Running head: Genomics of R-genes in Poaceae

Corresponding author:
Hanhui Kuang
Department of Vegetable Crops
College of Horticulture and Forestry
Huazhong Agricultural University
Wuhan, P.R. China, 430070
e-mail: kuangfile@gmail.com
phone: 86-27-87280752

Journal research area: full-length research article;
Research category: Genetics, Genomics, and Molecular Evolution
Dynamic Nucleotide-Binding-Site and Leucine-Rich-Repeat-encoding Genes in the Grass Family

Sha Luo2, Yu Zhang2, Qun Hu, Jiongjiong Chen, Kunpeng Li, Chen Lu, Hui Liu, Wen Wang, Hanhui Kuang*

Key Laboratory of Horticulture Biology, Ministry of Education, and Department of Vegetable Crops, College of Horticulture and Forestry, Huazhong Agricultural University, Wuhan, P.R. China, 430070 (S.L., Y.Z., Q.H., K.L., J.C., H.K.); Kunming Institute of Zoology, Chinese Academy of Sciences (H.L., W.W.)
This work was supported by the “973” National Key Basic Research Program grant No. 2009CB119000, China Transgenic grant No. 2009ZX08009-045B, Natural Science Foundation of China (NSFC) grant No. 30921002, an open project from Key Lab of Plant Germplasm Enhancement and Specialty Agriculture, Chinese Academy of Sciences (to J.C.), and Ministry of Agriculture of China grand No. 2009ZX08009-021B (to W.W.).

These authors contributed equally to the article

*Corresponding author; e-mail kuangfile@gmail.com

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Hanhui Kuang (kuangfile@gmail.com)
Proper use of resistance genes (R-genes) requires a comprehensive understanding of their genomics and evolution. We analyzed genes encoding nucleotide-binding-site and leucine-rich-repeat in the genomes of rice, maize, sorghum and Brachypodium distachyon. Frequent deletions and translocations of R-genes generated prevalent presence/absence polymorphism between different accessions/species. The deletions were caused by unequal crossover, homologous repair, non-homologous repair or other unknown mechanisms. R-gene loci identified from different genomes were mapped onto the chromosomes of rice cultivar Nipponbare using comparative genomics, resulting in an integrated map of 495 R-loci. Sequence analysis of R-genes from the partially sequenced genomes of an African rice cultivar and ten wild accessions suggested that there are many additional R-gene lineages in the AA genome of Oryza. The R-genes with chimeric structures (termed Type I R-genes) are diverse in different rice accessions, but only account for 5.8% of all R-genes in Nipponbare genome. In contrast, the vast majority of R-genes in rice genome are Type II R-genes, which are highly conserved in different accessions. Surprisingly, pseudogene-causing mutations in some Type II lineages are often conserved, indicating that their conservations were not due to their functions. Functional R-genes cloned from rice so far have more Type II R-genes than Type I R-genes, but Type I R-genes are predicted to contribute considerable diversity in wild species. Type I R-genes tend to reduce micro-synteny of their flanking regions significantly more than Type II R-genes, and their flanking regions have slightly but significantly lower G/C content than those of Type II R-genes.
INTRODUCTION

The grass family (Poaceae) provides the main food for human being directly or indirectly. The major threat to the crops in the grass family is diseases caused by various pathogens including fungi, bacteria and viruses. A reliable, economical and environmentally friendly approach for disease control is to use resistant cultivars. As a matter of fact, resistance trait is an important goal in modern crop breeding programs. The proper use of resistance resources for breeding requires a comprehensive understanding of the genetics, genomics, diversity and evolution of genes controlling disease resistance.

Thousands of resistance traits have been identified from different plant species and more than 100 genes controlling disease resistance have been cloned (Liu et al., 2007a). Most of the cloned resistance genes encode nucleotide-binding-site (NBS) and leucine-rich-repeat (LRR) domains. For example, 21 of the 22 cloned functional resistance genes against rice blast and one of the seven against rice bacterial blight encode NBS-LRR domains (Yoshimura et al., 1998; Liu et al., 2010a; Chen et al., 2011; Okuyama et al., 2011; Yuan et al., 2011; Zhai et al., 2011). For convenience, the NBS-LRR-encoding genes are referred to as \( R \)-genes hereafter. The \( R \)-genes belong to a large gene family in plant genome, with more than 400 copies in the genomes of rice, poplar and potato, respectively (Zhou et al., 2004; Kohler et al., 2008; Xu et al., 2011b). The proteins encoded by these \( R \)-genes can be further divided into two classes depending on their N terminal structures: either with a toll interleukin receptor (TIR) domain or a coiled-coil (CC) domain. However, monocotyledonous plants have only CC-NBS-LRR encoding genes (Bai et al., 2002; Meyers et al., 2003). Previous studies identified 464 and 483 CC-NBS-LRR encoding genes (or their partial homologues) from the sequenced genomes of rice cultivars Nipponbare and 93-11, respectively (Yang et al., 2006). Genome-wide comparison between the \( R \)-genes in the two sequenced rice genomes showed that many \( R \)-gene loci are conserved in the two cultivars (Yang et al., 2006). The conservation of \( R \)-gene loci provided foundations of
in silico map-based cloning. For example, the rice blast resistance genes \textit{Pi36}, \textit{Pi37} and \textit{Pik-p} were isolated from different rice genotypes through \textit{in silico} mapping and cloning (Lin et al., 2007; Liu et al., 2007b; Yuan et al., 2011).

Some \textit{R}-genes were present in one genotype but absent in another, exhibiting presence/absence (P/A) polymorphism (Grant et al., 1998; Henk et al., 1999; Kuang et al., 2004; Shen et al., 2006; Yang et al., 2006; Luo et al., 2011). The P/A polymorphism of \textit{R}-genes impedes their \textit{in silico} mapping and cloning. An integrated linkage/physical map of \textit{R}-gene loci from the genomes of different species of a plant family will greatly facilitate their mapping and cloning.

\textit{R}-genes with P/A polymorphism have various frequencies in a population or a species, ranging from a few percent to nearly fixed (Kuang et al., 2004; Shen et al., 2006; Yang et al., 2006). These genes may have highly conserved sequences in genotypes where they are present. Conserved \textit{R}-genes, regardless of their frequencies in a population/species, are usually divergent from and do not have sequence exchanges with their paralogues. These independently evolving and highly conserved \textit{R}-genes were termed Type II \textit{R}-genes (Kuang et al., 2004). The mechanism for the high conservation of Type II \textit{R}-genes remains unclear.

In contrast to the Type II \textit{R}-genes, Type I \textit{R}-genes have frequent sequence exchanges among homologues and consequently show extensive chimeric structures (Kuang et al., 2004; Kuang et al., 2008a). Each lineage of Type I \textit{R}-genes may have a large number of distinct homologues in a population/species, and Type I \textit{R}-genes are postulated to determine its major resistance diversity (Kuang et al., 2006). Due to its awkward complexity, Type I \textit{R}-genes were often excluded from previous studies on population genetics and evolution of \textit{R}-genes. Instead, Type II \textit{R}-genes have been preferentially selected for such studies simply because they were readily PCR amplified and sequenced (Bakker et al., 2006; Yang et al., 2006). Apparently, the exclusion of Type I \textit{R}-genes from such studies may give a biased conclusion on the evolution, diversity and genomics of \textit{R}-genes. To comprehensively understand the genetics, genomics and evolution of \textit{R}-genes, it is critical to investigate which \textit{R}-genes in a genome exhibit the evolutionary patterns of Type I \textit{R}-genes or Type II
It is also very likely that some \textit{R}-genes in a genome have evolutionary pattern unlike that of Type I or Type II \textit{R}-genes. The population genetics of Type I \textit{R}-genes and the mechanism for the differentiation of Type I and Type II \textit{R}-genes are also interesting topics on \textit{R}-genes. With the rapid progress of genome sequencing and re-sequencing, \textit{R}-gene sequences from multiple genomes are or will be available for species such as rice. Comparative analysis of \textit{R}-genes from the genomes of several genotypes or closely related species may shed lights on the evolution of \textit{R}-genes.

In this study, we re-annotated \textit{R}-genes from all five sequenced genomes in the grass family, including two rice cultivars and one genotype each of \textit{Brachypodium distachyon}, maize (\textit{Zea mays}) and sorghum (\textit{Sorghum bicolor}) (Goff et al., 2002; Yu et al., 2002; Paterson et al., 2009; Schnable et al., 2009; Vogel et al., 2010). An integrated \textit{R}-gene map was constructed for the grass family using comparative genomics. \textit{R}-gene sequences were also obtained from a panel of 19 rice cultivars using PCR and from ten wild rice accessions using next generation sequencing. The diversity and evolution of \textit{R}-genes from rice were extensively studied. Methylation, small RNAs, G/C content, density of transposable elements and synteny at \textit{R}-gene loci as well as expression of \textit{R}-genes were analyzed. The mechanism for the differentiation of Type I and Type II \textit{R}-genes was discussed.
RESULTS

Comparative Analysis of R-gene Homologues from Rice Cultivars Nipponbare and 93-11

Several approaches were taken to identify R-genes from the genomes of rice cultivars Nipponbare and 93-11, and all identified R-genes were annotated manually (see MM section). A total of 623 R-genes (including partial ones) were identified from the genome of rice cultivar Nipponbare, and 725 from the partially sequenced genome of rice cultivar 93-11 (Supplemental Table S1). Compared with the 464 and 483 R-genes identified previously (Yang et al., 2006), this study obtained 159 and 242 additional R-genes from the two cultivars, respectively. The main difference between our approaches and previous ones is on partial genes. Previous studies considered a gene as R-gene if it has both NBS and LRR domains. However, many members of the R-gene family are partial ones. In this study, partial genes that are highly homologous to NBS-LRR encoding genes were also considered as R-genes (pseudogenes) and included for further analysis. The 1,348 (623+725) R-genes from the two rice cultivars were empirically classified into 771 lineages so that nucleotide identities between genes from different lineages are lower than 85% with only a few exceptions (see below).

Manual annotation showed that 347 (55.7%) R-genes in Nipponbare and 345 (47.6%) R-genes in 93-11 are pseudogenes (Table 1). More than half of the pseudogenes were caused by large deletions (i.e. partial genes), while the others were caused by nonsense point mutations or small frame-shift insertions/deletions (indels).

The majority (74.3%) of the R-genes in Nipponbare are organized in clusters (i.e. two or more R-genes separated by no more than 8 non-R-genes). For convenience, each R-gene cluster is considered as a multiple-copy R-locus (Richly et al., 2002), and non-clustering R-gene is referred to as single-copy R-locus. The Nipponbare genome has a total of 279 R-loci, including 160 single-copy loci and 119 multiple-copy loci,
with the largest R-locus containing 19 homologues. The partially sequenced 93-11 genome has 323 R-loci as well as 15 homologues that could not be mapped due to lack of enough information on their flanking regions.

**The Presence/absence Polymorphism of R-gene Loci in Rice**

Of the 323 R-gene loci in cultivar 93-11, 84 loci are absent in the Nipponbare genome, exhibiting presence/absence (P/A) polymorphism between the two cultivars. On the other hand, at least 40 R-loci in Nipponbare are absent in 93-11. All together, there are at least 363 R-loci in the two rice genomes and 124 of them exhibit P/A polymorphism. Most of the R-loci with P/A polymorphism are single-copy loci, and only 17 are multiple copy R-loci, i.e. a cluster of R-genes were deleted in one of the two cultivars.

The 84 R-loci present in 93-11 but absent in Nipponbare were further investigated to understand their underlying genetic mechanisms. Sequence comparison between presence haplotype (with an R-gene) and absence haplotype (without an R-gene) could pinpoint the endpoints of deletions at 27 of the 84 R-loci. At nine loci, the two endpoints are homologous to each other and the deletions were apparently generated by unequal crossovers (Fig. 1a). The deletion endpoints at another seven loci were most likely ligated through homologous repair since the endpoints and the region between them are homologous to a continuous sequence elsewhere in the genome (Fig. 1b). The other eleven loci have up to 33.0 kb filler DNA insertion at the deletion haplotype (Fig. 1c), suggesting that the absence haplotypes were generated by double strand DNA break followed by non-homologous DNA end joining (Gorbunova and Levy, 1997). Compared with the presence haplotype, the absence haplotype at above 27 loci also missed 1.0 to 100.4 kb flanking sequences besides the R-gene(s). At the other 57 R-loci with P/A polymorphism, no obvious endpoints of deletions could be determined (Fig. 1d), indicating that complex processes were involved in the deletions or that the haplotypes in the two cultivars had diverged considerably prior to the deletions.
The \textit{R}-genes in five accessions of wild rice \textit{O. rufipogon} and five accessions of \textit{O. nivara}, which were partially (10\texttimes coverage) sequenced using next generation sequencing technology, were identified using homologous search. A total of 1,333 to 5,421 \textit{R}-gene sequences were discovered from the ten genomes. Note that the number of \textit{R}-genes should be much lower because one gene may be divided into several sequences due to sequencing gap. Using 85\% nucleotide identity as threshold (see above), 9 to 62 novel \textit{R}-gene sequences were found in the ten wild accessions. Using similar method, 35 novel \textit{R}-gene lineages were discovered from the partially sequenced genome of African cultivated rice (http://www2.genome.arizona.edu/genomes/rice). We conclude that the \textit{R}-genes identified from this study should cover the vast majority of \textit{R}-gene lineages present in the AA genome of \textit{Oryza} though additional \textit{R}-gene lineages will be discovered with more and more genomes sequenced.

\textbf{An Integrated \textit{R}-locus Map for the Grass Family}

The study of P/A polymorphism was extended to different species in the grass family. First, \textit{R}-genes from the sequenced genomes of \textit{B. distachyon}, sorghum and maize were identified using HMM and BLAST search. A total of 451 (187 \textit{R}-loci), 460 (207 \textit{R}-loci) and 216 (123 \textit{R}-loci) \textit{R}-genes were discovered from these three genomes, respectively. The numbers of \textit{R}-genes are considerably more than those identified previously from \textit{B. distachyon} (239), sorghum (245) and maize (129) (Li et al., 2010a). Nevertheless, the number of \textit{R}-genes in different genomes of the grass family varies dramatically, from 216 copies in maize to 725 in rice cultivar 93-11.

To generate an integrated \textit{R}-locus map for the grass family, all identified \textit{R}-genes in different species were mapped onto the Nipponbare chromosomes based on synteny (Supplemental Table S2). Consequently, a virtual map with 495 \textit{R}-loci was obtained (Fig. 2). The \textit{R}-loci show a remarkable P/A polymorphism between different species. Only 16 \textit{R}-loci were shared among all five genomes, while 222 loci were present in only one genome.
Frequent Translocation of R-genes

To investigate if translocation also played an important role in the observed high presence/absence polymorphism, we first identified allelic/orthologous pairs of R-genes in different genomes. To be alleles/orthologues, the two genes should be reciprocal best hit in BLAST search (Li et al., 2003). To exclude false-positive caused by deletions, alleles/orthologues must have high-scoring segment pairs (HSPs) with greater than 500 bp in length and nucleotide identities higher than 95%, 85%, 80% if they are from different subspecies, genus and subfamily, respectively. Approximately 63.6%, 13.6%, and 12.5% R-genes could find obvious alleles/orthologues from different subspecies, genus and subfamily, respectively. The lack of allelic/orthologous relationships for some R-genes were due to deletions (P/A polymorphism) or sequence exchanges (see below).

The locations of allelic/orthologous pairs were compared to investigate if they have been translocated. At least 49 allelic pairs between Nipponbare and 93-11 are not located at syntenic region, suggesting translocations. For example, gene Os01g20720 is located on chromosome 1 of Nipponbare, but its “allele” BGIOSGA012346 which has 99.7% nucleotide identity with Os01g20720 is located on chromosome 3 of 93-11. Pairwise analysis of R-genes from the five genomes found 9-35% alleles/orthologues have been translocated (Table 2). Therefore, like deletion, frequent translocation is also a major genetic mechanism for the prevalent P/A polymorphism of R-genes.

Poor Synteny around R-gene Loci

To investigate the effects of R-genes on genome stability, the synteny around R-loci was compared with that of non-R-gene loci. First, 20 genes flanking an R-gene were used to identify their syntenic region in another genome. Then the number of allelic/orthologous pairs among the 20 genes of the two genomes was counted. The average number of allelic/orthologous pairs flanking R-genes are 17.2, 8.9, 6.8, 6.5 between Nipponbare and 93-11, B. distachyon, sorghum and maize, respectively. In comparison, the average number of allelic/orthologous pairs flanking all non-R-genes
in the genomes are 18.1, 13.6, 10.9, 10.6, respectively. \( T\)-test shows that the number of orthologous pairs flanking \( R\)-loci is significantly lower than that flanking non-\( R\)-genes \((p<0.0001)\), indicating that \( R\)-loci tend to reduce micro-synteny in their flanking regions. \( T\)-test also shows that the number of orthologous pairs flanking multiple-copy \( R\)-loci is significantly lower than that flanking single-copy \( R\)-loci \((p<0.0001)\).

**\( R\)-gene Expression and Small RNAs Derived from \( R\)-genes or Targeting \( R\)-genes**

The expressions of \( R\)-genes were analyzed using RNA-seq data of the rice leaves (Lu et al., 2010). A total of 222 \( R\)-genes in 93-11 and 383 \( R\)-genes in Nipponbare have more than 0.15 RPKM (reads per kilobase of exon model per million mapped reads; Mortazavi et al., 2008), which corresponds to a false-discovery rate (FDR) and false-negative rate of 8\%. In addition, 290 and 203 \( R\)-genes from 93-11 and Nipponbare, respectively, have RPKM of less than 0.15, whereas the other 213 and 37 \( R\)-genes were not covered by any RNA-seq reads in 93-11 and Nipponbare, respectively. Furthermore, the expressed \( R\)-genes are among the genes with lowest expression, with only 1/9 of the expression of an average gene.

To investigate if the \( R\)-genes are located in low expression regions, the expression of 20 genes flanking \( R\)-loci was compared with that of genes away from \( R\)-loci. \( T\)-test found no difference between the expression of genes flanking \( R\)-loci and genes away from \( R\)-loci. Therefore, \( R\)-genes are not preferentially located in low expression regions in plant genome.

The average number of small RNAs (27.3) derived from one \( R\)-gene is significantly more than that (11.8) of all non-\( R\)-genes in rice genome \((t\)-test, \( p<0.001)\). However, the difference is not significant when gene length is considered. We also investigated the number of small RNAs targeting \( R\)-genes. The number of small RNAs targeting \( R\)-genes is significantly more than those targeting non-\( R\)-genes no matter whether gene length was considered or not. Therefore, the low expression of \( R\)-genes may partially attribute to small RNA regulations besides their weak
promoters.

**A Small Proportion of R-genes Exhibiting Chimeric Structures (i.e. Type I R-genes)**

*R*-gene sequences from each lineage were aligned and sequence exchanges were studied. Eleven lineages (a total of 38 genes) exhibit evolutionary patterns typical of Type I *R*-genes, i.e., with frequent sequence exchanges. Their nucleotide identities between members within a lineage vary from 87% to 99%. To confirm their Type I evolutionary patterns, PCR primers were designed for four randomly chosen lineages, and were used to amplify PCR fragments from four rice cultivars. PCR products were cloned and sequenced. The obtained sequences were combined with corresponding genes from Nipponbare and 93-11. As expected, no obvious allelic relationships were found for these genes. The lack of allelic relationship was due to frequent sequence exchanges between paralogues, confirming their Type I evolutionary patterns (Fig. 3).

Most members of the same Type I lineage are located at the same locus. The only exception is gene *Os12g33160*, which is in the same lineage as genes *Os12g36690*, *Os12g36720* and *Os12g36730* but is located 2.4 Mb away from them.

A total of 420 lineages of genes are highly conserved between rice cultivars Nipponbare and 93-11 and show apparent allelic relationships. Their average nucleotide identity is 99.46%, which is similar to that (99.45%) of 30,417 pairs of non-*R*-genes in the two genomes. Sequence exchanges rarely occurred between genes from different lineages. To test if these genes had evolutionary patterns of Type II *R*-genes, 43 of them were randomly chosen for further study. Of them, 25 are from single copy loci and the other 18 are from multiple copy loci. First, specific primers were designed for each lineage (a potential pair of alleles in this case) and used to amplify PCR products from 19 rice cultivars. The PCR products were sequenced directly. The sequences amplified from the 19 rice cultivars using primers specific to a lineage are highly conserved (98%-100% nucleotide identity) and no sequence exchanges with paralogues were found. Furthermore, they are also highly conserved
in at least one of the ten wild rice accessions that were partially sequenced. We conclude that the vast majority (89.4%) of \( R \)-genes in rice genome had evolutionary patterns of Type II \( R \)-genes.

The remaining 10 lineages do not have typical evolutionary patterns of either Type I or Type II \( R \)-genes characterized previously (Kuang et al., 2004). Similar to Type I \( R \)-genes, genes within a lineage exhibit nucleotide identity of 86-99% and they do not show obvious allelic relationships. However, unlike Type I \( R \)-genes, they are not extensive chimeras though one or two sequence exchanges occurred within a lineage.

**The Small Proportion of Type I \( R \)-genes in Rice Genome Contribute Considerable Diversity**

The Type II \( R \)-genes have very low nucleotide diversity in cultivated rice. The 43 lineages of Type II \( R \)-genes studied above have nucleotide diversity of only 0 to 0.009, with an average of 0.003 in the 21 rice cultivars (including cultivars Nipponbare and 93-11). This number is similar to The main polymorphic sites within each lineage are point mutations. For each Type II lineage, there are very few (1-8) distinct alleles (varying at least one non-synonymous point mutation) in the 21 rice cultivars. As shown in their distance trees, genes within a lineage are highly conserved but genes from different lineages vary considerably (Fig. 4).

Type I \( R \)-genes, though accounted a small proportion of all \( R \)-genes in a genome, were predicted to contribute the major diversity of the \( R \)-gene family in a population/species (Kuang et al., 2006). To test this hypothesis, a lineage of Type I \( R \)-genes was chosen for further analysis. This lineage has three copies in Nipponbare (\( Os11g42580, Os11g42590 \) and \( Os11g42770 \)) and three copies in 93-11. Using PCR primers specific to this lineage, a total of 47 sequences were obtained from the 21 rice cultivars, with 1-4 copies from each cultivar. Of them, 34 genes have at least one non-synonymous point mutation. Their average nucleotide diversity is \( \pi=0.038 \), which is significantly \( (t\text{-test}, p<0.00001) \) higher than that \( (\pi=0.003) \) for each lineage of Type
II $R$-genes. It is worth noting that the comparison between diversity of a lineage of Type I genes and a lineage of Type II genes may have limited meaning in the view of evolutionary analysis since a lineage of Type I genes may have paralogues and frequent gene conversions have occurred between them. The topology of distance tree for these Type I $R$-genes is in striking contrast to that for Type II $R$-genes (Fig. 4).

**Functional $R$-genes can be either Type I or Type II $R$-genes, Generated through Point Mutations or Sequence Exchanges**

To understand how resistance function was generated, the evolutions of all 22 functional $R$-genes cloned from rice were analyzed (Table 3). The homologues (alleles in many cases) of 14 resistance genes ($Pib$, $Pid3$, $Pi9$, $Pi25$, $Pi36$, $Pi37$, $Pikm1-TS$, $Pikm2-TS$, $Pikp-1$, $Pikp-2$, $Pik-1$, $Pik-2$, $Pi-ta$ and $Piz-t$) from multiple genotypes were obtained from previous studies and retrieved from GenBank (Zhou et al., 2007; Huang et al., 2008; Lee et al., 2009a; Shang et al., 2009; Costanzo and Jia, 2010; Dai et al., 2010; Liu et al., 2010b; Luo et al., 2011). For the other eight genes, their homologues were PCR amplified from the 19 rice cultivars and retrieved from the partially sequenced genomes of the ten wild rice accessions. Homologues of each functional $R$-gene were compared to understand their evolution and to infer how their functions were generated.

Of the 22 functional $R$-genes, six genes ($Pi2$, $Pi9$, $Pb1$, $Pib$, $Piz-t$ and $Pi37$) exhibit the evolutionary patterns of Type I $R$-genes. The gene $Pi37$ was generated through a unequal crossover followed by four point mutations (Luo et al., 2011). Sequence exchanges were also detected in genes $Pi2$, $Pib$, $Pb1$, $Piz-t$ and $Pi9$, suggesting that sequence exchanges might have contributed to the functions of these resistance genes.

The other 16 functional $R$-genes have evolutionary patterns of Type II $R$-genes. Genes $Pikm1-TS$, $Pikm2-TS$, $Pik-1$, $Pik-2$, $Pikp-2$, $Pi-ta$, $Pi36$ and $Pit$ are highly similar to their susceptible alleles, only with some short indels and/or a few point mutations. Compared with the susceptible alleles, the functional $R$-genes $Pi5-1$ and
*Pia-1* are diverse in the 5’ region and *Pikp-1* had a 1,237 bp gene conversion in the middle of the gene. Resistance genes *Pi25* and *Pid3*, both encoding resistance against *Magnaporthe oryzae*, are allelic and vary in a single nucleotide (Chen et al., 2011). Therefore, single point mutation may abolish or change the specificity of a resistance gene. Changes in promoter region, such as insertion of transposable elements may also contribute the function of some *R*-genes (Hayashi and Yoshida, 2009; Lee et al., 2009a).

As expected, some functional Type II *R*-genes have null alleles (i.e. absence of the *R*-gene) or pseudogene alleles in susceptible genotypes. For example, genes *Pi5-2* and *Pia-2* do not have allelic sequences in some susceptible cultivars (Lee et al., 2009b; Okuyama et al., 2011). Screening the panel of 19 cultivars showed that 12 (*Pikm1-TS, Pikm2-TS, Pik-1, Pik-2, Pikp-1, Pikp-2, Pi5-1, Pi5-2, Pia-2, Pit, Pi25 and Pid3*) of the 16 functional Type II *R*-genes exhibit P/A polymorphisms, with frequencies of 31.6% to 94.7%. The susceptible allele of *Pid3/Pi25* contains a nonsense point mutation resulting in a truncated protein. This nonsense mutation was found in 29 of the 32 *japonica* cultivars (Shang et al., 2009).

**Some Pseudogenes Are Highly Conserved**

At least 279 potential Type II *R*-genes in rice cultivars Nipponbare are apparent pseudogenes. Of them, 54 were caused by frame-shift indels, and 41 by nonsense point mutations, while others are caused by large deletions (i.e. partial genes). To better understand the evolution of pseudogenes, four lineages with frame-shift mutations and five lineages with nonsense point mutations were randomly chosen for further study. PCR primers spanning the pseudogene-causing mutations were designed and used to amplify sequences from the panel of 19 rice cultivars. Sequences were successfully obtained from 17-19 cultivars for each lineage. Like the 43 lineages of Type II *R*-genes analyzed above, the nine lineages of Type II pseudogenes are also highly conserved, with an average of pairwise nucleotide identity of 99.5%. The nucleotide identity of pseudogenes is slightly lower than that (99.7%) of Type II
$R$-genes with ORF, but considerably higher than that (96.2%) of Type I $R$-genes. Strikingly, the frame-shift indels and nonsense mutations were also present in most accessions (Fig. 5). All of the pseudogene-causing mutations were found in at least two of the ten partially sequenced wild rice accessions. The conservation of pseudogene-causing mutations in Type II $R$-genes indicate that the lack of sequence exchanges and high conservation of Type II $R$-genes were not due to their critical functions.

**Genome-wide Comparison of Type I and Type II $R$-gene Loci in Rice**

To investigate the mechanism underlying the differentiation of Type I and Type II $R$-genes, the 38 Type I $R$-genes were compared with the 557 Type II $R$-genes in the genome of Nipponbare. No differences were found on density of repetitive sequences, number of small RNAs, gene expression or methylation level between the flanking regions of Type I and Type II $R$-loci, indicating that these factors should not contribute to the differentiation of Type I and Type II $R$-genes. However, the G/C content at the Type II $R$-loci is slightly (1.2%) but significantly higher than that at the Type I $R$-loci ($t$-test, $p<0.05$). The synteny around Type I loci is significantly lower than that around Type II loci ($t$-test, $p<0.01$).
DISCUSSION

An Integrated R-gene Map for the Grass Family

With genetic maps and genome sequences available, in silico mapping and cloning are becoming an efficient approach for genetic analysis of many traits. To use this approach in studies of disease resistance also requires a comprehensive understanding of the genomics and evolution of R-genes. In this study, comparative genomics was applied to R-genes in rice, B. distachyon, maize and sorghum, and an integrated map with 495 R-gene loci was constructed for the grass family. The integrated map shows prevalent presence/absence polymorphisms between different species. Only 16 loci are present in all five genomes included in this study and 222 of the 495 R-loci are found in only one genome. In addition, 546 R-genes from the other four genomes could not be mapped onto the Nipponbare chromosomes because their flanking regions do not have synteny with any part of the Nipponbare genome. Such high presence/absence polymorphism between different species suggest that many additional R-loci are likely to be discovered with more genomes sequenced in future, and the R-gene map will be consequently enriched. Indeed, many novel R-gene fragments were found in the partially sequenced genomes of the African cultivated rice and the 10 wild accessions. Nevertheless, the integrated R-gene map constructed in this study can serve as a useful reference for future in silico mapping and cloning of R-genes in Poaceae species.

Frequent Deletions and Translocations of R-genes

One of the mechanisms for presence/absence polymorphism of R-genes is deletions. A small proportion of such deletions in rice genome were resulted from unequal crossovers between homologous sequences flanking an R-gene locus. Such unequal crossover removed all sequences (including the R-genes) between the two homologous sequences. Such unequal crossovers should also be responsible for the
tandem duplications of R-genes. Comparison between presence haplotypes and absence haplotypes suggested that some deletions might have occurred during repairs of double-strand breaks. We found evidence that some deletions of R-genes were caused by homologous repairs or gene conversions, while other deletions might have been resulted from non-homologous end joining after double-strand breaks (DSBs). Comparison between the presence and absence haplotypes showed that long stretch of DNA sequences including the R-gene(s) was deleted. The genetic or evolutionary mechanisms for the R-genes to be prone to deletions remain unclear.

Another mechanism for high P/A polymorphism of R-genes is frequent translocations. To identify translocations of genes requires verification of their allelic/orthologous relationships in two genomes. Several programs have been developed to identify orthologues from two or more species (Remm et al., 2001; Lee et al., 2002; Li et al., 2003). Orthologous relationship was assumed if two genes are reciprocal best hits in BLAST search. However, missing true orthologues due to deletions or incomplete genome were not well considered and a large proportion of the detected orthologues for R-genes might be false positive. Using above method, the detected orthologous R-gene pairs in A. thaliana and A. lyrata vary dramatically in synonymous divergence (dS: 0.08-0.74) (Guo et al., 2011). The extreme divergence (such as dS=0.74) unlikely occurred between true orthologues that separated only 10 MYA (Ossowski et al., 2010). We used a conservative method to determine orthologous relationship. Besides two genes from two species being reciprocal best hits, they should also have more than 80-95% nucleotide identity, depending on the phylogenetic relationship of the two species. Using above criteria, only 20.4% of genes showed orthologous relationship between rice cultivar Nipponbare and B. distachyon. The lack of orthologous relationship for many genes was mainly due to frequent deletions.

Comparative analysis showed that some R-gene orthologues are located at different loci of two species, indicating that translocation may have played important roles in rearrangement of R-gene loci in plant genomes. Unlike deletions which may result in a missing lineage of R-genes in a genome, translocations do not lose any
lineage but generate P/A polymorphism at two loci (the original locus and the destination locus). The extremely dynamic \( R \)-genes is inconsistent with the high colinearity between the genomes of different Poaceae species (Gale and Devos, 1998). The mechanism for such unusual rearrangement of \( R \)-genes remains unknown.

**Poor Synteny around \( R \)-genes**

Besides deletions and translocations, \( R \)-gene loci also had unusually frequent internal and adjacent chromosomal re-arrangements (Leister et al., 1998; Ramakrishna et al., 2002). Genome-wide analysis in this study showed that \( R \)-genes affect the synteny of their internal and adjacent regions. Unequal crossover between two homologous \( R \)-genes in a cluster inevitably reduces the synteny of \( R \)-genes’ internal region (Sun et al., 2001; Luo et al., 2011). As a matter of fact, \( R \)-gene clusters usually have poor internal synteny (Parker et al., 1997; Kuang et al., 2004; Kuang et al., 2005). In this study, we showed that \( R \)-genes significantly reduced the synteny of their flanking regions as well. How \( R \)-genes affect the synteny of their flanking regions remains unclear.

**Diversity Contributed by Type I \( R \)-genes**

Unequal crossovers between two homologous \( R \)-genes generate chimeras that have a bipartite structure. Their 5’ part is identical (or highly similar) to one gene while its 3’ part is identical (or highly similar) to another gene. In contrast, gene conversions generate chimeras in which a stretch of sequences from the middle were replaced by another gene. Most chimeras in rice genome do not show bipartite structure and we hypothesize that gene conversions rather than unequal crossovers mainly resulted in the chimeric structure of Type I \( R \)-genes.

Only 5.8% of \( R \)-genes in rice genome exhibit the evolutionary patterns typical of Type I \( R \)-genes. Type I \( R \)-genes, though in small proportion of all \( R \)-genes, are predicted to contribute considerable diversity of a species (Kuang et al., 2006; Kuang et al., 2008b). This hypothesis is at least partially supported by the variation of a Type
I lineage in the 21 rice cultivars included in this study. The panel of 21 rice cultivars contain at least 34 distinct Type I \(R\)-genes, which vary considerably compared with Type II \(R\)-genes. We predict that the variations of Type I \(R\)-genes in wild accessions would be more profound, and the proportion of Type I \(R\)-genes in all cloned functional resistance genes is going to increase steadily.

Type II \(R\)-genes within a lineage are nearly identical with only a few point mutations. Analysis of three lineages of Type II \(R\)-genes showed that each lineage encodes only 1-8 distinct proteins in the 21 rice cultivars, and they vary only one or a few amino acids. It is worth noting that a minor change of the Type II \(R\)-genes might generate a new specificity (Chen et al., 2011). In addition, the majority (89.4%) of \(R\)-genes in the rice genome are Type II \(R\)-genes. Therefore, Type II \(R\)-genes can still encode a large number of functional \(R\)-genes. As a matter of fact, the majority of cloned functional \(R\)-genes in rice are Type II \(R\)-genes. However, cultivated species and its close wild relatives should have identical or nearly identical Type II sequences, and the relative contribution of Type II \(R\)-genes might not be so significant as more and more accessions are included for conservation or analysis.

The Mechanism Underlying the Differentiation of Type I and Type II \(R\)-genes

The principal differences between Type I and Type II \(R\)-genes are nucleotide identity and frequency of sequence exchanges. Type II \(R\)-genes are highly conserved and do not exchange sequences with other paralogues. Gene functions were ruled out as the causes of high conservation of Type II \(R\)-genes (Kuang et al., 2004). Consistent with this hypothesis, we found that most pseudogene-causing mutations in Type II \(R\)-genes are highly conserved in different accessions and wild species.

Instead, genomic structures around the \(R\)-genes were hypothesized to play important roles on the differentiation of Type I and Type II \(R\)-genes (Luo et al., 2011). However, studies of the \textit{RPP8} family in \textit{A. thaliana} showed that genomic proximity is not the major cause for the differentiation of Type I and Type II \(R\)-genes. \textit{RPP8} homologues separated by 2.2 Mb had frequent sequence exchanges while homologues
only 4 kb apart rarely had sequence exchanges (Kuang et al., 2008a). Nevertheless, genome-wide analysis showed that cluster size is significantly correlated with the frequency of sequence exchanges in *A. thaliana* (Guo et al., 2011). Genes closely linked and with high sequence similarity are more likely to have gene conversions (Mondragon-Palomino and Gaut, 2005). Consistent with above observations, all but one Type I *R*-gene in rice genome is from a gene cluster.

Comparison of Type I and Type II *R*-loci in rice found no significant difference on density of repetitive sequences, number of small RNAs, gene expression or methylation levels. However, Type I *R*-loci showed significantly low synteny than Type II *R*-loci. The low synteny in Type I *R*-loci should not be the causes of frequent sequence exchanges between Type I genes, but the consequences of frequent shuffling between *R*-genes. Interestingly, the G/C content at the Type II *R*-loci is slightly (1.2%) but significantly higher than that at the Type I *R*-loci (*t*-test, *p*<0.05). It remains unclear if and how G/C content affects gene conversions or unequal crossovers.
MATERIALS AND METHODS

Plant Materials

Ten wild rice accessions and 19 rice cultivars were chosen to investigate the evolution and diversity of \( R \)-genes in rice. The ten wild rice include five accessions (P46, 105958, 105960, Nepal and YJ) of \( O. \) \textit{rufipogon} and five accessions (80470, 89215, 105327, 106105 and 106154) of \( O. \) \textit{nivara}. The panel of 19 rice cultivars (Wukezhan, Teqing, Zhonghan18, Yanshuichi, H94, Hemichan, Nanjing11, Minhui63, Zhenshan97, KDML105, IRAT109, Laohuzhong, Shennong606, wuyugeng-3, Qiuguang, Zidao, Muxiqiu, Zhonghua11, 58N, SLG) were randomly chosen from the core collections of rice germplasms to represent both subspecies \textit{indica} and \textit{japonica}.

Genome Sequences

The genome sequence and gene model of rice cultivar 93-11, GLEAN genes, were downloaded from http://rice.genomics.org.cn/rice/link/download.jsp (Yu et al., 2002). The genome sequence and gene model of rice cultivar Nipponbare, MSU RGAP v6.1, were downloaded from http://rice.planthbiology.msu.edu/ (Goff et al., 2002). The genome sequence and gene model of \( B. \) \textit{distachyon}, Brachypodium Assembly v1.0 and Annotation v1.2, were downloaded from http://www.brachypodium.org/ (Vogel et al., 2010). The genome sequence and gene model of sorghum, Sorghum Assembly v1.0 and annotation Sbi1_4, were downloaded from http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html (Paterson et al., 2009). The genome sequence and gene model of maize, MaizeSequence 5b.60, were downloaded from http://www.maizesequence.org/index.html (Schnable et al., 2009). The genome sequence of Africa rice was downloaded from http://www2.genome.arizona.edu/genomes/rice.

The 10x coverage genome sequences of five accessions of rice \( O. \) \textit{rufipogon} and five accessions of \( O. \) \textit{nivara} were obtained using next generation sequencing.
technology. And those reads were assembled using SOAPdenovo (default parameters) (Li et al., 2010b), obtaining 10 scaffolds with an average N50 of 1.6 kb and an average size of 268M (Xu et al., 2011a)

**Identification and Annotation of R GENES**

HMM and BLAST search were used to identify R-genes in the genomes of rice, *B. distachyon*, sorghum and maize. First, protein sequences in each genome were searched against the model of NB-ARC domain (Pfam PF00931) using hmmer3.0 (http://hmmer.org/) with default parameters. To validate the result of HMM search, a validating database was constructed, which contains 5,158 protein sequences retrieved from NCBI (http://www.ncbi.nlm.nih.gov) by using key word “NBS LRR, NB ARC”, as well as 3,110 proteins using key word “ATP binding cassette” and 6,979 proteins using key word “LRR kinase”. Potential R protein sequences obtained using HMM search were used as query to BLAST the validating database. Only sequences with best hit of “NBS LRR, NB ARC” protein were considered as seed R proteins sequences (E-value cutoff 1e-10). To identify homologues (such as diverse or partial ones) missed in above step, the obtained seed sequences were used as query to tBLASTn the entire genome. Again all significant hit (E-value cutoff 1e-10) were validated using the same method as above. The seed sequences and their homologous sequences were all regarded as candidate R proteins for further study.

Because many R-genes are partial or have frameshift/non-sense point mutations, automatic annotation are often problematic (Meyers et al., 2003). The gene model of all candidate R-genes identified from rice cultivars Nipponbare and 93-11 were verified manually. First, each candidate R-gene sequence were used to BLASTX the non-redundant protein database in GenBank. One or more best hits must be a NBS-LRR protein. From the best hits, one well-characterized functional R-gene was chosen as reference for gene model annotation. If no hit on functional R-genes was found for a candidate R-gene, then two or more best hit genes, which have identical gene model, were used as references for annotation. Genes closely related with their
reference genes should have the same gene model (intron/exon structure). If a stop
codon or frame-shift mutation occurs in the coding region, the gene is considered as
pseudogene. If more than 500 coding sequence is missing in any part of the gene, it is
considered as partial one. If sequences containing the start codon are deleted
(regardless of the length of the deletion), the gene is also considered as partial.

The coding sequences of R-genes identified from Nipponbare and 93-11 were
used as query to BLASTX search of the partially sequenced genomes of the ten wild
rice accessions. All HSPs as well as their 3 kb flanking sequences were retrieved and
a database of R-gene sequences was constructed for further analysis.

**Presence/Absence Polymorphism**

Presence/absence (P/A) polymorphism is assumed if an R-gene is present at one
locus in a genome but absent in its syntenic region in another genome. If two R-genes
are separated by no more than eight genes, they are considered as an R-gene cluster
(Richly et al., 2002). An R-gene cluster is also called a multiple-copy R-locus in this
manuscript. An absent R-locus refers to deletion of all R-genes at the locus. To
investigate the genetic mechanism underlying P/A polymorphism, endpoints of
deletions were determined by aligning the sequences of presence haplotype with those
of absence haplotype. If the two endpoints are homologous, the deletion is considered
to have been generated by unequal crossover. If the endpoints and the filler DNA in
the absence haplotype are homologous to a continuous sequence elsewhere in the
genome, it is considered as deletion followed by homologous repair. If the endpoints
and the filler DNA in the absence haplotype has no continuously homologous
sequences elsewhere in the genome, the deletion haplotype is assumed to be caused
by double strand breaks followed by non-homologous DNA end joining (Gorbunova
and Levy, 1997).

**Synteny and Orthologues**

The synteny between Nipponbare and the other four genomes was identified
using a two-step analysis as described previously (Wicker et al., 2010). The first step used gene models of a species to BLASTN the gene models of Nipponbare and identify the best hit (e-value cutoff 1e-10) on its syntenic chromosome for each query gene sequence (Soderlund et al., 2006; Abrouk et al., 2010; Vogel et al., 2010). The best hit pairs are considered as syntenic alleles/orthologues if at least 3 of the 20 genes (10 on each side) flanking the gene of interest in one genome have best hit with genes located in 40 genes (20 on each side) flanking the gene of interest in Nipponbare. The number of best hit pairs among the 20 genes flanking a pair of alleles/orthologues in two genomes is considered as the synteny value of the allelic/orthologous pair. For example, gene X in genome I is an syntenic orthologue of gene Y in genome II, and 10 of the 20 genes flanking gene X have best hits with genes located near (within 20 genes) gene Y in genome II, the synteny value of gene pair X/Y is 10. Synteny values were calculated for all R-genes and non-R-genes, and their difference was compared using t-test.

The R-allelic/orthologous pairs were identified using bi-directional BLASTN method to analyze the translocation of R-genes. To be alleles/orthologues, the two genes must be reciprocal best hit in BLAST search (Li et al., 2003). To exclude false-positive for R-genes due to frequent deletions, allelic/orthologous R-genes must have HSP length of greater than 500 bp and weighted average of HSP nucleotide identity higher than 95%, 85%, 80% if they are from different subspecies, genus and subfamily, respectively. The threshold for nucleotide identity was empirically determined by plotting nucleotide identities of all best hit pairs from two genomes.

**Integrated R-gene Map**

An integrated R-gene map was constructed using Nipponbare chromosomes as template. All R-gene loci identified from other genomes were mapped into their syntenic regions on Nipponbare chromosomes. Syntenic region is defined as the region with the highest synteny value (see above). If the highest synteny value for a R-gene is lower than 3, this gene is not mapped onto the Nipponbare chromosomes.
In the integrated map, two neighboring loci were merged into one if they are separated by no more than eight genes (Richly et al., 2002). Finally, a virtual integrated mapped was draw using perl GD module.

**Analysis of Small RNAs, Methylation and Gene Expression of R-genes**

Analysis of small RNAs and expression of R-genes was similar to Lu et al (Lu et al., 2012). Paired-end RNA-Seq datasets for rice cultivars Nipponbare and 93-11 were downloaded from the EMBL Sequence Read Archive (SRA) under accession number ERA000212 (Lu et al., 2010).

The RNA-seq reads were mapped to all R-gene sequences identified in this study. The RNA sequence read numbers for each R-gene were normalized using RPKM (reads per kilobase of exon model per million mapped reads) (Mortazavi et al., 2008), and background was calculated as described previously (Ramsköld et al., 2009). The normalized data for R-genes were compared with those for non-R-genes using $t$-test.

The RNA sequences were aligned to all R-gene sequences identified in this study. The RNA sequencing read numbers for each R-gene were first normalized using RPKM (reads per kilobase of exon model per million mapped reads), and the normalized data for R-genes were compared with those of non-R-genes using $t$-test.

The 781,886 rice small RNAs were downloaded from http://csrdb.ucdavis.edu/smrnas/ and miRbase. All R-gene sequences identified from rice genome were BLASTed with the small RNA sequences. If a small RNA has perfect match with an R-gene sequence, it is considered to be R-gene derived. If a small RNA has 1-4 nucleotide mismatches with an R-gene, it is believed to target the R-gene. The number of small RNAs for a gene is normalized with its length. $T$-test was performed to compare the difference between the number of small RNAs for R-genes and for non-R-genes.

The datasets of methylation were download form the National Institutes of Health Gene Expression Omnibus database under accession number GSE19602 (He et al., 2010). The reads were mapped to the genome of Nipponbare using SOAP v2.21 (Li et
al., 2009) and then were further aligned to the $R$-genes and other gene-models annotated by MSU RGAP v6.1 using R package girafe (Toedling et al., 2010). The methylation level was estimated by the number of reads located in a gene and normalized it with its length (He et al., 2010).

**PCR Amplification and Sequencing of $R$-genes**

Candidate Type I $R$-gene lineages in rice were randomly chosen to verify their evolutionary patterns. Three $R$-gene lineages were investigated in four rice cultivars (Wukezhan, Yanshuichi, IRAT109 and Laohuzhong), while one lineage was studied in the panel of 19 rice cultivars. First, PCR primers were designed for each lineage (Supplemental Table S3). PCR products were amplified using EX Taq (Takara, Dalian, China), with a reaction of 50 µl, 5 min at 94°C, followed by 32 cycles of 94°C for 30s, 55°C for 2 min, and 72°C for 10 min. The ~1 Kb PCR products were cloned into pEASY-T1 Vector (Transgen, Beijing, China), and individual clones were sequenced.

Forty-three lineages of candidate Type II $R$-genes in rice were randomly chosen to study their variation in the panel of the 19 rice cultivars. Specific primers were designed in the LRR region for each lineage to amplify approximately 1.5 kb fragments (Supplemental Table S3). PCR products were treated with Exo-SAP (Fermentas) before being sequenced directly.

To investigate the diversity of pseudogenes, PCR primers flanking the pseudogene-causing mutations were designed for ten lineages of pseudogenes. PCR products of approximately 500 bp were amplified from the panel of 19 rice cultivars. They were sequenced using the same method as for Type II $R$-genes (see above).

Homologues of 13 functional $R$-genes were obtained from previous studies and were retrieved from GenBank (Zhou et al., 2007; Huang et al., 2008; Yang et al., 2008; Lee et al., 2009a; Shang et al., 2009; Yoshida and Miyashita, 2009; Costanzo and Jia, 2010; Dai et al., 2010; Liu et al., 2010b; Luo et al., 2011). Homologues of six functional $R$-genes (Pi5-1, Pi5-2, Pia-1, Pia-2, Pii) were amplified from the 19 rice cultivars using the same method as described for Type II $R$-genes (Supplemental Table
Homologues of resistance gene Pb1 were amplified from four rice cultivars using the same method as described for Type I R-genes (Supplemental Table S3).

**Sequence Analysis**

Sequences were aligned using ClustalX (Thompson et al., 1997) and manually modified in GeneDoc (http://www.nrbsc.org/gfx/genedoc/). All R-genes from rice cultivars Nipponbare and 93-11 were combined and subfamilies (lineages) were empirically divided using nucleotide identity 85%. Neighbor-joining distance trees were constructed and bootstrap values were calculated using MEGA 4.0 (Tamura et al., 2007). Nucleotide identity between two sequences was calculated using a perl script. The GC contents of the genes were analyzed by a perl script. Sequence exchanges were detected by Geneconv and visual inspection (Sawyer, 1989). One hundred kb sequences flanking R-genes (50 kb each side) were used to analyze the frequency of repetitive sequences around R-loci.

Sequence data from this article have been deposited in the GenBank under the following accession numbers: JQ656330-JQ657221.

**Supplemental Data**

The following materials are available in the online version of this article

**Supplemental Table S1.** Annotation of the R-genes in nipponbare and 93-11.

**Supplemental Table S2.** R-genes in the integrated map.

**Supplemental Table S3.** Oligonucleotide primers used in this study.
LITERATURE CITED


understanding of the molecular mechanisms of the rice-\textit{Magnaporthe oryzae} interaction. Mol Plant Pathol \textbf{11}: 419-427


bacterial inoculation. Proc Natl Acad Sci USA 95: 1663-1668


Figure Legends

Figure 1. Comparison of the presence and absence haplotypes of R-gene. The black boxes represent R-genes, the grey boxes represent homologous sequences that are connected by shaded parallelogram between two haplotypes, and empty boxes represent unrelated sequences between two haplotypes. A, The deletion of R-gene in the absence haplotype was caused by unequal crossovers between two transposons (MuDR-5) resulting in a chimeric transposon. B, The absence haplotype was generated by a deletion followed by homologous repair using a homologous sequence from another chromosome as template; C, The deletion endpoints were likely ligated through non-homologous repair; D, Poor synteny between the presence and absence haplotypes.

Figure 2. An integrated map of R-loci in Poaceae. R-genes from different genomes were mapped onto the 12 chromosomes of rice cultivar Nipponbare. Prevalent P/A polymorphism between different genomes is evident.

Figure 3. Chimeric structure of Type I R-genes. Sequence exchange tracts are shaded. Only part of each exchange tract is shown. Consensus sequence is on the bottom. Dots represent nucleotides identical to consensus sequence. Numbers on top show nucleotide positions in gene BGIOSGA021935.

Figure 4. Contrasting topology of NJ distance trees for Type I and Type II R-genes. The numbers on nodes are bootstrap values, and values lower than 60 are not
shown. A, NJ distance trees for three Type I R-genes from Nipponbare and their close homologues from other cultivars. B, NJ distance tree for three Type II R-genes from Nipponbare and their close homologues from other cultivars. Genes from Nipponbare are marked with a filled cycle.

**Figure 5.** Nonsense mutations conserved in a Type II lineage. The left column are names of the cultivars and wild accessions from which the sequences were obtained. The numbers on top show nucleotide positions in gene *Os06g33360*. Resistance gene *Pi36* is included for comparison and its amino acid sequence is shown in the bottom. Nineteen cultivars with sequences identical to *Os06g33360/BG IOSGA023023* were not shown.
Table I. Summary of R-genes in rice cultivars Nipponbare and 93-11

<table>
<thead>
<tr>
<th>Cultivar</th>
<th>Full-length genes</th>
<th>Pseudogenes</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Partial</td>
<td>Frameshift/ nonsense point mutations</td>
<td>Un-determined</td>
</tr>
<tr>
<td>Nipponbare</td>
<td>276</td>
<td>209</td>
<td>138</td>
<td>-</td>
</tr>
<tr>
<td>93-11</td>
<td>223</td>
<td>199</td>
<td>146</td>
<td>157(^a)</td>
</tr>
</tbody>
</table>

\(^a\) Not fully covered by the genome assembly
Table II. Translocated R-genes among the five genomes included in this study

<table>
<thead>
<tr>
<th>Genome pair</th>
<th>Total number of R-genes</th>
<th>Number of orthologous pair</th>
<th>Number of translocated orthologues</th>
<th>Number of syntenic orthologues</th>
<th>Not determined</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-M</td>
<td>451-216</td>
<td>30</td>
<td>6</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>B-S</td>
<td>451-460</td>
<td>58</td>
<td>16</td>
<td>27</td>
<td>15</td>
</tr>
<tr>
<td>I-B</td>
<td>725-451</td>
<td>82</td>
<td>24</td>
<td>36</td>
<td>22</td>
</tr>
<tr>
<td>I-M</td>
<td>725-216</td>
<td>37</td>
<td>11</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>I-S</td>
<td>725-460</td>
<td>85</td>
<td>20</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>J-B</td>
<td>623-451</td>
<td>88</td>
<td>31</td>
<td>38</td>
<td>19</td>
</tr>
<tr>
<td>J-I</td>
<td>623-725</td>
<td>432</td>
<td>49</td>
<td>349</td>
<td>34</td>
</tr>
<tr>
<td>J-M</td>
<td>623-216</td>
<td>40</td>
<td>12</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>J-S</td>
<td>623-460</td>
<td>88</td>
<td>21</td>
<td>40</td>
<td>27</td>
</tr>
<tr>
<td>S-M</td>
<td>460-216</td>
<td>46</td>
<td>4</td>
<td>24</td>
<td>18</td>
</tr>
</tbody>
</table>

*a* B: *B. distachyon*; M: maize, S: sorghum, I: *indica*, J: *japonica*. *b* Their relative positions could not be determined due to poor synteny in flanking sequence. *c* Total of R-genes in the two genomes in comparison.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Pathogen</th>
<th>Type I/Type II</th>
<th>P/A polymorphism</th>
<th>Comparison between resistant and susceptible alleles</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pi5-1</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Diverse in 5' region</td>
<td>Lee et al., 2009</td>
</tr>
<tr>
<td>Pi5-2</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>P/A polymorphism</td>
<td>Lee et al., 2009</td>
</tr>
<tr>
<td>Pikm1-TS</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Point mutations</td>
<td>Ashikawa et al., 2008</td>
</tr>
<tr>
<td>Pikm2-TS</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Point mutations</td>
<td>Ashikawa et al., 2008</td>
</tr>
<tr>
<td>Pikp-1</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Gene conversion</td>
<td>Yuan et al., 2011</td>
</tr>
<tr>
<td>Pikp-2</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Point mutations</td>
<td>Yuan et al., 2011</td>
</tr>
<tr>
<td>Pik-1</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Point mutations</td>
<td>Zhai et al., 2011</td>
</tr>
<tr>
<td>Pik-2</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Point mutations</td>
<td>Zhai et al., 2011</td>
</tr>
<tr>
<td>Pi25</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Nonsense point mutation</td>
<td>Chen et al., 2011</td>
</tr>
<tr>
<td>Pid3</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>Nonsense point mutation</td>
<td>Shang et al., 2009</td>
</tr>
<tr>
<td>Pia-1</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>-</td>
<td>Diverse in 5' region</td>
<td>Okuyama et al., 2011</td>
</tr>
<tr>
<td>Pia-2</td>
<td><em>M. oryzae</em></td>
<td>Type II</td>
<td>+</td>
<td>P/A polymorphism</td>
<td>Okuyama et al., 2011</td>
</tr>
<tr>
<td><em>Xa1</em></td>
<td><em>Xanthomonas oryzae</em></td>
<td>Type II</td>
<td>Unkown</td>
<td>Unkown</td>
<td>Yoshimura et al., 1998</td>
</tr>
<tr>
<td>Pi-ta</td>
<td><em>M. grisea</em></td>
<td>Type II</td>
<td>-</td>
<td>Point mutations</td>
<td>Bryan et al., 2000</td>
</tr>
<tr>
<td>Pit</td>
<td><em>M. grisea</em></td>
<td>Type II</td>
<td>+</td>
<td>Point mutations and an insertion of retrotransposon in upstream</td>
<td>Hayashi et al., 2009</td>
</tr>
<tr>
<td>Pi36</td>
<td><em>M. grisea</em></td>
<td>Type II</td>
<td>-</td>
<td>Point mutations</td>
<td>Liu et al., 2007</td>
</tr>
<tr>
<td>Pi2</td>
<td><em>M. grisea</em></td>
<td>Type I</td>
<td>-</td>
<td>Sequence exchanges</td>
<td>Zhou et al., 2006</td>
</tr>
<tr>
<td>Pi9</td>
<td><em>M. grisea</em></td>
<td>Type I</td>
<td>N/A</td>
<td>Sequence exchanges</td>
<td>Qu et al., 2006</td>
</tr>
<tr>
<td>Pi-zt</td>
<td><em>M. grisea</em></td>
<td>Type I</td>
<td>N/A</td>
<td>Sequence exchanges</td>
<td>Zhou et al., 2006</td>
</tr>
<tr>
<td>Pi-b</td>
<td><em>M. grisea</em></td>
<td>Type I</td>
<td>N/A</td>
<td>Sequence exchanges</td>
<td>Wang et al., 1999</td>
</tr>
<tr>
<td>Pi37</td>
<td><em>M. grisea</em></td>
<td>Type I</td>
<td>N/A</td>
<td>Point mutations</td>
<td>Lin et al., 2007</td>
</tr>
<tr>
<td>Pb1</td>
<td><em>M. grisea</em></td>
<td>Type I</td>
<td>N/A</td>
<td>Sequence exchanges</td>
<td>Hayash et al., 2010</td>
</tr>
</tbody>
</table>

*Presence/absence polymorphism is not applicable for Type I R-genes.*
Position 1396 1431
Os06g33360 TATGAGGTGTGAGAGGAGTTACTTCAAACGAGCTTGTA
BG10SGA023023 TATGAGGTGTGAGAGGAGTTACTTCAAACGAGCTTGTA
O. nivara_106154 TATGAGGTGTGAGAGGAGTTACTTCAAACGAGCTTGTA
O. nivara_106105 TATGAGGTGTGAGAGGAGTTACTTCAAACGAGCTTGTA
O. nivara_89215 TATGAGGTGTGAGAGGAGTTACTTCAAACGAGCTTGTA
O. rufipogon_Nepal TATGAGCTTGGGAGAGTTACTTCAAATGAGCTTGTA
Pi36 Y E L G E S Y F N E L I