Long-Range and Targeted Ectopic Recombination between the Two Homeologous Chromosomes 11 and 12 in Oryza Species

J. Jacquemin,*†1 C. Chaparro,1 M. Laudie,1 A. Berger,2 F. Gavory,2 J. L. Goicoechea,3 R. A. Wing,3 and R. Cooke1

1Laboratoire Génome et Développement des Plantes, Unité Mixte de Recherche Centre National de la Recherche Scientifique/Institut de Recherche pour le Développement/Université de Perpignan Via Domitia, Université de Perpignan, 66860 Perpignan-Cedex, France
2Genoscope-Centre National de Séquençage, Evry, France
3Arizona Genomics Institute, The University of Arizona
†Present address: Arizona Genomics Institute, The University of Arizona

*Corresponding author: E-mail: juliej@cals.arizona.edu.
Associate editor: Manolo Gouy

Abstract

Whole genome duplication (WGD) and subsequent evolution of gene pairs have been shown to have shaped the present day genomics of most, if not all, plants and to have played an essential role in the evolution of many eukaryotic genomes. Analysis of the rice (Oryza sativa ssp. japonica) genome sequence suggested an ancestral WGD ~50–70 Ma common to all cereals and a segmental duplication between chromosomes 11 and 12 as recently as 5 Ma. More recent studies based on coding sequences have demonstrated that gene conversion is responsible for the high sequence conservation which suggested such a recent duplication. We previously showed that gene conversion has been a recurrent process throughout the Oryza genus and in closely related species and that orthologous duplicated regions are also highly conserved in other cereal genomes. We have extended these studies to compare megabase regions of genomic (coding and noncoding) sequences between two cultivated (O. sativa, Oryza glaberrima) and one wild (Oryza brachyantha) rice species using a novel approach of topological incongruency. The high levels of intraspecies conservation of both gene and nongene sequences, particularly in O. brachyantha, indicate long-range conversion events less than 4 Ma in all three species. These observations demonstrate megabase-scale conversion initiated within a highly rearranged region located at ~2.1 Mb from the chromosome termini and emphasize the importance of gene conversion in cereal genome evolution.

Key words: comparative genomics, duplication, gene conversion, Oryza genus, recombination hot spot.

Introduction

The availability of genome sequences from closely related species, such as yeasts (reviewed in Dujon 2010) or Drosophila (Hahn et al. 2007), has led to considerable advances in our understanding of genome evolution. In plants, the Oryza Map Alignment Project (OMAP, Wing et al. 2005), articulated around the reference Oryza sativa ssp. japonica c.v. Nilpponbare genome sequence (hereafter RefSeq) has developed resources aimed at characterizing rice genome evolution. In a genus containing two cultivated and 22 wild species, molecular resources have been created representing the ten genome types and which provide the means of studying short-term evolutionary dynamics in plants. This has allowed deep comparative analysis of these closely related species at specific loci (Lu et al. 2009; Sanyal et al. 2010).

The importance of duplications in the evolution of plant genomes has been emphasized by the analysis of several complete genome sequences (Vanth de Peer et al. 2009). Preliminary analysis of the rice RefSeq suggested a whole genome duplication (WGD), probably common to all grasses and a more recent segmental duplication of ~2–3 Mb in the distal region of the short arms of chromosomes 11 and 12 (Yu et al. 2005; Rice Chromosomes 11 and 12 Sequencing Consortia 2005). More recent studies by ourselves (Jacquemin et al. 2009) and others (Paterson et al. 2009) demonstrated that this duplicated block is not specific to the Oryza genus, as its presumed age suggested, and this is confirmed by its presence in two other model cereal genomes, Sorghum bicolor and Brachypodium distachyon. As chromosomes 11 and 12 result from the WGD at the base of the Poaceae, this strongly suggests that this duplication has the same origin. Wang et al. (2007), comparing 278 gene pairs along the whole 11–12 block in the RefSeq and the indica subspecies sequence, proposed a stochastic evolution of gene pairs in this region, in which gene conversion acts as an occasional, sometimes frequent interruption to independent evolution of paralogs. Our study (Jacquemin et al. 2009) on a wider sampling of species within and closely related to the Oryza genus rather indicated recurrent concerted evolution affecting the same gene pairs in all species, at least in the immediate
subtelomeric region and suggested a breakpoint in colinearity at ~2 Mb from the telomeres.

Gene conversion is the nonreciprocal transfer of genetic information between homologous sequences, leading to homogenization during meiotic or mitotic recombination (Szostak et al. 1983). Four pathways to repair DNA double strand breaks (DSBs) through homologous recombination (HR) are generally grouped under the term of gene conversion (reviewed in Chen et al. 2007; Llorente et al. 2008; De Muyt et al. 2009; Duret and Galtier 2009): double-strand break repair (DSBR), double-Holliday junction (HJ) dissolution, synthesis-dependent strand annealing (SDSA), and break-induced replication (BIR). Ectopic gene conversion involves dispersed duplicated sequences rather than sister chromatid cohesion (reviewed in Chen et al. 2007; Llorente et al. 2008; De Muyt et al. 2009; Duret and Galtier 2009), double-strand break repair (DSBR), double-Holliday junction (HJ) dissolution, synthesis-dependent strand annealing (SDSA), and break-induced replication (BIR). Ectopic gene conversion involves dispersed duplicated sequences rather than sister chromatid cohesion.

Materials and Methods
A detailed version of all Methods is available in supplementary text S1 (Supplementary Material online).

Sequencing, Assembly, and Contig Annotation
BAC contigs were defined using SyMAP (Soderlund et al. 2006) and refined manually. Lengths of assembled contigs are reported in table 1. Annotation was carried out using available tools and in-house Perl scripts, gene models being refined in Artemis (Rutherford et al. 2000). Overall statistics are presented in table 1.

Comparative Structural Analysis
Sequence conservation and rearrangement was analyzed with Dotter (Sonnhammer and Durbin 1995) using default parameters and with the Artemis comparison tool (ACT, Carver et al. 2008) for small rearrangements.

Inference of Paralogous Pairs and Homologous Sextets
BlastN (Altschul et al. 1990) alignment was used to identify paralogous pairs for each species, with a cutoff e value of 1 × e−10, and homologous sextets using O. glaberrima chromosome 11 CDSs as query sequences and retaining the best hit on each chromosome with minima of 60% identity and 10% length coverage. These criteria were defined empirically to take into account widely divergent genes and potential anomalies in annotation of poorly supported gene models. Corresponding CDSs were translated, amino acid sequences aligned with ClustalW (Thompson et al. 1994) and CDS aligned with bp_mrtrans (Stajich J, jasonatbioperl.org).

Whole Contig Alignments
 Finished contigs were aligned with Mauve (Darling et al. 2004), using minimum Locally Colinear Block (LCB) weight and backbone size at 100 and 50, respectively. Homologous colinear sequence blocks were aligned with ClustalW, as were intervening sequences. These data set were joined together and the resulting alignment split into 500 bp segments (including gaps). In total, 1539 blocks with six homologous sequences were analyzed. Gap information was coded with the simple indel coding (SIC) method (Simmons and Ochoterena 2000) using Indelcoder (Ogden and Rosenberg 2007).

Evolutionary Distances, Phylogenetic, and GENECONV Analysis
For all paralogous gene pairs, pairwise synonymous (dS) and nonsynonymous (dN) substitution rates and nonsynonymous/synonymous (ω) substitution ratios were calculated with the basic maximum likelihood (ML) method of Goldman and Yang (1994). In order to detect functional constraint on both copies in paralogous gene pairs, we determined if the ω values were significantly lower than 0.5 using the likelihood ratio test (LRT, Yang 1998; Betran et al. 2002). For genes in homologous sextets, random-site codon substitution models (Nielsen and Yang 1998), which allow the ω to vary among codons, were implemented in CODEML (PAML 4.3, Yang 2007) and tested with the LRT (M0 vs. M3, M1 vs. M2, M7 vs. M8). Phylogenetic trees were reconstructed by ML and Bayesian inference (BI) methods. The DNA substitution model was selected using the Data-monkey webserver (Kosakovsky Pond and Frost 2005), with all sequences fitting the Hasegawa–Kishino–Yano (HKY85) model. ML was implemented with PhyML 3.0.
Table 1. General Features of Contigs of *Oryza glaberrima* and *Oryza brachyantha* and Orthologous Segments on the MSU Rice Genome Annotation v6.1 Pseudomolecules of *Oryza sativa* ssp. *japonica* (RefSeq).

<table>
<thead>
<tr>
<th></th>
<th><em>O. glaberrima</em></th>
<th><em>O. brachyantha</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ch11</td>
<td>Ch12</td>
</tr>
<tr>
<td>Length (bp)</td>
<td>1,090,000</td>
<td>1,200,000</td>
</tr>
<tr>
<td>Genbank accession</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Coordinates (kb)*</td>
<td>1.42–2.51</td>
<td>1.34–2.54</td>
</tr>
<tr>
<td>Number of genes</td>
<td>180</td>
<td>168</td>
</tr>
<tr>
<td>Density (genes/kb)</td>
<td>0.165</td>
<td>0.139</td>
</tr>
<tr>
<td>%GC</td>
<td>42.76</td>
<td>43.41</td>
</tr>
<tr>
<td>Coding %</td>
<td>37.4</td>
<td>32.4</td>
</tr>
<tr>
<td>TE%</td>
<td>15.3</td>
<td>33</td>
</tr>
<tr>
<td>Class I TE</td>
<td>18</td>
<td>47</td>
</tr>
<tr>
<td>Class II TE</td>
<td>43</td>
<td>36</td>
</tr>
<tr>
<td>MITEs</td>
<td>153</td>
<td>136</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

**Note.**—Numbers of genes do not include alternative splicing forms and CDS with TE-related annotations.

* Coordinates are relative to the RefSeq.

(Guindon and Gascuel 2003) and BI with MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001), with Nst = 2, Rates = Invgamma. For the 500 bp blocks of the whole contig analysis, nucleotide distances were inferred by Bl. The data were partitioned according to data type: DNA (HKY85 model) and binary gap information. Statistical analysis and graph construction was performed using the R software (R Development Core Team 2011). GENECONV (Sawyer 1989) was used with the default settings.

Results

Global Structural Analysis

Despite the high level of conservation between each 1 Mb paralogous segment pair, comparison of orthologous chromosomes shows the extensive divergence of this region. For similar Blast minimal criteria, an ACT comparison emphasizes the strong divergence in the distal region between the orthologs in AA and FF genomes compared with the paralogous pairs (fig. 1). *Oryza sativa* and *O. glaberrima* orthologous contigs display weaker divergence than with the FF species. A complete list of large structural variants (paralogous chromosome pairs or in syntenic positions). Comparison with the two AA genomes identifies three large indels for a total contraction of 272.2 kb in *O. brachyantha*. On the first half of the largest insertion (~160 kb, RefSeq coordinates: 1,925,654–2,117,228), seven genes were annotated on the RefSeq (between Os12g04720 and Os12g04850), of which at least three are expressed, and four on *O. glaberrima*. The proximal region is composed of transposable elements in the RefSeq and is reshuffled in *O. glaberrima* (supplementary table S1, Supplementary Material online). Sequence analysis of the non-TE region showed significant nucleotide conservation only with sequences from AA genome species, suggesting that the genes may be de novo genes specific to the AA complex.

Of particular interest in the context of potential conversion are species-specific rearrangements shared by chromosomes 11 and 12 (supplementary table S1, Supplementary Material online). We observed four large events shared by paralogous chromosome pairs or in syntenic positions. *Oryza brachyantha* chromosomes 11 and 12 have insertions of ~20 kb on chromosome 11 and ~32 kb on...
chromosome 12 in common and a tandem duplication spanning ~16 kb. The latter contains two pairs of annotated genes. Construction of phylogenetic trees of the coding sequences using the AA genome sequences as out-group (supplementary fig. S1, Supplementary Material online) clearly shows a topology of (Ob11-1, Ob12-1), (Ob11-2, Ob12-2), indicative of gene conversion rather than independent duplication. The four AA lineage chromosomes share two expansions compared with *O. brachyantha*, the first varying from 10 to 38 kb and the second covering approximately 29 kb (supplementary table S1, Supplementary Material online). The most parsimonious explanation for these rearrangements conserved between paralogous chromosomes, but which are specific to the two lineages, is concerted evolution since their divergence at the time of the WGD, after speciation events.

We found 65 CDSs conserved on all chromosomes (sex-tets: see fig. 1 and supplementary table S2, Supplementary Material online). A further 20 were absent only on *O. brachyantha* chromosome 12, consistent with the observed deletions. Six were observed in the AA genomes, but not in *O. brachyantha*, whereas one was absent only in *O. glaberrima*. Only seven CDSs were specific to orthologous chromosomes 11 and six to chromosomes 12, all except one located at the proximal end, confirming the widespread homogenization of the distal ends of the duplicated blocks. Three, 9, 3, 4, 29, and 27 genes are specific to Og11, Og12, Ob11, Ob12, Os11, and Os12, respectively. The greater number for the RefSeq sequences can be explained by our stringent annotation for the wild species, as at least nine and six of the CDSs on Os11 and Os12, respectively, are TE related, although they are not annotated as such.

**Fig. 1.** Graphical representation of synteny between the orthologous and paralogous 11 and 12 contigs in the RefSeq, *Oryza glaberrima*, and *Oryza brachyantha*. Coordinates are indicated in kilobases. The segments for the RefSeq correspond to 1.42–2.51 Mb on chromosome 11 and 1.34–2.54 Mb on chromosome 12. Lines represent sequence similarity comparison by BlastN, with blue lines representing inverted matches. The minimum score and size of matches are 300 and 300 bp, respectively. The CDS composition of each contig is shown, with a color code indicating their presence/absence on the six homologous chromosomes.
Gene Conversion between Paralogous Coding Sequences

We applied a topological incongruency approach (Gao and Innan 2004; Lin et al. 2006) to the sextets. Fifteen contained redundant sequences, resulting from local duplication on one or several of the six chromosomes and were excluded from the analysis. Figure 2 shows the topologies expected under different evolutionary schemes. Topology 0 is the null hypothesis, indicative of no conversion events. Topology 2, where all paralogous pairs are grouped together, is expected if gene conversion has occurred separately in all lineages since their divergence. Topology 1, in which O. sativa and O. glaberrima orthologs group together and O. brachyantha copies form an intermediate branch, indicates conversion specific to O. brachyantha. In topology 1M, one orthologous O. sativa/O. glaberrima pair (11 or 12) forms a terminal node with one of the paralogous genes, whereas the other is more distant in the tree. This topology, indicative of conversion in O. brachyantha, is not informative on the relationships between O. sativa and O. glaberrima, as several hypotheses can explain it.

Using BI methods, 24 of 50 sextets present topology 1 and 15 topology 1M (table 3 and supplementary table S2, Supplementary Material online). For two 1M sextets (Os11g04200 and Os11g04650), the distances between the four O. sativa and O. glaberrima sequences are too weak to distinguish the relationships clearly and for four (Os11g04274, Os11g04360, Os11g4570, and Os11g04650), one of the sequences is highly divergent, putatively indicative of pseudogenization. For the last nine 1M sextets, the topology and distances observed could indicate conversion of one of the two paralogous pairs or a greater divergence in one pair.

We found no topology 2 trees and only seven sextets indicated lack of conversion (topology 0), all located in the proximal region of the contigs, after sextet Os11g04980. However, this region also contains three sextets showing conversion in O. brachyantha. Finally, four showed uninterpretable topology 3. Eight trees were congruent between Bayesian and ML methods, most moving between topologies 1 and 1M. These results suggest widespread conversion in O. brachyantha since its divergence from the AA lineage, notably in the distal region.

Nongenic Conversion

Recombination is not exclusively observed in intragenic regions (Mézard 2006). The availability of megabase-sized sequences from closely related species allows the identification of conversion on a large scale, in both gene and nongene regions. We first tested the frequently used program

![Diagram](image-url)
GENECONV on the CDS sextet data set (see Results in supplementary table S2, Supplementary Material online). Among the 27 sextets where conversion tracts were detected, 19 display topology 1 or 1M. For seven of these, GENECONV found converted fragments only for *O. sativa* and *O. glaberrima* pairs, although we also expected conversion for *O. brachyantha* copies. More surprisingly, GENECONV did not detect conversion tracts for *O. brachyantha* in the remaining 21 sextets with topology 1 and 1M. This apparent contradiction with the topological incongruency analysis may be explained by the failure of GENECONV to detect conversion events when the duplicated region is highly homogenized (McGrath et al. 2009). This confirms the prediction of Mansai and Innan (2010) that GENECONV detects few regions in the case of large-scale gene conversion and can only give indications on events which are both local and relatively recent.

As GENECONV proved to be an unsatisfactory tool, we adapted a topological approach, incorporating indel coding, to look for evidence of conversion throughout the 1 Mb region. Mauve alignment was used to identify conserved blocks between the six genomic sequences, choosing 500 bp segments for topological analysis as gene conversion tracts described in the literature range from a few base pairs to 3 kb (Kuang et al. 2004; Mondragon-Palomino and Gaut 2005; Chen et al. 2007; Xu et al. 2008; Benovoy and Drouin 2009). This approach inevitably

---

**Table 3.** Topology Data for Sextets of CDS and Whole Contig Blocks (Divided into Three Zones).

<table>
<thead>
<tr>
<th>Topology</th>
<th>1</th>
<th>1M</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS sextets</td>
<td>24</td>
<td>15</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>Whole sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zone 1</td>
<td>515</td>
<td>136</td>
<td>1</td>
<td>1</td>
<td>147</td>
<td>800</td>
</tr>
<tr>
<td>Intermediate zone</td>
<td>32</td>
<td>1</td>
<td>30</td>
<td>0</td>
<td>135</td>
<td>198</td>
</tr>
<tr>
<td>Zone 2</td>
<td>5</td>
<td>0</td>
<td>404</td>
<td>0</td>
<td>132</td>
<td>541</td>
</tr>
<tr>
<td>Total</td>
<td>552</td>
<td>137</td>
<td>435</td>
<td>1</td>
<td>414</td>
<td>1539</td>
</tr>
<tr>
<td>% Zone 1</td>
<td>93.3</td>
<td>99.3</td>
<td>0.2</td>
<td>100</td>
<td>35.5</td>
<td></td>
</tr>
<tr>
<td>% Intermediate zone</td>
<td>5.8</td>
<td>0.7</td>
<td>6.9</td>
<td>0</td>
<td>32.6</td>
<td></td>
</tr>
<tr>
<td>% Zone 2</td>
<td>0.9</td>
<td>0</td>
<td>92.9</td>
<td>0</td>
<td>31.9</td>
<td></td>
</tr>
</tbody>
</table>

---

**Fig. 3.** Spatial distribution of synonymous substitution rates (ds) between paralogous gene pairs computed with the basic ML codon model, plotted against the number of pairs (a) and BI nucleotide distances between paralogous 500 bp fragments of the whole contig alignment, plotted against the chromosome 11 coordinates (kb) for the three species (b).
produces a number of uninformative alignments and, among the 1539 trees examined, those with strongly divergent branches were classified as topology 3 (table 3). The distribution of the tree topologies along the 1 Mb sequence is not random, defining three regions (table 3 and supplementary table S2, Supplementary Material online).

The distal region (zone 1), where more than 80% of the trees display topologies 1 (515) or 1M (136), extends to block 1,047,501–1,048,000 (total 800 blocks), corresponding to 2,108,257 and 2,151,747 bp on O. sativa chromosomes 11 and 12, respectively. As for the CDS, topologies 1M are mainly indicative of very weak distances between the four O. sativa and O. glaberrima contigs. Only one block in this region has topology 0 (856,501–857,000) and only one (663,001–663,500) suggests independent conversion in both O. sativa and O. glaberrima (topology 2). For the 652 blocks displaying topologies 1, 1M, and 2 in the first zone, 337 (~106,000 bp), 414 (135,000 bp), and 415 (126,000 bp) are located in intergenic regions for O. sativa, O. glaberrima, and O. brachyantha, respectively, whereas 315 (120,000 bp), 238 (94,000 bp), and 237 (93,000 bp) overlap protein-coding sequences.

The proximal region (zone 2) extends from block 1182001–1182500 to the end and covers RefSeq chromosome 11 from 2,195,478 bp and chromosome 12 from 2,214,633 bp. Most trees in this region show topology 0 (404; 75%) with only five isolated topology 1 alignments. Nine CDS sextets were found in this area (beginning after sextet Os11g05050), all classified as topology 0 except for two showing topology 1 (Os11g05320 and Os11g05370). However, we did not find topology 1 in the 500 bp blocks corresponding to these two loci (1511501_1512000 to 1516501_1517000 and 1551501_1552000 to 1552500). This could be explained by the presence of introns and coding of gaps in the whole contig analysis, suggesting rather local conversion events limited to CDSs. The intermediate zone displays a balanced ratio of topologies 1 and 0, and a high percentage of topologies 3 (135, 68%), indicating considerable rearrangement.

The uniformity of conservation of large tracts of both coding and noncoding sequences in the distal regions is indicative of long-range mechanisms rather than small and repetitive recombination events. Nonetheless, our GENECONV analysis and observations of topologies 1M in phylogenetic analysis confirm that regular small-scale conversion may have occurred since the divergence of the AA species but no extensive homogenization. In the proximal regions, we found 23, 16, and 9 paralogs in O. sativa, O. glaberrima, and O. brachyantha. This conservation of isolated coding sequences after the breakpoint of conservation could be due to local conversion events but may simply reflect slowly diverging gene pairs, generated by older conversion events.

Finding the Limits and Dating the Conversion Events

Figure 3 displays the synonymous substitution rates (dS) resulting from ML analysis for all paralogous gene pairs and the nucleotide distances inferred by the BI between pairs of fragments from the whole contig analysis, plotted against their positions on the contigs. There is a clear rupture in the distribution in all three species, values being low in the first two-thirds of the region, increasing clearly in the proximal region. The breakpoint in the whole contig analysis is located between 2,100,000 and 2,106,000 bp on O. sativa chromosome 11, corresponding to 2,120,000–2,128,000 bp on O. sativa chromosome 12, in agreement with the topological analysis on sextets. It is at syntenic locations in O. glaberrima, between 591,500–597,000 bp and 599,500–606,000 bp on O. glaberrima contigs 11 and 12, respectively. The O. brachyantha breakpoint is slightly more proximal, between 518,000–519,000 bp and 374,000–375,000 bp on contigs 11 and 12.
(2,118,000 and 2,155,000 bp on RefSeq chromosome 11 and 12, respectively). These breakpoints all map to the intermediate region as described above.

The distributions of nucleotide distance values for the paired 500 bp fragments show a bimodal distribution, with the first peak corresponding to zone 1 (fig. 4). Distributions of distance values for zone 1 (fig. 4, small histograms) indicate that these regions of ~0.6 Mb were homogenized at the same time either by one unique conversion event or by several concomitant long-range events. The first peak is at 0.03–0.04 for the AA species and 0.01–0.02 for O. brachyantha, indicative of more recent conversion in the FF genome. Furthermore, the mean distance between O. brachyantha contig pairs (0.07) is lower than that of the AA pairs (0.17) (supplementary table S2, Supplementary Material online). The second peak represents the distances between the sequences in the nonconverted contig ends (1.25–1.26 for the AA species and 1.04–1.05 for O. brachyantha). Distributions of dS rate for the paired genes display a unimodal distribution with peaks at 0.02–0.04, 0.04–0.06, and 0.02–0.04 for O. sativa, O. glaberrima, and O. brachyantha, respectively, consistent with the whole contig analysis (results not shown).

Based on a divergence time of 15 Ma for O. brachyantha in the genus and ~0.8 Ma for the divergence of O. sativa and O. glaberrima, we estimated the relative time of the last conversion event for each paralogous pair using the median dS and nucleotide distance values among the orthologs and paralogs (supplementary table S2, Supplementary Material online) using the formulas:

\[ x(p_{11}, p_{12}) = \frac{\text{median}(d(p_{11}, p_{12})) \times 0.8}{\text{mean}(\text{median}(d(Os_{11}, Og_{11})), \text{median}(d(Os_{12}, Og_{12})))} \]

\[ x(p_{11}, p_{12}) = \frac{\text{median}(d(p_{11}, p_{12})) \times 15}{\text{mean}(\text{median}(d(Os_{11}, Ob_{11})), \text{median}(d(Os_{12}, Ob_{12})))} \]

where \( p_{11} \) and \( p_{12} \) are the paralogous pair considered and \( d(a,b) \) either the dS or the BI distance.

Considering only zone 1, the last conversion events were dated between 2.5 and 4.0 Ma for the AA species and 1.5–3.5 Ma for O. brachyantha, much lower than previous estimations, from 5 to 21 Ma, given for the whole region in O. sativa but based only on coding sequences (Goff et al. 2002; Rice Chromosomes 11 and 12 Sequencing Consortium 2005; Wang et al. 2005; Salse et al. 2008). Using pairs from zone 2, we calculate 15–55 Ma for the AA species and 20–50 Ma for O. brachyantha. Age estimations for the WGD event are somewhat greater (50–90 Ma, Chaw et al. 2004; Yu et al. 2005) but the difference is easily explained by the small size of the region, local conversion events since the duplication or traces of older conversion events.

Paralog Divergence after Conversion

Large-scale conversion events as described here reset the evolutionary clock and harmonize both coding and essential noncoding regions. We have analyzed the divergence and selection pressure on the 11 and 12 paralogous copies, because we thought that could indicate, indirectly, the role of this recurrent homogenization. If paralogous functionally redundant copies are conserved identically, we should see purifying selection, whereas if the copies are evolving toward pseudogenization, subfunctionalization, or neofunctionalization, we would expect to observe signals of neutral evolution or positive selection (Innan and Kondrashov 2010). Studies using tiling arrays (Li et al. 2005) or microarrays (Throude et al. 2009) did not detect significantly different expression patterns between gene pairs in the 11–12 duplication. However, Yim et al. (2009) observed that between 50.9% and 67.3% of 55 gene pairs in the block may have diverged in their expression, so no clear conclusion can be drawn. We compared the nonsynonymous/synonymous ratios (\( \omega \)) for paralogs in the three species and tested for selection pressures.

We found 122, 76, and 67 paralogous pairs in the RefSeq, O. glaberrima and O. brachyantha sequences, respectively and eliminated those with null dS values. The \( \omega \) ratio, calculated by the method of Goldman and Yang (1994), ranged from 0.001 to 1.042 (mean 0.3 ± 0.02), 0.001 to 1.282 (0.25 ± 0.02), and 0.001 to 1.560 (0.34 ± 0.03), in Os, Og, and Ob, respectively. Only two pairs in O. sativa displayed \( \omega = 1 \) (neutrality level) and one pair for each other species displayed \( \omega > 1 \) (indicator of positive selection). Under the LRT, among 245 paralogous pairs, 112 showed an \( \omega \) value that was significantly lower than 0.5 with \( P < 0.05 \) (71 pairs with \( P < 0.001 \)), indicating that duplicated copies are both under purifying selection. The Benjamini–Hochberg procedure for controlling the false discovery rate in multiple comparisons was implemented at the \( \alpha = 0.05 \) level, and ratios for 103 paralogous pairs were still significantly <0.5 at \( P < 0.05 \) (45 for the RefSeq, 38 for O. glaberrima, and 20 for O. brachyantha).

Random-site codon substitution models were applied to sextets in order to test the presence of positive Darwinian selection at individual sites. The one-ratio model (M0) gives the average \( \omega \) over all sites and branches for each data set and this ranged from 0.004 to 0.57, still indicating the overwhelming role of purifying selection. The LRT indicates that M3 fits the data significantly better than M0 for 36 sextets (degrees of freedom [df] = 4, \( P = 0.05 \)), indicating significant variation in selective constraints among sites. For 22 sextets, both models M2 and M8, which allow the \( \omega \) ratio to exceed 1, fit the data significantly better (df = 2, \( P = 0.05 \)) than models M1 and M7 (supplementary table S2, Supplementary Material online). The number of sites with \( \omega > 1 \) varied from 5 to 142.

Thus, a certain fraction of duplicated pairs (42%, 52%, and 32% in Os, Og, and Ob, respectively) are under purifying selection in the region under study suggesting they could tend to diverge slowly after conversion, whereas only 22 pairs common to all three species display positive selection on a fraction of codons.

Discussion

We have demonstrated that the duplicated blocks between 1.5 and 2.1 Mb on the RefSeq chromosomes 11 and 12, and orthologous regions in O. glaberrima and O. brachyantha,
are uniformly homogenized by long-range recombination mechanisms. Our observation of syntenic breakpoints of conservation in the AA (O. sativa and O. glaberrima) and FF (O. brachyantha) lineages suggests that conversion is recurrently initiated around this point (2.1 Mb on the RefSeq), indicative of a putative hot spot of recombination. This is coherent with the fact that, in Poaceae, recombination increases with relative distance from the centromere (Wu et al. 2003; Anderson et al. 2004; Kao et al. 2006) and is greater in gene-dense regions near the telomeres (Mézard 2006).

Two studies provide estimations of recombination rates along the 12 chromosomes in rice, and both support our hypothesis (Rizzon et al. 2006; Tian et al. 2009). Indeed, both chromosomes 11 and 12 display a high recombination rate (\(~12\) cM/Mb and >12 cM/Mb, respectively in Tian et al. 2009) between 2 and 3 Mb from the short arm telomere. The peak is more striking for chromosome 12 compared with the surrounding regions.

The extent of gene conversion depends on the recombination process involved, but we have no evidence allowing us to favor one particular mechanism. Nonetheless, we can exclude noncrossover DSBR and SDSA as they generally yield small conversion tracts, less than a few kilobases (Mancera et al. 2008). Two mechanisms could potentially explain the large conversion tracts observed. A DSBR event associated with half crossing over between the short arm ends of these two chromosomes would lead to reciprocal exchange between the two chromatids. This could generate gametes with conversion tracts depending on how the chromatids segregate. The second process is BIR, which is initiated as DSBR, following a DSB where just one of the two ends can undergo homology-dependent strand invasion (Llorente et al. 2008). It continues with a processive replication fork, and DNA synthesis proceeds to the end of the donor chromosome (Llorente et al. 2008). BIR have been implicated in homogenization of subtelomeric regions in yeast (Bosco and Haber 1998) and their relative frequency increases toward telomeric regions, in which their consequences are less deleterious than in other regions of the chromosomes (Ricchetti et al. 2003). The 11–12 duplicated block extends beyond the limit of the subtelomeric regions (~500 kb from the distal end, Fan et al. 2008), but the underlying mechanisms of BIR (reviewed for the yeast model in Lydeard et al. 2007; Llorente et al. 2008) do not limit the size of the fragment which is reconstructed. These two mechanisms are described as putative models of formation of segmental duplications (Koszul and Fischer 2009), which was the first hypothesis proposed for the 11/12 duplication (Goff et al. 2002).

We propose that conversion events have recurrently replaced large segments of one chromosome with homologous sequences from another, which implies the recurrence of meiotic pairing of nonhomologous chromosomes 11 and 12 since their formation by polyploidization, certainly facilitated by the maintenance of redundancy in their telomeric and subtelomeric regions which obscure true homologous relationships.

Whatever the mechanism leading to this duplication, it has not occurred independently in the two AA species since their divergence. This extends the observations of Wang et al. (2007) on the O. sativa subspecies who found very few partial-gene conversion events and only two whole-gene conversions, both in O. sativa ssp. japonica. To our knowledge, the 11–12 duplication and its orthologs in sorghum and Brachypodium (Wang et al. 2011) represent the first described example of such long-term conservation of two duplicated segments in plants.

Based on our calculation of selective pressure on paralogous gene pairs, we can not exclude the possibility that the presence and maintenance of the recombination hot spot and long-range gene conversion are selected themselves for the benefits of buffering crucial functionality. However, no particular class of genes have been identified in the segments. The Rice Chromosomes 11 and 12 Sequencing Consortia (2005) came to the conclusion that chromosomes 11 and 12 are enriched in disease resistance gene clusters, but these are not preferentially located to the distal 2 Mb of the chromosomes and are rather known for their variability. No significant bias of Pfam domain composition or GO categories was found in the converted genes in rice and sorghum genomes (Wang et al. 2009 and our unpublished observations).

Our comparative study highlights considerable divergence, not only between the AA and FF genomes but also between the two AA genomes, including de novo gene formation. If we consider only interspecific rearrangements larger than 10 kb with genes involved, we observe one inversion (five genes) specific to the RefSeq, one expansion for the AA lineage compared with O. brachyantha (four genes), and two tandem duplication, one for O. brachyantha (involving ten genes) and one for the AA species (two genes), all on the chromosome 11 1 Mb segment. On chromosome 12, we observed one inversion (seven genes), one expansion (two genes) specific to O. glaberrima, and one expansion (five genes) on O. sativa. Contractions compared with the RefSeq (6, 2, and 19 genes) were particularly striking on O. brachyantha chromosome 12. Genome expansions and contractions in the 11–12 duplicated region (15 and 12, respectively) in a short evolutionary time frame, involving up to one-third of the genome sequence, are strikingly different from the highly conserved gene colinearity observed in the comparative studies of MONOCULM1 orthologous regions (2.4 Mb, chr6) in 14 Oryza genomes (Lu et al. 2009). This latter region is disrupted by only three rearrangements (a three-gene segment translocation in O. coarctata, a three-gene segment insertion in O. sativa, and a single gene tandem duplication in O. granulata).

Wang et al. (2011) recently showed that ectopic concerted evolution acting on the duplicated blocks in rice chromosomes 11 and 12 and homologous sorghum chromosomes 5 and 8 has significantly increased gene divergence between lineages compared with the genome-wide average, particularly in the more distal ends of these blocks which show the greatest intragenomic similarity. Whereas these studies concerned gene content and
divergence, our studies on structural rearrangements lead to the same conclusion. Two segments derived from the initial duplication event will diverge independently and accumulate structural variants. Subsequent interspecies divergence will depend on the timing of speciation and conversion events, as well as on the direction of conversion. After speciation (species A and B), if conversion occurs from chromosome 11 to 12 in A, and from chromosome 12 to 11 in B, the comparison between A11–B11 or A12–B12 represents the divergence since the duplication and not since the speciation. Repetitive cycles of divergence and alternative conversion will increase the distance between orthologous pairs.

Gene-scale conversion is already incorporated in the classical models of the evolution of duplicated genes (Teshima and Innan 2004; Gay et al. 2007; Innan 2009; Innan and Kondrashov 2010) and the occurrence of conversion between homeologous genes during polyploid formation and divergence (Udall et al. 2005; Salmon et al. 2009), or between the two LTR of a retrotransposon (Kijima and Innan 2010) have also been discussed. However, the story of conversion in the 11–12 distal ends is currently unique in genome evolution. Further comparative genomic and genetic studies within and outside the Oryza genus will be useful to confirm our hypothesis and clear up the mystery of possible functionality and benefits of this genome redundancy.

Supplementary Material
Supplementary text S1, tables S1 and S2, and figure S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
We thank Michael Rosenberg for supplying his program Indelcoder, and Yves Desdevises for his helpful advice on the Bayesian method. This work was supported by the Centre National de la Recherche Scientifique (Cooke laboratory). Sequencing was financed by Génoscope CNS.

References


