Different Types and Rates of Genome Evolution Detected by Comparative Sequence Analysis of Orthologous Segments From Four Cereal Genomes

Wusirika Ramakrishna,* Jorge Dubcovsky,† Yong-Jin Park,* Carlos Busso,† John Emberton,* Phillip SanMiguel‡ and Jeffrey L. Bennetzen*‡

*Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, †Department of Agronomy and Range Science, University of California, Davis, California 95616 and ‡Purdue University Genomics Core, WSLR, Purdue University, West Lafayette, Indiana 47907

Manuscript received April 18, 2002
Accepted for publication July 17, 2002

ABSTRACT

Orthologous regions in barley, rice, sorghum, and wheat were studied by bacterial artificial chromosome sequence analysis. General microcolinearity was observed for the four shared genes in this region. However, three genetic rearrangements were observed. First, the rice region contains a cluster of 48 predicted small nucleolar RNA genes, but the comparable region from sorghum contains no homologous loci. Second, gene 2 was inverted in the barley lineage by an apparent unequal recombination after the ancestors of barley and wheat diverged, 11–15 million years ago (mya). Third, gene 4 underwent direct tandem duplication in a common ancestor of barley and wheat 29–41 mya. All four of the shared genes show the same synonymous substitution rate, but nonsynonymous substitution rates show significant variations between genes 4a and 4b, suggesting that gene 4b was largely released from the strong purifying selection that acts on gene 4a in both barley and wheat. Intergenic retrotransposon blocks, many of them organized as nested insertions, mostly account for the lower gene density of the barley and wheat regions. All but two of the retrotransposons were found in the regions between genes, while all but 2 of the 51 inverted repeat transposable elements were found as insertions in genic regions and outside the retrotransposon blocks.

The grass (Poaceae) family of plants, including barley, maize, millets, oat, rice, rye, sorghum, and wheat, contributes ~60% of the world’s food production. Beyond their agronomic importance, these cereals also serve as a model system for comparative genetics (Bennetzen and Freeling 1993; Freeling 2001), wherein several species have highly developed molecular and genetic tool kits plus a long history of detailed physiological, developmental, and genetic characterization.

One of the most significant differences in grass genomes is their nuclear DNA content, ranging from <200 to >80,000 Mb (Bennetzen and Smith 1976; Bennett et al. 2000). Among the cereals, the approximate haploid DNA contents of rice, sorghum, maize, barley, and diploid wheat are 430, 750, 2500, 4800, and 5700 Mb, respectively (Arumuganathan and Earle 1991). In maize, we know that most (>60%) of the nuclear genome is composed of retrotransposons, often arranged as nested insertions within insertions (SanMiguel et al. 1996; San-Miguel and Bennetzen 1998; Meyers et al. 2001). In barley and wheat, it is also clear that much of the genome is composed of repetitive DNA and that much of this repetitive DNA consists of retrotransposons (Pans-truga et al. 1998; Shirasu et al. 2000; Stein et al. 2000; Wicker et al. 2001; Sanmiguel et al. 2002), but too few genic segments have been investigated to identify any consistent patterns of arrangement. The smaller grass genomes, including rice and sorghum, appear to have both a lower amount of repetitive DNA and fewer retrotransposons inserted between genes (Chen et al. 1997, 1998; Tikhonov et al. 1999; Klein et al. 2000; Tarchini et al. 2000). Further comparative DNA sequence analysis is needed to determine how these genomes differ in gene density and in the nature/organization of repetitive DNAs in genic regions.

Comparative genetic mapping of several cereal genomes using DNA markers has shown that all can be depicted as simple variants of a single genetic map (Moore et al. 1995; Gale and Devos 1998). However, a small number of large genomic rearrangements (commonly full arm translocations or inversions) differentiate many of the genomes, and a certain percentage of DNA markers (perhaps as high as 30%) do not fit into any clear colinear pattern (Bennetzen 2000). Many of the large rearrangements mark specific lineages, as in the case of three translocations that occurred in the Panicoideae ancestor that gave rise to maize, sorghum,
pearl millet, foxtail millet, and finger millet after their divergence from a common ancestor with barley, wheat, and rice \(\sim 50-60\) million years ago (mya; \textit{Gale} and \textit{Devos} 1998).

Despite the general colinearity of their genetic maps, we do not yet know the frequency or nature of the small rearrangements that differentiate cereal genomes. Any rearrangements smaller than a few centimorgans (e.g., several megabases) would have been missed by standard comparative genetic maps. In the grasses, comparative genomic sequencing studies that involved genomic segments >30 kb have been limited to maize, sorghum, and rice for the \(sh2/al\) region (\textit{Chen} et al. 1997, 1998), maize and sorghum for the \(adl\) region (\textit{Tikhonov} et al. 1999), and barley and rice for a region near the \(Vrn1\) gene (\textit{Dubcovsky} et al. 2001). These studies indicated that only the genes were conserved in these segments, but that some small rearrangements involving one or two genes had occurred also. Because so few segments and species were investigated in each study, no overall patterns could be discerned.

Comparative sequence analysis of small genomic regions (14–23 kb) at the \(Lrk/Tak\) loci in wheat and homologous regions from barley, maize, and rice uncovered a high gene density and numerous rearrangements (\textit{Feuillet} and \textit{Keller} 1999). This observation is not surprising for disease resistance genes of this type, which are organized in tandem clusters that tend to undergo rapid reorganization.

In all grasses studied, gene density has been higher than that predicted for a random dispersal of genes. Because investigators have always begun their studies by selecting a clone that contained a gene, an inherent bias was present in the gene-density outcome. Regardless, the observed gene densities in barley and wheat sequences (15–22 and 5–42 kb, respectively) are much lower than the 200–250 kb/gene that would be predicted by random dispersal (\textit{Keller} and \textit{Feuillet} 2000; \textit{Shirasu} et al. 2000; \textit{Dubcovsky} et al. 2001; \textit{Feuillet} et al. 2001; \textit{Wicker} et al. 2001). Cytogenetic studies also strongly support the existence of gene-rich islands in barley and wheat (\textit{Gill} et al. 1996a, b; \textit{Faris} et al. 2000; \textit{Kunzel} et al. 2000).

In dicotyledonous plants, comparative sequence analyses involving large genomic segments are restricted to comparisons with \textit{Arabidopsis thaliana}, a species that has undergone a very high frequency of ancestral rearrangements involving small chromosomal segments, primarily genic deletions (\textit{Lagercrantz} 1998; \textit{Blanc} et al. 2000; \textit{Ku} et al. 2000; \textit{O’Neill} and \textit{Bancroft} 2000; \textit{Vision} et al. 2000). Extensive chromosomal rearrangement is observed between \textit{Arabidopsis} and \textit{Brassica} species (\textit{O’Neill} and \textit{Bancroft} 2000). However, significant residual microcolinearity is observed between \textit{Arabidopsis} and its close relative, \textit{Capsella}, two species that diverged 6–10 mya (\textit{Rossberg} et al. 2001).

Genes in plants, and in other eukaryotes, can evolve at very different rates, perhaps due to differences in selective pressure or to different local rates of mutation (\textit{Wolfe} et al. 1987, 1989a, b; \textit{Gaut} 1998; \textit{Zhang} et al. 2001). For instance, \(adl\) was found to evolve faster than \(adh2\) at nonsynonymous sites and at similar rates at synonymous sites (\textit{Gaut} et al. 1999). Overall, little is known about substitution rate variations among nuclear genes of grass genomes and how this phenomenon may be related to local genome organization.

In this study, we present comparative sequence analysis of an orthologous chromosomal region from barley, rice, sorghum, and wheat. Our analyses indicate the frequency, nature, and lineages of several different types of genome evolution. These include variation in gene order and number, differences in the local rates of nucleotide variation, and patterns of transposable element accumulation. These combined studies provide a first indication of the value of a multi-species analysis of local genome structure and evolution in the \textit{Poaceae}.

**MATERIALS AND METHODS**

**BAC selection, restriction mapping, and sequencing:** Restriction fragment length polymorphism (RFLP) marker \textit{WG644} was used to screen the Morex barley artificial chromosome (BAC) library (\textit{Yu} et al. 2000), the \textit{HindIII} BAC library made from \textit{Nipponbare} rice DNA (http://www.genome.clemson.edu/orders/lib_desc/nippon.html), the \textit{HindIII} BAC library made from \textit{Bxs623} sorghum DNA (http://www.genome.clemson.edu/orders/lib_desc/SB_BBc.html), and the \textit{HindIII} BAC library made from \textit{DV} diploid wheat DNA (\textit{Ljajetzky} et al. 1999). The diploid wheat and sorghum libraries were also screened with probes corresponding to different genes of barley and rice BACs (\textit{Dubcovsky} et al. 2001) to identify the orthologous BAC. Positive BACs were fingerprinted with restriction enzyme \textit{HindIII}, transferred to nylon membranes, and hybridized to confirm that all BACs contained the homologous locus.

Restriction maps of \textit{36I5} (rice), \textit{635P2} (barley), \textit{116F2} (diploid wheat), \textit{115G1} (diploid wheat), and \textit{170F8} (sorghum) were constructed to experimentally validate computer sequence assembly. This experimental confirmation was important to determine the effect of large retroelements with large direct repeats on the assembly algorithms. BACs were individually digested with the 8-bp specificity restriction enzymes Asd, \textit{Nod}, \textit{Pac}, \textit{Pme}, and \textit{Sool}. All possible single and double digestions were analyzed for restriction enzymes with one or more sites within the mapped BAC. Restriction fragments were separated by pulsed-field electrophoresis, as described earlier (\textit{Dubcovsky} et al. 2001).

Preparation of shotgun libraries, sequencing, and analysis were as described by \textit{Dubcovsky} et al. (2001). For completing BAC sequences, gaps were closed by a combination of different approaches, including the use of different sequence chemistries, the thermodmelase enzyme, PCR amplification of gaps, shotgun sequencing of transposon-inserted subclones that span a gap, and direct sequencing of BAC template. When gaps were due to repetitive regions, subclones that either started or ended in unique regions with the remaining portion in the repetitive region were assembled separately and inserted into the main assembly.

**Sequence analysis:** Annotation and sequence analysis were performed as described earlier (\textit{Dubcovsky} et al. 2001).
FGENESH (http://www.softberry.com/nucleo.html) with the maize training set was used for gene prediction in addition to GENSCAN (http://genes.mit.edu/GENSCAN.html) and GeneMark.hmm (http://genemark.biology.gatech.edu/GeneMark/).

Estimation of nucleotide substitution rates and phylogenetic reconstructions: Genes were aligned using CLUSTALX (THOMPSON et al. 1997). Rates of nucleotide substitution were estimated using the distance measures of NEI and GOJOBORE (1986) and the Jukes-Cantor correction as implemented in the MEGA2 (molecular evolutionary genetic analysis) package (KUMAR et al. 2001). Phylogenetic reconstructions were performed by the neighbor-joining method, with synonymous sites for the analyses. Synonymous and nonsynonymous substitution rates were estimated as described by GAUT et al. (1996). Divergence and duplication times (T) were estimated for all genes, except those shown by relative rate tests to be evolving at different rates, using $k = K_s/2T$. $k$ is the absolute rate of synonymous substitution/site/year; $K_s$ is the estimated number of synonymous substitutions per site between homologous sequences and from the neighbor-joining trees generated using the MEGA2 package.

Relative rate tests: Relative rate tests were used to assess heterogeneity in the numbers of substitutions per site estimated from the one-parameter method (Jukes and CANTOR 1969). Tajima’s relative rate test was used as implemented in MEGA2 for testing the molecular clock hypothesis because it can be applied even when the substitution rate varies among different sites (TAJIMA 1993). We used this test as implemented in MEGA2 to assess the molecular clock hypothesis. It compares two sequences with an outgroup sequence and counts the number of unique substitutions in both lineages. When one of the two sequences accumulates a significantly larger number of substitutions, it indicates that the molecular clock hypothesis can be rejected.

RESULTS

Isolation and sequencing of orthologous barley, rice, sorghum, and wheat BACs: To study genomic organization near the vrn1 locus, BACs Hv635P2, Os36I5, Sb170F8, and Tm115G1/Tm116F2 were selected by their hybridization to DNA marker WG644. WG644 had been mapped to orthologous regions in rice, barley, and wheat (KLEINHOFS et al. 1993; DUBCOVSKY et al. 1998; SARMA et al. 1998). Colinearity between the long arm of rice chromosome 3 and the long arm of homeologous group 5 in the Triticeae has been studied extensively because of the presence of vernalization and frost tolerance genes Vrn1 and Frl in wheat and barley and the heading date gene hso6 in rice (DUBCOVSKY et al. 1998; SARMA et al. 1998; KATO et al. 1999; SUTKA et al. 1999). The DNA on the sorghum BAC has not been mapped to the sorghum genetic map, but all selected sorghum clones formed a single contiguous set, indicating that there was only one homologous region in sorghum.

Previously, we reported the sequence of one of these BACs (barley clone Hv635P2) and part (50 kb) of the rice BAC Os36I5 (DUBCOVSKY et al. 2001). Transposable element composition, dinucleotide arrangement, and recombination frequencies have also been reported for the wheat region (SANMIGUEL et al. 2002). The three BACs (two wheat, one sorghum) and an extra 21.3 kb of the rice BAC were sequenced (accession nos. AY013245, AY013245, AY099491, AF053433, and AF459639) by a shotgun approach like that described earlier (DUBCOVSKY et al. 2001). The final error rate was <1 bp/10 kb and the consensus sequence was of high quality (PHRED value of >25). The two diploid wheat (Triticum monococcum) BACs formed a 215-kb contiguous sequence with 20,573 bp of overlap. The genome sequence from BACs Hv635P2, Sb170F8, and Tm115G1/116F2 are 102,433, 142,376, and 215,220 bp, respectively. The insert in BAC Os36I5 is ~75 kb. This BAC has two regions of contiguous sequence of 65.5 and 5.8 kb, with a gap of ~3.7 kb that we have been unable to close (Figure 1). In every case, restriction maps of the BAC clones agree completely with the sequence assemblies.

Gene content and organization in the sequenced clones: As previously reported (DUBCOVSKY et al. 2001), barley BAC Hv635P2 contains five predicted genes, four long terminal repeat (LTR) retrotransposons, a solo LTR of BAGY-2 (SHIRASU et al. 2000), a LINE retroponson, at least nine miniature inverted repeat repeat transposable elements (MITEs), and a Mutator-like transposable element (Figure 1). Two inverted repeats flank barley gene 2 and a novel transposable element, Inysub, is inserted in the rightward inverted repeat.

Rice BAC Os36I5 has eight predicted genes, one LTR retrotransposon, one Mutator transposable element, and at least 12 MITEs. Gene 6 is a hypothetical gene with no significant similarity to any known gene, protein, or expressed sequence tag (EST), but has been postulated to be a gene by all three gene prediction programs. The predicted protein product of gene 7 shows highest homology to an Arabidopsis unknown protein (NP189619, 1e-42). Gene 7 of rice is orthologous to sorghum gene 6. The predicted gene 8 protein product shows highest homology to an Arabidopsis ribosomal protein (NP199657, 3e-45) and was also conserved as sorghum gene 7. However, gene 8 appears to be a pseudogene in rice because it has a stop codon in the first predicted coding exon. In addition, exon 4 (the last exon) is missing in the rice BAC. Downstream of this region, a single 1.9-kb segment is repeated four times (94.8–99% sequence identity). A fifth repeat is truncated at the end and extends into the unsequenced gap. This 1.9-kb repeat shows homology to rice small nucleolar RNA (snoRNA) genes that are organized in a cluster (AJ310377). Each cluster in the rice BAC, Os36I5, has 10 snoRNA-like genes that show homology (93–99%) to snoRNA genes ranging from 79 bp (AJ507952) to 188 bp (AJ320263). The last cluster contains eight candidate snoRNA genes. snoRNAs are small in size and essential for processing ribosomal RNAs (LEADER et al. 1997, 1999; BROWN et al. 2001). Plant snoRNA gene clusters are transcribed as a polycistronic pre-snoRNA transcript (LEADER et al. 1997, 1999). A comparable snoRNA gene cluster is not present in the orthologous region in the Triticeae.
Figure 1.—Microcolinearity of an orthologous region in barley, rice, sorghum, and diploid wheat. Arrows indicate genes, plus their size and proposed direction of transcription. Nested insertions of retrotransposons are shown in barley and diploid wheat. Asterisks indicate simple sequence repeats that contain tandem repeat numbers of eight or more. MITEs are indicated by small triangles. In the rice BAC, a dashed line represents a region with homology to a cluster of rice small nucleolar RNA genes.
TABLE 1

Predicted genes in BACs Os36I5 and Sb170F8

<table>
<thead>
<tr>
<th>Gene Homology Accession no.</th>
<th>E value</th>
<th>ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice BAC Os36I5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Hypothetical gene (A. thaliana) NP189619</td>
<td>2e-56</td>
</tr>
<tr>
<td>7</td>
<td>Unknown protein (A. thaliana) NP199657</td>
<td>3e-45</td>
</tr>
<tr>
<td>8</td>
<td>60S ribosomal protein L13a (A. thaliana) NP186862</td>
<td>e-101</td>
</tr>
<tr>
<td>Sorghum BAC Sb170F8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Unknown protein (A. thaliana) NP189619</td>
<td>3e-57</td>
</tr>
<tr>
<td>6</td>
<td>Unknown protein (A. thaliana) NP199657</td>
<td>e-101</td>
</tr>
<tr>
<td>7</td>
<td>Putative protein kinase (A. thaliana) NP186862</td>
<td>2e-56</td>
</tr>
<tr>
<td>8</td>
<td>Putative protein kinase (A. thaliana) NP186862</td>
<td>3e-45</td>
</tr>
<tr>
<td>9</td>
<td>Expresssed protein (A. thaliana) NP564059</td>
<td>2e-25</td>
</tr>
<tr>
<td>10</td>
<td>Expresssed protein (A. thaliana) NP566244</td>
<td>4e-70</td>
</tr>
<tr>
<td>11</td>
<td>Hypothetical protein (rice) AAK16187</td>
<td>2e-64</td>
</tr>
<tr>
<td>12</td>
<td>Unknown protein (rice) AAK16184</td>
<td>1e-25</td>
</tr>
<tr>
<td>13</td>
<td>Putative glucosyl transferase (rice) AAK16180</td>
<td>1e-170</td>
</tr>
<tr>
<td>14</td>
<td>Putative glucosyl transferase (rice) AAK16175</td>
<td>1e-155</td>
</tr>
<tr>
<td>15</td>
<td>Putative glucosyl transferase (rice) AAK16175</td>
<td>1e-155</td>
</tr>
<tr>
<td>16</td>
<td>Putative glucosyl transferase (rice) AAK16175</td>
<td>1e-155</td>
</tr>
<tr>
<td>17</td>
<td>Putative glucosyl transferase (rice) AAK16172</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>Putative glucosyl transferase (rice) AAK16194</td>
<td>e-115</td>
</tr>
</tbody>
</table>

Predicted protein products of barley, rice, sorghum, and wheat genes 1, 2, 3, and 4 show homology to mitochondrial carrier protein, cleavage stimulation factor subunit 1, serine/threonine kinase, and ABC-transporter-like protein, respectively (Dubcovsky et al. 2001). Genes 7 and 8 of rice are colinear with genes 6 and 7 of sorghum. Genes 11–20 of sorghum are colinear with a segment of a rice BAC (accession no. AC079887) that maps close to Os36I5. Rice BAC Os36I5 has five clusters with homology to snoRNA genes. Each cluster has 10 snoRNA-like genes except the last cluster, which has 8 snoRNA-like genes.

* Hypothetical gene.
* Truncated genes.
* Stop codon in the middle of the truncated gene.
* Partial gene located at the end of the BAC.

The sorghum genomic segment in BAC Sb170F8 contains 20 predicted genes, two LTR retrotransposons, and at least 20 MITEs (Figure 1). At the most-leftward end of the BAC is a partial retrotransposon (Figure 1). The predicted gene 5 of sorghum shows 100% identity to a sorghum EST (BE592631). The sorghum BAC also has a cluster of putative glucosyl transferase genes (genes 13–19). Interestingly, all but genes 8–11 are in the same transcriptional orientation on this BAC. In fact, for all BACs investigated, only these 4 genes and gene 2 on the barley BAC are in the opposite transcriptional organization from all other genes (Figure 1).

The contiguous sequence contained on diploid wheat BACs Tm115G1 and Tm116F2 harbors five predicted genes and 21 intact or partially deleted LTR retrotransposons. As in barley, gene 4 is tandemly duplicated. Within the 215-kb wheat sequence, the five genes are arranged in clusters of 13 and 29 kb, containing two and three genes, respectively (Figure 1). Within these genomic regions, the average gene density is one gene per 8.4 kb.

Mobile DNA organization: The majority of the retrotransposons in wheat were similar to those previously identified in barley and wheat (Panstruga et al. 1998; Shirasu et al. 2000; Wicker et al. 2001), although we discovered and named six new elements (Eway, Latidu, Miuse, Nusif, Veju, and Wham). We also discovered new retrotransposons in barley (Inav, Ikeros, and Sedef), rice (Alulu), and sorghum (Pyrubu and Unum). Over one-half of the retrotransposons in the wheat BACs are organized in a nested fashion like that seen in the maize genome (Sanmiguel et al. 1996). The 21 LTR retrotransposons comprise ~160 kb of the 215-kb region. These retroelements were named by the approach described in Sanmiguel et al. (2002). Nested insertions of LTR retrotransposons were found only on wheat BACs, except in the case of barley BAC Hv635P2, where one BARE-1 element was inserted in the opposite orientation into another BARE-1 element.
BARE-1-like retrotransposons have been named Angela in wheat (Wicker et al. 2001) and constitute ~11% of the wheat region. Solo LTRs were not found in any of the sequenced regions except a BAGY-2 solo LTR in barley. Subrina-like and Wham LTR retrotransposons constitute 10 and 9% of the sequenced wheat region, respectively. Overall, there are nine largely intact LTR retrotransposons in addition to several partial elements in the wheat region. Angela elements appear to show the most recent insertional activity, because none of them contain another inserted LTR retrotransposon. Five of the identified retroelements (not including the two partial Angela elements at the ends of the BAC contig) were partly deleted. Sorghum BAC SB170F8 contains two novel LTR retrotransposons, Unum and Pyrubu, and a partial LTR retrotransposon that together account for ~15% of the 142-kb BAC insert. Rice BAC Os36I5 contains a retrotransposon, Atalu. All retrotransposons except one (Unum inserted in intron 13 of sorghum gene 3) were found in the intergenic regions.

MITEs are equally numerous, relative to genes, in all regions. Identified MITEs include 9 in barley, 12 in rice, 20 in sorghum, and 10 in wheat. On average, there are 2 MITEs per gene in barley, rice, and wheat. In sorghum, the ends of at least eight more MITE-like elements could not be determined. With two exceptions (one each in barley and wheat), all MITEs were found as insertions in the generic regions. A MITE is inserted in the LTR retrotransposon Inav in barley, and a second MITE is inserted in a partial element (downstream of the first partial Angela) in wheat that has homology to RIRE2.

Comparison of orthologous barley, rice, sorghum, and wheat regions: Four predicted genes are conserved across the four genomes in this study. However, they are distributed across 102 kb in barley, 30 kb in rice, 35 kb in sorghum, and 215 kb in diploid wheat (Figure 2).

Adjacent to the four conserved genes, the sorghum BAC shares two additional genes (6 and 7) with rice BAC Os36I5 (Figure 2). In addition, the predicted protein product (116 amino acids) of a divergent gene (gene 6) in rice has a small stretch of similarity (63% identity in a region of 19 amino acids) to the predicted protein product (117 amino acids) of sorghum gene 5. Sorghum genes 11–20 are collinear with a segment of a rice BAC (AC079887) in the GenBank database. This BAC is downstream from Os36I5 and includes RFLP marker R2404, which is 0.3 cM from marker R2311 located in Os36I5 on chromosome 3. The cluster of putative glucosyl transferase genes (genes 13–19) in the sorghum BAC, Sb170F8, is also present in the rice BAC (AC079887). Sorghum has seven of these genes, compared to five in rice.

Gene 2 is in inverted orientation in barley relative to rice, sorghum, and wheat. This indicates that the inversion occurred in the barley lineage after wheat and barley ancestors diverged. Gene 4 is duplicated in both barley and wheat, suggesting that this duplication occurred before the divergence of barley and wheat ancestors but after their divergence from rice and sorghum ancestors. Alternatively, the duplicated gene may have been lost from the rice and sorghum lineages.

Intron-exon and exon-intron boundaries were analyzed for the four genes that were common among the four genomes. These four genes were also compared to the most closely related genes in Arabidopsis. In the grasses, exon-intron boundaries predominantly exhibit the sequence GT/A (56%), followed by GT/G (26%) as the next most frequent. This compares to 62% GT/A and 14% GT/G in Arabidopsis. Intron-exon boundaries are C/AG (78%) or T/AG (22%) in the grass genomes, compared to 63% C/AG and 29% T/AG in the putative orthologous genes in Arabidopsis. Hence, the grasses show a narrower range of variation in intron splicing sites than is seen in Arabidopsis.

Expansion of intergenic spaces was largely caused by LTR retrotransposon insertion in barley and wheat (Figure 2). Retrotransposons account for ~60, 8, 15, and 70% of the sequenced barley, rice, sorghum, and wheat regions, respectively. However, none of the retrotransposons, MITEs, or other mobile elements were in orthologous locations, except a small truncated element of 752 bp in Hv635P2, which was conserved in Tm115G1/116F2 (84% identity; Figure 1) and was similar (84–89% identities) to parts of the LTR of the Barbara element (AF326781; Wicker et al. 2001). Hence, almost all these elements were inserted after the ancestors of these species diverged from each other.

Detailed comparison of the homologous barley and wheat regions: Exons and introns in the five common genes between wheat and barley were well conserved, facilitating a comprehensive analysis of these sequences. The coding regions of the five genes were very similar, varying from 95.9 to 97.5% identity at the DNA level. The five predicted exon-intron boundaries found within the sequenced regions except a small truncated element of 10 MITEs or MITE-like elements, barley and wheat, suggesting that this duplication occurred before the divergence of barley and wheat ancestors but after their divergence from rice and sorghum ancestors. Alternatively, the duplicated gene may have been lost from the rice and sorghum lineages.

A different proportion of transitions and transversions was observed between introns and exons. In the exons, 66% of the 273 point mutations were transitions and 34% were transversions. In the introns, 58.3% of the 1284 point mutations were transitions and 41.7% were transversions. An analysis of variance of the transition proportions using the five genes as replicates indicated that this difference was significant at $P < 0.05$.

The unaligned portions of the introns (20%) were caused by insertions of 10 MITEs or MITE-like elements, 262 small indels (<15 bp), 19 intermediate indels (15–
Figure 2.—Comparison of an orthologous region in barley, rice, sorghum, and diploid wheat. The compactness of rice and sorghum genomes is evident compared to barley and diploid wheat genomes. Lines connect the genes that are common among the BACs. Gene 5 of sorghum and gene 6 of rice (connected by a dotted line) show limited sequence homology and may not be truly orthologous. Large transposons are shown as blocks. Arrows indicate genes and their transcriptional orientation.

200 bp), two large indels (395 and 1532 bp), and three regions of gene 4b (introns 1, 2, and 16) that showed unusually low levels of conservation and could not be accurately aligned. The presence of 10 MITEs or MITE-like elements in these five genes indicates an average insertion rate of approximately one insertion/gene/evolutionary lineage/10 million years. The predicted MITEs (or remainders of partially deleted MITEs) varied in length from 17 to 226 bp and all showed duplication of short host sequences (generally TA) and the presence of perfect or imperfect inverted repeats. These apparent insertions accounted for an increase of 1249 bp in the size of the introns.

Most of the other indels were small, including 1 bp (32%), 2 bp (22%), 3 bp (10%), or 4 bp (8%) events. However, the largest amount of indel size variation in these 66 introns was provided by the 19 intermediate (831 bp total) and 2 large (1927 bp total) indels. To understand the possible origin of these indels, we compared their flanking regions. Forty-four percent of the indels of three or more base pairs included perfect short direct repeats in one border of the indel and in the opposite border of the paired region between barley and wheat. This proportion increased to 65% with the inclusion of direct repeats with 1 bp difference among repeats (>4 bp) or 1–2 bp away from the exact border of the indel.

Nucleotide substitution rates: Synonymous ($K_s$) and nonsynonymous ($K_a$) substitution rates within a gene are often correlated (Wolfe and Sharp 1993). The $K_a/K_s$ ratio indicates the level of selective constraint acting on proteins. For most of the comparisons for genes 1–4, $K_s$ and $K_a$ were also correlated ($r = 0.82$, $P < 0.0001$; Table 2). The data show that gene 4 is highly constrained in its nonsynonymous substitution rate in rice and sorghum, but only gene 4a is similarly constrained in barley and wheat (Figure 3). In fact, gene 4a shows a decrease in its nonsynonymous substitution rate in the barley-wheat comparison, resulting in a $K_a/K_s$ ratio of 0.03 that indicates an increase in purifying selection.

To estimate whether the four genes conserved among the four genomes are evolving at different rates, two-tailed t-tests were performed for different gene pairs. $P$ values were significant at a 5% level only for comparisons of $K_s$ values of gene 4b with genes 1, 3, and 4a after correction for multiple tests ($P < 0.002, \alpha = 0.0051$). Comparisons of $K_s$ for different gene pairs showed 5 out of 10 comparisons to be statistically significant at a 5% level after correction for multiple tests ($P < 0.0037, \alpha = 0.0051$). This suggests more variable nonsynonymous substitution rates than synonymous substitution rates among these four genes.

To estimate the time of duplication of gene 4 and the divergence of wheat and barley, the synonymous substitution rates must not differ significantly and the duplicated sequences must evolve at similar rates after duplication. Tajima’s relative rate test (Tajima 1993)
was used to scan for substitution rate heterogeneity in the four genes conserved in the four grass genomes. Rice was taken as an outgroup to estimate relative rates of genes 1–4 in barley and wheat. Relative rate tests showed that synonymous substitution rate variations were not significant for genes 1–4. However, highly significant variations in nonsynonymous substitution rates were observed between genes 4a and 4b. Therefore, synonymous substitutions were used to estimate divergence times. Relative rate tests and nucleotide substitution rates show that gene 4b is evolving at a faster rate compared to gene 4a (Figure 3 and Table 3).

Rice fossils dated to ~40 mya were described by Stebbins (1981). Molecular data indicate that the divergence of the rice lineage from the lineages giving rise to barley, sorghum, and wheat occurred 50–70 mya (Wolfe et al. 1989a). Therefore, 60 mya was taken as the divergence time between rice and barley/wheat. We used average synonymous substitution rates for the rice and barley/wheat divergence to estimate absolute synonymous substitution rates. The average synonymous substitution rate of the adh 1 and adh 2 genes in grasses was reported to be $6.5 \times 10^{-9}$ substitutions/synonymous site/year (Gaut et al. 1996). The average for six nuclear genes (adh 1, adh 2, waxy, shrunken 1, gapC, and chalcone synthase) between maize and wheat or barley was reported to be $6.1 \times 10^{-9}$ substitutions/synonymous site/year (Wolfe et al. 1989b). Our estimate is $4.2 \times 10^{-9}$ substitutions/synonymous site/year based on genes 1–4 of rice and wheat and a divergence time of 60 mya. On the basis of our estimates, the ancestors of barley and wheat diverged from each other 11–15 mya, which is in a similar range (10–14 mya) to that estimated by Wolfe et al. (1989a).

Using our calculated rate of sequence divergence, we determined that the duplication of gene 4 in a Triticeae ancestor occurred 29–41 mya, many millions of years after this lineage diverged from ancestral lineages that gave rise to rice or sorghum. Hence, our data are more consistent with a duplication of gene 4 that occurred specifically in a shared ancestor of barley and wheat than it is with a model proposing deletion of one gene 4 copy from the rice/sorghum or rice and sorghum lineages.

### Discussion

#### Types and times of genic rearrangement:

Comparative sequence analysis can provide a wealth of information about the nature of sequence arrangement and evolution, including gene content, order, and orientation. Microcolinearity among grass genomes has been shown by the sequencing of genomic segments from orthologous loci from rice, maize, sorghum, barley, and wheat (Chen et al. 1997; Tikhonov et al. 1999; Dubcovsky et al. 2001; Feuillet et al. 2001). In the present study, comparative sequence analysis of orthologous genomic segments of barley, rice, sorghum, and wheat revealed two small genic rearrangements. Gene 2 was inverted in barley relative to the other three genomes studied, probably by unequal recombination between inverted repeats flanking the gene, after the ancestors of barley and wheat diverged from each other. It is likely that the duplication of gene 4 occurred before the divergence of barley and wheat, because the duplicated genes are present in these two genomes but absent in rice and sorghum.

One dramatic variation between the studied regions is a difference in the presence of an entire duplicated gene family in the comparison of rice and sorghum.
Rice BAC Os36I5 contains at least 48 candidate snoRNA genes in five clusters, none of which are present in the comparable sorghum region. This lack of colinear location is also observed with rDNA repeats, storage protein gene clusters, and tandem disease resistance genes (Dubcovsky and Dvorak 1995; Leister et al. 1998; Richly et al. 2002; J. Messing, personal communication). It is not clear why tandem gene clusters should be less stable in genomic location. However, when this property is shared by several different types of clustered genes, it suggests that this is a consistent and evolved property of higher plant genomes.

Distribution and density of genes and repeats: Gene density in species from the Triticeae tribe with large genomes such as barley and wheat is of immense interest. The large difference between expected and observed gene density in barley and wheat (Keller and Feuillet 2000) and studies based on GC composition (Barakat et al. 1997) and cytogenetics (Gil et al. 1996a,b; Faris et al. 2000) support the hypothesis that gene-rich regions exist in these large genomes. Previous estimates of gene density of the wheat genome were mostly based on disease resistance genes (Stein et al. 2000; Feuillet et al. 2001; Wicker et al. 2001), which are largely exceptional in their organization and lack of stability (Leister et al. 1998; Hulbert et al. 2001; Richly et al. 2002). The map position of Vrn1 places it in a region (between breakpoints in deletion lines 5AL-6 and 5AL-17) that undergoes a moderate level of recombination (Gill et al. 1996b) and, thus, it should be an average genic region of the wheat genome. In our study, sorghum had the highest gene density of one gene per 7 kb, while rice (excluding the snoRNA loci), barley, and diploid wheat had gene densities of one gene per 9, 20, and 43 kb, respectively. These observed gene densities compare to predicted respective gene densities of one per 25, 15, 160, and 190 kb for a random gene dispersal model if each of these species contains ~30,000 genes. Hence, all species show higher-than-random gene densities, but this effect is much more pronounced for the large genomes. This result suggests that most of the larger plant genomes are composed of largely gene-free...
regions, like paracentromeric heterochromatin, as is also suggested by cytogenetic studies (Gill et al. 1996a,b; Faris et al. 2000).

We identified numerous novel LTR retrotransposons in the sequenced regions and found that retrotransposons constitute >70% of the 215-kb region sequenced in wheat. Retrotransposons and genes are organized in separate clusters in wheat and to a lesser extent in barley. The most striking features in the wheat region and the nested insertions of retrotransposons. Hence, the preferential insertion of the abundant classes of retrotransposons into each other reported in maize (Sanmiguel et al. 1996) appears to be a general characteristic of the wheat genome. In addition, we found that MITEs in wheat and barley also show the preferential accumulation near genes that has been reported in maize (Bureau and Wessler 1994; Tikhonov et al. 1999). As in maize (Sanmiguel et al. 1998), all of the LTR retrotransposons discovered in this study appear to be of recent origin, partly evidenced by the lack of orthologous elements inserted in the barley and wheat regions. This may reflect a recent explosion in LTR retrotransposon activity in these lineages (accounting for the large overall genome size) or may indicate relatively rapid removal of older transposon sequences (Devos et al. 2002).

Mutations, selection, and drift: Comparison of barley and wheat genic regions revealed several interesting features. The ubiquitous presence of direct tandem repeats in the indels suggests that replication slippage, illegitimate recombination (Devos et al. 2002), and/or transposable element excision were responsible for these indels. If all the indels >4 bp and flanked by perfect repeats were considered scars from transposon activity, the previous estimate of the rate of transposon insertion would change from 1 to 2.2 (or 2.6 if imperfect repeats are also included) insertion per gene per evolutionary lineage per 11 million years. If the indels (including the two large indels of uncertain origin) are assumed to be deletions, they would have removed at least 3.5 kb of an 18.5-kb region of aligned intron sequences in the last 11–15 million years. These deletions could provide an adequate counterbalance to the increase in size originated by transposon insertion in barley, wheat, and other plant species.

In the present study, we compared the modes and rates of evolution of four orthologous genes that are next to each other in the four cereal genomes. Our data indicate that the synonymous substitutions evolve at a more uniform rate than nonsynonymous substitutions, not only among different genes but also in different lineages.

Rice-sorghum and wheat-barley (Triticeae) are separately derived from ancestors that diverged 50–70 mya (Wolfe et al. 1989a). We used this date and synonymous substitutions to estimate divergence and duplication times. Genes 1–4 show 4.2 × 10⁻⁷ substitutions per synonymous site per year, which is close to that reported previously for grass genomes (Wolfe et al. 1989b; Gaut et al. 1996). Our data suggest that the duplication of gene 4 occurred 29–41 mya. This conclusion is also supported by the absence of the duplicated gene 4 in the orthologous region in both rice and sorghum. However, the duplication may have happened earlier, if it was followed by rounds of unequal gene conversion.

Nonsynonymous substitutions tend to be uninformative over short evolutionary time periods, partly because they are subject to positive selection. We found that the nonsynonymous substitution rates deviated significantly from clock-like behavior after duplication of gene 4, as revealed by relative rate tests. Similar variations were observed between the two duplicated grass adh loci (Gaut et al. 1996). In contrast, several of the 26 groups of duplicated genes analyzed in zebrafish showed significant differences in the rate of evolution as measured by both $K_s$ and $K_a$ (Van de Peer et al. 2001). In nine genomes previously studied (including Arabidopsis and rice), duplicated genes were found to arise at a rate averaging 0.01 duplications/gene/million years (Lynch and Conery 2000). However, these duplicated genes often have a short half-life, 23.4 million years, for instance, in Arabidopsis (Lynch and Conery 2001). Biological innovation and functional diversification can lead to evolution of new functions for duplicated genes. In fact, gene 4a seems to be the most conserved locus of the four genes compared, while gene 4b is the least conserved. Thus, these duplicated genes appear to be on different evolutionary routes that may lead (or may have already led) to a new role for gene 4b.

Small-scale gene rearrangements have been found in other eukaryotic genome comparisons and can be a major factor in speciation and genome evolution (Wolfe and Shields 1997; Wagner 2001). The very high rate of genic rearrangement observed in this study and other

### Table 3

Tajima’s relative rate tests

<table>
<thead>
<tr>
<th></th>
<th>Syn⁴</th>
<th>Nonsyn⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>0.17</td>
<td>0.5</td>
</tr>
<tr>
<td>Gene 2</td>
<td>0.18</td>
<td>0</td>
</tr>
<tr>
<td>Gene 3</td>
<td>2.38</td>
<td>0</td>
</tr>
<tr>
<td>Gene 4a</td>
<td>0.42</td>
<td>0</td>
</tr>
<tr>
<td>Gene 4b</td>
<td>0.51</td>
<td>1</td>
</tr>
<tr>
<td>Barley G4a-b</td>
<td>4.31*</td>
<td>21.49*</td>
</tr>
<tr>
<td>Wheat G4a-b</td>
<td>3.85</td>
<td>23.2</td>
</tr>
</tbody>
</table>

Barley and wheat are ingroups with rice as an outgroup except in the last two cases where the ingroups are barley genes 4a and 4b and wheat genes 4a and 4b, respectively. Syn⁴, synonymous substitution rate; Nonsyn⁴, nonsynonymous rate.

* Significant at 5% level.

** Significant at 0.1% level.
plant genome comparisons (reviewed in BENNETZEN and RAMAKRISHNA 2002) indicates that plants have much more unstable genomes than those seen in mammals. This instability can provide the extensive haplotype variability characteristic of most plants, providing broad opportunities for the action of natural selection. Further studies are warranted to provide a more comprehensive indication of the nature, mechanisms, rates, and linkages of genome rearrangement in plants.

This work was supported by the National Science Foundation Plant Genome Program (grant no. 9975793) and United States Department of Agriculture-National Research Initiative grant no. 2000-1678.

LITERATURE CITED


KATO, K., H. MIURA and S. SAWADA, 1999 Comparative mapping of the wheat Vrn1 region with the rice Hd-6 region. Genome 42: 204–209.


LAGERCRANTZ, U., 1998 Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetcs 150: 1217–1228.

LEADER, D. J., G. P. CLARK, J. WATTERS, A. F. BEVEN, P. J. SHAW et al., 1997 Clusters of multiple different small nuclear RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. EMBO J. 16: 5742–5751.


Sanmiguel, P., and J. L. Bennetzen. 1998  Evidence that a recent increase in maize genome size was caused by the massive amplification of intergenic retrotransposons. Ann. Bot. 82: 37–44.


Communicating editor: J. A. Birchler