Mode of Amplification and Reorganization of Resistance Genes During Recent *Arabidopsis thaliana* Evolution

Erik Richly, Joachim Kurth, and Dario Leister

The NBS-LRR (nucleotide-binding site plus leucine-rich repeat) genes represent the major class of disease resistance genes in flowering plants and comprise 166 genes in the ecotype Col-0 of *Arabidopsis thaliana*. NBS-LRR genes are organized in single-gene loci, clusters, and superclusters. Phylogenetic analysis reveals nine monophyletic clades and a few phylogenetic orphans. Most clusters contain only genes from the same phylogenetic lineage, reflecting their origin from the exchange of sequence blocks as a result of intralocus recombination. Multiple duplications increased the number of NBS-LRR genes in the progenitors of *Arabidopsis*, suggesting that the present complexity in Col-0 may derive from as few as 17 progenitors. The combination of physical and phylogenetic analyses of the NBS-LRR genes makes it possible to detect relatively recent gene rearrangements, which increased the number of NBS-LRR genes by about 50, but which are almost never associated with large segmental duplications. The identification of 10 heterogeneous clusters containing members from different clades demonstrates that sequence sampling between different resistance gene loci and clades has occurred. Such events may have taken place early during flowering plant evolution, but they generated modules that have been duplicated and remobilized also more recently.

Introduction

In flowering plants, four families of disease resistance (*R*) genes confer gene-for-gene resistance to a wide array of pathogens by recognizing the products of the corresponding pathogen avirulence genes (Staskawicz et al. 1995; Baker et al. 1997; Hammond-Kosack and Jones 1997). Recognition triggers oxidative bursts, thickening of the cell wall, induction of defense gene expression, and rapid cell death at the infection site (Morel and Dangl 1997). The largest group of *R* genes encodes proteins with nucleotide-binding site (NBS) (Traut 1994) and leucine-rich repeat (LRR) (Kobe and Deisenhofer 1993) domains, and includes at least 15 functional *R* genes from six plant species. Homologues of NBS-LRR *R* genes have since been isolated by PCR using degenerate primers (Leister et al. 1996; Yu, Buss, and Maroof 1996; Kanazin, Marek, and Shoemaker 1996; Lagudah, Moullet, and Appels 1997; Leister et al. 1998; Speulman et al. 1998; Spielmeyer et al. 1998) or have been identified by sequence database searches (Meyers et al. 1999; Pan, Wendel, and Fluhr 2000), and most may encode functional *R* proteins (Young 2000). Two subclasses of NBS-LRR proteins can be distinguished based on their amino-terminal sequences; one contains a TIR domain with homology to the innate immunity factor Toll and interleukin receptor-like genes (Whitham et al. 1994; Hammond-Kosack and Jones 1997), and the other contains a putative coiled-coil (CC) structure (Pan, Wendel, and Fluhr 2000). The absence of TIR-type genes in cereal genomes has been interpreted as indicating the loss of this subclass following diversification of the monocots (Meyers et al. 1999; Pan, Wendel, and Fluhr 2000). NBS-LRR genes are often found in tandem arrays (clusters), which are assumed to serve as reservoirs of variation for the generation of new *R* gene specificities (Michelmore and Meyers 1998).

Comparative mapping in grasses and Solanaceae suggests that reorganization of NBS-LRR genes can occur rapidly (Leister et al. 1998; Pan et al. 2000). In extreme cases, copy number can vary widely among varieties of a particular species. Moreover, in different grass species, syntenic loci are frequently lost (Leister et al. 1998), although plant genomes may harbor clusters of highly dissimilar NBS-LRR genes, the so-called mixed clusters (Leister et al. 1998, 1999; Pan et al. 2000). Both interlocus recombination and divergent selection acting on duplicated genes have been suggested as major factors in the generation of gene diversity within existing clusters (Leister et al. 1998; Michelmore and Meyers 1998; Parniske and Jones 1999). Unlike these mechanisms, unequal crossing-over and gene conversion should cause sequence homogenization, and thus lead to concerted evolution of *R* genes within clusters (Michelmore and Meyers 1998; Parniske and Jones 1999; Young 2000).

The genome of the ecotype Col-0 of *Arabidopsis thaliana*—the first genome of a flowering plant to be completely sequenced—contains more than 150 NBS-LRR genes organized as isolated single genes and in tandem arrays (The Arabidopsis Genome Initiative 2000). We have reconstructed the mode of relatively recent *R* gene evolution in this plant by combining data on the genomic organization of the entire set of NBS-LRR genes with their phylogenetic analysis.

Materials and Methods

Identification of NBS-LRR *R* Genes

A set of 25,462 nonredundant amino acid sequences was retrieved from the MIPS *A. thaliana* DataBase (MATDB, release of January 18, 2001). BLAST searches (Altschul et al. 1997) were carried out with both complete and partial (from P-loop to GLPLAL domain) sequences of known *R* genes. The resulting BLAST hits

Key words: *Arabidopsis thaliana*, evolution, gene cluster, genome, resistance gene.

were then used in subsequent iterative *R* gene homologue searches. Homologues identified were then submitted to PEDANT analysis (http://pedant.mips.biochem.mpg.de) to identify amino-terminal CC or TIR domains, to determine EST coverage, and to verify that these proteins contained the motifs and signatures specific to NBS-LRR sequences.

Sequence Mapping and Physical Clustering

Complete *A. thaliana* BAC clone sequences and the BAC sequence status tables were retrieved from MATDB and used to assemble pseudochromosomes containing the contiguous genomic sequence of the ecotype Col-0. The MIPS annotations of NBS-LRR proteins were used to extract their genomic DNA sequences and subsequently BLASTed against all five pseudochromosomes for physical mapping. Linked NBS-LRR genes were grouped into clusters when they were not interrupted by more than eight other open reading frames (ORFs) encoding non–NBS-LRR proteins.

Phylogeny and Sequence Analyses

Cluster-based Approach

Protein BLAST searches were carried out among all members of individual clusters, and the highest e-value found was assigned to this cluster and used as a threshold similarity level for a BLAST search against all other NBS-LRR protein sequences. The resulting BLAST matrix was then used to identify NBS-LRR protein sequences homologous to cluster members. Lists of related NBS-LRR genes for each of the 40 clusters were compared and grouped into clades. Sequence similarities on the amino acid and nucleotide levels were determined within and among gene clades using the Genetics Computer Group (GCG) (Devereux, Haeberli, and Smithies 1984) package.

Universal Phylogenetic Tree

All 166 NBS-LRR protein sequences were aligned by CLUSTALW (Thompson, Higgins, and Gibson 1994), bootstrapped, and then subjected to parsimony and distance-matrix (observed differences and neighbor-joining) analyses (PAUP, V4b5 for Unix; Swofford 2000).

Phylogenetic Analyses of NBS-LRR Protein Sequences

Protein sequences present in the clades, identified by the two approaches, were aligned. Trees were inferred using protein maximum-likelihood with PROTML (MOLPHY; Adachi and Hasegawa 1996), using the JTT-F matrix with the neighbor-joining tree of ML distances as the starting topology and RELL bootstrapping (10^4). Members of different clades with similar N-termini (TIR or CC) were selected as outgroups.

Modeling of Random Sequence Sampling

The expectation values t- for the generation of heterogeneous clusters by random sequence sampling were determined according to the following equation:

$$\text{t -} = (25501 - n)p(1 - q^{n-1});$$

$$\text{with } p = \frac{166}{25500}, \qquad q = \frac{25334}{25500}$$

where (25,501 − n) is the number of possible overlapping windows with a size of n genes, p is the probability that an NBS-LRR gene is present in this window, (1 − q_{n-1}) is the probability that at least one of the remaining (n − 1) genes is also an NBS-LRR. The number 25,500 refers to the total number of genes in the Col-0 genome with 166 NBS-LRR genes and 25,334 non–NBS-LRR genes. Note that the calculated expectation value t- slightly overestimates the number of heterogeneous clusters generated by random sequence sampling because it also contains the number of randomly sampled (and less frequent) homogeneous clusters.

Results

Genomic Organization of NBS-LRR Genes

A total of 166 NBS-LRR genes, distributed among 91 loci, are present in the genome of the *A. thaliana* ecotype Col-0. Fifty-one represent single-gene loci, whereas 40 consist of between 2 and 10 tightly linked genes (clusters) (table 1 and fig. 1*a*). Cluster size varies between 5.95 kbp (cluster 13) and 79.03 kbp (cluster 21), the mean size being 18.47 kbp. The average distance between NBS-LRR genes within clusters is 3.61 kbp, with a lower limit of 121 bp (cluster 2) and an upper limit of 42.3 kbp (cluster 15). Single-gene loci are dispersed in the genome, and clusters are distributed unevenly over the five chromosomes, as highlighted by the presence of two clusters of clusters (superclusters) on chromosomes 1 (clusters 4–10) and 5 (clusters 27–39), which cover 2.53 and 4.46 Mbp, respectively. Chromosome 2, in contrast, contains only five single-gene loci and one cluster. TIR-type genes predominate in the Col-0 genome (108 TIR- vs. 40 CC-type genes; table 1). Strikingly, CC- and TIR-type genes never occur together in a cluster. Fifteen NBS-LRR genes have no obvious TIR or CC domain, and three genes do not code for a domain N-terminal to the NBS (fig. 1*b*). Thirty-three truncated NBS-LRR genes have been found, most of which lack the LRR domain, and they are often located adjacent to complete NBS-LRR genes (fig. 1*a*).

NBS-LRR Genes Derive from a Few Ancestors

NBS-LRR genes were classified into distinct clades based on the following analyses. First, the degree of similarity between the genes within each cluster was determined based on protein sequences. This data set was then used to search for genes outside each specific cluster that showed the same degree of sequence similarity to any of its members as the genes within that cluster. Eight major clades and a few orphans were identified in this way. In the second approach, the entire set of NBS-LRR genes was subjected to a phylogenetic analysis. Parsimony- and distance-matrix–based calculations yielded very similar results (data not shown), allowing the construction of a universal NBS-LRR protein se-

Table 1
Classification of *Arabidopsis* NBS-LRR Genes According to Their Coded N-terminal Protein Domains and Genomic Organization

	CHROMOSOME					
	I	II	III	IV	V	SUM
Total.	50	7	17	32	60	166
N-terminus						
CC.	20	0	5	5	10	40
TIR	27	6	9	22	44	108
Undefined	3	1	3	4	4	15
None	0	0	0	1	2	3
Genomic organization						
Single-gene loci	16	5	5	9	16	51
Clusters[a]	11 (34)	1 (2)	6 (12)	6 (23)	16 (44)	40 (115)

[a] Numbers of genes contained in clusters are given in parentheses.

quence tree (fig. 2*a*). The eight clades identified by the first approach were clearly separated in the tree, but one family was subdivided into two phylogenetic lineages (dark green and light green), yielding a total of nine gene clades. Seven gene clades contained only TIR-type protein sequences, and two clades contained CC-type proteins. This reflects the well-known ancient differentiation of NBS-LRR genes into two major groups (Meyers et al. 1999; Pan, Wendel, and Fluhr 2000). Orphan lineages tended to branch deeply near the center of the tree, suggesting that they represent relatively ancient gene lineages. The CC-type orphans formed four small groups of related genes, suggesting a limited expansion of this class during evolution. The few TIR-type orphans defined at least three distinct lineages (fig. 2*a*).

Maximum-likelihood analysis of each of the nine gene clades using members of different clades as outgroups (fig. 2*b*–*j*) was largely in agreement with the phylogenies obtained by parsimony- and distance-matrix–based calculations. Protein sequence identities within clades varied between 44.1% (yellow clade) and 66.8% (orange clade), whereas the average sequence identity among members of different clades was 32.9%. At the nucleotide level, average sequence identities within clades ranged from 61.3% (yellow) to 79.1% (orange).

Clusters Can Sample Genes from Different Clades

The data obtained from the mapping of NBS-LRR genes and from phylogenetic analyses were combined, visually summarizing all information on relative gene location in the physical map, gene orientation, gene clustering, and assignment of cluster members to clades (fig. 3). Most gene clusters were found to contain members of only one clade, but 10 clusters were made up of genes from up to three different clades (heterogeneous clusters). Twenty-six clusters contained only two gene copies in either head-to-head or head-to-tail orientation. Clusters containing members of the same clade exhibited modules of common origin, such as the head-to-tail orientation of genes within the pink clusters 6, 7, 8, and 32, and head-to-head orientation within the dark-green/light-green clusters 12, 18, 24, and 33. Furthermore, large clusters, such as the pink cluster 1 with four genes and the dark-green/light-green cluster 34 with six genes, were easily interpreted as amplification products of two-gene modules. The three clusters 11, 21, and 37 had up to 10 gene copies, mostly organized in a head-to-tail orientation. Seven heterogeneous clusters contained a module comprising a dark-green/light-green gene pair oriented head-to-head. The heterogeneous cluster 28 contained a yellow head-to-head module and a red gene, whereas cluster 38 consisted of one orphan gene next to two genes from the pink clade.

Heterogeneous Clusters May Not Be Derived from Random Sequence Sampling

Six heterogeneous clusters were identified which were not derived from duplication of heterogeneous progenitor clusters (clusters 18, 20, 22, 25, 28, and 38). Out of these, clusters 18, 20, and 22 were not interrupted by other genes, 38 contained one non-*R* gene, and 28 was totally made up of five genes (three NBS-LRR and two others).

Expectation values for the number of heterogeneous clusters derived from random sequence sampling were calculated as described in *Materials and Methods.* Three different cluster sizes having decreasing numbers of total genes in the cluster (i.e., 10, 5, and 2) were tested. The prediction was that 9.5 heterogeneous clusters should exist with 10 genes, or 4.3 with a total of 5 genes, or 1.1 with 2 genes. The actual numbers in the Col-0 genome were 6, 5, and 3, respectively, implying that most heterogeneous clusters containing solely NBS-LRR genes were not derived from random sampling. On the basis of these results, we conclude either that a mechanism selectively sampling NBS-LRR genes of different clades is active or that positive selection acts on heterogeneous clusters generated by random sequence sampling.

Phylogeny Meets Genomics

To reconstruct the evolution of NBS-LRR genes, we considered all recent gene duplication events revealed by terminal branchings and, in addition, some

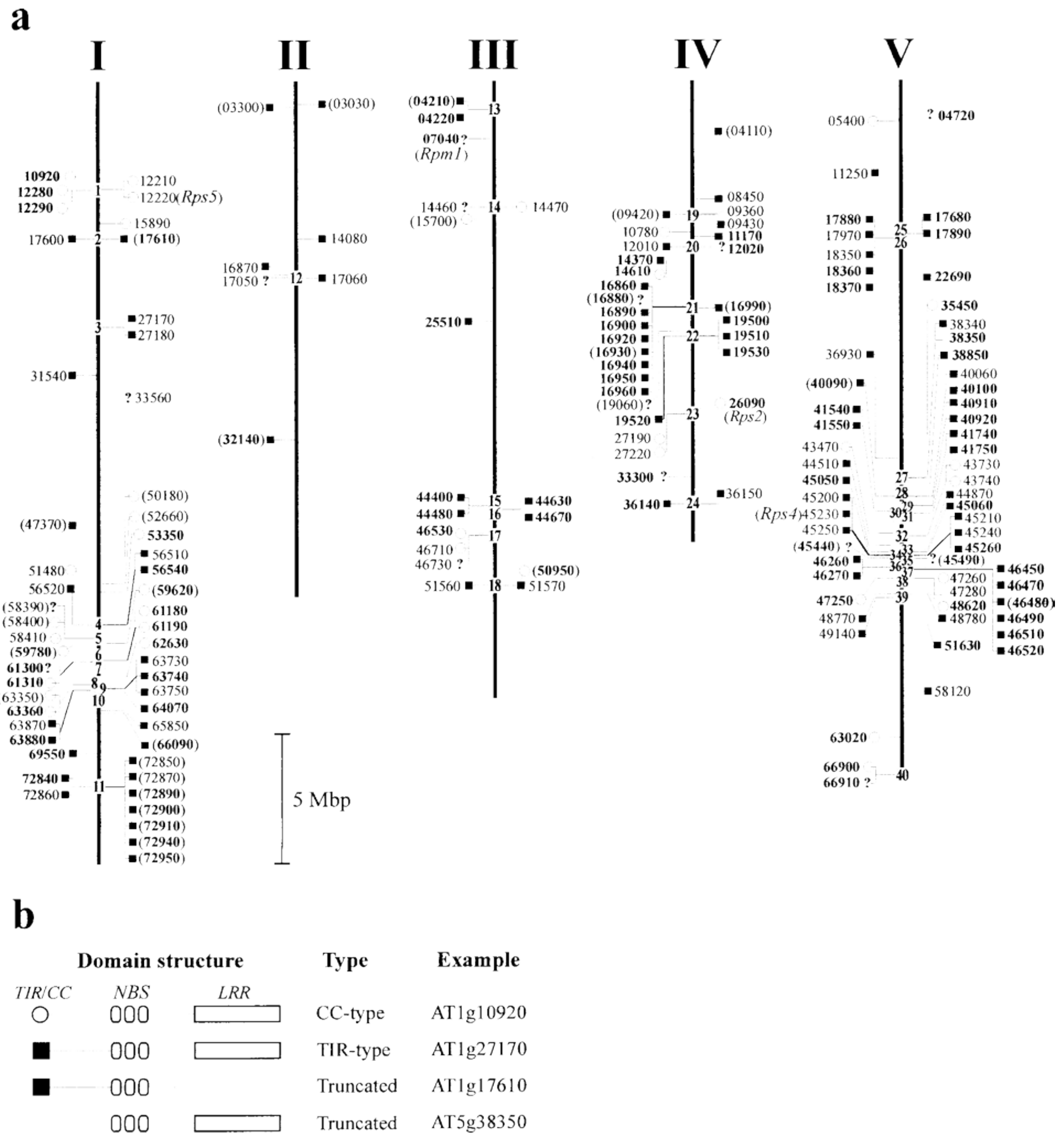


FIG. 1.—Genomic organization of NBS-LRR genes in the *A. thaliana* ecotype Col-0. (*a*) Chromosome lengths are drawn to scale based on the pseudochromosome sequences. NBS-LRR genes are designated according to their MIPS database protein entry codes, omitting the first four original digits (i.e., 10920 on chromosome 1 refers to the MIPS protein entry *AT1g10920*). Genes in forward (reverse) orientation are placed to the right- (left) hand side of chromosomes. Solid squares indicate genes coding for TIR-type proteins, whereas open circles indicate the presence of a CC domain. Question marks indicate NBS-LRR genes without an obvious TIR or CC domain, whereas the absence of squares, circles, or question marks indicates NBS-LRR genes that do not code for any domain N-terminal to the NBS. Gene designations in brackets correspond to truncated NBS-LRR fragments, whereas bold designations indicate that ESTs are known for these genes. Known resistance gene specificities (*Rps2*, *Rps4*, *Rps5*, and *Rpm1*) are listed at the corresponding loci. (*b*) Domain structures of different types of NBS-LRR genes. Note that the NBS domain is defined by a tripartite motif.

recent duplications indicated by subterminal branchings in the trees of figure 2*b–j*. These events were also included in the combined map of figure 3, as indicated by black highlighting and designation of the duplicated copy in parentheses or by a line joining the genes concerned. Presumably recent intra- or interchromosomal rearrangements that increased gene number were indicated in 50 cases. Sixteen were recent rearrangements of single-gene loci involving duplication coupled to remobilization of the gene to another locus. Twenty-six duplication events solely concerned gene clusters, indicating recent enlargement of these loci. In some cases, duplication and remobilization of entire clusters was evident. Thus, the yellow cluster 19 was formed by duplication of a part of cluster 11 or vice versa (fig. 2*j*). Duplication and remobilization is also strongly supported by the fact that the relative orientation of these three genes is the same in both clusters ([tail-head$_{\text{gene 1}}$][head-tail$_{\text{gene 2}}$][tail-head$_{\text{gene 3}}$]; fig. 3). Further examples of duplication and remobilization of entire clusters include the yellow clusters 2 and 28, and the heterogeneous clusters 33 and 34. Precise reconstruction of the phylogeny of clusters 12, 18, 22, and 34 is hampered because they are closely related. The heterogeneous clusters 22 and 25 contain genes most closely related to members of other adjacent clusters (fig. 3), suggesting that they originated from recent intrachromosomal recombination events.

NBS-LRR Gene Rearrangements Break Gene Order

The positions of NBS-LRR gene rearrangements were correlated with the duplicated chromosomal segments present in the Col-0 genome—proposed to be the result of an ancestral duplication of the progenitor genome that took place about 112 Myr ago (Ku et al. 2000), which was followed by subsequent chromosomal rearrangements (The Arabidopsis Genome Initiative 2000). The 24 large duplicated segments of ≈100 kbp make up 65.6 Mbp (58%) of the genome (Lin et al. 1999; Mayer et al. 1999; Blanc et al. 2000; The Arabidopsis Genome Initiative 2000) and harbor 126 of the 166 NBS-LRR genes detected. Recent NBS-LRR gene rearrangements could not be accounted for based on the positions and extents of segmental duplications. However, such large segmental duplications may have given rise to as many as 30 NBS-LRR rearrangements (data not shown). To identify additional, smaller, segmental duplications, we performed pairwise sequence similarity searches for the next 10 genes (five on each side) flanking each rearranged NBS-LRR gene locus. Only four recent events on chromosomes 1, 3, and 4 (as indicated by gray highlighting or by a dotted line joining the genes concerned in fig. 3), involving duplications of at least one gene tightly linked to NBS-LRR loci, were detected (cluster 8 and the single-gene locus 1g62630, single-gene loci 1g10920 and 1g53350, single-gene loci 1g52660 and 3g15700, one gene copy each from clusters 20 and 22).

Discussion

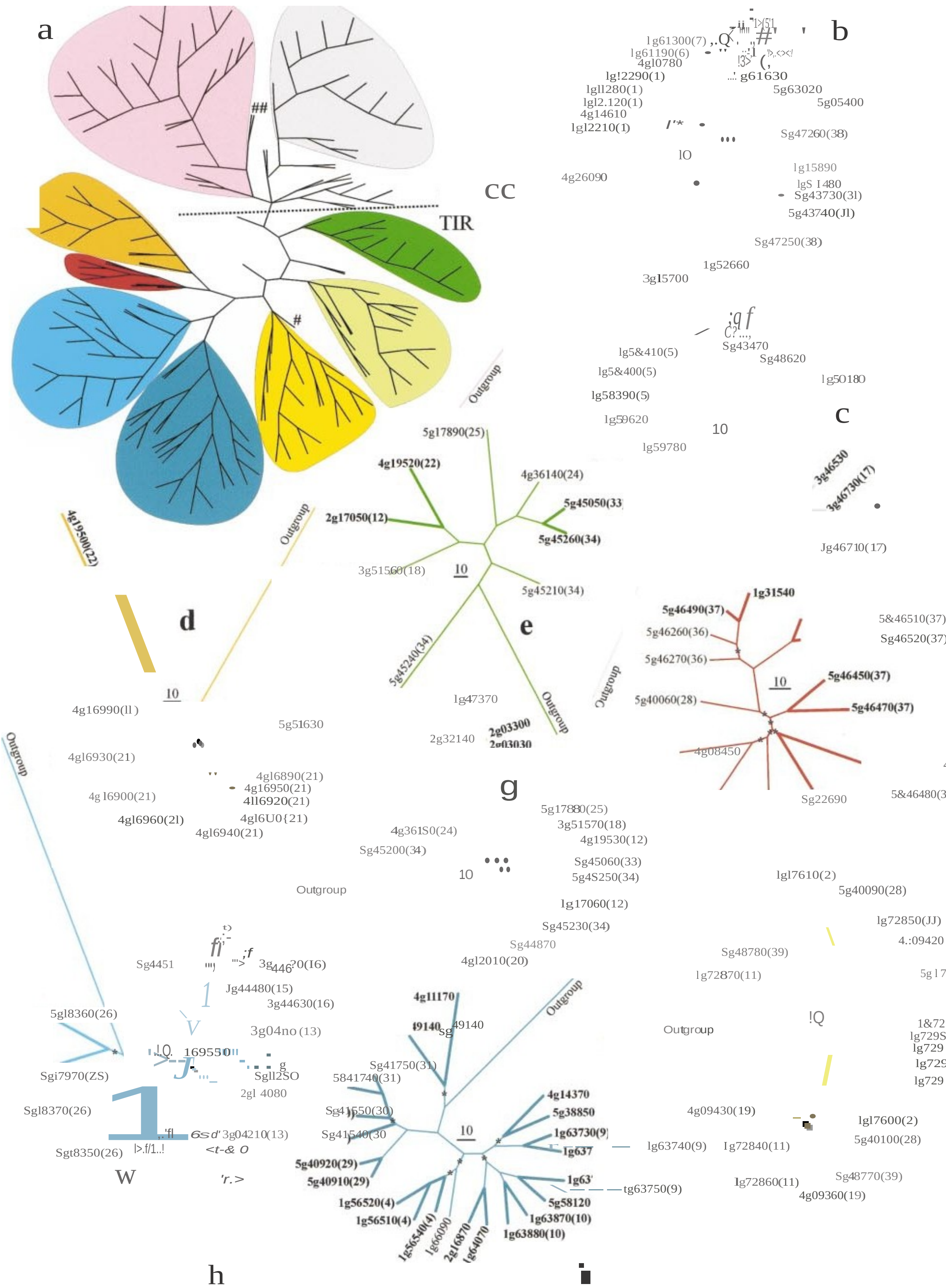
The genetic organization of the *R*-like genes in higher plant genomes has received a great deal of attention (Staskawicz et al. 1995; Hammond-Kosack and Jones 1997; Ronald 1998; Ellis, Dodds, and Pryor 2000). In *A. thaliana*, a total of 166 such genes have been identified in the genomic sequence. An ancient duplication of the entire genome and subsequent chromosomal rearrangements (Ku et al. 2000; The Arabidopsis Genome Initiative 2000) contributed to the early increase in the copy number of NBS-LRR genes in the progenitors of *A. thaliana*, but the more recent stages in the evolution of this gene complement are poorly understood. These late events, in fact, appear to be independent of the segmental duplications, which occurred about 112 Myr ago (Ku et al. 2000; The Arabidopsis Genome Initiative 2000).

Gene amplification followed by unequal crossing-over was previously postulated to be the major mechanism involved in the generation of tandem and dispersed *R* gene families (Sheperd and Mayo 1972; Hammond-Kosack and Jones 1997; Holub 1997; Hulbert et al. 1997; Michelmore and Meyers 1998; Ellis, Dodds, and Pryor 2000), but no comprehensive treatment of this large gene family as the product of cycles of repeated gene rearrangements (The Arabidopsis Genome Initiative 2000) has yet been attempted. We provide such an analysis. The combination of physical and phylogenetic analyses of the NBS-LRR genes of the Col-0 ecotype of *A. thaliana* makes it possible to detect, besides ancient events, relatively recent gene rearrangements. The analysis confirms that NBS-LRR genes are organized in single-gene loci, clusters, and superclusters, and that—as described previously (Meyers et al. 1999; Pan, Wendel, and Fluhr 2000)—mixed clusters containing both TIR- and CC-type genes do not occur. Nine NBS-LRR gene clades and a few phylogenetic orphans can be recognized.

Their phylogeny is reflected in the physical organization of clusters: about three-quarters of the 40 clusters contain only genes from the same phylogenetic lineage; clusters made up of similar genes almost always have identical structures, and large gene clusters most probably originated from simpler modules (figs. 1–3). These events have previously been interpreted in terms of exchange of sequence blocks as a result of intralocus recombination (McDowell et al. 1998; Noel et al. 1999; Ellis, Dodds, and Pryor 2000)—a mechanism, which has the capacity to alter the numbers of genes in clusters and to generate paralogous chromosomal loci with a highly variable number of LRR units (Noel et al. 1999).

The pattern of *R* gene organization, summarized in figures 1–3, suggests that after an ancient event which generated CC- and TIR-type classes, a few ancestral genes underwent local amplification, leading to tandem gene pairs, which could have been broken up by chro-

FIG. 2.—Phylogenetic analyses of NBS-LRR protein sequences of *A. thaliana* Col-0. (*a*) Universal tree of all NBS-LRR protein sequences based on distance-matrix phylogeny calculations. The dotted line indicates the strict separation of CC- and TIR-type protein sequences. The symbol, #, indicates original clade members that were removed to adjust the data sets for optimal calculations of phylogenies displayed in (*b*) through (*j*). Please note that this tree is unrooted. (*b*) through (*j*) Maximum-likelihood–based phylogenetic trees for the clades identified in (*a*). Branch colors correspond to those used in (*a*) to highlight clades. Local bootstrap proportions (BPs) below 95% estimated by RELL are indicated by single asterisks, whereas two asterisks denote BPs between 50% and 60%. Branches not marked by asterisks showed BPs above 95%. Branch lengths reflect the estimated number of substitutions per 100 sites. Thick lines and gene designations in bold indicate terminal or subterminal branchings considered for the reconstruction of recent NBS-LRR gene rearrangements as shown in figure 3.


a
##
#
TIR
CC
4g19500(22)
Outgroup
b
1g61300(7)
1g61190(6)
4g10780
1g12290(1)
1g11280(1)
1g12120(1)
4g14610
1g12210(1)
g61630
5g63020
5g05400
5g47260(38)
10
4g26090
1g15890
1g51480
5g43730(31)
5g43740(31)
5g47250(38)
1g52660
3g15700
5g43470
5g48620
1g58410(5)
1g58400(5)
1g50180
1g58390(5)
1g59620
10
1g59780
c
3g46530
3g46730(17)
3g46710(17)
Outgroup
5g17890(25)
4g19520(22)
4g36140(24)
5g45050(33)
2g17050(12)
5g45260(34)
3g51560(18)
10
5g45210(34)
d
e
5g45240(34)
1g47370
Outgroup
2g32140
2g03300
2g03030
1g31540
5g46490(37)
5g46510(37)
5g46260(36)
5g46520(37)
5g46270(36)
10
5g46450(37)
Outgroup
5g40060(28)
5g46470(37)
4g08450
5g22690
5g46480(37)
10
4g16990(11)
5g51630
Outgroup
4g16930(21)
4g16890(21)
4g16950(21)
4g16920(21)
4g16900(21)
4g16960(21)
4g16U0(21)
4g16940(21)
g
5g17880(25)
3g51570(18)
4g19530(12)
4g36150(24)
5g45200(34)
5g45060(33)
10
5g45250(34)
Outgroup
1g17060(12)
5g45230(34)
5g44870
4g12010(20)
1g17610(2)
5g40090(28)
1g72850(11)
4g09420
5g48780(39)
1g72870(11)
Outgroup
10
1g72900
1g72920
1g72930
1g72940
1g72950
4g09430(19)
1g17600(2)
5g40100(28)
1g72840(11)
5g48770(39)
1g72860(11)
4g09360(19)
5g44510
3g44630(16)
3g44670(16)
3g44480(15)
3g04210(13)
1g69550
5g18360(26)
5g17970(26)
5g11250
2g14080
5g18370(26)
3g04210(13)
5g18350(26)
h
4g11170
4g19140
4g19140
Outgroup
5g41750(31)
5g41740(31)
4g14370
5g38850
5g41550(30)
1g63730(9)
5g41540(30)
10
1g63740(9)
1g63750(9)
1g63860
5g40920(29)
5g40910(29)
1g63880
1g63750(9)
5g58120
1g56520(4)
1g63870(10)
1g56510(4)
1g63880(10)
1g56540(4)
1g66090
2g16870
4g64070
i

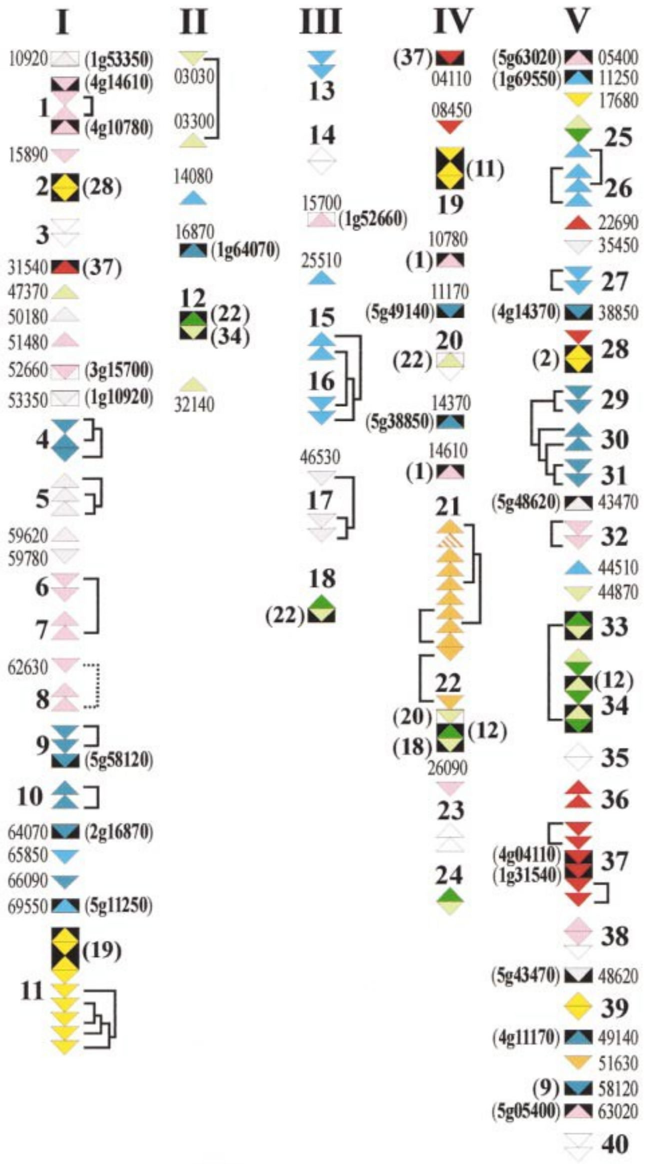


FIG. 3.—Combined physical-phylogenetic map. NBS-LRR genes are depicted as arrowheads directed toward the 3' ends of genes. Clusters are depicted as contiguous arrowheads. Clusters and single-gene loci are designated as in figure 1, and the colors indicate membership of the corresponding gene clades in figure 2. Orphan clusters are indicated as white arrowheads, whereas orphan single-gene loci are not shown. Recent gene duplications are symbolized by highlighting the genes involved against a black background (the duplicated copy is indicated in parentheses) or by lines joining the genes concerned. Duplication events involving additional genes flanking NBS-LRR loci are symbolized by highlighting against a light-gray background or by dotted lines joining the genes concerned. Member 4g16880 of cluster 21 (indicated by hatching) could not be assigned unambiguously to the orange clade because its sequence is too short.

mosomal translocations or by other types of gene relocation (see discussion that follows). Tandem gene pairs appear to have been amplified to form larger clusters or novel cluster loci, but only nine lineages expanded significantly, thus generating the nine contemporary gene clades. The other lineages had a more limited expansion capacity: in four cases, expansion stopped after the first duplication, leading to phylogenetically isolated tandem gene pairs (orphan clusters 3, 14, 23, 35, and 40), whereas the remaining genes developed into, or were maintained as, single-gene loci. Concomitantly, a contraction phase might have reduced the sizes of specific clades. Some orphan genes are expressed, and one orphan single-gene locus is functional (*Rpm1*; Grant et al. 1995), demonstrating that presumably ancient genes can still be active. Because the 20 orphans can be assigned to three TIR- and five CC-type lineages, and the nine ancient progenitors of the large contemporary NBS-LRR gene clades comprise seven TIR- and two CC-type lineages, the entire set of NBS-LRR genes in Col-0 may derive from as few as 10 TIR- and 7 CC-type progenitors.

Although the large segmental duplications in the *Arabidopsis* genome increased the number of NBS-LRR genes by about 30, the 50 rearrangements of NBS-LRR gene loci considered in our analyses are by definition recent, and thus are not associated with the duplications of large chromosomal segments. Those recent gene rearrangements increased the number of NBS-LRR genes by about 50, with about 20 genes generated by intralocus rearrangements and 30 by duplication followed by translocation. Indeed, such recent NBS-LRR rearrangements interrupt the colinearity of gene order in duplicated chromosomal fragments, and this is compatible with the lack of synteny of NBS-LRR gene loci that is also observed when different species are compared (Leister et al. 1998). However, previous analyses of *R* gene clusters revealed that genes separated by speciation but occupying allelic positions within clusters can be more similar than duplicated sequences within a cluster (Michelmore and Meyers 1998). This was interpreted as evidence for a birth-and-death model (Michelmore and Meyers 1998), claiming that (1) divergent selection acting on duplicated genes is the major mechanism underlying the generation of *R* gene variation and (2) in the generation of *R* gene variation, intergenic unequal crossing-over and gene conversions are not the primary mechanisms. In contrast, analyses of the genomic organization of cereal NBS-LRR genes (Leister et al. 1998, 1999) and of the *Hcr9* gene cluster in tomato (Parniske and Jones 1999) suggested that ectopic recombination (Leister et al. 1998) (facilitated by molecular mechanisms poorly understood as yet) and interlocus recombination events (Parniske and Jones 1999) have been important contributors to *R* gene cluster heterogeneity. The latter types of rearrangements are difficult to accommodate within the context of conventional evolutionary concepts, but our analyses demonstrate that such mechanisms, which allow sequence sampling between different *R* gene loci, must exist.

Besides homogeneous clusters, we have found 10 heterogeneous ones, and the sequences within these clusters belong to the major gene clades. The possibility that heterogeneous clusters all derive from diversification within homogeneous clusters must be rejected because it would imply that heterogeneous progenitor clusters gave rise to the present clades. This cannot be the case, simply because orphans appear phylogenetically older than genes grouped in clusters (fig. 2*a*). Our data rather show that heterogeneous clusters probably derive from chromosomal translocation or gene-cluster remobilization events that brought together sequences from different clades. This type of gene reorganization is difficult to explain other than in terms of a positive selection for cluster complexity, operative in the presence of pathogens. Some of these events may have taken place early during higher plant evolution, but they nevertheless generated modules that have been duplicated and remobilized more recently. This process resulted in the contemporary situation in which cluster heterogeneity is a major component of NBS-LRR gene complexity, possibly providing the starting material for the generation of new resistance specificities by recombination of diverse NBS-LRR genes. For large clusters, a role for divergent selection acting on duplicated genes in the generation of cluster heterogeneity remains possible, but our data demonstrate that it cannot be considered as the only relevant mechanism in this process.

Is the scenario of amplification and rearrangement described earlier unique for NBS-LRR genes, or does it reflect common principles in genomic evolution in *A. thaliana*? Because other gene families within the *A. thaliana* genome exhibit a similarly complex organization (examples include the genes encoding receptor-like kinases, the genes for cytochrome P450-like proteins, and the other classes of *R*-like genes [The Arabidopsis Genome Initiative 2000]), it will be interesting to test whether similar mechanisms underlie their evolution.

Acknowledgments

We acknowledge Francesco Salamini for stimulating discussions and comments on the manuscript. Paul Hardy, Gü nter Theissen, Thomas Mü nster, Bill Martin, and the two anonymous reviewers are gratefully acknowledged for their critical comments on the manuscript. This work was supported by the award of a Habilitation Stipend of the Deutsche Forschungsgemeinschaft to D.L.

LITERATURE CITED

Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. **28**:1–150.

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

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

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12**:1093–1101.

Devereux, J., P. Haeberli, and O. A. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.

Ellis, J., P. Dodds, and T. Pryor. 2000. Structure, function and evolution of plant disease resistance genes. Curr. Opin. Plant Biol. **3**:278–284.

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

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

Holub, E. B. 1997. Organization of resistance genes in Arabidopsis. Pp. 5–26 *in* I. R. Crute, E. B. Holub, J. J. Burdon, eds. The gene-for-gene relationships in plant–parasite interactions. British Society for Plant Pathology, CAB International, Oxon, U.K.

Hulbert, S., T. Pryor, G. Hu, T. Richter, and J. Drake. 1997. Genetic fine structure of resistance loci. Pp. 27–43 *in* I. R. Crute, E. B. Holub, and J. J. Burdon, eds. The gene-for-gene relationships in plant–parasite interactions. British Society for Plant Pathology, CAB International, Oxon, U.K.

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

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

Ku, H. M., T. Vision, J. Liu, and S. D. Tanksley. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. USA **97**:9121–9126.

Lagudah, E. S., O. Moullet, and R. Appels. 1997. Map-based cloning of a gene sequence encoding a nucleotide-binding domain and a leucine-rich region at the *Cre3* nematode resistance locus of wheat. Genome **40**:659–665.

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

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

Leister, D., J. Kurth, D. A. Laurie, M. Yano, T. Sasaki, A. Graner, and P. Schulze-Lefert. 1999. RFLP- and physical mapping of resistance gene homologues in rice (*O. sativa*) and barley (*H. vulgare*). Theor. Appl. Genet. **98**: 509–520.

Lin, X., S. Kaul, S. Rounsley et al. (37 co-authors). 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402**:761–768.

Mayer, K., C. Schuller, R. Wambutt et al. (229 co-authors). 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402**:769–777.

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

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

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

Morel, J. B., and J. L. Dangl. 1997. The hypersensitive response and the induction of cell death in plants. Cell Death Differ. **4**:671–683.

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

PAN, Q., Y. S. LIU, O. BUDAI-HADRIAN, M. SELA, L. CARMEL-GOREN, D. ZAMIR, and R. FLUHR. 2000. Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: tomato and *Arabidopsis*. Genetics **155**:309–322.

PAN, Q., J. WENDEL, and R. FLUHR. 2000. Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. J. Mol. Evol. **50**:203–213.

PARNISKE, M., and J. D. JONES. 1999. Recombination between diverged clusters of the tomato Cf-9 plant disease resistance gene family. Proc. Natl. Acad. Sci. USA **96**:5850–5855.

RONALD, P. C. 1998. Resistance gene evolution. Curr. Opin. Plant Biol. **1**:294–298.

SHEPHERD, K. W., and G. M. E. MAYO. 1972. Genes conferring specific plant disease resistance. Science **175**:375–380.

SPEULMAN, E., D. BOUCHEZ, E. B. HOLUB, and J. L. BEYNON. 1998. Disease resistance gene homologs correlate with disease resistance loci of *Arabidopsis thaliana*. Plant J. **14**: 467–474.

SPIELMEYER, W., M. ROBERTSON, N. COLLINS, D. LEISTER, P. SCHULZE-LEFERT, S. SEAH, O. MOULLET, and E. S. LAGUDAH. 1998. A superfamily of disease resistance gene analogs is located on all homologous chromosome groups of wheat (*Triticum aestivum*). Genome **41**:782–788.

STASKAWICZ, B. J., F. M. AUSUBEL, B. J. BAKER, J. G. ELLIS, and J. D. JONES. 1995. Molecular genetics of plant disease resistance. Science **268**:661–667.

SWOFFORD, D. L. 2000. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

THE ARABIDOPSIS GENOME INITIATIVE. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 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**:4673–4480.

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.

WHITHAM, S., S. P. DINESH-KUMAR, D. CHOI, R. HEHL, C. CORR, and B. BAKER. 1994. The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. Cell **78**:1101–1115.

YOUNG, N. D. 2000. The genetic architecture of resistance. Curr. Opin. Plant Biol. **3**:285–290.

YU, Y. G., G. R. BUSS, and M. A. MAROOF. 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.