Comparative Sequence Analysis of the Sorghum Rph Region and the Maize Rp1 Resistance Gene Complex

Wusirika Ramakrishna, John Emberton, Phillip SanMiguel, Matthew Ogden, Victor Llaca, Joachim Messing, and Jeffrey L. Bennetzen*

Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907 (W.R., J.E., M.O., J.L.B.); Purdue University Genomics Core, Purdue University, West Lafayette, Indiana 47907 (P.S.M.); and Waksman Institute, Rutgers University, Piscataway, New Jersey 08854 (V.L., J.M.)

A 268-kb chromosomal segment containing sorghum (Sorghum bicolor) genes that are orthologous to the maize (Zea mays) Rp1 disease resistance (R) gene complex was sequenced. A region of approximately 27 kb in sorghum was found to contain five Rp1 homologs, but most have structures indicating that they are not functional. In contrast, maize inbred B73 has 15 Rp1 homologs in two nearby clusters of 250 and 300 kb. As at maize Rp1, the cluster of R gene homologs is interrupted by the presence of several genes that appear to have no resistance role, but these genes were different from the ones found within the maize Rp1 complex. More than 200 kb of DNA downstream from the sorghum Rp1-orthologous R gene cluster was sequenced and found to contain many duplicated and/or truncated genes. None of the duplications currently exist as simple tandem events, suggesting that numerous rearrangements were required to generate the current genomic structure. Four truncated genes were observed, including one gene that appears to have both 5’ and 3’ deletions. The maize Rp1 region is also unusually enriched in truncated genes. Hence, the orthologous maize and sorghum regions share numerous structural features, but all involve events that occurred independently in each species. The data suggest that complex R gene clusters are unusually prone to frequent internal and adjacent chromosomal rearrangements of several types.

Disease resistance (R) genes in plants provide a major mode of defense against a wide variety of pathogens and pests. The most abundant class of R genes encodes proteins with a nucleotide-binding site (NBS) and a Leu-rich repeat (LRR) region. The genome of the model plant, Arabidopsis, contains more than 120 NBS-LRR genes (Arabidopsis Genome Initiative, 2000), but the possible R specificities of these candidate R genes are only known for a handful of loci. From the few NBS-LRR genes that have received functional analysis, it appears that the LRR region provides the specificity for recognition of a pathogen gene product (Ellis et al., 2000a), thereby leading to the initiation of a signal transduction cascade that activates several defense pathways.

Most of the R gene loci in plants are highly complex, containing numerous nearby R genes (Ellis et al., 2000b). Unequal recombination within an R gene cluster can result in variation in the size and organization of a complex locus. New R genes with novel pathogen-recognition specificities can be generated by this unequal recombination process (Richter et al., 1995). Some R genes appear to be the obvious products of unequal crossing over and/or conversion between R genes in the same tandem array, thereby providing a new “chimeric” locus with a possible new recognition specificity. Hence, complex R gene loci can undergo rapid internal reorganization. In addition, R gene loci often show a lack of synteny in closely related species (Leister et al., 1998), suggesting that they are also unstable in their gross chromosomal locations. For instance, neither the Lr1 R gene of wheat (Triticum aestivum) nor the Rpg1 rust resistance gene of barley (Hordeum vulgare) appears to have an orthologous locus in rice (Oryza sativa; Kilian et al., 1997; Gallego et al., 1998; Brueggeman et al., 2002). In contrast, most of the other types of genes in rice, wheat, and barley are found in colinear order on syntenic chromosomes (Gale and Devos, 1998; Bennetzen, 2000).

Because pathogens can easily become virulent by loss of features that allow their recognition, it is expected that R genes must evolve rapidly to keep pace with evolving pathogen populations (Bennetzen and Hulbert, 1992). Under disease pressure from a pathogen that has evolved to escape detection, selection on the plant for the few progeny with a new and appropriate R specificity would be very strong. Most of the R genes studied in plants exhibit the effects of diversifying selection, where non-synonymous substitutions exceed synonymous substitutions within an LRR region (Parmiske et al., 1997; Song et al., 1997; Meyers et al., 1998). The rapid evolution of R genes is also suggested by the high degree of haplotype diversity within a species, as shown in sequence anal...
yses of Rpm1 and Rps2 alleles from different ecotypes of Arabidopsis (Caicedo et al., 1999; Stahl et al., 1999).

In addition to selected changes in gene sequence and in the number, order, or chimeric origins of R loci, other structural variations have been detected in R gene clusters. Transposable elements are often inserted within complex R loci, sometimes causing insertional inactivation of an R gene (Song et al., 1997; Multani et al., 1998; Noel et al., 1999; Dodds et al., 2001; Ramakrishna et al., 2002b). In addition, inserted transposons can provide new regulatory sequences for any adjacent gene (Wessler et al., 1995). Moreover, two nearby transposons belonging to the same family could provide the homology needed for R gene duplication or deletion by unequal exchange. Numerous small insertions or deletions are often seen within R genes or adjacent sequences, but their mode(s) of origin and biological significance are not clear. Because so few complex R loci have been analyzed for structural variation (Parniske and Jones, 1999; Noel et al., 1999), we do not know the relative importance of these numerous rearrangement mechanisms for generating new R gene capabilities.

Rp1 is a complex R locus in maize (Zea mays) conferring race-specific resistance to a fungal pathogen, common leaf rust (Puccinia sorghi). Rp1 is an exceptionally unstable locus, even for complex R genes (Bennetzen et al., 1988; Sudupak et al., 1993; Richter et al., 1995). Both unequal crossing over and unequal conversion have been associated with changes in R specificity in the region (Sudupak et al., 1993; Hu and Hulbert, 1994; Richter et al., 1995). Small DNA segments containing most of the Rp1 homologs from the maize Rp1-D haplotype have been cloned and sequenced, providing evidence for chimeric gene generation and diversifying selection (Collins et al., 1999; Sun et al., 2001). We have recently sequenced 95- and 99-kb segments of the maize Rp1 cluster from the rust-susceptible inbred B73 (Ramakrishna et al., 2002b), uncovering a complex mixture of duplicated R genes and transposable elements. Most interesting, we found that the Rp1 complex in B73 contains at least one large (43 kb) non-tandem duplication and also several clusters of truncated genes with no homology to any known R gene (Ramakrishna et al., 2002b). Rp1 gene homologs have also been cloned and sequenced from the barley genome, although their function is not known (Ayliffe et al., 2000).

For this study, we sequenced a contiguous 268 kb of the Rp1-orthologous region in sorghum (Sorghum bicolor) to determine structural variation for an R gene cluster that has diverged at least since the ancestral divergence of maize and sorghum about 15 to 20 million years ago (mya; Gaut and Doebley, 1997). Comparison with the maize Rp1 complex indicates major differences in the current structures, but also provides evidence for common processes that create diversity in these complex R gene regions.

**RESULTS**

Isolation and Mapping of Rp1-Orthologous Clones from Sorghum

Ten bacterial artificial chromosome (BAC) clones that hybridized to a probe from the 5′ end of the Rp1-D gene (Collins et al., 1999) were identified from the sorghum BTx623 BAC library (http://hbz.tamu.edu/bacindex.html). Restriction enzymes MluI, NcoI, and NotI were used to make a physical map based on overlapping restriction fragments. Gel-blot hybridization analysis with the Rp1 5′ probe was used to confirm the physical map, generating a contiguous series (contig) of overlapping clones. The number of Rp1 homologs in the contig was determined by hybridization of NcoI-digested BAC DNAs with the Rp1 5′ probe. Sun et al. (2001) previously suggested that the number of NcoI-hybridizing fragments should correspond to the number of Rp1 homologs in maize, but we found that some Rp1 homologs within a cluster are so similar in sequence that they give identical fragment sizes with most restriction enzymes (Ramakrishna et al., 2002b). In addition, more than one hybridizing band per gene would be observed if there were an NcoI restriction site in the region of Rp1 covered by the probe. Taking these factors into account, we identified five Rp1 homologs that mapped to a 27-kb region in a 350-kb contig (Fig. 1). This contig mapped to linkage group H of sorghum (P.E. Klein, personal communication). Two other fingerprinted BACs that hybridized to the probe from the 5′ region of Rp1-D revealed one unique band on an NcoI gel blot. However, these clones did not hybridize to a probe from the Rp1 3′ region, suggesting that the sequences were 3′-truncated paralogs and/or not closely related to Rp1. In agreement with this prediction, these two clones were unlinked to the Rp1-orthologous region, forming a contig that mapped to linkage group G of sorghum (P.E. Klein, personal communication; Klein et al., 2000).

**Figure 1.** Physical map of the rph1 region from sorghum inbred BTx623. Each circle represents an Rp1 homolog (i.e. rph1 gene). Physical distances and names of BACs are represented by lines above and below the positions of the rph1 genes. The names of BACs in parentheses are from the naming convention used in fingerprint analysis (Klein et al., 2000; P.E. Klein, personal communication).
Sequence Analysis of Sorghum Rp1-Orthologous BACs

The genomic organization at the Rp1-orthologous region in sorghum was studied by sequencing two sorghum BACs, Sb95A23 and Sb98N08. The two clones that were chosen contain all five Rp1 homologs, and more than 200 kb of sorghum DNA that is 3' to the duplicated Rp1 homologs (Fig. 1). These two BACs were fully sequenced by our standard shotgun approach (Dubcovsky et al., 2001; Song et al., 2001; Ramakrishna et al., 2002a). Sb95A23 (97,616 bp) and Sb98N08 (237,576 bp) overlap by 67,759 bp and together cover 268,433 bp (accession no. AY144442) of the sorghum genome.

Sequence analysis by our standard approach, using a combination of gene-prediction programs, homology searches, and manual annotation indicates that this 268-kb region contains 31 candidate genes, including four that appear to be truncated. Numerous mobile DNAs are also found in the region (Fig. 2). Identified retrotransposon sequences constitute about 20% of the region, a relatively high percentage for the sorghum genome. We recently sequenced a different sorghum region of 425 kb that contains only 4% retrotransposon sequences (Song et al., 2002).

Sequences annotated as genes on these BACs are described in Table I. Genes were classified as truncated when homology was detected only to a part of a protein entry in the GenBank database using the BLASTX program. The homology in these truncated genes was limited only to portions of the protein-encoding regions that was confined to either N-terminal, C-terminal, or central regions of the putative protein.

The Sorghum rph1 Gene Cluster

Because the sequenced genes are highly homologous to the Rp1 genes of maize and because they map on sorghum linkage group H at a syntenic position with the short arm of maize chromosome 10 where Rp1 is found (Peng et al., 1999), we have named these genes rph1-1 through rph1-5 (Fig. 2). The portion of the Rp1-orthologous gene cluster that we sequenced covers approximately 27 kb. This region contains three apparent full-length Rp1 homologs (rph1-1, rph1-2, and rph1-4) and two truncated Rp1 homologs (rph1-3 and rph1-5, truncated at exactly the same nucleotide). Four other predicted genes are found within the sequenced Rp1-orthologous region. These four predicted genes are present as duplicated pairs, with two between rph1-1 and rph1-2 (genes 1A and 2A) and the other two between rph1-3 and rph1-4 (genes 1B and 2B). One cDNA has been sequenced from sorghum that has 93% and 90% identities to genes 1A and 1B, respectively, but homology was not identified for either gene 1 or 2 in any other species, nor can any function be predicted for these genes.

Although genes rph1-1, rph1-2, and rph1-4 are all largely intact, only rph1-2 appears to encode a predicted full-length Rp1 protein like that encoded by the maize Rp1-D gene (Fig. 3). Gene rph1-1 contains a stop codon at the position corresponding to amino

![Figure 2](https://example.com/figure2.png)

**Figure 2.** Sequence organization of the sorghum genomic segment as represented by overlapping BACs Sb95A23 and Sb98N08. The five Rp1 homologs were named rph1-1 to rph1-5. All other genes are given simplex numbers from 1 to 19. Arrows indicate genes or truncated genes, plus their size and proposed direction of transcription. A through C, Multiple copies of the same gene family. Asterisks, SSRs that contain tandem repeat numbers of eight or more. Retrotransposons are indicated by rectangles without arrows on top. Novel full-length LTR-retrotransposons were named as described by SanMiguel et al. (2002). MITEs are indicated by small triangles. Arrowheads with bars between genes 15C and 17 indicate direct repeat sequences of 468 and 470 bp that are 91% identical to each other and separated by 6,112 bp. Gene 19p is given its unique designation because the sequenced region only contains the 3' portion (p) of the gene.
A novel LTR retrotransposon, *Kaema*-1, is inserted 752 bp downstream of the predicted translation initiation codon of *rph1-4*, which should also result in a truncated protein product. The 5' regions that would encode the N-terminal 840 amino acids are deleted from *rph1-3* and *rph1-5*. This presumably indicates that the deletion was a single event that predated the duplication of *rph1-3* and *rph1-5* from a common ancestral gene. This interpretation is supported by the observation that the retained sequences of *rph1-3* and *rph1-5* are more closely related to each other than to the same region in any of the other *rph1* candidate genes in this region.

A 4,077-bp region from the end of *rph1-1* to the start of *rph1-2* and a 4,256-bp region from the end of *rph1-3* to the start of *rph1-4* exhibit 96.2% sequence identity, ignoring insertions and deletions (indels). These are the regions that contain the predicted genes 1A, 2A, 1B, and 2B. The very high sequence identity within this area suggests that the region was duplicated within the last three million years, probably by unequal recombination sited within the *rph1* genes themselves.

### Table 1. Predicted genes on sorghum BACs Sb95A23 and Sb98N08

<table>
<thead>
<tr>
<th>Gene</th>
<th>Homology</th>
<th>Accession No.</th>
<th>e Value</th>
<th>Sorghum ESTs (Identity %)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>rph1-1</em></td>
<td>Rust resistance-like protein, RP1-3 (maize)</td>
<td>AAM03018</td>
<td>0</td>
<td>AW923446 (100%)</td>
</tr>
<tr>
<td>1A</td>
<td></td>
<td></td>
<td></td>
<td>AW923525 (100%)</td>
</tr>
<tr>
<td>1B</td>
<td></td>
<td></td>
<td></td>
<td>AW923525 (99%)</td>
</tr>
<tr>
<td><em>rph1-2</em></td>
<td>Rust resistance RP1D-like protein (maize)</td>
<td>AAK18308</td>
<td>0</td>
<td>AW923446 (99%)</td>
</tr>
<tr>
<td><em>rph1-3</em></td>
<td>Rust resistance-like protein, RP1-2 (maize)</td>
<td>AAM03016</td>
<td>e=148</td>
<td>AW923446 (91%)</td>
</tr>
<tr>
<td>2A</td>
<td></td>
<td></td>
<td></td>
<td>AW923525 (95%)</td>
</tr>
<tr>
<td>2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>rph1-4</em></td>
<td>Rust resistance-like protein, RP1 (maize)</td>
<td>AAM03014</td>
<td>0</td>
<td>AW923446 (93%)</td>
</tr>
<tr>
<td>3</td>
<td>Signal transducer of phototropic response, RPT2 (Arabidopsis)</td>
<td>AAF33112</td>
<td>2e=62</td>
<td>BM324312 (94%)</td>
</tr>
<tr>
<td>4A</td>
<td>Hypothetical protein (Arabidopsis)</td>
<td>AAG51364</td>
<td>9e-32</td>
<td>BE599243 (100%)</td>
</tr>
<tr>
<td>5</td>
<td>Hypothetical protein (Arabidopsis)</td>
<td>AAG51355</td>
<td>e=167</td>
<td>BG241140 (84%)</td>
</tr>
<tr>
<td>4B</td>
<td>Hypothetical protein (Arabidopsis)</td>
<td>AAG51364</td>
<td>1e-36</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7A</td>
<td><em>Xa1</em>-like (rice)</td>
<td>AAK2518</td>
<td>e=179</td>
<td></td>
</tr>
<tr>
<td>8*</td>
<td>Putative protein (Arabidopsis)</td>
<td>CAB96668</td>
<td>3e-25</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Putative protein (Arabidopsis)</td>
<td>CAB82970</td>
<td>2e-48</td>
<td>BE356244 (99%)</td>
</tr>
<tr>
<td>7B*</td>
<td><em>Xa1</em>-like (rice)</td>
<td>AAK2518</td>
<td>8e-20</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Expressed protein (Arabidopsis)</td>
<td>AAB63084</td>
<td>4e-42</td>
<td>BE599243 (86%)</td>
</tr>
<tr>
<td>11</td>
<td>Cysteine proteinase (Arabidopsis)</td>
<td>CAB41164</td>
<td>6e-17</td>
<td>BG053332 (99%)</td>
</tr>
<tr>
<td>12</td>
<td>Cytochrome P450 (sorghum) (Obtusifoliol 14-α demethylase)</td>
<td>AAC49659</td>
<td>0</td>
<td>BM327831 (100%)</td>
</tr>
<tr>
<td>13</td>
<td>Cyclin-dependent kinase (Tobacco [Nicotiana tabacum])</td>
<td>AAG01533</td>
<td>7e-47</td>
<td>BG487760 (98%)</td>
</tr>
<tr>
<td>14</td>
<td>Hypothetical protein (rice)</td>
<td>BAB62578</td>
<td>3e-32</td>
<td>BE125790 (99%) BE359797 (100%)</td>
</tr>
<tr>
<td>15A</td>
<td>Hypothetical protein (rice)</td>
<td>BAB62578</td>
<td>1e-30</td>
<td>BM032442 (97%) AW680740 (82%)</td>
</tr>
<tr>
<td>16A</td>
<td>Ankyrin-like protein (Arabidopsis)</td>
<td>AAF04911</td>
<td>2e-76</td>
<td>BE125790 (92%)</td>
</tr>
<tr>
<td>15B</td>
<td>Hypothetical protein (rice)</td>
<td>BAB62578</td>
<td>4e-69</td>
<td></td>
</tr>
<tr>
<td>16B</td>
<td>Ankyrin-like protein (Arabidopsis)</td>
<td>AAF04911</td>
<td>1e-33</td>
<td>BE359797 (90%)</td>
</tr>
<tr>
<td>15C</td>
<td>Hypothetical protein (rice)</td>
<td>BAB62578</td>
<td>1e-54</td>
<td>BG047811 (100%) BM3302000 (99%)</td>
</tr>
<tr>
<td>17</td>
<td>ATPase-like (Arabidopsis)</td>
<td>BAA96915</td>
<td>4e-14</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Hypothetical protein (Arabidopsis)</td>
<td>NP_178129</td>
<td>1e-26</td>
<td>AW283596 (100%) BE358200 (99%)</td>
</tr>
<tr>
<td>19-p</td>
<td>Unknown protein (rice)</td>
<td>BAB93427</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Truncated genes. **Predicted as a gene by two or more gene prediction programs but does not have significant homology to any entries in EST or protein databases of GenBank.*

acid 197 in Rp1-D. A novel LTR retrotransposon, *Kaema*-1, is inserted 752 bp downstream of the predicted translation initiation codon of *rph1-4*, which should also result in a truncated protein product. The 5' regions that would encode the N-terminal 840 amino acids are deleted from *rph1-3* and *rph1-5*. This presumably indicates that the deletion was a single event that predated the duplication of *rph1-3* and *rph1-5* from a common ancestral gene. This interpretation is supported by the observation that the retained sequences of *rph1-3* and *rph1-5* are more closely related to each other than to the same region in any of the other *rph1* candidate genes in this region.

A 4,077-bp region from the end of *rph1-1* to the start of *rph1-2* and a 4,256-bp region from the end of *rph1-3* to the start of *rph1-4* exhibit 96.2% sequence identity, ignoring insertions and deletions (indels). These are the regions that contain the predicted genes 1A, 2A, 1B, and 2B. The very high sequence identity within this area suggests that the region was duplicated within the last three million years, probably by unequal recombination sited within the *rph1* genes themselves.
Non-Tandem Gene Duplications Near the rphI Cluster

We identified 31 possible genes within the more than 268 kb of genomic DNA that we sequenced from sorghum. Five of these genes are RphI homologs. Duplicated genes 1A and 1B are predicted by three independent gene-finding programs (FGENESH, GeneMark.hmm, and GENSCAN), but only part of gene 1 has similarity to a cDNA sequence (BE362870: 90% identity over a 260-bp region). Duplicated genes 2A and 2B are annotated as genes based only on gene prediction programs. The degrees of homology and predicted functions (based exclusively on these homologies) for the other 22 predicted genes in this region are provided in Table I. These genes encode a wide variety of predicted proteins, ranging from ankyrin-like proteins, to a cytochrome P-450 enzyme, to a protein that is involved in signal transduction of the phototropic response.

Many of the genes in this region are duplicated, but none are duplicated in a simple tandem manner. Genes 4A and 4B are in an inverted orientation, with predicted gene 4 in between. Gene 7A is homologous to a Xa1-like putative disease R gene from rice and has a truncated copy (gene 7B) upstream of gene 9. Gene 7B has an N-terminal deletion compared with gene 7A (3,855 bp) and the rice Xa1-like gene (4,206 bp; OSJNA0056A20.14), leaving only about 0.9 kb of gene-like sequence. Gene 15 has three copies (genes 15A–15C) with predicted protein products that show homology to a rice hypothetical protein (BAB62578). Comparison of gene 15A with gene 15C shows 91% identity over a 681-bp region. In contrast, gene 15B has respective 73% and 69% identities with genes 15A and 15C. This suggests that genes 15A and 15C are the more recent duplication products. An alternative hypothesis, that gene 15B is evolving more rapidly, seems unlikely because the $K_s$ to $K_a$ ratios for all three gene 15 comparisons are very similar (0.57–0.67; data not shown). Gene 16 also has duplicated copies, genes 16A and 16B, that are 84% identical. Genes 16B and 15C are in an inverted orientation compared with genes 15A, 16A and 15B, suggesting that a minimum of four rearrangements (duplication of 15A/16A and 15B/16B, followed by inversion of 16B, insertion of Leviathan, and duplication/insertion of 15C from 15A) would have been required to generate the current genomic structure in this region. A more complex series of rearrangements is more likely, however, given the apparent relative dates of the duplication events.

Truncated Genes

Several of the predicted genes in the sequenced region appear to be truncated relative to more intact copies in this same region, and sometimes also to other homologs in the databases. Genes rphI-3, rphI-5, and 7B mentioned above all contain deletions that remove a major portion of their single coding exons. Gene 8 has only a single copy in this region but also appears to be truncated, from both the 5' and 3' ends. The predicted protein that could be encoded by gene 8 shows homology to a part (amino acids 183–252) of an Arabidopsis putative protein (CAJB96668). Gene 8 also shows 87% nucleotide identity, over the identical 210-bp region, to the interior portion of a wheat expressed sequence tag (BG313327).

Mobile DNAs and Other Structural Features

Five largely intact retroelements were identified in the 268-kb sorghum region. Four of these elements are retrotransposons with long terminal repeats (LTRs), whereas one element is a LINE. Only one of the LTR-retrotransposons, Leviathan, had been previously identified in sorghum (Bennetzen, 1996; Chen et al., 1998). We discovered and named two new elements (Kaema and Pifi). These retroelements were given designations using the approach described by SanMiguel et al. (2002). Pifi was found upstream of rphI-1. About 123 kb from the region where Kaema-1 is inserted in rphI-4, there is another copy of this retrotransposon between genes 11 and 12. A full-length Leviathan retrotransposon of 14,522 bp is present between genes 15B and 16B. In addition, there are many truncated retrotransposons in this region. Miniature inverted-repeat transposable elements (MITEs) were abundant (at least 21) in the sorghum BACs. There were no large inverted repeat (class II) transposable elements found in this region. In comparison with the rphI region, the 425-kb sorghum region described earlier has only three retroelements (Song et al., 2002).
Several simple sequence repeats (SSRs) were annotated in the region. Most of these SSRs were found within or near genes, but one was identified within an LTR retrotransposon fragment. These hypervariable sequences can serve as excellent markers to follow the segregation of this region, including in future tests to see whether any disease resistance genes in sorghum might map at this location.

Molecular Analysis of Rpis Homolog Origins and Evolution

To investigate the evolutionary history of the Rpis-like genes in the rph1 region, all of the Rpis homologs sequenced in the present study and those in the GenBank database (Collins et al., 1999; Ayliffe et al., 2000; Sun et al., 2001; Ramakrishna et al., 2002b) were used to construct a neighbor-joining tree. Because the Rpis homologs show diversifying selection mostly in the LRR region, where non-synonymous substitutions exceed synonymous substitutions (Sun et al., 2001; Ramakrishna et al., 2002b), synonymous substitutions were used to construct gene trees. Neighbor-joining trees with different topologies were obtained for different domains, indicative of mosaics of shared gene segments (chimeric genes) that could have resulted from unequal crossing over and/or gene conversion events. However, most of the mosaic patterns were observed in the C-terminal region of the gene (Ramakrishna et al., 2002b). Therefore, the N-terminal regions were used to make neighbor-joining trees using Kimura two-parameter distance estimates and synonymous substitutions separately (Ramakrishna et al., 2002b), but one such mobile insertion time is very recent, presumably after the duplication of the segments between the rph1 genes that also duplicated genes 1 and 2. By this same reasoning, it appears that Leviathan also inserted very recently (<0.5 mya), long after the duplications of genes 15 and 16.

Retrotransposon Insertion Times

The insertion times of full-length sorghum retrotransposons was estimated from the divergence of their LTRs (Table II) as described by SanMiguel and coworkers (2002). The synonymous substitution rate for adh1 and adh2 genes of grasses (6.5 × 10^{-9} substitutions year^{-1} site^{-1}) was used to estimate the insertion times of retrotransposons. Under these criteria, the analyses indicated that three of four sorghum LTR-retrotransposons inserted within the last one million years. Our analyses indicated that Kaema-2 is the most ancient insertion in this region, but it still dates back to only about 2.5 mya. The Kaema-1 insertion time is very recent, presumably after the duplication of the segments between the rph1 genes that also duplicated genes 1 and 2. By this same reasoning, it appears that Leviathan also inserted very recently (<0.5 mya), long after the duplications of genes 15 and 16.

Comparison of Orthologous rp1/rph1 Loci from Maize and Sorghum

Two segments of the maize genome containing Rpis homologs have been sequenced, and these 95- and 99-kb maize BAC inserts (AF466931 and AF466932) mapped to the syntenic location with the sorghum rph1 cluster (Ramakrishna et al., 2002b). The entire Rpis R gene cluster was mapped to two segments of 250 and 300 kb (separated by an unknown distance), containing approximately 15 Rpis homologs. In sorghum, the orthologous locus is much smaller (about 27 kb) and contains only five Rpis homologs, rph1-1 through rph1-5. Most of the difference in regional size appears to be attributable to the greater number of LTR retrotransposons inserted in the maize region (Ramakrishna et al., 2002b), but one such mobile DNA (Kaema-1) is found in the sorghum region. The recent insertion times of both the maize and sorghum LTR retrotransposons in these regions indicates that they landed at these locations more than 10 million years after the divergence of maize and sorghum from a common grass ancestor.

In both maize and sorghum, many of the Rpis homologs appear to be nonfunctional. However, despite the absence of any known R phenotype specified by either the maize rp1 region (from inbred B73) or the sorghum rph1 region, cDNAs have been identified in their respective species, indicating that at least one gene in the region is expressed.

Figure 4. A neighbor-joining tree of barley, maize, and sorghum Rpis homologs. A neighbor-joining tree was constructed based on synonymous substitutions in the N-terminal region of all the barley, maize, and sorghum Rpis gene family members from the present study and those present in the GenBank database. Bootstrap values based on 1,000 replicates are indicated at the nodes.
As in the rp1 complex of maize, there are additional genes within the rph1 R gene cluster. None of these genes has any known association with disease resistance activity. Despite the presence of potentially functional non-R genes in each cluster, the same “extra” genes found within the rp1 complex were not seen in the sorghum rph1 complex, nor were the extra genes from the rph1 complex observed in the maize rp1 complex.

**DISCUSSION**

**The Evolution of R Gene Clusters**

Many previous studies have shown that comparative analysis of DNA sequences from orthologous chromosomal regions can provide important and unique insights into genome evolution (for review, see Bennetzen, 2000). Loci containing R genes have not been subjected to this analysis, partly because the first few attempts have yielded a lack of conserved R gene presence/absence between the compared species (Kilian et al., 1997; Feuillet and Keller, 1999; Brueggeman et al., 2002). This is not an entirely surprising result, given that comparative recombinational mapping in grass genomes indicated a rapid reorganization of R gene homologs, often leading to a complete lack of synteny for these candidate R genes (Leister et al., 1998).

In this study, we have determined the sequence of the complex locus that we call rph1, plus more than 200 kb of contiguous downstream sequence. The rph1 complex maps in the orthologous position relative to the complex Rp1 locus of maize, and contains duplicated copies of Rp1-like genes. Comparison of the gene sequences suggests that the Rp1-homologous loci from maize, sorghum, and barley have all arisen independently from a single Rp1-like gene that was present in their common ancestor. However, this conclusion may be incorrect because the high frequency of known and predicted unequal recombination events in this region (Sudupak et al., 1993) can cause the R genes within a cluster to evolve as a concerted set. The exception to this prediction of concerted evolution is present in the LRR region, where natural selection has apparently enriched for sequence diversity in this pathogen-recognition component.

Compared with the maize rp1 complex from maize inbred B73, the rph1 cluster of sorghum inbred BTx623 has a significantly lower number of Rp1-homologous loci (15 versus five). However, various haplotypes of Rp1 in maize contain different numbers of Rp1 homologs, ranging from possibly as few as 1 to as many as 20 in known alleles (Collins et al., 1999; Sun et al., 2001; Ramakrishna et al., 2002b; Webb et al., 2002). Hence, the sorghum rph1 complex of BTx623 is within the range of maize Rp1 variability. In both maize and sorghum, it is expected that unequal interchromatid exchange will cause exactly as many decreases in Rp1 homolog number as increases, and that only selection for gene function within the duplicated regions will allow maintenance of a high copy number. It is not known whether the gene fragments might have any role, but they could serve as a reservoir of R gene variability for later use in creating a new Rp1 gene by unequal exchange or conversion.

Unlike most other complex disease resistance loci, the maize Rp1 region contains at least two apparently intact genes within the cluster that have no obvious role in disease resistance (Ramakrishna et al., 2002b). One of these genes exhibits homology to Arabidopsis and rice hypothetical proteins with no similarity to any known cDNAs. The second gene shows homology to rice unknown and hypothetical proteins with 91% and 96% identity to maize cDNAs (H89387 and AW424864) over a 440-bp region. We did not find homologs of these genes in the sorghum rph1 complex. In fact, only the Rp1 homologs are conserved between the sequenced maize and sorghum BACs. The sorghum rph1 region does contain some non-R genes, and these were apparently duplicated as part of an unequal recombination event that duplicated rph1 genes.

The precise orders of the events that gave rise to the current rph1 complex is ambiguous, given that multiple events have apparently occurred. The simplest series of events that we can predict would require a minimum of five steps. The first step would have been an ancient duplication of a single-copy rph1 gene to give rise to rph1-1 and rph1-2, either by a mechanism that allowed them to be separated by two predicted genes (1A and 2A), or followed by insertions of these two genes in this region. It may seem unlikely that genes frequently insert into new locations, but we have determined that the adh1 gene was inserted as a single functional locus onto a new chromosome in a common ancestor of maize and sorghum (Tikhonov et al., 1999) and that two genes were inserted near the adh2 locus of sorghum after its divergence from a common ancestor with maize (K. Ilic and J. Bennetzen, unpublished data). The second step would have been the direct duplication of the rph1-3 gene from rph1-2, to create a [rph1-1-gene 1A-gene 2A-rph1-2-rph1-3] complex. The third step would have been the 5′ truncation of rph1-3. We know that this step preceded the next duplication because the identical 5′ truncation is seen in rph1-5. The fourth step was the direct tandem

**Table II. Predicted retrotransposon insertion times**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sites</th>
<th>K</th>
<th>st</th>
<th>Mya</th>
<th>st</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pifi</td>
<td>235</td>
<td>0.009</td>
<td>0.006</td>
<td>0.66</td>
<td>0.47</td>
</tr>
<tr>
<td>Kaema-1</td>
<td>457</td>
<td>0.002</td>
<td>0.002</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>Kaema-2</td>
<td>377</td>
<td>0.024</td>
<td>0.008</td>
<td>1.87</td>
<td>0.63</td>
</tr>
<tr>
<td>Leviathan</td>
<td>3,922</td>
<td>0.004</td>
<td>0.001</td>
<td>0.33</td>
<td>0.08</td>
</tr>
</tbody>
</table>
duplication of the [gene 1A-gene 2A-rph1-2-rph1-3] portion of the complex to create the current order and content of R gene homologs. This may have occurred by unequal recombination, apparently sited just 3' to rph1-1 and 3' to rph1-3. We do not see any mobile DNAs or other repeats at these sites now that may have provided homology for such unequal events, but the rapid decay of repeats in higher plants makes it likely that such legacies would no longer be detectable (SanMiguel et al., 1998; Devos et al., 2002). One additional structural rearrangement must then have occurred, the insertion of Kaema-1. The predicted nature and order of these events are depicted in Figure 5. We cannot predict the precise timing of these proposed events. However, the degree of sequence divergence between rph1-1, rph1-2, and rph1-4 (Fig. 4) can be used to predict approximate dates for the duplications of these genes. Comparing the maize and sorghum Rp1 orthologs indicates that the synonymous substitution rate at this locus is 2-fold higher than the synonymous substitution rate (6.5 × 10^{-9} substitutions year^{-1} site^{-1}) at the adh locus of grasses (Gaut et al., 1996). Taking this into account, the initial rph1 duplication into rph1-1 and rph1-2 would have occurred about 8 to 10 mya, and the duplication of the [rph1-1-gene 1A-gene 2A-rph1-2-rph1-3] complex would have taken place within the last two million years.

In some R genes of the NBS-LRR class, unequal recombination is often sited in the repeats in the LRR region, as observed in mutants at the M locus in flax and RPP5 in Arabidopsis (Anderson et al., 1997; Noel et al., 1999). Unequal intragenic recombination has been shown to be frequent at the maize Rp1 locus (Collins et al., 1999; Sun et al., 2001), although rearrangements with breakpoints outside the Rp1 homologs may have been under-represented in these studies. Haplotypes of the Cf-4/Cf-9 region primarily show unequal recombination events sited at a specific location that is not within a R gene (Parniske et al., 1997; Parniske and Jones, 1999). Our data do not allow us to determine the precise breakpoints of the unequal recombinational events that gave rise to the current rph1 complex in BTx623, but they do suggest that the recombination occurred just 3' to rph1-1, before gene 1. Interestingly, our analysis of the rph1 complex from maize inbred B73 detected similar end points for a large segmental duplication: just 3' to an Rp1 homolog (Ramakrishna et al., 2002b).

Nearby Rearrangements: Non-Tandem Gene Duplication and Truncation

Of the rph1 R gene homologs, only one of them (rph1-2) contains an uninterrupted reading frame like that seen in the maize Rp1-D gene. Although sorghum does not have any R gene mapped to this region, rph1-2 does show 99% to 100% identity over a 700-bp region with cDNAs from sorghum inbred BTx623 (AW923446 and AW923525). This indicates that at least one Rp1-homologous gene in this complex is expressed in the inbred that was the source of the sequenced BACs.

Most of the Rp1 homologs in the rph1 complex are defective, two by apparent nonsense mutations and two by 5' truncations. All of the sorghum rph1 genes that we sequenced are in the same transcriptional orientation. Several other genes downstream from the rph1 complex also show nearby duplicated copies and/or 5' truncation.

Three duplicated gene pairs and one gene triplication are found in the more than 200 kb of sorghum DNA that we sequenced downstream of the rph1 cluster. One (gene 7) of the duplicated gene pairs is in direct orientation, whereas two (genes 4 and 16) are in an inverted orientation. Of the triplicated genes, 15A and 15B are directly repeated, whereas 15C is inverted relative to the two other copies. In every case, the duplicated or triplicated loci are separated by additional predicted genes. Hence, none of these rearrangements appear to have arisen from a single simple event like an unequal exchange. It seems likely that the current gene content, order, and orientation of this region is attributable to multiple independent rearrangements within the last few million years. Hence, the area adjacent to the rph complex of sorghum is unusually unstable, compared with other analyzed regions of the sorghum genome (Chen et al., 1998; Tikhonov et al., 1999; Ramakrishna et al., 2002a; Song et al., 2002).

Figure 5. A model for the origin of the current state of the sequenced rph1 complex from a single region with only one rph gene. Horizontal arrows designate the positions, orientations, and approximate transcript sizes for the indicated genes.
The gene truncations that we observed in this region are, in contrast, mostly simple events. Genes *rph1*-3, *rph1*-5, and 7B all contain only 5′ deletions of the single predicted exon that they originally encoded. There are several processes that could account for these deletions, including illegitimate recombination (Devos et al., 2002). Gene 8 exhibits apparent deletions that have removed all homology to previously sequenced Arabidopsis and wheat genes from both the 5′ and 3′ ends. Only a central 210-bp segment of a single exon remains in the sequence we have generated for gene 8. Although two independent deletions are possible, it is also possible that the gene 8 exon fragment arrived in this location in the same way that similarly truncated genes were found to cluster within the *rp1* complex (Ramakrishna et al., 2002b). Although the precise mechanism of gene fragment insertion is not known, illegitimate repair of a double stranded DNA break seems a likely candidate (Wessler et al., 1990; Salomon and Puchta, 1998). The gene insertions, inversions, duplications, and fragmentations observed in the *rph1*-adjacent region have all been seen as rare events in other regions of various plant genomes. However, they are greatly enriched in this sequenced region of sorghum. Further studies are needed to determine whether close juxtaposition to an unstable R gene complex might increase the instability of nearby loci. At the very least, it is clear that numerous genic rearrangements have occurred in this small region within the last eight million years, the approximate date of two of the gene duplication events (genes 7A/B and 15A/C) in this small chromosomal segment.

Mobile DNA Accumulation Specificities

As we and others have observed previously (Tikhonov et al., 1999; Mao et al., 2000; Ramakrishna et al., 2002a), the small class II transposable elements in the sequenced region (MITEs) are primarily associated with genes. None is inserted within a known retroelement. In contrast to other regions of the sorghum genome that have been sequenced (Tikhonov et al., 1999; Ramakrishna et al., 2002a; Song et al., 2002), there are quite a few retrotransposons in the sequenced DNA, although none are in an obvious nested organization like that seen in maize (San-Miguel et al., 1996). Many of these retroelements are also truncated, suggesting that they have undergone the same type of illegitimate recombination that has removed most of the retrotransposon sequences from Arabidopsis (Devos et al., 2002).

Orthologous *rp1* and *rph1* Regions of Sorghum and Maize: Similarities and Contrasts

The *Rp1* region in maize has been placed near the telomere of the short arm of chromosome 10 by both cytogenetic and recombinational mapping (Rhoades, 1935; Saxena and Hooker, 1968; Hulbert and Bennetzen, 1991). Although recombination rates in the region can vary a great deal in different haplotype combinations, the complex *Rp1* locus usually is found to cover about 0.4 centiMorgans. *Rp1* is flanked by markers bn3.04 and ksu3/4, separated by about 4 centiMorgans on the maize genetic map (Jiang et al., 1996). Most of the sorghum BACs harboring *Rp1* homologs mapped to a 27-kb region close to marker bn3.04 on sorghum linkage group H. Few markers near *Rp1* in maize have also been mapped in sorghum, but markers bn3.04, umc130 and rz561 flank *Rp1* in maize and flank the region of linkage group H in sorghum that we sequenced (Wilson et al., 1999; Klein et al., 2000; http://www.gramene.org/cmap). Therefore, the sequenced region in sorghum appears to be syntenic to the *Rp1* region mapped on the short arm of chromosome 10. Also, because no other confirmed BAC clones other than those we sequenced exhibited strong homology to *Rp1* probes, all evidence suggests that *rph1* and *rp1* are orthologous complex loci.

In both sorghum and maize, the *Rp1* homologs in these syntenic regions are arrayed as nearby duplicated copies, but with some other gene candidates found interrupting the duplicated regions. In maize, a large number of the non-R gene sequences found in the *Rp1* region were present as truncated gene fragments. Thirteen of the 16 truncated genes that we found in the maize *rp1* complex were in three clusters, suggesting that they arose from multiple rounds of illegitimate break repair at the same sites or from single repairs at each of these sites with multiple unlinked DNA templates (Ramakrishna et al., 2002b). We do not know how frequently genes are truncated in plants, but they tend to occur after gene amplification (Llaca and Messing, 1998). However, the frequency of truncated genes near the *rph1* region of sorghum suggests that there is something about disease resistance gene clusters that attracts this type of DNA rearrangement. Further genome analysis is needed to investigate this point.

It is not clear why the *rp1* complex of maize and the *rph1* complex of sorghum have retained syntenic/orthologous locations on their respective genetic maps, given that conserved syntenic positions are often not seen with R genes (Leister et al., 1998; Feuillet and Keller, 1999; Brueggeman et al., 2002). In several other ways, the *rp1* and *rph1* complex loci share common properties. They have repeated R genes that are mostly in direct orientation, they have non-R genes within the complex that have been duplicated along with the R genes, they have created direct repetitions of large segments of the R gene region that contain more than one R gene with breakpoints sited just 3′ to *Rp1* homologs, they have recently inserted retrotransposons within the cluster, and they have nearby truncated gene fragments. Although all of these events appear to have occurred
independently in each lineage, their common general outcomes suggest that similar mechanisms and/or selective forces have continued to act on these loci for many millions of years.

MATERIALS AND METHODS

BAC Selection and Restriction Mapping

An Rpl 5′ probe, previously described (Ramakrishna et al., 2002b), was used to screen a BAC library from sorghum (Sorghum bicolor) inbred B7362 (http://bhz.tamu.edu/bacindex.html). A total of 10 BACs with homology to this probe were identified, and these clones were mapped using different combinations of MlaI, Ncol, and NotI. Fragments were size fractionated by field inversion gel electrophoresis, transferred to nylon membranes, and hybridized with the Rpl 5′ probe.

Detailed restriction maps of BACs Sb95A23 and Sb98N08 were constructed to experimentally validate sequence assembly. BACs were digested with restriction enzymes AscI, NotI, PstI, PmlI, SfiI, and Sool. The DNAs in the gels were transferred to nylon membranes and hybridized with the Rpl 5′ probe. Fragments observed on the agarose gel were compared between BACs to identify common fragments.

Sequencing and Sequence Analysis

Shotgun libraries for BACs Sb95A23 and Sb98N08 were made as described previously (Dubcovsky et al., 2001; Song et al., 2001). For finishing the sequences of the BACs, gaps were closed as described by Ramakrishna et al., (2002a). Annotation and sequence analysis were performed as described earlier (Dubcovsky et al., 2001; Song et al., 2001; Ramakrishna et al., 2002a). FGENESH (http://www.sofberry.com/berry.pl.htm) with the monocot training set was used for gene prediction in addition to GENSCAN (http://genes.mit.edu/GENSCAN.html) and GeneMark hmm (http://opal.biology.gatech.edu/GeneMark/eukhmnn.cgi). The criteria used to define a gene were, first, a match to a sequence in a protein database using BLASTX (Altschul et al., 1997), second, a match to cDNAs or, third, prediction of a gene were, first, a match to a sequence in a protein database using GeneMark.hmm (http://genes.mit.edu/GENSCAN.html) and GeneMark (http://opal.biology.gatech.edu/GeneMark/eukhmnn.cgi). The criteria used to define a gene were, first, a match to a sequence in a protein database using BLASTX (Altschul et al., 1997), second, a match to cDNAs or, third, prediction of a gene were, first, a match to a sequence in a protein database using GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/eukhmnn.cgi). The criteria used to define a gene were, first, a match to a sequence in a protein database using GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/eukhmnn.cgi).


Llaca V, Messing J (1998) Amplification of maize zein genes are conserved within genic but expanded and constricted in intergenic regions. Plant Cell 10: 211–220


**Sequence Analysis of the Sorghum None Rph1 Complex Locus**
Ramakrishna et al.


