinapidou, K. Williams, and J.L. Beynon, unpublished data). Overexpression by demethylation of one gene of unknown function (At4g16890) constitutively activates defense responses in the absence of a pathogen (Stokes et al., 2002). Therefore, the current data are consistent with all of the CNL and TNL proteins being involved in disease resistance. However, it is still possible that some of *CNL* or *TNL* genes may have evolved to confer functions other than disease resistance, particularly in the more divergent clades that currently lack a known *R* gene product.

Homologs of plant NBS-LRR proteins also have been identified in animals. However, genes that encode CNL and TNL proteins have been amplified preferentially in plants, and the defense response triggered by these proteins has become the primary defense mechanism. The mammalian Apaf-1 and CED-4 proteins, which regulate apoptotic cell death, include an NBS similar to that in plant CNL and TNL proteins, suggesting an ancient relationship between the programmed cell death of the plant hypersensitive response and the mammalian caspase-induced apoptosis (Dangl et al., 1996; van der Biezen and Jones, 1998b). Apaf-1 and CED-4 lack LRR domains; however, several mammalian genes have been identified that encode NBS-LRR proteins. These include the Nod and the PYRIN-containing PYPAF families (Inohara and Nunez, 2001; Wang et al., 2002). The ~18 NBS-LRR proteins in the Nod and PYPAF families all contain conserved motifs in an NBS variously referred to as NB-ARC (van der Biezen and Jones, 1998b), Ap-ATPase (Aravind et al., 1999), NACHT (Koonin and Aravind, 2000), or NOD (Inohara and Nunez, 2001). In addition to the NBS and LRR, all of these mammalian proteins contain N-terminal domains that play critical roles in the formation of signaling complexes and the activation of downstream immune responses. Natural mutations in these proteins have been implicated in autoinflammatory diseases, suggesting that NBS-LRR proteins may be involved directly in the regulation of programmed cell death and innate immune responses in animals (Hoffman et al., 2001; Hugot et al., 2001; Miceli-Richard et al., 2001; Ogura et al., 2001).

The functional equivalence of CNL and TNL proteins is unknown. Also, the consequences of the variation in frequencies of TNL versus CNL proteins between species is unclear, particularly in rice, which lacks TNL proteins. CNL and TNL proteins may activate different but overlapping downstream signaling pathways (reviewed by Glazebrook, 2001). Mutations in EDS1 and NDR1 differentially affect some but not all CNL and TNL proteins (McDowell et al., 2000; Glazebrook, 2001). However, mutations in SGT1b and RAR1 indicate that CNL and TNL proteins also may share signaling components (Austin et al., 2002; Tor et al., 2002). Variation in the domains and in the motifs within the domains described here may reflect different levels of control or sensitivity, interactions with different proteins in macromolecular signaling complexes, or identity by descent with little functional relevance. The greatest difference between CNL and TNL proteins was the result of the large and variable C-terminal domains present only in TNL proteins; this domain may confer functions that are lacking in CNL proteins. A mutation that removes the C-terminal domain causes a loss of function in the flax TNL P2 (Dodds et al., 2001). The N-terminal domain contains the TIR and CC sequences that distinguish the CNL and TNL groups. These sequences also are present in proteins that lack LRRs. The ratio of TX and TN proteins to CX and CN proteins is far greater than the ratio of TNL to CNL proteins. The ~50 TX and TN proteins potentially could interact with the ~100 TNL proteins; however, there are only ~5 *CN* and *CX* genes compared with ~55 *CNL* genes. Therefore, the stoichiometry or specificity of interactions between these proteins, if they occur, must be very different. Extensive intergenomic comparisons combined with structure-function studies now are needed to demonstrate the relationship between the diversity in domains and motifs and the types of molecules that are recognized by CNL and TNL proteins, the mechanisms by which recognition occurs, and the resistance phenotypes that these proteins confer.

METHODS

Similarity Searches for Sequences That Encode NBS Motifs Characteristic of R Proteins

BLAST (Basic Local Alignment Search Tool) version 2.0.3 (Altschul et al., 1997) was used to search the *Arabidopsis thaliana* genomic sequence using servers available from MIPS (http://mips.gsf.de) and TAIR (http://www.arabidopsis.org). Initial searches were conducted using the entire predicted protein sequences of the Arabidopsis genes identified by Meyers et al. (1999). BLASTX and TBLASTN searches were repeated using novel sequences obtained during the initial rounds of analysis. BLAST searches were performed using sequences available during the period from April 2000 to June 2002. The threshold expectation value was set to 10^{-4}, a value determined empirically to filter out most of the spurious hits. Other numerical options were left at default values. Sequences found multiple times in the output were identified and removed based on identical names and sequence comparisons (each sequence removed was checked by hand). The complete file of sequences is available at http://niblrrs.ucdavis.edu. The sequence files and annotations were obtained from TIGR, using release 2.0 or 3.0 of the ATH1 annotation (http://www.tigr.org); modifications were made to the annotation of these sequences, as described in the text.

Alignment and Phylogenetic Analysis of Sequences

For the alignment of the NBS domain, complete predicted protein sequences for the CNL, TNL, and related proteins were trimmed at ~10 amino acids N terminal to the first Gly in the P-loop motif and ~30 amino acids beyond the MHDV motif. Sequences then were aligned using CLUSTAL W (Thompson et al., 1994) with default options, and the alignment was corrected manually using the alignment editor in GeneDoc (Nicholas et al., 1997). Software packages for automated improvement of the alignments (Notredame et al., 2000) could not be used because the quantities and lengths of the sequences in our data set exceeded the limits of our computing capacity. In the resulting alignments, the conserved motifs are likely to have been aligned accurately, whereas the more variable sequences between motifs might have contained minor ambiguous alignments. This alignment is available at http://www.niblrrs.ucdavis.edu.

Phylogenetic analyses, including distance, parsimony, and bootstrap analyses, were performed using PAUP*4.0 (Swofford, 2000). Bootstrapping provided an estimate of the confidence for each branch point. Both the CNL and TNL trees were rooted using a sequence from *Streptomy-*

ces as an outgroup; nonplant proteins Apaf-1 and CED-4 were not used in the phylogenetic analysis because they are more distantly related to plant NBS-encoding R proteins than the *Streptomyces* sequence (data not shown).

Analysis of Conserved Motif Structures

hmmpfam and hmmsearch were run locally to identify known protein motifs in all domains (Sonnhammer et al., 1997; Bateman et al., 2002). SSPro was performed on full-length protein sequences using default parameters (Pollastri et al., 2002).

MEME (Multiple Expectation Maximization for Motif Elicitation) (Bailey and Elkan, 1995) was used to analyze conserved motif structures among CNL and TNL sequences. MEME is based on expectation maximization and identifies motifs in unaligned sequences with no a priori assumptions about the sequences or their alignments (Bailey and Elkan, 1995). The output of MEME consists of a profile that is a mathematical description of the conserved sequence pattern. An individual profile describing amino acid frequencies is generated for each motif. Each position in the profile describes the probability of observing each amino acid at that position. Matches between the profile and individual sequences are scored by the program for each amino acid along the width of the profile.

To compare LRR motifs found in both CNL and TNL sequences, some genes had to be removed in the first round of MEME analysis because of the limitations of the software. A second round of MEME motif analysis was performed on each group separately containing all of either the CNL or the TNL sequences. Multiple MEME analyses were performed with settings designed to identify 20, 25, 30, or 50 motifs; increasing the number of motifs simultaneously separates related motifs in different subgroups (less desirable) while identifying motifs present in smaller groups of sequences (more desirable). The program MAST (Bailey and Gribskov, 1998) was used to assess correlations between MEME motifs in the distance matrix; we empirically chose the MEME analysis parameters that recognized the greatest number of nonoverlapping motifs (see MEME and MAST outputs in the supplemental data online).

Individual repeats within the LRR were recognized inefficiently by protein domain analysis programs such as hmmpfam and hmmsearch (Sonnhammer et al., 1997) and SMART (Schultz et al., 1998) (data not shown). We were able to manually identify individual repeat units in all CNL and TNL proteins by combining the identification of the R protein LRR consensus sequence (Jones and Jones, 1997) with predictions of the E4C5 core of secondary structure (Mondragon-Palomino et al., 2002). This analysis is displayed for all CNL and TNL proteins at http://niblrrs.ucdavis.edu. These conditions were appropriate to define the LRRs because BLAST searches with individual LRR units matched multiple sites within the putative LRR of other proteins (data not shown), confirming that the predicted LRR was part of a repeated pattern. By contrast, sequences predicted to be non-LRR regions matched only regions in identical positions in BLAST searches (relative to the NBS and LRR), indicating that these were unique and not repeating motifs. Positions of the identified motifs were compared with described *R* gene LRR regions to identify non-LRR motifs in the C terminus and to identify previously defined LRR regions (Jones and Jones, 1997; Botella et al., 1998; McDowell et al., 1998; Warren et al., 1998; Gassmann et al., 1999; van der Biezen et al., 2002).

Sequence of Arabidopsis Landsberg *erecta* Clusters

Regions homologous with the Columbia cluster of At5g48610 to At5g48640 were obtained by PCR amplification and sequenced using cycle sequencing chemistry (Applied Biosystems, Foster City, CA).

Upon request, all novel materials described in this article will be made available in a timely manner for noncommercial research purposes.

Accession Numbers

The GenBank accession numbers for the sequences mentioned in this article are as follows: AV441399 and AV545928 (two Arabidopsis ESTs), P25941 (*Streptomyces* sequence), and AF089710 (*Ler RPP8* cluster).

ACKNOWLEDGMENTS

We gratefully acknowledge Steve Edberg for bioinformatics support, Sam Lee for technical assistance, Muriel Guittet for analysis of patterns in the LRR, Xiaoping Tan for insightful discussions, and Gianluca Pollastri of the University of California, Irvine, for assistance with SSPro. This work was supported by National Science Foundation Plant Genome Grant 9975971.

Received November 21, 2002; accepted February 13, 2003.

REFERENCES

Aarts, M.G., te Lintel Hekkert, B., Holub, E.B., Beynon, J.L., Stiekema, W.J., and Pereira, A. (1998). Identification of R gene homologous DNA fragments genetically linked to disease resistance loci in *Arabidopsis thaliana*. Mol. Plant-Microbe Interact. **11,** 251–258.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25,** 3389–3402.

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.

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

Aravind, L., Dixit, V.M., and Koonin, E.V. (1999). The domains of death: Evolution of the apoptosis machinery. Trends Biochem. Sci. **24,** 47–53.

Austin, M.J., Muskett, P., Kahn, K., Feys, B.J., Jones, J.D., and Parker, J.E. (2002). Regulatory role of SGT1 in early R gene-mediated plant defenses. Science **295,** 2077–2080.

Bai, J., Pennill, L.A., Ning, J., Lee, S.W., Ramalingam, J., Webb, C.R., Zhao, B., Sun, Q., Nelson, J.C., Leach, J.E., and Hulbert, S.H. (2002). Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res. **12,** 1871–1884.

Bailey, T.L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol. Biol. **3,** 21–29.

Bailey, T.L., and Gribskov, M. (1998). Methods and statistics for combining motif match scores. J. Comput. Biol. **5,** 211–221.

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

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic Acids Res. **30,** 276–280.

Bendahmane, A., Farnham, G., Moffett, P., and Baulcombe, D.C. (2002). Constitutive gain-of-function mutants in a nucleotide binding site-leucine rich repeat protein encoded at the Rx locus of potato. Plant J. **32,** 195–204.

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.

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. (1998). Three genes of the Arabidopsis RPP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. Plant Cell **10,** 1847–1860.

Bourne, H.R., Sanders, D.A., and McCormick, F. (1991). The GTPase superfamily: Conserved structure and molecular mechanism. Nature **349,** 117–127.

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

Cannon, S.B., Zhu, H., Baumgarten, A.M., Spangler, R., May, G., Cook, D.R., and Young, N.D. (2002). Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. J. Mol. Evol. **54,** 548–562.

Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Kesseli, R.V., Lavelle, D.O., and Michelmore, R.W. (2001). Recombination and spontaneous mutation at the major cluster of resistance genes in lettuce (*Lactuca sativa*). Genetics **157,** 831–849.

Collins, N.C., Webb, C.A., Seah, S., Ellis, J.G., Hulbert, S.H., and Pryor, A. (1998). The isolation and mapping of disease resistance gene analogs in maize. Mol. Plant-Microbe Interact. **11,** 968–978.

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

Dangl, J.L., Dietrich, R.A., and Richberg, M.H. (1996). Death don't have no mercy: Cell death programs in plant–microbe interactions. Plant Cell **8,** 1793–1807.

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

Deng, Z., Huang, S., Ling, P., Chen, C., Yu, C., Weber, C., Moore, G., and Gmitter, F., Jr. (2000). Cloning and characterization of NBS-LRR class resistance-gene candidate sequences in citrus. Theor. Appl. Genet. **101,** 814–822.

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

Dinesh-Kumar, S.P., and Baker, B.J. (2000). Alternatively spliced N resistance gene transcripts: Their possible role in tobacco mosaic virus resistance. Proc. Natl. Acad. Sci. USA **97,** 1908–1913.

Dodds, P., Lawrence, G., and Ellis, J. (2001). Six amino acid changes confined to the leucine-rich repeat β-strand/β-turn motif determine the difference between the *P* and *P2* rust resistance specificities in flax. Plant Cell **13,** 163–178.

Donald, T., Pellerone, F., Adam-Blondon, A.-F., Bouquet, A., Thomas, M., and Dry, I. (2002). Identification of resistance gene analogs linked to a powdery mildew resistance locus in grapevine. Theor. Appl. Genet. **104,** 610–618.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755–763.

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.

Fitzgerald, K.A., et al. (2001). Mal (MyD88-adapter-like) is required for Toll-like receptor-4 signal transduction. Nature **413,** 78–83.

Flor, H.H. (1956). The complementary genic systems in flax and flax rust. Adv. Genet. **8,** 29–54.

Flor, H.H. (1971). Current status of the gene-for-gene concept. Annu. Rev. Phytopathol. **9,** 275–296.

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.

Glazebrook, J. (2001). Genes controlling expression of defense responses in Arabidopsis: 2001 status. Curr. Opin. Plant Biol. **4,** 301–308.

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.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. (2002). Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol. **3,** RESEARCH0029.

Halterman, D., Zhou, F., Wei, F., Wise, R.P., and Schulze-Lefert, P. (2001). The MLA6 coiled-coil, NBS-LRR protein confers AvrMla6-dependent resistance specificity to *Blumeria graminis* f. sp. *hordei* in barley and wheat. Plant J. **25,** 335–348.

Hammond-Kosack, K.E., and Jones, J.D. (1996). Resistance gene-dependent plant defense responses. Plant Cell **8,** 1773–1791.

Heath, M.C. (2000). Hypersensitive response-related death. Plant Mol. Biol. **44,** 321–334.

Hoffman, H.M., Mueller, J.L., Broide, D.H., Wanderer, A.A., and Kolodner, R.D. (2001). Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. Nat. Genet. **29,** 301–305.

Holub, E.B. (2001). The arms race is ancient history in Arabidopsis, the wildflower. Nat. Rev. Genet. **2,** 516–527.

Hu, Y., Benedict, M.A., Ding, L., and Nunez, G. (1999). Role of cytochrome *c* and dATP/ATP hydrolysis in Apaf-1-mediated caspase-9 activation and apoptosis. EMBO J. **18,** 3586–3595.

Hugot, J.P., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature **411,** 599–603.

Hulbert, S.H., Webb, C.A., Smith, S.M., and Sun, Q. (2001). Resistance gene complexes: Evolution and utilization. Annu. Rev. Phytopathol. **39,** 285–312.

Hwang, C.F., Bhakta, A.V., Truesdell, G.M., Pudlo, W.M., and Williamson, V.M. (2000). Evidence for a role of the N terminus and leucine-rich repeat region of the Mi gene product in regulation of localized cell death. Plant Cell **12,** 1319–1329.

Inohara, N., and Nunez, G. (2001). The NOD: A signaling module that regulates apoptosis and host defense against pathogens. Oncogene **20,** 6473–6481.

Inohara, N., Ogura, Y., and Nunez, G. (2002). Nods: A family of cytosolic proteins that regulate the host response to pathogens. Curr. Opin. Microbiol. **5,** 76–80.

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.

Jones, D.A., Thomas, C.M., Hammond-Kosack, K.E., Balint-Kurti, P.J., and Jones, J.D. (1994). Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. Science **266,** 789–793.

Kanazin, V., Marek, L.F., and Shoemaker, R.C. (1996). Resistance gene analogs are conserved and clustered in soybean. Proc. Natl. Acad. Sci. USA **93,** 11746–11750.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16,** 111–120.

Koonin, E.V., and Aravind, L. (2000). The NACHT family: A new group of predicted NTPases implicated in apoptosis and MHC transcription activation. Trends Biochem. Sci. **25,** 223–224.

Kopp, E.B., and Medzhitov, R. (1999). The Toll-receptor family and control of innate immunity. Curr. Opin. Immunol. **11,** 13–18.

Kozik, A., Kochetkova, E., and Michelmore, R. (2002). GenomePixelizer: A visualization program for comparative genomics within and between species. Bioinformatics **18,** 335–336.

Lahaye, T. (2002). The *Arabidopsis RRS1-R* disease resistance gene: Uncovering the plant's nucleus as the new battlefield of plant defense? Trends Plant Sci. **7,** 425–427.

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.

Leister, D., Kurth, J., Laurie, D.A., Yano, M., Sasaki, T., Devos, K., Graner, A., and Schulze-Lefert, P. (1998). Rapid reorganization of resistance gene homologues in cereal genomes. Proc. Natl. Acad. Sci. USA **95,** 370–375.

Long, M., and Deutsch, M. (1999). Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol. Biol. Evol. **16,** 1528–1534.

Luck, J.E., Lawrence, G.J., Dodds, P.N., Shepherd, 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.

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science **252,** 1162–1164.

Martin, G.B., Brommonschenkel, S.H., Chunwongse, J., Frary, A., Ganal, M.W., Spivey, R., Wu, T., Earle, E.D., and Tanksley, S.D. (1993). Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science **262,** 1432–1436.

McDowell, J.M., Cuzick, A., Can, C., Beynon, J., Dangl, J.L., and Holub, E.B. (2000). Downy mildew (*Peronospora parasitica*) resistance genes in Arabidopsis vary in functional requirements for *NDR1*, *EDS1*, *NPR1* and salicylic acid accumulation. Plant J. **22,** 523–529.

McDowell, J.M., Dhandaydham, M., Long, T.A., Aarts, M.G., 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.

Medzhitov, R., Preston-Hurlburt, P., Kopp, E., Stadlen, A., Chen, C., Ghosh, S., and Janeway, C.A., Jr. (1998). MyD88 is an adaptor protein in the hToll/IL-1 receptor family signaling pathways. Mol. Cell **2,** 253–258.

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

Meyers, B.C., Dickerman, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sobral, B.W., and Young, N.D. (1999). Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J. **20,** 317–332.

Meyers, B.C., Morgante, M., and Michelmore, R.W. (2002). TIR-X and TIR-NBS proteins: Two new families related to disease resistance TIR-NBS-LRR proteins encoded in Arabidopsis and other plant genomes. Plant J. **32,** 77–92.

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

Miceli-Richard, C., Lesage, S., Rybojad, M., Prieur, A.M., Manouvrier-Hanu, S., Hafner, R., Chamaillard, M., Zouali, H., Thomas, G., and Hugot, J.P. (2001). CARD15 mutations in Blau syndrome. Nat. Genet **29,** 19–20.

Michelmore, R.W., and Meyers, B.C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. **8,** 1113–1130.

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.

Moffett, P., Farnham, G., Peart, J., and Baulcombe, D.C. (2002). Interaction between domains of a plant NBS-LRR protein in disease resistance-related cell death. EMBO J. **21,** 4511–4519.

Mondragon-Palomino, M., Meyers, B.C., Michelmore, R.W., and Gaut, B.S. (2002). Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Res. **12,** 1305–1315.

Nicholas, K.B., Nicholas, H.B.J., and Deerfield, D.W.I. (1997). GeneDoc: Analysis and visualization of genetic variation. EMBNEW.NEWS **4,** 14.

Noel, L., Moores, T.L., van der Biezen, E.A., Parniske, M., Daniels, M.J., Parker, J.E., and Jones, J.D. (1999). Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of Arabidopsis. Plant Cell **11,** 2099–2112.

Noir, S., Combes, M.C., Anthony, F., and Lashermes, P. (2001). Origin, diversity and evolution of NBS-type disease-resistance gene homologues in coffee trees (*Coffea* L.). Mol. Genet. Genomics **265,** 654–662.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302,** 205–217.

Ogura, Y., et al. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature **411,** 603–606.

Pan, Q., Wendel, J., and Fluhr, R. (2000). Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. J. Mol. Evol. **50,** 203–213.

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

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins **47,** 228–235.

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.

Saraste, M., Sibbald, P.R., and Wittinghofer, A. (1990). The P-loop: A common motif in ATP- and GTP-binding proteins. Trends Biochem. Sci. **15,** 430–434.

Sawant, S.V., Kiran, K., Singh, P.K., and Tuli, R. (2001). Sequence architecture downstream of the initiator codon enhances gene expression and protein stability in plants. Plant Physiol. **126,** 1630–1636.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. Proc. Natl. Acad. Sci. USA **95,** 5857–5864.

Sharp, P.A. (1981). Speculations on RNA splicing. Cell **23,** 643–646.

Shen, K.A., Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Lavelle, D.O., Wroblewski, T., Meyers, B.C., and Michelmore, R.W. (2002). *Dm3* is one member of a large constitutively expressed family of nucleotide binding site-leucine-rich repeat encoding genes. Mol. Plant-Microbe Interact. **15,** 251–261.

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

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **99,** 13627–13632.

Song, R., Llaca, V., and Messing, J. (2002). Mosaic organization of orthologous sequences in grass genomes. Genome Res. **12,** 1549–1555.

Song, W.-Y., Wang, G.-L., Chen, L.-L., Kim, H.-S., Pi, L.-Y., Holsten, T.E., Gardner, J., Wang, B., Zhai, W.-X., Zhu, L.-H., Fauquet, C., and Ronald, P.C. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. Science **270,** 1804–1806.

Sonnhammer, E.L., Eddy, S.R., and Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins **28,** 405–420.

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

Stokes, T.L., Kunkel, B.N., and Richards, E.J. (2002). Epigenetic variation in Arabidopsis disease resistance. Genes Dev. **16,** 171–182.

Swofford, D. (2000). PAUP*: Phylogenetic Analysis Using Parsimony. (Sunderland, MA: Sinauer).

Tai, T.H., Dahlbeck, D., Clark, E.T., Gajiwala, P., Pasion, R., Whalen, M.C., Stall, R.E., and Staskawicz, B.J. (1999). Expression of the *Bs2* pepper gene confers resistance to bacterial spot disease in tomato. Proc. Natl. Acad. Sci. USA **96,** 14153–14158.

Tameling, W.I., Elzinga, S.D., Darmin, P.S., Vossen, J.H., Takken, F.L., Haring, M.A., and Cornelissen, B.J. (2002). The tomato R gene products I-2 and MI-1 are functional ATP binding proteins with ATPase activity. Plant Cell **14,** 2929–2939.

Tao, Y., Yuan, F., Leister, R.T., Ausubel, F.M., and Katagiri, F. (2000). Mutational analysis of the Arabidopsis nucleotide binding site–leucine-rich repeat resistance gene *RPS2*. Plant Cell **12,** 2541–2554.

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

Tor, M., Gordon, P., Cuzick, A., Eulgem, T., Sinapidou, E., Mert-Turk, F., Can, C., Dangl, J.L., and Holub, E.B. (2002). Arabidopsis SGT1b is required for defense signaling conferred by several downy mildew resistance genes. Plant Cell **14,** 993–1003.

Tornero, P., Chao, R.A., Luthin, W.N., Goff, S.A., and Dangl, J.L. (2002). Large-scale structure-function analysis of the Arabidopsis RPM1 disease resistance protein. Plant Cell **14,** 435–450.

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.

van der Biezen, E.A., Freddie, C.T., Kahn, K., Parker, J.E., and Jones, J.D. (2002). Arabidopsis *RPP4* is a member of the *RPP5* multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. Plant J. **29,** 439–451.

van der Biezen, E.A., and Jones, J.D. (1998a). Plant disease-resistance proteins and the gene-for-gene concept. Trends Biochem. Sci. **23,** 454–456.

van der Biezen, E.A., and Jones, J.D. (1998b). The NB-ARC domain: A novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Curr. Biol. **8,** R226–R227.

Vision, T.J., Brown, D.G., and Tanksley, S.D. (2000). The origins of genomic duplications in Arabidopsis. Science **290,** 2114–2117.

Walker, J.E., Saraste, M., Runswick, M.J., and Gay, N.J. (1982). Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. **1,** 945–951.

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.

Wang, L., Manji, G.A., Grenier, J.M., Al-Garawi, A., Merriam, S., Lora, J.M., Geddes, B.J., Briskin, M., DiStefano, P.S., and Bertin, J. (2002). PYPAF7, a novel PYRIN-containing Apaf1-like protein that regulates activation of NF-kappa B and caspase-1-dependent cytokine processing. J. Biol. Chem. **277,** 29874–29880.

Warren, R.F., Henk, A., Mowery, P., Holub, E., 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.

Whitham, S., Dinesh-Kumar, S.P., Choi, D., Hehl, R., Corr, C., and Baker, B. (1994). The product of the tobacco mosaic virus resistance gene *N*: Similarity to Toll and the interleukin-1 receptor. Cell **78,** 1101–1115.

Xiao, S., Ellwood, S., Calis, O., Patrick, E., Li, T., Coleman, M., and Turner, J.G. (2001). Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by *RPW8*. Science **291,** 118–120.

Yu, Y.G., Buss, G.R., and Maroof, M.A. (1996). Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site. Proc. Natl. Acad. Sci. USA **93,** 11751–11756.

Zhu, H., Cannon, S.B., Young, N.D., and Cook, D.R. (2002). Phylogeny and genomic organization of the TIR and non-TIR NBS-LRR resistance gene family in *Medicago truncatula*. Mol. Plant-Microbe Interact. **15,** 529–539.

Final Reference	**B-34**
Presenting author	Jonathan Jones
E-mail	jonathan.jones@sainsbury-laboratory.ac.uk
Presentation type	Invited lecture
Session	Concurrent sessions invited speakers

RESISTANCE GENE-DEPENDENT DEFENCE ACTIVATION

Jonathan Jones
Sainsbury Lab, JIC, Norwich, NR4 7UH, UK

Tomato Cf- genes confer race-specific resistance to Cladosporium fulvum. Plant cell death is correlated with resistance, but at high humidity, cell death is prevented without loss of resistance. The tomato Rcr3 gene is required for the function of Cf-2, but not other Cf- genes, and is likely involved in Avr2 recognition. Rcr3 encodes a secreted cysteine protease. Alleles of Rcr3 provoke Avr2 independent, Cf-2 dependent cell death. Recent information on mechanisms of Rcr3 and Cf-2 function will be presented. In Cf-9-carrying tobacco cell cultures, provision of Avr9 results in cell death within 3–4 hrs. This cell death is prevented by proteasome inhibitors. Expression profiling of elicited cells has revealed several E3 ubiquitin ligase genes, and a protein kinase, that appear to be required for cell death. Further analysis of these functions will be presented.

Final Reference	**B-29**
Presenting author	Richard Michelmore
E-mail	rwmichelmore@ucdavis.edu
Presentation type	Invited lecture
Session	Concurrent sessions invited speakers

COMPARATIVE ANALYSES OF RESISTANCE GENE EVOLUTION

Richard Michelmore
Department of Vegetable Crops, University of California, Davis, CA 95616, USA

The outcome of a pathogen challenge is determined by interactions among multiple plant and pathogen components. Genes encoding these components are under antagonistic cycles of selection. We are taking comparative approaches to investigate the evolution of pathogen effector proteins, plant targets of these effectors, and plant resistance genes. A significant proportion of genes in plant genomes encode proteins involved in disease resistance. Bioinformatics and functional analyses of NBS-LRR-encoding genes in Arabidopsis have defined different classes of resistance genes. In addition to two previously-described groups of sequences, the TIR-NBS-LRR and CC-NBS-LRR, we identified genes encoding potential adapter proteins with configurations of TIR-X, TIR-NBS, and CC-NBS. Comparisons among predicted protein sequences indicate conserved motifs exist in all protein domains. A broad range of genetic mechanisms has been shown to influence the evolution of disease resistance genes. These include point mutations, insertion/deletions, intragenic and intergenic unequal crossing-over, and gene conversion. These mechanisms have been important at different times and influence different parts of the resistance protein and that resistance genes within the same cluster can exhibit heterogeneous rates of evolution. In the major cluster of resistance genes in lettuce, some genes evolve slowly as distinct lineages with little sequence exchange between paralogs. Orthologs of these genes are readily detectable in diverse germplasm. Deletion events have led to loss of certain lineages in some haplotypes. Other genes within the same cluster are evolving more rapidly with exchanges between paralogs and close orthologs are rare in germplasm. We are currently refining a 'birth-and-death' model of the evolution of plant disease resistance genes using data from lettuce, Arabidopsis, tomato and other plant species. The different rates of evolution may be indicative of different types of pathogen ligands detected.

Final Reference	**B-15**
Presenting author	John Mundy
E-mail	mundy@biobase.dk
Presentation type	Invited lecture
Session	Concurrent sessions invited speakers

HORMONE SIGNALING IN STRESS & PATHOGENESIS

John Mundy
Molecular Biology Institute, Copenhagen University, Copenhagen, Denmark

Plants respond to certain pathogens by the induction of the hypersensitive response (HR) and the development of salicylate (SA)-dependent systemic acquired resistance (SAR). Loss of function mutations in genes that encode negative regulators of SAR or HR may cause plants to exhibit accelerated cell death and/or constitutive SAR.

For example, the Arabidopsis MAP kinase 4 (mpk4) mutant exhibits SAR with elevated SA levels, increased resistance to virulent pathogens, and constitutive pathogenesis-related gene expression (Petersen et al. 2000 Cell 103, 1111-20). MPK4 kinase activity is required to repress SAR as an inactive MPK4 form fails to complement mpk4. Analysis of mpk4 expressing the SA hydroxylase NahG, and of mpk4/npr1 double mutants, indicates that SAR in mpk4 is dependent upon elevated SA levels, but is independent of NPR1. PDF1.2 and THI2.1 gene induction by jasmonate was blocked in mpk4 expressing NahG, suggesting that MPK4 is required for JA-responsive gene expression.

Similarly, the accelerated-cell-death11 mutant (acd11) constitutively expresses defense-related genes and also exhibits characteristics of animal apoptosis (programmed cell death) monitored by flow cytometry (Brodersen et al. 2002 Genes & Develop. 16, 490-502). The PCD and defense pathways activated in acd11 are (SA)-dependent, but do not require intact jasmonic acid or ethylene signaling pathways. Epistatic analysis showed that the SA-dependent pathways require two regulators of SA-mediated resistance responses, PAD4 and EDS1. Furthermore, acd11 PR1 gene expression, but not cell death, depends on the SA signal tranducer NPR1, suggesting that the npr1-1 mutation uncouples resistance responses and cell death in acd11.

Final Reference	**A-6**
Presenting author	Jeff Dangl
E-mail	dangl@email.unc.edu
Presentation type	Invited lecture
Session	Plenary sessions invited speakers

MOLECULAR SPECIFICITY IN PLANT DISEASE AND DISEASE RESISTANCE

Jeff Dangl
Dept. of Biology, Curriculum in Genetics and Department of Microbiology and Immunology, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC USA

Plants cannot move to escape environmental challenges. Biotic stresses result from a battery of potential pathogens: fungi, bacteria, nematodes and insects intercept the photosynthate produced by plants and viruses utilize replication machinery at the host's expense. Plants, in turn, have evolved sophisticated mechanisms to perceive such attacks, and to translate that perception into an adaptive response. Recognition is controlled by proteins in the plant that are structurally similar to mammalian Nod proteins- they are called R proteins. There are only 175 genes for the major class of R protein in the finished Arabidopsis genome sequence, which presents problems in terms of how large the pathogen recognition effective repertoire can be. R protein action is triggered by intracellular virulence factors produced by many extracellular bacterial and fungal pathogens. The bacterial pathogens deliver these virulence factors through the evolutionarily conserved type III secretion pilus, and the virulence factors are hence called type III effector proteins. I will review the current knowledge of recognition-dependent disease resistance in plants, with special emphasis on a model that may get around this repertoire problem. I will also describe our efforts to characterize the protein complex in the host cell that recognizes the pathogen encoded trigger. Finally, I will describe our genomics based efforts to identify all of the type III effector proteins produced by Pseudomonas syringae pathogens of plants and why this effort may help us define the number and nature of their host targets. I will highlight a few concepts to compare and contrast plant innate immunity from that more commonly associated with animals. There are appreciable differences, but also surprising parallels.

Work on these topics in my lab is funded by the NIH, DOE, NSF and USDA.

De "Claudio C. Ramirez" <clramirez@pehuenche.utalca.cl>

Fecha Wednesday, March 26, 2003 6:46 pm
Para regisl@123click.cl
Copia
Copia Oculta
Asunto

Midgut adaptation and digestive enzyme distribution in a phloem feeding insect, the pea aphid **Acyrthosiphon pisum**

Cristofoletti PT, Ribeiro AF, Deraison C, Rahbe Y, Terra WR

JOURNAL OF INSECT PHYSIOLOGY
49 (1): 11-24 JAN 2003

Document type: Article **Language:** English **Cited References:** 44 **Times Cited:** 0

Abstract:
Transmission electron micrographs of the pea **aphid** midgut revealed that its anterior region has cells with an apical complex network of lamellae (apical lamellae) instead of the usual regularly-arranged microvilli. These apical lamellae are linked to one another by trabeculae. Modified perimicrovillar membranes (MPM) are associated with the lamellae and project into the lumen. Trabeculae and MPM become less conspicuous along the midgut. The most active A. pisum digestive enzymes are membrane-bound. An aminopeptidase (APN) is described elsewhere. An alpha-glucosidase (alpha-Glu) has a molecular mass of 72 kDa, pH optimum 6.0 and catalyzes in vitro transglycosylations in the presence of an excess of the substrate sucrose. There is a major cysteine proteinase activity (CP) on protein substrates that has a molecular mass of 40 kDa, pH optimum 5.5, is inhibited by E-64 and chymostatin and is activated by EDTA+cysteine. The enzyme is more active against arbobenzoxy-Phe-Arg-4-methylcoumarin-7-amide (ZFRMCA) than against ZRRMCA. These features identify the purified CP as a cathepsin-L-like cysteine proteinase. Most CP is found in the anterior midgut, whereas alpha-Glu and APN predominate in the posterior midgut. With the aid of antibodies, alpha-Glu and CP were immumolocalized in cell vesicles and MPM, whereas APN was localized in vesicles, apical lamellae and MPM. The data suggest that the anterior midgut is structurally reinforced to resist osmotic pressures and that the transglycosylating alpha-Glu, together with CP and APN are bound to MPM, thus being both distributed over a large surface and prevented from excretion with honeydew. alpha-Glu frees glucose from sucrose without increasing the osmolarity, and CP and APN may process toxins or other proteins occasionally present in phloem. (C) 2003 Elsevier Science Ltd. All rights reserved.

Author Keywords:
cysteine proteinase, alpha-glucosidase, perimicrovillar membranes, transglycosylation, aminopeptidase, midgut ultrastructure, midgut function

KeyWords Plus:
DYSDERCUS-PERUVIANUS HEMIPTERA, PERIMICROVILLAR MEMBRANES, AUTOPHAGIC VACUOLE, BICINCHONINIC ACID, CELLS, PROTEINS, GLUCOSIDASE, MICROVILLAR, MECHANISMS, METABOLISM

Addresses:
Terra WR, Univ Sao Paulo, Inst Quim, Dept Bioquim, CP 26077, BR-05513970 Sao Paulo, Brazil
Univ Sao Paulo, Inst Quim, Dept Bioquim, BR-05513970 Sao Paulo, Brazil
Univ Sao Paulo, Inst Biociencias, Dept Biol, BR-05422970 Sao Paulo, Brazil
Inst Natl Sci Appl, INRA, UMR Biol Fonct Insectes & Interact, F-69621 Villeurbanne, France

Claudio C. Ramirez
Centro de Investigación en Biotecnologia Silvoagricola
Instituto de Biologia Vegetal y Biotecnologia
Universidad de Talca
Casilla 747
Talca-Chile
Fono: 56-71-200-289
Fax: 56-71-200-276