Theor Appl Genet (2004) 109: 1434–1447
DOI 10.1007/s00122-004-1758-x

ORIGINAL PAPER

B. Monosi · R. J. Wisser · L. Pennill · S. H. Hulbert

Full-genome analysis of resistance gene homologues in rice

Received: 18 February 2004 / Accepted: 16 June 2004 / Published online: 10 August 2004

Abstract The availability of the rice genome sequence enabled the global characterization of nucleotide-binding site (NBS)–leucine-rich repeat (LRR) genes, the largest class of plant disease resistance genes. The rice genome carries approximately 500 NBS–LRR genes that are very similar to the non-Toll/interleukin-1 receptor homology region (TIR) class (class 2) genes of *Arabidopsis* but none that are homologous to the TIR class genes. Over 100 of these genes were predicted to be pseudogenes in the rice cultivar Nipponbare, but some of these are functional in other rice lines. Over 80 other NBS-encoding genes were identified that belonged to four different classes, only two of which are present in dicotyledonous plant sequences present in databases. Map positions of the identified genes show that these genes occur in clusters, many of which included members from distantly related groups. Members of phylogenetic subgroups of the class 2 NBS–LRR genes mapped to as many as ten different chromosomes. The patterns of duplication of the NBS–LRR genes indicate that they were duplicated by many independent genetic events that have occurred continuously through the expansion of the NBS–LRR superfamily and the evolution of the modern rice genome. Genetic events, such as inversions, that inhibit the ability of recently duplicated genes to recombine promote the divergence of their sequences by inhibiting concerted evolution.

Communicated by Q. Zhang

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-004-1758-x

B. Monosi · L. Pennill · S. H. Hulbert (✉)
Department of Plant Pathology, Kansas State University,
Manhattan, KS, 66506-5502, USA
e-mail: shulbrt@ksu.edu

R. J. Wisser
Department of Plant Pathology, Cornell University,
Ithaca, NY, 14853, USA

Introduction

Plants use a variety of strategies to defend themselves against microbial attack. One important defense mechanism is the plant's ability to recognize the presence of specific pathogens and initiate defense responses. Pathogen recognition is mediated by resistance genes, most of which belong to an ancient family encoding proteins with nucleotide-binding site (NBS) and leucine-rich repeat (LRR) domains (Bai et al. 2002; Cannon et al. 2002; Michelmore and Meyers 1998; Young 2000). They control resistance to a wide variety of pathogens and pests including viruses, bacteria, fungi, nematodes, and insects (Dangl and Jones 2001). NBS–LRR genes are abundant in plant genomes, with approximately 150 described in the Colombia ecotype of *Arabidopsis* (Meyers et al. 2003) and many more estimated in the rice genome (Bai et al. 2002; Goff et al. 2002).

NBS–LRR genes of dicotyledonous plants can be subdivided into two distinct classes based on the structure of the N terminus of the protein, upstream of the NBS domain. To some extent, the classes are thought to correspond to the defense signaling pathways that the gene controls (Aarts et al. 1998). The first class codes for a TIR domain, so called because of sequence homology to the intracellular domain of the *Drosophila* Toll and mammalian interleukin-1 receptors in the N terminus (Pan et al. 2000b). The second class typically codes for a domain with a predicted coiled-coil (CC) motif, sometimes in the form of a leucine zipper, at the N terminus. Variant types with different structures exist in both these classes (Meyers et al. 2003), but they can generally be classified into the TIR or non-TIR (nT) classes based on their NBS-coding sequences. For example, genes of both classes exist that are lacking LRR-coding domains. While the TIR class genes account for most of the NBS–LRR genes in *Arabidopsis*, they have not been found in cereal sequences. Most NBS–LRR proteins of cereals are very similar to the non-TIR class genes of dicots. Many of these genes do not code for predicted CC motifs, but their N-terminal domains have good homology to those of dicots.

The cereal genes typically code for a conserved nT motif that is also present in the dicot genes (Bai et al. 2002). Examination of cereal sequences identified a third class that codes for an NBS domain and an amino terminus with homology to those of the nT class. No homologues of this class were identified in *Arabidopsis* or other dicot sequences in the database. It therefore appears that the resistance gene arsenals of dicot and cereal species have diverged considerably during their independent evolution since monocot and dicot lineages diverged approximately 200 million years ago (Wolfe et al. 1989).

The nearly complete sequence of the rice genome (Goff et al. 2002; Feng et al. 2002; Sasaki et al. 2002) allows a close inspection of the diversity of resistance gene sequences, including searches for classes that may be different from those in dicots. Since the Rice Genome Project (RGP) has now anchored nearly all of this sequence to a fine-structure genetic map, it is also possible to examine the genomic distribution of the genes. Comparing the genetic relationships of the different genes, together with their genetic proximity on rice chromosomes, provides a global view of the types of evolutionary events that allowed this gene superfamily to amplify and diverge. In this study, we identify and examine the NBS–LRR and related genes in the rice cultivar Nipponbare. The genes are classified into classes and subgroups based on sequence similarity and their map positions compared. The analysis provides a picture of how the NBS–LRR superfamily became the premier pathogen surveillance system in cereal genomes, accounting for approximately 1% of the genes.

Materials and methods

Database searches, gene prediction, and gene localization

Several searches of the Nipponbare rice genome sequence, produced by the Rice Genome Project Consortium, were carried out to identify NBS-encoding sequences. A variety of types of sequences were used in tBLASTN searches, using the NCBI and TIGR (http://tigrblast.tigr.org/euk-blast/index.cgi?project=osa1) databases. Initial searches were conducted with six NBS–LRR-coding sequences that were selected from different branches of a previously generated phylogenetic tree (Bai et al. 2002). Hidden Markov model (HMM) searches were also performed for sequences encoding NBS domains, using the NB-ARC Pfam HMM (PF00931). Additional searches of the genomic sequences were conducted with novel classes of genes related to NBS–LRR genes that showed limited homology in the initial searches. Once consensus sequences were identified for the novel gene classes, they were used to conduct additional searches. Final searches were conducted of the TIGR predicted protein database (September 2003), using an HMM search to identify recently generated sequences, but few new sequences were identified in the final 3 months of the analysis. The genomic sequences that were collected were in the form of completely or incompletely sequenced genomic (BAC or PAC) clones.

The collected genomic DNA sequences were analyzed using the gene prediction programs GENSCAN (Burge and Karlin 1997; http://genes.mit.edu/GENSCAN.html) and FGENESH (Salamov and Solovyev 2001; http://www.softberry.com). Default settings for *Arabidopsis* were used with GENSCAN and default settings for monocots with FGENESH. Gene predictions were also compared to GenBank annotations when available. Some genes were manually annotated by translation into the six reading frames, using the six-frame translation program from BCM sequence utilities site (http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html) and determining the proper frame(s) and orientation in which the motifs occur. Splice site positions previously (Bai et al. 2002) found to be common in rice NBS–LRR genes were used as guides to predict mRNA splice sites. Genes were predicted to be pseudogenes if they were found to contain one or more in-frame stop codons in sequences predicted to be coding by comparisons to closely related genes. Other genes were predicted to be pseudogenes if their alignments with closely related predicted genes indicated that they lacked one or more of the conserved motifs characteristic of that class of gene. Translations of the DNA sequence were examined for these sequences to ensure that the gene prediction programs did not delete these sequences by predicting intron splicing incorrectly. It was found that automatic annotations commonly inserted introns to remove stop codons or frameshift mutations. Genes thought to be correctly predicted by the gene-prediction programs and manual predictions were compiled into a database, using the BLAST server package from NCBI, which enables local searches to be carried out against the predicted genes. Further manual gene/protein predictions were facilitated by translated alignments (BLASTX) of the predicted genes against the genes in the database. Searches of the rice EST database in GenBank, including 28,000 full-length cDNAs (Kikuchi et al. 2003), were also used to identify transcripts and check gene predictions. The predicted proteins are available at http://www.oznet.ksu.edu/plantpath/hulbertlab/nbsPredictions.htm.

Sequences of the predicted NBS-containing proteins were compared to each other by BLASTP searches of the local database, allowing the identification of redundant entries from overlapping genomic clones. Clones with similar map positions were examined to determine whether they carried overlapping genomic fragments to determine the physical distances between the closely linked genes. Map positions (in centiMorgans) corresponding to individual BAC or PAC clones were determined using marker-based physical maps of each rice chromosome, developed by the International Rice Genome Sequencing Project (http://rgp.dna.affrc.go.jp/IRGSP/download). While some BAC clones carried the actual sequences corresponding to genetically mapped DNA markers, the positions of others were inferred from the physical distances to the two closest flanking markers.

Estimated map positions for each of the predicted NBS-encoding genes were compiled in an Excel file (available at http://www.oznet.ksu.edu/plantpath/hulbertlab/nbsPositions.htm). Disease resistance genes that have been genetically mapped within 1–5 cM of DNA markers were correlated with the positions of NBS—LRR sequences by estimating their approximate map positions on the RGP map. The positions of DNA markers flanking the resistance genes were found using the Rice Genome Browser at the Gramene Web site (Ware et al. 2002).

Sequence alignment and phylogenetic tree construction

A multiple sequence alignment of the amino acid sequence of the predicted nT–NBS–LRR genes was performed and a phylogenetic tree constructed from the alignment. The sequences were trimmed so that the sequence region used in the multiple sequence alignment spanned from the nT sequence through the NBS-coding region to a motif before the start of the LRR, where the amino acids M, H, and D are partially conserved (MHD motif). The region spanned approximately 450–500 amino acids, depending on the gene. Where no clear nT sequence or MHD motif was identified in a predicted gene, homologous sequences were identified by alignments with related genes. The sequences were aligned utilizing the Clustal X, version 1.81, program (Jeanmougin et al. 1988), using default settings and edited using GeneDoc (Nicholas et al. 1997). A neighbor-joining distance tree was constructed using the tree-drawing application in Clustal X, also set at default settings. Bootstrap analysis was performed to evaluate the degree of support for each group in the tree.

Results

NBS–LRR genes in the Nipponbare genome

tBLASTN searches of the rice genome database and HMM searches of the TIGR predicted protein database (September 2003) identified 462 different rice genomic clones (BAC or PAC) related to NBS–LRR-encoding sequences. A nonredundant set of 581 potential NBS-encoding sequences were identified in the genomic sequences collected. Of these, 489 genes belonged to the nT class of NBS–LRR genes (class 2, Table 1) and none belonged to the TIR–NBS–LRR class (class 1, Table 1). Of the 489 class 2 genes, 100 (20%) were predicted to be probable pseudogenes; these either had in-frame stop codons in predicted exons or were missing conserved motifs that were present in closely related genes. Genes that had normal nT and NBS domains but little or no LRR-coding domain were not classified as probable pseudogenes, because these genes are common in dicots (Meyers et al. 2003) and have also been identified in cereal cDNAs (Collins et al. 1999). Predicted LRR-coding domains were not carefully scrutinized for the presence of consensus sequences because of the lack of conserved features in these domains in cereal NBS–LRR genes. The LRR of some of the cereal resistance gene proteins characterized to date are typically leucine-rich but sometimes have few regions where these leucines are arranged in the LxxLxLxxL signature (e.g., *Pi-ta*; Bryan et al. 2000).

To verify the manual gene predictions, a cDNA search was performed so that the sequences of transcripts could be compared to gene predictions to verify manual annotation, particularly for those gene predictions with multiple predicted splices or predictions in which the splice sites were not at conserved sites (Bai et al. 2002). Most of the rice EST sequences in databases prior to June 2003 were of limited use for checking gene predictions because of their short sequence length. However, the Rice Full-Length cDNA Consortium (Kikuchi et al. 2003) released the full sequences of over 28,000 cDNA clones in July 2003. From this set, 190 cDNAs were identified,

Table 1 Different classes of nucleotide-binding site (*NBS*)–leucine-rich repeat (*LRR*)-related genes in rice

Class	Features N terminus to C terminus	No. in rice[a]	No. in *Arabidopsis*[b]
1	TIR[c], NBS, LRR	0	117
2	nT[d], NBS, LRR[e]	392 (101)	60
3	Divergent nT, divergent NBS	45 (6)	0
4	Divergent nT, partial NBS, L-rich	21 (5)	0
5	Partial NBS, LRR	9 (1)	2
6	S-rich, divergent TIR, divergent NBS	5	2
7	TIR (no NBS or LRR)	1	30

[a]Number of genes identified in the rice cultivar Nipponbare genome sequences. Rice sequences predicted to be probable pseudogenes are not included in the totals and are listed in *parentheses*
[b]*Arabidopsis* numbers taken from Meyers et al. 2003
[c]*TIR* Toll/interleukin-1 homology region
[d]*nT* non-TIR
[e]Not all class 1 and class 2 proteins have LRR domains. Totals for rice include four cDNA sequences not found in the genomic sequences

representing 153 different predicted NBS-encoding genes, many of which corresponded to the full coding regions of the gene predictions. Only four cDNAs (GenBank accession nos. AK102672, AK102485, AK069428, and AK110806) were observed that were not present in Nipponbare genomic sequences (October 2003), indicating that the available sequences represented approximately 98% of the genome.

Full-length cDNAs were identified for all of the predicted classes of NBS-encoding genes. Most of the cDNAs (130/153) represented class 2 NBS–LRR genes. cDNAs were identified for many of the genes that were predicted to be pseudogenes, based on the genomic sequence and the aberrant sequences (e.g., stop codons) that were verified. Interestingly, approximately 20% (27/130) of the genes represented by the cDNAs were predicted to be pseudogenes, indicating that many of the pseudogenes are transcribed. The second largest class of sequences was the class 3 (Table 1) or nT-NBS class described by Bai et al. (2002). The Nipponbare genomic sequences included 51 different predicted class 3 genes, but at least five of these were predicted to be pseudogenes. Fully sequenced cDNAs were identified for 11 of these genes. cDNAs (accession nos. AK108636 and AK101847) were identified corresponding for two of the five predicted TIR-NBS (class 6) genes (on genomic accessions AP000364, AP003873, AP003256, AP005392, and AC109929) and one cDNA accession (AK108970) was identified which corresponded to the single TIR (class 7) gene identified in the genomic sequences (on accessions AP003866 and AP003932).

Most of the class 2 genes have structures very similar to characterized cereal resistance genes such as *Pi-ta* (Bryan et al. 2000), *Xa1* (Yoshimura et al. 1998), or members of the *Mla* (Zhou et al. 2001), *Rp1* (Collins et al. 1999), or *Rp3* (Webb et al. 2002) gene families. All but 25 of these class 2 genes coded for predicted nT domains at their N termini. The distance between the nT domain and the P-loop (phosphate-binding) motif at the beginning of the NBS domain was variable, with most genes having ~130 amino acids between the two domains. None of the predicted rice genes coded for large N termini like that of the PRF protein with more than 1,000 amino acids before the P-loop motif (Salmeron et al. 1996). Only one of the predicted rice proteins had an N terminus consisting of more than 350 amino acids before the P-loop motif (from accession AP003862), and this prediction included an intron splice, which might have been predicted incorrectly.

A

mtelaagavss**LLGVIRDEARLLGGVRGDVQFIKDEMESMNSFLxHL**arsapqggge**HDEQVRTWMKQVRELAYDCQNC**id
lylysgnpeihrtkgrlrrhlwwvywslrkmva**QHRAATRIRELKDRARDVGERR**lrygveipattkaaapdatggyvae
ddeeededdregqfavatptlahhsarwpvfeppslddyveakllewiggvpgnaivtlsiaivapdadnkevlaiahet
lvapnyyyrrsimvnvpavhldvlplrpkevlyyilrelereeaagsqkqptdqgeweeedpdpwqdyykkcgiyrskks
vlgkikrnikkmniyekldkiksdiregqhksnkllllqlqkkgadqvdlhvllqllvlqsqqdqaknkavdthklpewn
dnlieklamrlkdhmeadektkklneqtgveeetavrqggggereedekdergdgeeegkeerrdmekggeerkeqqee
qekegrkeeqnevrketegrkeqvageeeekedhdadndedsnddddeeeeeeddddeeepihlhedqyeqilrevft
knasskaqeqdklvaeqatktaattldeerikqmineakqdvlrelrgretdknqatgepdvppdknqatgqhavvldqn
eeayfeeveqkieeikqelkeqlkikwivdkikhhlqdqcp**liilkfd**qmmdgsrweeirkalsllelsadalifttgst
eqakgycypprepidhcslvglyyytvlkltskhknednqaqifrgileece**ghefc**mkifthavyanpkrsneelrkl
hstlqspkksfdtiakkmfm**FCYNDLPNEYKSCLWYLSIFPRGYKIRRSTLVRRWVAEGl**tfkedwpssvyqanrcfdal
irrwlvypddisat**GKVKSCTVNDPVHDFIT**aiarkqhivetrlshhlarhfsifndlrlrssdrigtffqglsrssrvs
LLKVLDLEGCQCLasknqrylkdicnkmll**LKYLSLRGTDITQLPKEINNLQELETLDIRQTKV**panatvhvlllklkrl
lagasqidptprnfv**STVRIPRGIGKMTNMEILSHVK**aqghdnledig**KLWQLRKLGVVI**dgkkshlgsllkaisdlhas
lrslsitipttttlevtpsspelqdiasrlkdh**PKLLESLSISG**akhlfplltkggnkklakvtlsntplnqddlkffaql
pmlqcvrlrhisctesvlnfkkddfkclkylliegsnltnitfedeaacelekmvlsstciesisgvhglpkfeelelns
sscgrllsscfynveriakltlrgtllkqgdlriiarelnicclvllensfdisqnqitfekeefiwlkllsvdcstitk
infitgsaprlkkivwssftslsginnlprlkelefngysvpndveeaiknnksinlkhnkp

B

mstpeaeaegeggcilfskyiddgiltpllsslqvigdaksfrgaesssdlaqldsirdmmrelqaifvkmgenerrivh
lfdpieelvddvlnslaaggeistlqpklagvgvqigiireaisgsykikiqeepspdrgrdvedsasaptmagirrcvr
dgeqmahlrravhg**LDTQLRQCLLCFAVFPENAVIKKRLLIHWWIGE**kfvssledgnelfqqlvdlrfvrtvrrrghcdt
ahact**vhp**wirrmlvavarssafleidpdgngasasnndfsrthraclhdggllqgsrfhpqlstiynvgqsyvklstaw
ftyrsqlgtvqlgqwrvadpvdqiayprkshieliddhhlkgigacknlrylslrgisrima**IPSSIGNLRQLVVLDLNA
CHNLEQLCAEITSLxKLEYLDMSECYLLEPMPKGI**nkmsqlqvlkgflvvnsnkrntcnlgelvslsnlrklsisvskkl
kraedelkvlanfa**ALESLTITWG**mvsprdaadesdaakfnlv**LPPNLEKLDLRCFP**srqfgtissdslkklyftggklh
sleiqngeckvevlrlrclkdlqfcweelrelypmlefveaqhcsgvanwpldnnkvwtkdaetsnaasaqseeptapdv
cpeiveeg

Fig. 1a, b Consensus sequences of the class 4 and class 5 genes. Conserved sequences, as determined using BLOCKMAKER, are shown in *boldface* and *uppercase*. *Underlined* sequences are those that align well with class 2 genes. **a** Consensus sequence of the class 4 genes, generated by aligning 20 genes from this class. The *shaded* sequence at the N terminus aligns with the non-Toll/interleukin-1 receptor homology region (nT) domain of the class 2 genes. Other *shaded* sequences correspond to the kinase 2, GLPL, and MHD (amino acids M, H, and D) motifs (N terminus to C terminus, respectively). The class 4 genes align with the class 2 genes at the first 80–150 amino acids at the N terminus and again starting at ~670 amino acids to near the end of the sequence (*underlined* region). **b** Consensus sequence of the class 5 genes generated by alignment of the nine genes from this class. The *shaded* sequence (*vhp*) aligns with the MHD motif that occurs in class 2 genes at the C terminus of the NBS domain, before the start of the leucine-rich repeat (LRR) domain. Leucine residues in the predicted LRR domain are also *shaded*

The majority of the class 2 genes had large LRR-coding regions, but some were small or nonexistent. Twenty-one predicted genes either coded for fewer than 100 amino acids after the conserved MHD motif (12 genes) or the coding regions ended before the MHD motif (nine genes). Other than the lack of an LRR, these genes were typically indistinguishable from the class 2 genes. Fourteen of the 21 clustered closely with normal NBS-LRR genes in the phylogenetic analysis (below).

Five predicted genes had a partially duplicated NBS structure that was first described for the rice *Pib* gene (Wang et al. 1999). Similar structures in NBS-coding regions of TIR–NBS–LRR (class 1) genes have also been found in *Arabidopsis* (Meyers et al. 2002). Three of the genes (on accession AP003862) are related to *Pib* but map to chromosome 8, while the fourth (on accessions AP004048 and AP004028) shows the highest sequence homology and maps to chromosome 2 like *Pib*. This *Pib* allele was predicted to be a pseudogene, illustrating how the resistance genes that are pseudogenes in one cultivar may represent functional genes in other germplasm. The fifth gene with a partially duplicated NBS structure was found on genomic accession AP005064 and matched cDNA clone AK073893. This gene was more distantly related to the other four, and its predicted protein showed only 30–35% amino acid identity to the other predicted proteins in the NBS-encoding region.

Identification of two novel gene classes related to NBS–LRR genes

Two novel classes of genes (classes 4 and 5, Table 1) were identified that have not yet been previously described. Members of both classes of genes were occasionally observed in database searches with class 1 or class 2 NBS–LRR genes, although homology was weak. Conserved-domain (CD) searches (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) with predicted proteins from both types of genes found homology to NBS domains, but homology was weak and confined to the C-terminal portion of the NBS domains. No homology was observed to the P-loop motif and the sequences aligning with other conserved motifs, such as the kinase 2 and Gly-Leu-Pro-Leu (GLPL) motifs, fit the consensus sequences poorly. The CD search also identified a small LRR domain in the proteins predicted for the class 5 genes, but not the class 4 genes (Fig. 1). The N termini of the proteins encoded by the class 4 genes had significant homology to the N termini of class 2 NBS–LRR proteins, including the conserved nT domain.

To better examine the protein structures of these classes of genes, consensus sequences of the proteins from each class were constructed using BLOCKMAKER (Henikoff et al. 2000). The predicted proteins within each class aligned well but showed variable levels of homology, as expected for very old gene families. Nine predicted proteins from class 5 genes, which varied from 58% to 95% amino acid identity, were used to construct a consensus sequence of the class 5 proteins (Fig. 1b). Twenty predicted proteins from class 4 genes, ranging from 52% to 90% amino acid identity, were used for the class 4 consensus sequence (Fig. 1a). Alignments of predicted proteins from several class 2 genes with the consensus sequence of class 4 predicted proteins demonstrated that they align well at amino acid positions 1–150 in consensus sequence and again at amino acid position 670 until the end of the sequence (Fig. 1). The first region of homology included the nT domain, and the second corresponded to most of the NBS domain, from the kinase 2 motif through the MHD motif, and continued through the LRR. Although an LRR structure was not predicted for this class of genes, it is likely derived from an LRR; the C-terminal 500-amino acid region comprises 16.6% leucine residues and aligns well with the LRR of many of the class 2 genes. Interestingly, the conserved motifs in the NBS regions of class 2 genes did not correspond to regions conserved in the class 4 genes. The lack of an apparent P loop and poor conservation of the other conserved domains indicates that these NBS-homologous regions may not function as NBS domains. Alternatively, the region corresponding to the nT domain is conserved in both the class 2 and class 4 genes.

The consensus sequence of the proteins from the class 5 genes aligns well with most class 2 NBS–LRR genes through the middle of its coding region (amino acids 165–496, Fig. 1). The corresponding region of homology in class 2 proteins reaches from the C-terminal end of the NBS domain (after the GLPL) into the LRR domain.

Rice cDNA sequences were identified for several of the class 4 and class 5 genes, providing support for their predicted structures. Sequences were identified for 6 of the 27 class 4 genes (GenBank accession nos. AK073869, AK103061, AK105315, AK100243, AK103156, and AK103545) and four of the nine class 5 genes (accession nos. AK069045, AK101164, AK106733, and AK066757). When the class 4 sequences were used to screen the EST and genomic sequences in databases at NCBI, no similar genes were identified from dicot species. EST sequences from similar genes were identified from wheat (e.g., accession BE400398) and sorghum (BG487725), but the most similar sequences in dicot genomes were typical class 2 genes like *Rpm1* of *Arabidopsis*. In contrast, two *Arabidopsis* genes (*At5g45500* and *At5g45520*) were identified that were similar to the rice class 5 genes (tBLASTN probabilities = $3e^{-54}$ and $2e^{-48}$). The two genes map within 20 kb of each other on *Arabidopsis* chromosome 5. When the predicted proteins from two genes were aligned with ten of the rice class 5 proteins, they showed regions of sequence similarity with the regions that were conserved within the class 5 genes, but the sequence identity between the genes was low, as would be expected of genes that have diverged long ago (Fig. 2). Both *Arabidopsis* genes formed a group with two of the rice class 5 genes in the phylogenetic analysis.

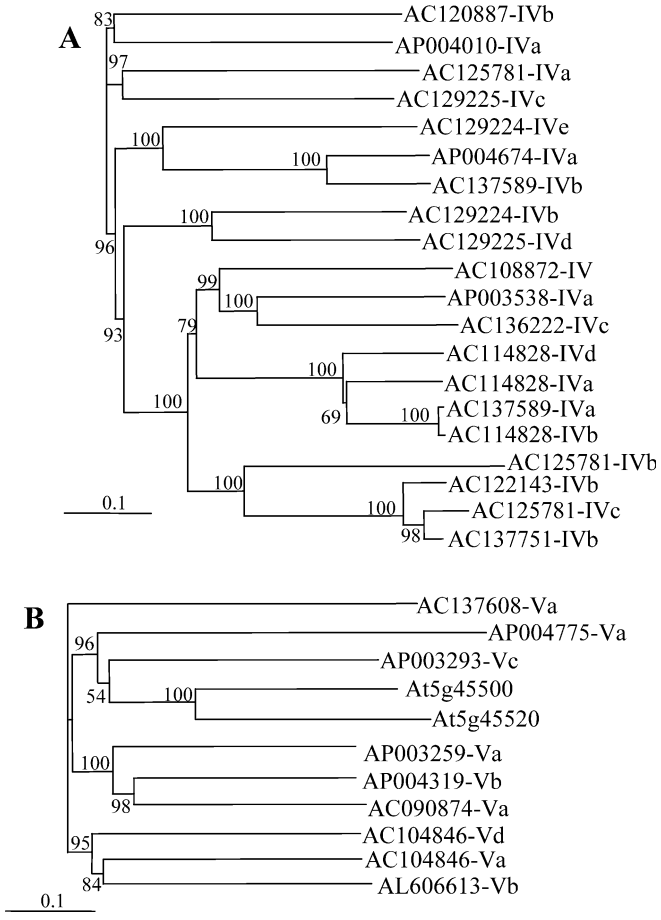

Fig. 2a, b Phylogenetic analysis of two distinct classes of nucleotide-binding site (NBS)-encoding genes. The *numbers* on the trees are bootstrap values (percentage of trees of 1,000 generated) that support each node. **a** Tree generated from alignment of 20 class 4 genes. No dicot sequences homologous to this class were identified. **b** Tree generated from alignment of nine class 5 genes together with two homologous predicted proteins from two *Arabidopsis* loci: *At5g45500* and *At5g45520*

Phylogenetic analysis of the class 2 NBS–LRR genes

A phylogenetic approach was taken to identify subfamilies and relationships between the class 2 genes. Amino acid sequences from the nT motif (consensus WVxxIRELAY-DIED) through the MHD motif were utilized, typically providing 450–500 amino acids for comparison. A total of 400 sequences were used with Clustal X-derived alignments to produce neighbor-joining trees with 1,000 bootstrap iterations (Electronic Supplementary Material, Fig. 4). Because of the large number of sequences in the analysis, genes that grouped together in more that 900 of 1,000 bootstrap cycles were represented as a single branch having the average branch length within that group. Members of groups with more than two members were given number designations (1–43) and are listed in Table 2 along with the bootstrap values for the group. The class 2 genes formed varying branch lengths on the phylogenetic tree reflecting the ancient divergence of these genes from each other. Genes that were not clustered together on the neighbor-joining tree showed very poor sequence identity to each other, typically less than 30%. Some of the genes did not cluster with any others and thus formed their own branches, indicating that these genes evolved independently of the rest of the class 2 genes for most of their evolutionary history.

Extension of the sequence region used for phylogenetic analysis from the nT to the end of the NBS (450–500 amino acids) improved the resolving power of the analysis, increasing the number of phylogenetically related groups over that achieved previously (Bai et al. 2002) when only the NBS region (~300 amino acids) was used for comparison. The use of longer sequence lengths enabled larger groups of evolutionarily related sequences to be identified with greater confidence (e.g., >90% bootstrap values) than what was possible with only ~300 amino acids. The 43 groups with more than two members were well supported by bootstrap analysis and separated by long branch lengths. The genes within each group are therefore presumed to have a common origin and arose by duplication from a single ancestral gene. The amplification of many of these groups occurred long ago as indicated by the low amino acid similarity within them (data not shown). The sequence divergence between members in a group ranged from ~36% (group 9) to ~69% (group 14). Some of the groups with more homogeneous members mapped to single clusters in the genome. For example, the three members of group 14 occurred in a cluster on chromosomes 2 and were found on sequence accession AC083751. The genes in such groups are likely derived from duplications of a single gene in recent evolutionary history.

Genome organization of rice NBS–LRR genes

To determine the distribution of rice NBS–LRR genes in the genome, a genetic map with the positions of the predicted genes was constructed (Fig. 3). Most of the Nipponbare sequence contigs in the database have been anchored on the high-density genetic map constructed by the RGP (Chen et al. 2002; Harushima et al. 1998; Sasaki et al. 2002). The approximate genetic map position of sequences carried on a single BAC clone was determined by identifying sequences corresponding to genetic markers on the BAC clones or on overlapping clones. Genes from a single BAC sequence were placed on the map at the position of the genetic marker residing on that BAC clone sequence. In genomic areas with high levels of recombination, some BAC clones that carried more than one marker spanned 1 cM or more. Gene(s) on these clones were then drawn at an average map position for the markers. Genes on two overlapping BAC sequences marked by the same genetic marker or genes on BACs with markers closer than 1 cM to each other were drawn as mapping to the same position.

Many of the groups of closely related genes included members that were located at different positions within one chromosome or on different chromosomes (Table 2). Some of the groups map to a surprising number of

Table 2 Partial list of class 2 NBS–LRR genes that group together in phylogenetic analysis (complete table in Electronic Supplementary Material)

Group no.	Bootstrap %[a]	Range of Identities (%)[b]	Genes designated by GenBank accession numbers[c]	C[d]
1	100	36–96	AP003219-1	1
			AP005777-2, AP005924-2, AP005924-1	2
			AL606669	4
			AC098834	5
			AP004010	7
			AP003617, AP003621-1 −2, −3	6
			AC079128-2	10
			AL713909-1, −2	12
5	100	42–92	AP005694-1, −2, −3, ***-4***	2
			AC104282-1, ***-2***, −3, −4	4
10	91	30–98	AP002540-1, −2	1
			AP005696	2
			AC097277-1, −2	3
			AC137001, AC104274-2	5
			AP003839-1, −2	7
			AP005151, AP004256	8
			AP005782-1, AC090057, AP005876, AP005926-3	9
			AC092388, AC078948	10
			AC134922-2, −3, −4, −5, AC135643-1, −2, −3, −4, AL713947-1	11
			AC135460, ***AC114012-5***	11
			AC134045	12
13	99	36–84	AC107314-1	10
			AC135644-4, −5, −6, −8, AC135190-2, AC133005-2, AC119072-1, −2	11
			AL772421	12
29	99	30–68	AP003859-1, AP004137-2	8
			AC092548, AC107314-2	10
			AC137113-1, AC119072-3, AC119073-1, AC133005-1	11
			AC135190-3, −5, AC135644-7, −9	11
			AL772419-2, AL954153-2	12
30	100	54–78	AP003276	1
			AC074283, AC093093	10
			AC135643-5, AC136905	11
38	94	24–94	AP003226, AP003345	1
			AC134234, AC105928-1, −2	3
			AL606441	4
			AP005700, AP005547, AP005879, AP005586-1, −2, −3	9
			AC114011, AC108871	11
			AL731630, BX000346-3, AL935067-1, −2	12
39	99	38–97	AP003275-3, −4, −5	1
			AP005009	2
			AC092559-1, −2, −3	3
			AC105767-1, −2, AC136222-2	5
			AP003914-1, −2, −3, −4	8
			AC122147-1, −2, −4	10

[a]Bootstrap % refers to the percentage of trees in which the members formed a clade in the phylogenetic analysis

[b]Refers to the range in predicted amino acid identities between the members of the group in the nT–NBS coding region

[c]Gene designations in *boldface italics* do not code for LRR domains

[d]Chromosome on which the genes map

different positions. For example, groups 1 and 10 had members that mapped to eight and ten different chromosomes, respectively. Similarly, groups 38 and 39 both had members mapping to six different chromosomes. In addition, genes from a single phylogenetic group that lay on the same chromosome did not always cluster at a single locus. For example, the genes in group 38 on chromosome 9 map approximately 2, 21, 29, and 35 map units from the top of the chromosome on the RGP map. The distribution of the duplicated genes around the genome indicates that the events that duplicated the genes from the different phylogenetic groups occurred for the most part independently. There were, however, genes that were apparently duplicated as a group. For instance, unrelated genes from

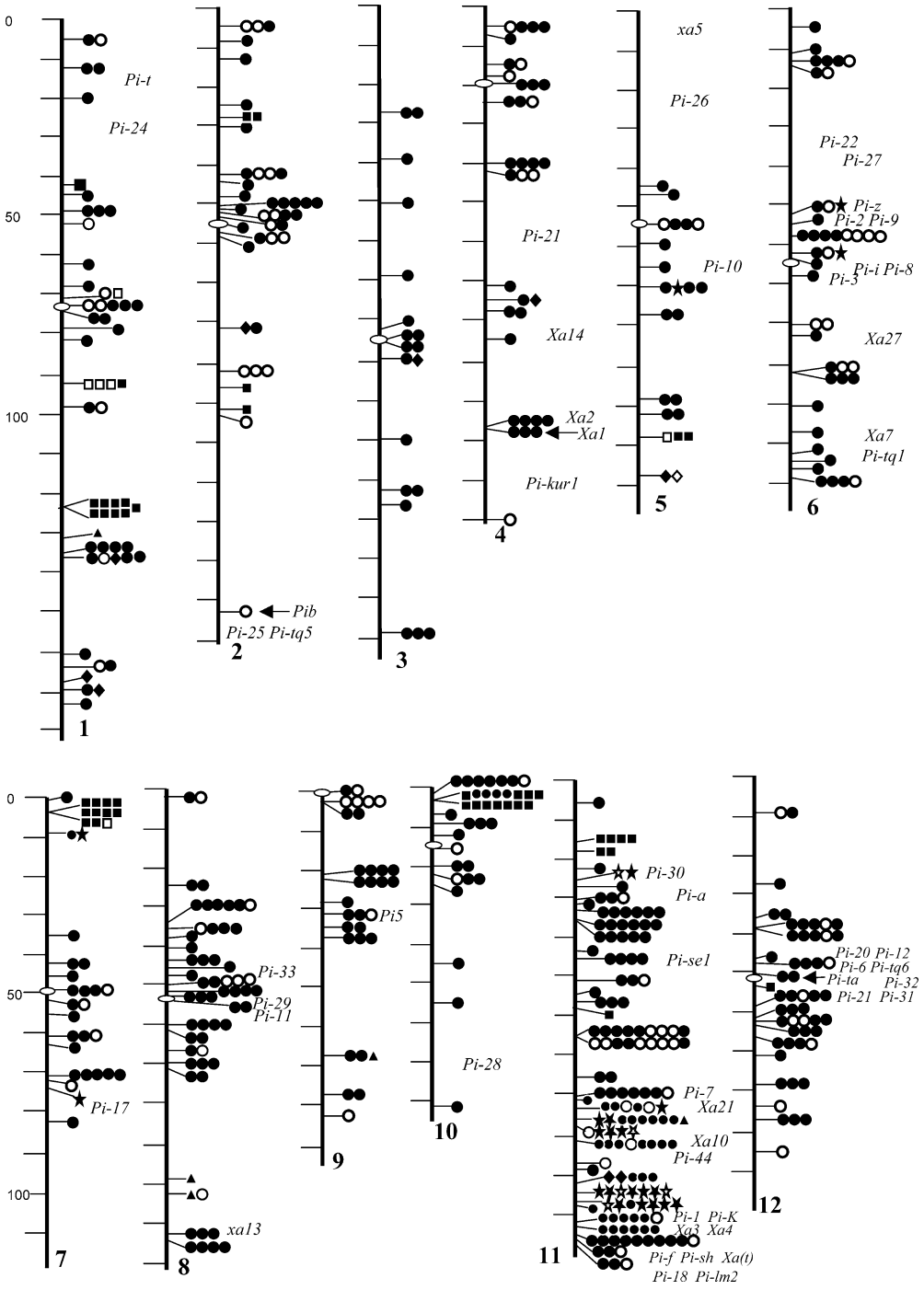

Fig. 3 Distribution of NBS-encoding genes in the rice genome. Genetic map positions of NBS–LRR and related genes were based on relative positions of the genomic clones to genetic markers on the Rice Genome Project genetic map (Chen et al. 2002). Genes whose sequences are on the same BAC or PAC clone or predicted to map within 1 cM were drawn as mapping to the same location. Genes are represented as shapes to the right of the chromosomes (*vertical lines*) as follows: *circles* class 2, *squares* class 3, *stars* class 4, *diamonds* class 5, and *triangles* class 6. Predicted pseudogenes from the different classes are shown as *open shapes*. *Open ovals* represent centromeres. The locations of three characterized rice NBS–LRR resistance genes, *Pib*, *Pi-ta*, and *Xa1*, are designated with *arrows*. The approximate map positions of other bacterial blight (designated *Xa*-) and blast (designated *Pi*-) resistance genes that have been mapped by phenotype by various groups (Berruyer et al. 2003; Causse et al. 1994; Conaway-Bormans et al. 2003; Gu et al. 2004; Jeon et al. 2003; Jiang and Wang 2002; Li et al. 1999; Liu et al. 2002; Porter et al. 2003; Sallaud et al. 2003; Tabien et al. 2002; Wang et al. 1994; Yoshimura et al. 1995) are shown to the *right* of the chromosomes

groups 1, 5, 10, and 39 occurred together on chromosome 2 and on chromosome 5, where they are clustered together at the 40- to 50-cM region in both chromosomes. Similarly, genes from groups 29 and 13 are interspersed in two distant locations of chromosome 11. One cluster, composed of genes from accessions AC136998, AC135190, and AC135644, mapped at position 115 while the other, composed of genes from AC133005 and AC119072, mapped near position 35 (Electronic Supplementary Material, Fig. 5). The simplest explanation for this type of arrangement is an event that duplicated a chromosome segment carrying at least one member of each of these groups. Duplications of blocks of NBS–LRR genes were also predicted to be important in the amplification of the NBS–LRR genes in *Arabidopsis* (Baumgarten et al. 2003).

If the cytological events that duplicated the members of these families to different chromosomal regions all occurred at one time, the amount of divergence between the duplicated members would be expected to be similar. Alternatively, if duplication and dispersion is a continuous process in the rice genome, then a range of divergences would be expected. A very broad range of similarities was observed between the members of groups when the nT–

NBS portions of their coding regions were compared, even between group members that were duplicated on different chromosomes. Many groups have members mapping to different chromosomes that have as little as 30% amino acid sequence identity (Table 2). At the other extreme, some groups have members that appear to be duplicated to different chromosomes relatively recently. For example, gene AC135643-5 on chromosome 11 (group 30, Table 2) was 78% identical to gene AC093093 on chromosome 10 and 69% identical to AP003276 on chromosome 1. The five genes in this group map to three different chromosomes, and none of the members are within 15 cM of each other, yet the most dissimilar ones are 55% identical. Some groups have both distantly and closely related members dispersed throughout the genome. For example, group 10 has members with only 30% identity but also includes genes like AC134045 on chromosome 11 and AC104274-2 on chromosome 5, which are 73% identical.

The NBS–LRR genes were noticeably clustered in the rice genome. The numbers of genes in these clusters ranged from 2–12 or more (Fig. 3). A few of the smaller gene clusters were composed of a single family of closely related genes, but most of these are within 1–2 cM of unrelated NBS–LRR genes. Thus, a surprising number of clusters are composed of class 2 genes from two or more phylogenetic groups or even two different gene classes. These "mixed class" clusters are typically composed of class 2 genes together with genes from one of the other classes. Most of the class 4 and class 5 genes occur in close proximity to class 2 genes. This arrangement probably reflects the original pattern of duplication where the ancestral class 4 or class 5 gene arose by an ancient local duplication of a class 2 gene followed extensive divergence. These original mixed class clusters were then duplicated together as chromosome segments to other regions of the genome. The observation that certain groups of class 2 genes are associated with the class 4 and class 5 genes supports this model of segmental duplication and divergence. Class 2 genes from phylogenetic group 38 are clustered with class 5 genes at two positions on chromosome 1 (map positions ~137 and ~170). Another class 5 gene is flanked by group 38 genes on chromosome 4 (positions ~70–78).

Some of the clustered NBS–LRR regions extend over several centiMorgans and average more than two NBS-encoding genes per centiMorgan. A few of these clusters are in regions of low recombination, as in the vicinity of the centromeres of chromosomes 1 and 5. The genes appear closely clustered genetically, but cover a large physical distance. Other clusters do have a higher density of NBS-encoding genes. Some of the largest clusters map to chromosome 11, where they cover much of the chromosome. Over 25% of the physically mapped NBS-encoding genes we identified (153/592) mapped to chromosome 11.

Sequence divergence among linked NBS–LRR family members

Some of the resistance genes characterized in maize are members of large families of closely linked, highly homologous (>90% amino acid identity) NBS–LRR genes (Sun et al. 2001; Webb et al. 2002). Such large families of closely related sequences were not found to be common among the class 2 genes in the rice genome. The relative homology of linked NBS–LRR genes was assessed as percent amino acid identity in alignments of the deduced sequences from their amino terminal portions (nT through NBS domains). Linked genes with high levels of sequence identity were not uncommon, but there were usually only two to four highly similar genes in these clusters (Electronic Supplementary Material, Fig. 5). Larger families of linked genes were not uncommon, but they typically showed lower levels of sequence homology. The most homogeneous large families included four genes on AC104282 (chromosome 4) that showed 79–92% identity and six genes on accession AC104848 (chromosome 11) with from 59% to 87% sequence identity. Interestingly, a gene on chromosome 7 (AP005258) was more similar to five of the six chromosome 11 genes than was the sixth gene in the cluster. Neither of the clusters of related genes was arranged as a simple tandem array of genes; unrelated genes were predicted between at least some of the family members. Thus, the closely related genes in linked clusters were not always close in proximity. In one extreme example, a gene on accession AL713900 was approximately 9 cM from a closely related (up to 94% identical) cluster of three genes on accession AL954871.

Most of the class 3 genes in rice were found in three tightly linked clusters on chromosomes 1 (on accession AP003293), 7 (AP003810), and 10 (AC079843 and AC074283). As with the class 2 clusters, the members of the class 3 clusters were not all closely related. The chromosome 1 cluster carried nine genes with less than 74% predicted amino acid sequence identity between members, and with less than 40% between some members. The chromosome 7 cluster was very similar: ten predicted genes (plus one pseudogene) with 74% identity among the most similar and less than 40% among the least similar. The chromosome 10 cluster carried 11 predicted class 3 genes. Again, the least closely related had less than 40% amino acid sequence identity, but one pair of adjacent genes showed 84% sequence identity, and another three were also more than 80% identical to each other. These gene clusters all had members that are tandemly arranged but they were not simple tandem arrays, because they have other predicted genes interspersed with them. The chromosome 10 cluster had one class 2 gene within the array.

The rate and extent of divergence of linked gene families is thought to be affected by the extent to which they recombine with each other (Baumgarten et al. 2003; Hulbert et al. 2001). The orientation of linked genes affects whether they can mispair and crossover, since the products of unequal crossing over between inverted genes

are dicentric and acentric chromosomes. To examine the effect of orientation on divergence, the extent of divergence between adjacent class 2 NBS–LRR genes was therefore compared for pairs of genes in the same or opposite orientations. Genes with close physical linkages were identified by selecting BAC or PAC clones, or adjacent overlapping clones that carried two or more NBS-encoding genes (Electronic Supplementary Material, Fig. 5). While many of the clusters had genes in only a single orientation, 33 of the fragments had NBS-encoding genes in both orientations (excluding pseudogenes). This allowed for the comparison of 92 pairs of adjacent genes in the same orientation to 55 gene pairs in opposite orientation. Orientation of genes with respect to each other had a pronounced effect on the extent of sequence homology between them. The NBS-coding regions (nT to MHD motifs) of the genes in the same orientation averaged 63% amino acid identity, while those in the opposite orientation averaged only 40%. The distance between the duplicated genes appeared to have much less, if any, effect on their divergence. When gene pairs in the same orientation were classified by whether their NBS-encoding sequences were more or less than 20 kb apart, the genes within 20 kb of each other (63 gene pairs) averaged 64% identity, while the genes more than 20 kb apart (29 pairs) averaged 62%. This similarity was surprising, considering that NBS–LRR genes more than 20 kb apart typically had unrelated genes predicted between them, while the closer genes frequently did not.

A second possible explanation for lower levels of sequence identity between oppositely oriented genes is if gene clusters with such genes are typically formed differently than clusters with genes in the same orientation. For example, if two genes in opposite orientation were brought together by a chromosomal rearrangement that brings genes from distant locations together, they would be expected to be unrelated. Many of the divergent genes linked in opposite orientation are members of the same phylogenetic group, however, indicating that they arose by local duplications from a single progenitor gene. Examples include group 2 genes on accession AP005305 (43% identical) and group 10 genes on AP003839 (36%) and AC097277 (45%). The results therefore suggest that many, if not most, of the closely linked gene clusters arose by divergence from one or a few progenitor genes and that duplications that invert the orientation of the genes promote divergence by inhibiting recombination.

Homology with NBS–LRR genes from other cereals

Bai et al. (2002) found that many of the wheat, barley, and maize NBS-encoding sequences available in GenBank were highly homologous (up to 85% identical in amino acid sequence) to the NBS domains of rice proteins. Other NBS-encoding genes showed much lower homology between genera. Ten of 28 families of NBS-encoding genes from other cereals showed 60% or less amino acid identity in their predicted NBS domains to the most similar rice sequence available at that time. Since the rice databases were nearly complete, the genomic sequences in GenBank were searched (June 2003) to determine whether more similar homologues were available. While sequences were identified that were slightly more homologous to two of the families, eight gene families were still only 60% or less identical to the closest rice sequences. The result explains why many of the NBS–LRR genes cross-hybridize poorly between cereals in gel blot hybridization experiments (Bai et al. 2002; Leister et al. 1998).

Association of the NBS–LRR genes with major genes for disease resistance

A large number of disease resistance genes have been described in rice, and many of these have been roughly placed on the genetic maps. At least three NBS–LRR resistance genes have been isolated and sequenced (Bryan et al. 2000; Wang et al. 1999; Yoshimura et al. 1998). The rice blast resistance genes *Pib* and *Pi-ta* are allelic to Nipponbare sequences on chromosomes 2 and 12, respectively (Fig. 3), and the bacterial blight resistance gene *Xa1* is allelic to a gene on chromosome 4. The approximate map positions of many other blast resistance (*Pi*) and blight resistance (*Xa*) genes suggest that the NBS–LRR genes may be candidates for some of them. For example, most of the *Pi* and *Xa* genes on chromosome 11 are closely associated with clusters of NBS–LRR genes. Several other resistance genes are not near NBS–LRR genes, like the *xa5* and *Pi-26* genes at the top of chromosome 5. It is possible that these resistances are conferred by other types of genes. However, it is also possible there are NBS–LRR genes in some regions that have not yet been identified, since the genomic sequence is not complete and gaps in the physical map remain. Two NBS–LRR genes on genomic sequences (on accession AC145395) and four cDNAs were identified for which the map locations are not yet known.

Discussion

The characterization of the NBS-encoding genes in the rice genome provides an initial view of a large component of the defense gene arsenal in cereal genomes, a view that is somewhat different from that provided by the *Arabidopsis* genome. The identification of 585 predicted NBS-encoding genes verifies earlier predictions (Bai et al. 2002; Meyers et al. 2002) that rice carries many more of these sequences than *Arabidopsis* (Meyers et al. 2003), and that this class of genes probably accounts for approximately 1% or more of the genes in rice (Feng et al. 2002; Goff et al. 2002; Yu et al. 2002). The lack of class 1 (TIR–NBS–LRR) genes, the largest class of NBS–LRR genes in *Arabidopsis*, is apparently compensated for by a greatly amplified class 2 complement with over 500 predicted members (including pseudogenes). In addition, rice carries two additional classes of genes related to

NBS–LRR genes, totaling approximately 80 genes, with no homologues in *Arabidopsis* or other dicot sequences in databases. As in dicots, several class 2 genes have demonstrated functions in cereals as resistance genes (Bryan et al. 2000; Collins et al. 1999; Wang et al. 1999; Yoshimura et al. 1998; Zhou et al. 2001), and no other phenotypes have been associated with them. No phenotypes have yet been associated with the class 3 through class 6 NBS-encoding genes. The largest of these classes is the class 3 genes, with more than 50 members in rice. These are the most similar in sequence homology to the class 1 and class 2 NBS–LRR genes and appear to be confined to monocot genomes. The lack of an LRR-encoding domain may be an indication that they are not directly involved in pathogen recognition. Alternatively, the class 4 and class 5 genes both lack an apparent P-loop motif and, therefore, do not likely have a functional nucleotide-binding domain. Mutagenesis studies have shown that the NBS domain is critical for function in signaling defense responses by NBS–LRR genes (Axtell et al. 2001; Dinesh-Kumar and Baker 2000; Tao et al. 2000; Tornero et al. 2002). Therefore, if the class 4 and class 5 genes are involved in defense signaling, they may perform this function differently than other NBS–LRR genes. Like the class 3 genes, the class 4 genes do not have homologues in *Arabidopsis*. Two class 5 genes were identified in the *Arabidopsis* genome, but no phenotypes have yet been associated with them.

Approximately 20% of the NBS–LRR genes in the Nipponbare genome were predicted to be pseudogenes. A similar estimate was derived when only transcribed sequences were considered. While this seems very high, it may actually be an underestimate of the number of genes that are nonfunctional, since only fragmented genes or those with obvious mutations such as in-frame stop codons were classified as pseudogenes. There are, however, functional alleles for at least some of these genes in other rice germplasm. An example is *Pib*; the allele in the *indica* cultivar Engkatek confers resistance to isolates of the rice blast fungus *Magnaporthe grisea*, but the allele from Nipponbare is an apparent pseudogene. Meyers et al. (2003) found that nearly 10% of the NBS–LRR genes in the Columbia ecotype of *Arabidopsis* were apparent pseudogenes. Pan et al. (2000a) also noticed a high frequency of pseudogenes in tomato, for which seven of 70 NBS-homologous gene fragments cloned had in-frame stop codons. A few of the clusters of related NBS–LRR genes in rice carried multiple pseudogenes (Fig. 3; Electronic Supplementary Material, Fig. 5). This has also been observed at the *Rpp5* gene cluster in *Arabidopsis*, where most of the *Rpp5* paralogues in the *Ler* and Colombia haplotypes were predicted to be pseudogenes. The high frequency of pseudogenes is consistent with the idea that these genes perform no useful function in environments without specific pathogens or even specific pathogen races, and the functional alleles may even confer a fitness cost (Tian et al. 2003).

The near completion of the rice genome sequencing efforts and integration with the genetic map has allowed a global view of the distribution of NBS-encoding genes. Coupled with a phylogenetic analysis of these sequences, the genome distribution provides a picture of how the NBS-encoding genes expanded in the rice genome. The presence of monophyletic groups of genes at multiple map positions shows that the extensive duplication events that amplified them also spread them through the genome. The majority (31/42=74%) of groups of class 2 NBS–LRR genes with more than two members were mapped to multiple chromosomes. All 12 rice chromosomes carried members of at least five different dispersed groups. Segmental duplication of chromosomal fragments carrying multiple NBS-encoding genes accounted for some of the duplications, but many appeared duplicated independently. Baumgarten et al. (2003) estimated that most of the duplications moving NBS–LRR genes to ectopic chromosomal locations were the same chromosomal segmental duplications that amplified much of the *Arabidopsis* genome. The history of chromosomal duplications that occurred during the evolution of the rice genome is still unclear, but the picture emerging is somewhat different from the patterns of duplications observed here for NBS–LRR genes (Vandepoele et al. 2003). A major fraction of the previously detected segmental (block) duplications in the rice genome involved chromosome 2. In contrast, chromosomes 11 and 12 included the most members of different multi-chromosome groups of NBS–LRR genes, each carrying members of 15 different groups that mapped to multiple chromosomes. As with the segmental duplications previously described in the rice genome, the NBS–LRR gene duplication events occurred independently many times, as evidenced by the broad range of homologies between different duplicated sequences. Both segmental duplications and duplications of individual NBS–LRR genes were clearly important in dispersing these genes in the rice genome. Whether they were dispersed individually more frequently than other classes of rice genes will not be clear until a very thorough analysis of gene duplication through the whole rice genome is performed.

The most frequent duplications of NBS–LRR genes are to adjacent genomic positions. It is these localized duplications that cause the clustering of NBS-encoding genes in the rice genome, as is also observed in the *Arabidopsis* genome (Baumgarten et al. 2003; Meyers et al. 2003; Richly et al. 2002; Young 2000) and inferred by mapping experiments in other plants. The most surprising aspects of the clustering in the rice genome are the diversity of the different genes in the clusters and the lack of large homogeneous arrays of genes. Many of the clusters are composed of genes from different classes as well as different phylogenetic groups of class 2 genes. The genes in the *Rp1* and *Rp3* complexes of maize typically code for proteins with approximately 90–99% sequence identity. Different maize lines carry different numbers of genes in their haplotypes because they recombine unequally, but haplotypes with a dozen or more genes are common, and some *Rp1* haplotypes carry more than 50 tightly clustered family members (S. Smith and S. Hulbert,

unpublished data). Rice also carries clusters of closely related genes, but they typically are small or have more divergent members. It is not clear whether this represents a difference in the organization of NBS–LRR genes in the maize and rice genomes, because is it not known how representative *Rp1* and *Rp3* are of the genes in the maize genome.

The extent to which recently duplicated genes can recombine influences how rapidly they diverge. Recombination between linked genes is thought to homogenize their sequences so they evolve in a concerted fashion (Arnheim 1983; Dover 1982). Duplications to unlinked locations allow the genes largely to escape these effects, although rare recombination events may still occur (Parniske and Jones 1999). Comparisons of closely linked NBS–LRR genes in the rice genome indicated that localized duplications that invert the orientation of the genes also allow the genes to evolve more independently and diverge from the other members of the cluster. Unequal crossing over between inverted repeats would effectively be inhibited, since the products cause chromosomal aberrations. Once members of a gene family have diverged, they are less likely to recombine with each other, since sequence heterogeneity tends to reduce recombination (Dooner and Martinez-Ferez 1997). Previous estimates of copy number of specific NBS–LRR gene families in different rice lines implied that unequal crossover events are not common in most rice families. Only one of 73 NBS probes hybridized detected noticeable differences in copy numbers among the different rice cultivars (Bai et al. 2002). This is consistent with the idea that most of these families are similar enough to cross-hybridize in gel-blot experiments but divergent enough that they mispair infrequently.

Some of the NBS–LRR gene duplication events in the rice genome occurred relatively recently, possibly after many of the grass genera diverged from each other. These recent duplication events may partly explain the poor synteny that has been observed in comparisons of positions of NBS–LRR genes between cereal genomes (Leister et al. 1998; Ramalingam et al. 2003). A possibly more important contributor to this apparent reduced synteny is resistance gene loss. This is supported by the relatively poor homology exhibited by many cereal NBS sequences when compared to the closest homologues in the rice genome. The evolution of many NBS–LRR genes is influenced by diversifying selection, which might cause different genes to evolve at different rates. However, this is mainly true of the LRR domain (Mondragon-Palomino et al. 2003). The sequences used in the present analysis, including the NBS domain, are generally the most conserved regions of these gene families, and it is not clear why these sequences would evolve at noticeably different rates. The simplest explanation for the poor intergenus conservation for many of the NBS sequences in cereal genomes is the loss of genes from specific grass lineages. The poorly homologous sequences under comparison are likely paralogues and not true orthologues. The rice genome is now considered an important tool for map-based gene identification of genes controlling important traits in other cereals (Bennetzen and Freeling 1993). The map positions of all the rice NBS–LRR genes could therefore be very useful in predicting whether related genes correspond to resistance genes in other cereals if syntenic regions can be identified in rice. The utility of this approach is often limited by breakdown of microsynteny caused by small-scale rearrangements, duplications, and deletions (Bennetzen and Ramakrishna 2003; Kilian et al. 1999). More comparative analysis between the cereals is needed to determine whether this approach will be less effective for resistance genes than for genes controlling other phenotypes.

The detailed map positions of the NBS–LRR encoding genes in the rice genome provide candidate genes for many simply inherited resistances that have been or will be placed on the genetic map. Most resistance genes characterized to date belong to the NBS–LRR family (Hulbert et al. 2001). Until we better understand the nature of genes that contribute to quantitative levels of resistance, the NBS–LRR genes are also viable candidates for these genes. The integration of the annotated genome sequence and the genetic map of rice will allow the molecular dissection of both simply inherited and complicated traits in a crop plant with unprecedented efficiency. The localization of all the genes known to be involved in disease resistance is an important step in this direction.

Acknowledgements The authors wish to thank James C. Nelson, Rebecca Nelson, and Jianfa Bai for valuable suggestions for the data analysis and manuscript preparation. This work was supported by National Science Foundation grants 9975971 and 0090883. This is contribution number 04-197-J from the Kansas Agricultural Experiment Station.

References

Aarts N, Metz M, Holub E, Staskawicz BJ, Daniels MJ, Parker JE (1998) 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 USA 95:10306–10311

Arnheim N (1983) Concerted evolution of multigene families. In: Nei M, Koehn RK (eds) Evolution of genes and proteins. Sinauer, Sunderland, Mass., pp 38–61

Axtell MJ, McNellis TW, Mudgett MB, Hsu CS, Staskawicz BJ (2001) Mutational analysis of the *Arabidopsis RPS2* disease resistance gene and the corresponding *Pseudomonas syringae avrRpt2* avirulence gene. Mol Plant Microbe Interact 14:181–188

Bai J, Pennill L, Ning J, Lee SW, Ramalingam J, Webb CA, Zhao B, Sun Q, Nelson JC, Leach JE, Hulbert SH (2002) Diversity of nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res 12:1871–1884

Baumgarten A, Cannon S, Spangler R, May G (2003) Genome-level evolution of resistance genes in *Arabidopsis thaliana*. Genetics 165:309–319

Bennetzen JL, Freeling M (1993) Grasses as a single genetic system: genome composition, colinearity and compatibility. Trends Genet 9:259–261

Bennetzen JL, Ramakrishna W (2003) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol Biol 48:821–827

Berruyer R, Adreit H, Milazzo J, Gaillard S, Berger A (2003) Identification and fine mapping of *Pi33*, the rice resistance genes corresponding to the *Magnaporthe grisea* avirulence gene *ACE1*. Theor Appl Genet 107:1139–1147

Bryan GT, Wu K-S, Farrall L, Jia Y, Hershey HP, McAdams SA, Donaldson GK, Tarchini R, Valent B (2000) A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. Plant Cell 12:2033–2046

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Cannon SB, Zhu H, Baumgarten AM, Spangler R, May G, Cook DR, Young ND (2002) Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS–LRR resistance gene subfamilies. J Mol Evol 54:548–562

Causse MA, Fulton TM, Cho YG, Ahn SN, Chunwongse J, Wu K, Xiao J, Yu Z, Ronald PC, Harrington SE, Second G, McCouch SR, Tanksley SD (1994) Saturated molecular map of the rice genome based on an interspecific backcross population. Genetics 138:1251–1274

Chen M, Presting G, Barbazuk W, Goicoechea JL, Blackmon B, Fang G, Kim H (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

Collins N, Drake J, Ayliffe M, Sun Q, Ellis J, Hulbert S, Pryor T (1999) Molecular characterization of the maize *Rp1-D* rust resistance haplotype and its mutants. Plant Cell 11:1365–1376

Conaway-Bormans CA, Marchetti MA, Johnson CW, McClung AM, Park WD (2003) Molecular markers linked to the blast resistance gene *Pi-z* in rice for use in marker-assisted selection. Theor Appl Genet 107:1014–1020

Dangl JL, Jones JD (2001) Plant pathogens and integrated defense responses to infection. Nature 411:826–833

Dinesh-Kumar SP, Baker BJ (2000) Alternatively spliced N resistance gene transcripts: their possible role in tobacco mosaic virus resistance. Proc Natl Acad Sci USA 97:1908–1913

Dooner HK, Martinez-Ferez IM (1997) Recombination occurs uniformly within the *bronze* gene, a meiotic recombination hotspot in the maize genome. Plant Cell 9:1633–1646

Dover GA (1982) Molecular drive: a cohesive mode of species evolution. Nature 199:111–117

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X, Jia P, Zhang Y, Zhao Q, Ying K, Yu S, Tang Y, Weng Q, Zhang L, Lu Y, Mu J, Lu Y, Zhang LS, Yu Z, Fan D, Liu X, Lu T, Li C, Wu Y, Sun T, Lei H, Li T, Hu H, Guan J, Wu M, Zhang R, Zhou B, Chen Z, Chen L, Jin Z, Wang R, Yin H, Cai Z, Ren S, Lv G, Gu W, Zhu G, Tu Y, Jia J, Zhang Y, Chen J, Kang H, Chen X, Shao C, Sun Y, Hu Q, Zhang X, Zhang W, Wang L, Ding C, Sheng H, Gu J, Chen S, Ni L, Zhu F, Chen W, Lan L, Lai Y, Cheng Z, Gu M, Jiang J, Li J, Hong G, Xue Y, Han B (2002) Sequence and analysis of rice chromosome 4. Nature 420:316–320

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft seqeunce of the rice genomes (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gu K,Tian D, Yang F, Wu L, Sreekala C, Wang D, Wang G-L, Yin Z (2004) High-resolution genetic mapping of *Xa27(t)*, a new bacterial blight resistance gene in rice, *Oryza sativa* L. Theor Appl Genet 108:800–807

Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin SY, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS, Sasaki T (1998) A high-density rice genetic linkage map with 2,275 markers using a single F_2 population. Genetics 148:479–494

Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the Blocks Database servers. Nucleic Acids Res 28:228–230

Hulbert SH, Webb CA, Smith SM, and Sun Q (2001) Resistance gene complexes: evolution and utilization. Annu Rev Phytopathol 39:285–312

Jeanmougin F, Thompson JD, Guoy M, Higgins DG, Gibson TJ (1988) Multiple sequence alignment with Clustal X. Trends Biochem Sci 10:403–405

Jeon J-S, Chen D, Yi G-H, Wang GL, Ronald PC (2003) Genetic and physical mapping of Pi5(t), a locus associated with broad-spectrum resistance to rice blast. Mol Gen Genet 269:280–289

Jiang J, Wang S (2002) Identification of a 118-kb DNA fragment containing the locus of blast resistance gene *Pi-2(t)* in rice. Mol Genet Genomics 268:249–252

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu R, Ueda M, Matsubara K, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. Science 301:376–379

Kilian HF, Chen JP, Kudrna D, Steffenson B, Yamamoto K, Matsumoto T, Sasaki T, Kleinhofs A (1999) Sequence analysis of a rice BAC covering the syntenous barley *Rpg1* region. Genome 42:1071–1076

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

Li ZK, Luo LJ, Mei HW, Paterson AH, Zhao XH, Zhong DB, Wang YP, Yu XQ, Zhu L, Tabien R, Stansel JW, Ying CS (1999) A "defeated" rice resistance gene acts as a QTL against a virulent strain of *Xanthomonas oryzae* pv *oryzae*. Mol Genet Genomics 261:58–63

Liu G, Lu G, Zeng L, Wang G-L (2002) Two broad-spectrum blast resistance genes, *Pi9(t)* and *Pi2(t)*, are physically linked on rice chromosome 6. Mol Genet Genomics 267:472–480

Meyers BC, Morgante M, Michelmore RW (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 BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS–LRR-encoding genes in *Arabidopsis*. Plant Cell 15:809–834

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 8:1113–1130

Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS (2003) Patterns of positive selection in the complete NBS–LRR gene family of *Arabidopsis thaliana*. Genome Res 12:1305–1325

Nicholas KB, Nicholas HB, Deerfield DW (1997) GeneDoc: analysis and visualization of genetic variation. Embnet News 4:1–4

Pan Q, Liu Y-S, Budai-Hadrian O, Sela M, Carmel-Goren L, Zamir D, Fluhr R (2000a) 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, Wendel J, Fluhr R (2000b) Divergent evolution of plant NBS–LRR resistance gene homologues in dicot and cereal genomes. J Mol Evol 50:203–213

Parniske M, Jones JDG (1999) Recombination between diverged clusters of the tomato *Cf-9* plant disease resistance gene family. Proc Natl Acad Sci USA 96:5850–5855

Porter BW, Chittoor JM, Yano M, Sasaki T, White FF (2003) Development of mapping markers linked to the rice bacterial blight resistance gene *Xa7*. Crop Sci 43:1484–1492

Ramalingam J, Vera Cruz CM, Kukreja K, Chittoor JM, Wu JL, Lee SW, Baraoidan MR, George ML, Cohen M, Hulbert SH, Leach JE, Leung H (2003) Candidate resistance genes from rice, barley, and maize and their association with qualitative and quantitative resistance in rice. Mol Plant Microbe Interact 16:14–24

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

Salamov A, Solovyev V (2001) *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res 10:516–522

Sallaud C, Lorieux M, Roumen E, Tharreau D, Berruyer R, Svestasrani P, Garsmeur O, Ghesquiere A, Notteghem JL (2003) Identification of five new blast resistance genes in the highly blast-resistant rice variety IR64 using a QTL mapping strategy. Theor Appl Genet 106:794–803

Salmeron JM, Oldroyd GED, Tommens CMT, Scofield SR, Kim H-S, Lavelle DT, Dahlbeck D, Staskawicz BJ (1996) Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. Cell 86:123–133

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, Antonio BA, Kanamori H, Hosokawa S, Masukawa M, Arikawa K, Chiden Y, Hayashi M, Okamoto M, Ando T, Aoki H, Arita K, Hamada M, Harada C, Hijishita S, Honda M, Ichikawa Y, Idonuma A, Iijima M, Ikeda M, Ikeno M, Ito S, Ito T, Ito Y, Ito Y, Iwabuchi A, Kamiya K, Karasawa W, Katagiri S, Kikuta A, Kobayashi N, Kono I, Machita K, Maehara T, Mizuno H, Mizubayashi T, Mukai Y, Nagasaki H, Nakashima M, Nakama Y, Nakamichi Y, Nakamura M, Namiki N, Negishi M, Ohta I, Ono N, Saji S, Sakai K, Shibata M, Shimokawa T, Shomura A, Song J, Takazaki Y, Terasawa K, Tsuji K, Waki K, Yamagata H, Yamane H, Yoshiki S, Yoshihara R, Yukawa K, Zhong H, Iwama H, Endo T, Ito H, Hahn JH, Kim HI, Eun MY, Yano M, Jiang J, Gojobori T (2002) The genome sequence and structure of rice chromosome 1. Nature 420:312–316

Sun Q, Collins NC, Ayliffe M, Smith SM, Drake J, Pryor A, Hulbert SH (2001) Recombination between paralogues at the *rp1* rust resistance locus in maize. Genetics 158:423–438

Tabien E, Li Z, Patterson AH, Marchetti A, Stansel W, Pinson M (2002) Mapping QTLs for field resistance to the rice blast pathogen and evaluating their individual and combined utility in improved varieties. Theor Appl Genet 105:313–324

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

Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J (2003) Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. Nature 423:74–77

Tornero P, Chao RA, Luthin WN, Goff SA, Dangl JL (2002) Large-scale structure-function analysis of the *ArabidopsisRPM1* disease resistance protein. Plant Cell 14:435–450

Vandepoele K, Simillion C, Van de Peer Y (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell 15:2192–2202

Wang G, Mackill DJ, Bonman JM, McCouch SR, Nelson RJ (1994) RFLP mapping of genes conferring complete and partial resistance to blast resistance in a durably resistant rice cultivar. Genetics 136:1421–1434

Wang ZX, Yano M, Yamanouchi U, Iwamoto M, Monna L, Hayasaka H, Katayose Y, Sasaki T (1999) The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. Plant J 19:55–64

Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, Cartinhour S, Stein LD, McCouch SR (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Res 30:103–105

Webb CA, Richter TE, Collins NC, Nicolas M, Trick HN, Pryor T, Hulbert SH (2002) Genetic and molecular characterization of the maize *rp3* rust resistance locus. Genetics 162:381–394

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86:6201–6205

Yoshimura S, Yoshimura A, Iwata N, McCouch SR, Abenes ML, Baraoidan MR, Mew TW, Nelson RJ (1995) Tagging and combining bacterial blight resistance genes in rice using RAPD and RFLP markers. Mol Breed 1:375–387

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

Young ND (2000) The genetic architecture of resistance. Curr Opin Plant Biol 3:285–290

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L ssp. *indica*). Science 296:79–92

Zhou F, Kurth J, Wei F, Elliot C, Vale G, Yahiaoui N, Keller B, Somerville S, Wise R, Shulze-Lefert P (2001) Cell-autonomous expression of barley *Mla1* confers race-specific resistance to the powdery mildew fungus via a *Rar1*-independent signaling pathway. Plant Cell 13:337–350