Rapid Evolution in a Pair of Recent Duplicate Segments of Rice

HUIFENG JIANG1,2, DONGYUAN LIU3, ZHENGONG GU4, AND WEN WANG1*
1CAS-Max Planck Junior Research Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (CAS), Kunming, Yunnan 650223, China
2Graduate School of Chinese Academy Sciences, Beijing 100039, China
3Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China
4Stanford Genome Technology Center, Stanford University, Palo Alto, California 94304

ABSTRACT Gene duplication has been considered the most important way of generating genetic novelties. The subsequent evolution right after gene duplication is critical for new function to occur. Here we analyzed the evolutionary pattern for a recently duplicated segment between rice chromosomes 11 and 12. This duplication event was estimated to occur about 6 million years ago, during the divergence of the B- and C-genome rice species. The duplicate segment in chromosome 12 has significantly higher frequency of sequence rearrangement rate than non-duplicated regions. The rearrangement rate is ~6.5 breakages/Mb per million years, about six times higher than the fastest rate ever reported in eukaryotes. The genes within both segments experienced accelerated nucleotide substitution rates revealed by synonymous (\(K_s\)) and non-synonymous divergence (\(K_a\)) between Oryza sativa indica and O. sativa japonica. Analysis using EST data also implicates rapid divergence in expression between these segmental duplicate genes. These overall rapid changes from different perspective for the first time provide evidence that relaxation of selection also occurs in large-scale duplications.


Gene duplication has been of great interest to geneticists because it is considered a major evolutionary source for new genes and functions. Because of redundancy between duplicate genes, one or both duplicate copies would experience relaxed selective constraint that result in elevated rates of evolution (Ohno, '70; Wagner, 2002), and sometimes one copy could be freed to evolve brand new function (Long and Langley, '93; Prince and Pickett, 2002; Wang et al., 2002, 2004; Zhang et al., 2002; Long et al., 2003; Loppin et al., 2005). Along with the accelerated sequence evolution, duplicate genes also undergo rapid divergence in expression (Gu et al., 2002b; Makova and Li, 2003). In spite of numerous reports on rapid evolution of duplicate genes, it is not yet to be tested if duplicated large segments as a whole would also experience rapid evolution. For example, it is still unknown whether relaxation of selection will allow more rearrangements in large duplications, which can efficiently create new genes by shuffling (Samonte and Eichler, 2002). Indeed, among different forms of gene duplications, large-scale duplication events, such as segmental and genome duplication, could provide a considerable amount of materials for evolution at once, and might have played important roles in several canonical steps of organism evolution (Ohno, '70; Wolfe and Shields, '97; Arabidopsis Genome initiative, 2000; Gu et al., 2002a; Mclysahit et al., 2002; Yu et al., 2005).

Grant sponsor: CAS-Max Planck Society Fellowship; Grant number: KSCX2-SW-121; Grant sponsor: NSFC award; Grant number: 30325016; Grant sponsor: NSFC key grant; Grant number: 30430400.
*Correspondence to: W. Wang, Kunming Institute of Zoology, East 32 Jiaochang Road, Kunming, Yunnan 650223, China. E-mail: wwang@mail.kiz.ac.cn
Received 1 March 2006; Revised 8 May 2006; Accepted 25 May 2006
Published online 12 July 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21122.
Two premises are required in order to understand evolution patterns in large segmental duplicates: availability of recently duplicated segments and genomic sequences of two closely related taxa (species or subspecies). In old duplicates, big divergences have obscured sequence information that is necessary to infer detailed molecular changes after duplication. Without closely related taxa, it is impossible to compare divergence rates between duplicated and non-duplicated segments. Fortunately, the opportunity to investigate evolution patterns of large-scale duplications has now emerged in domesticated rice.

Domesticated rice cultivars (Oryza sativa L.) belong to two major subspecies, japonica and indica. Two recent studies congruently estimated that the two subspecies separated about 0.4 million years ago (mya) (Ge et al., 2004; Ma and Bennetzen, 2004). Besides an ancient genome duplication, extensive and continuous gene duplications have been observed in rice genomes (Yu et al., 2005). More interestingly, a likely young segmental duplication event between chromosomes 11 and 12 was identified (Wu et al., '98; Rice Chromosomes 11 and 12 Sequencing Consortia, 2005; Yu et al., 2005). Therefore, the availability of the two rice subspecies’ genome sequences (Goff et al., 2002; Yu et al., 2002, 2005; International Rice Genome Sequencing Project, 2005) would allow us to conduct comparisons between duplicated and non-duplicated segments. In addition, Pevzner and Tesler (2003) recently developed a new visualization tool that made fine comparison of rearrangements between large segments possible. In this study we analyzed sequence rearrangement rate, nucleotide substitution pattern, and expression divergence rate within the recently duplicated genomic segments of rice by taking advantage of rich genomic information for rice.

**METHODS AND MATERIALS**

**Data resources**

Masked genomic sequences of indica and japonica were generated by BGI (Beijing genomics institute http://www.genomics.org.cn/bgi_new/english/index.htm). Full-length cDNAs of japonica were downloaded from KOME (http://cdna01.dna.afrc.go.jp/cDNA) (Kikuchi et al., 2003). ESTs were downloaded from NCBI dbEST (ftp://ftp.ncbi.nih.gov/repository/dbEST/).

**Identification of recent segmental duplication in distal regions of chromosomes 11 and 12**

As described by Yu et al. (2005), there is likely a recent segmental duplication between the distal regions of chromosomes 11 and 12. To further confirm the recent segmental duplication, we aligned chromosome 11 with chromosome 12 both for indica and japonica by BLAST (Altschul et al., '90). A certain piece of fragment in chromosome 11 may find homologous hit in chromosome 12. We plotted all the homologous hits based on their chromosomal positions in the two chromosomes. Both the scatter plots in indica and japonica indicate existence of segmental duplication between the distal regions of chromosomes 11 and 12.

**Detection of synteny blocks**

Because syntenic relationships between two genomes played an extremely important role in rearrangement studies (Pevzner and Tesler, 2003), we did careful alignments between homologous chromosomes of the two rice subspecies. To identify conserved segments (also called anchors), we divided each chromosome of indica into 500 bp fragments (average 419 bp after masked repeats using Repeatmasker: http://www.repeatmasker.org) and aligned them with japonica’s homologous chromosome sequences using BLAST algorithm. The non-overlapped best hit of each fragment was defined as an anchor. On a two-dimensional coordination system with the axis of X representing indica chromosomal location and Y representing japonica chromosomal location, each anchor is represented by its start points on both X and Y coordination. If the distance between two anchors is smaller than the threshold gap size (denoted as G), anchors were connected into a synteny block. A synteny block would be retained for the following analyses if its length (between the start point and the end point of the synteny blocks in the two-dimensional graph) is larger than the pre-defined minimum cluster size C (Pevzner and Tesler, 2003). The threshold values of G and C are variable depending on the systems in query. In this study, both the threshold values of G and C are set to be 10 kb based on two considerations. First, the divergence between the two rice subspecies is still very small and there is rarely huge rearrangement yet (Bennetzen and Ramakrishna, 2002). If G and C values are too big, such as 100 kb, it would be difficult to identify rearrange-
ment events. Second, because the average size of genes in rice is about 1.2 kb (Yu et al., 2005), if G and C values are too small, such as around 1 kb, existence of numerous duplicate genes in rice will make it difficult to accurately define syntenic relationships.

**Identification of rearrangement events**

Following Pevzner and Tesler’s (2003) methods, we first arranged the identified synteny blocks onto each chromosome in indica, and orderly labeled them with number according to their location in the chromosomes. Then we mapped these synteny blocks onto japonica chromosomes. If rearrangement happens in either indica or japonica, the order of labeling numbers will be disrupted in japonica. In order to simplify the algorithm, we first identify those long-distance rearrangement events involved movement of more than 20 synteny blocks. These rearrangements generally resulted from transpositions which involve movement of a chunk of DNA from one position to another. For the remaining synteny blocks, we put 20 continuous synteny blocks in each subspecies into a group to identify short-distance rearrangement events within a group based on block order. Because it is difficult to distinguish between short-distance transposition and inversion, to be conservative, we classed all short-distance rearrangements (less than 20 synteny blocks) into inversions (Blanchette et al., ’96; Coghlan and Wolfe, 2002). It is noteworthy that the arbitrary selection of 20 blocks does not change the total number of rearrangement events identified because we obtained the same results when we chose 10 or 50 blocks to do the analysis. Based on these transposition and inversion events, we calculated the number of breakpoints (three breakages for a transposition and two breakages for an inversion) (Pevzner and Tesler, 2003). The rearrangement rate in one segment is obtained by the calculation: number of breakpoints/(segment length x divergence time).

**Nucleotide substitution rates**

In order to investigate gene divergence in the recently duplicated segments, we mapped 26,373 full-length cDNAs of japonica which are longer than 300 bp (Kikuchi et al., 2003) onto the assembled japonica genome using blat algorithm (Kent et al., 2002). Then we searched their orthologous hits in the duplicated segment of indica chromosome 11 (IRD11) and indica chromosome 12 (IRD12), respectively. Non-synonymous ($K_a$) and synonymous ($K_s$) substitution divergences were estimated for the gene pairs in the duplicated and non-duplicated regions between japonica and indica according to the Li-93 method (Li, ’93, ’97). We used $F$-test to test variance of $K_a$ and $K_s$ between duplicate and non-duplicate blocks. The difference of means was tested by $t$-test.

**Expression divergence of duplicated genes**

To detect expression divergence for a duplicated gene pair between chromosomes 11 and 12 within a subspecies, we downloaded 229,966 rice ESTs from NCBI that were generated from six tissues and nine different developmental stages of indica or japonica, respectively. We mapped these ESTs into the genome using stringent cut-off (≥ 95% identities) to remove the possibility that one EST sequence was matched to more than one gene. If a duplicate gene pair has expression difference (absence/presence of EST) in one or more of the six tissues and nine developmental stages, they were tentatively considered as diverged at the expression level.

**RESULTS**

**Characterization of the recently duplicated segment**

Yu et al. (2005) briefly reported the identification of a segmental duplication between chromosomes 11 and 12 (referred to RD11 and RD 12, respectively). In order to characterize this duplication in detail, we further aligned chromosome 11 against 12 in indica and japonica, respectively. Overlapped hits and hits smaller than 100 bp were filtered out. Finally, there are 1,589 hits in japonica and 1,168 hits in indica. Both the scatter plots of hits in indica and japonica indicate the existence of segmental duplication between the distal regions of chromosomes 11 and 12 (Fig. 1). Based on the assembled genome sequences (Yu et al., 2005), the length and position of the segments on the indica and japonica chromosomes 11 and 12 are shown in Table 1. However, the length of duplicated segment (about 2 Mb) is smaller than Yu et al. described (6 Mb) but similar with a previous experimental result (2.5 Mb) (Wu et al., ’98). To clarify the discrepancy, we conducted further analysis. Yu et al. (2005) identified 80 duplicated gene pairs in the 6 Mb duplication regions in indica, but 59 of them are located in the
2 Mb duplication regions. The average $K_s$ of 80 gene pairs in the 6 Mb region is 0.22 but average $K_s$ of the 59 gene pairs in the 2 Mb region is only 0.08. If we exclude genes in the 2 Mb region, the average $K_s$ for genes in the remaining 4 Mb region is 0.64, which is similar with the average $K_s$ for genes resulting from the ancient genome duplication (Yu et al., 2005). This observation suggests that it is more appropriate to only consider the 2 Mb regions as the recently duplicated segment. In Table 1, we can also see that both the japonica segments are longer than those in indica, which is due to many large indels (larger than 10 kb) in japonica and indica. Almost half of the indels (11/24 for chromosome 11 and 11/25 for chromosome 12) are masked as large repeated sequences.

**Table 1. Recent duplicate segments between indica and japonica chromosomes 11 and 12**

<table>
<thead>
<tr>
<th>Chromosome 11</th>
<th>Start</th>
<th>End</th>
<th>Size (Mb)</th>
<th>Chromosome 12</th>
<th>Start</th>
<th>End</th>
<th>Size (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indica</td>
<td>502,167</td>
<td>2,202,793</td>
<td>1.71</td>
<td>30,738</td>
<td>1,832,118</td>
<td>1.81</td>
<td></td>
</tr>
<tr>
<td>Japonica</td>
<td>179,068</td>
<td>2,345,313</td>
<td>2.17</td>
<td>0</td>
<td>2,042,268</td>
<td>2.05</td>
<td></td>
</tr>
</tbody>
</table>

**High rearrangement rate in the duplicated segments**

We identified two inversions in RD11, and five inversions in RD12 between indica and japonica. For whole genome, we identified 39 transpositions and 406 inversions. Figure 2 shows number of breakages per Mb for each chromosome, average of the whole genome and each duplicate segment. The duplicate segment on chromosome 12 (RD12) has the highest rearrangement rate (2.59 breakages/Mb), significantly higher than the rate of 1.17 breakages/Mb for the whole genome ($P = 0$, two tails Z-test) (Fig. 2). However, the duplication segment RD11 retained a normal rearrangement rate which is comparable to other chromosomes. This result shows that the predicted relaxed
selective constraint by the traditional model (Ohno, '70) does display on one copy of the duplicate segments at the level of gene order and structural changes, at least for this particular case.

**Accelerated nucleotide substitution rates**

Using the 26,373 full-length cDNAs as queries, we identified all homologous sequences longer than 300bp in the four duplicated segments in *indica* and *japonica*. 127 cDNAs have homologous hits in all the four copies of duplication blocks. Among these 127 genes, 32 in *indica* RD11 (IRD11), 33 in *japonica* RD11 (JRD11), 18 in *indica* RD12 (IRD12), and 33 in *japonica* RD12 (JRD12) are pseudogenes which have frame-shift and/or non-sense mutations. IRD11 and JRD11 share 23 pseudogenes, and IRD12 and JRD12 share 15 pseudogenes, suggesting that the majority of pseudogenes died before the separation of the two rice subspecies although there are some duplicates that were lost specifically in one subspecies after the separation. Because we could only annotate possible functions for the intact ancestral genes of five pseudogenes based on gene ontology information, we cannot infer possible functional roles of these pseudolization events at this moment.

After excluding these pseudogenes, 85 intact gene pairs can be found in RD11, and 91 intact gene pairs can be found in RD12 (Table 2). We calculated $K_a$ and $K_s$ between *indica* and *japonica* orthologs for each of these gene pairs. At the same time, as a control we randomly picked 1,000 genes located on other chromosomal regions (non-duplicated regions) and calculated their $K_a$ and $K_s$ values between *japonica* and *indica*. Table 2 shows the average $K_a$ and $K_s$ values between the non-duplicated and duplicated regions. Obviously, as the whole genes in the duplicated regions have both significantly higher $K_a$ and $K_s$ values than genes in non-duplicated regions (Table 2). In addition, if we consider RD11 and RD12 separately, both the gene sets also have higher $K_a$ and $K_s$ values than genes in non-duplicated regions (Table 2), suggesting that the accelerated evolution at the substitution level was not restricted in one copy of the segments. Nevertheless, both $K_a$ and $K_s$ values are larger for genes in RD12 than in RD11 (Table 2). This observation, in combination with the previous result that more arrangements are observed in RD12, suggests that functional constraints could be more relaxed for genes in RD12 than in RD11.

**Majority of duplicated genes could have diverged at the expression level**

By mapping all the downloaded ESTs into the duplicated regions we found that 964 EST sequences were mapped into 70 intact gene pairs in JRD11 and JRD12, and 922 EST sequences were mapped into 61 intact gene pairs in IRD11 and IRD12. Among them, 56 gene pairs (80%) between JRD11 and JRD12 and 52 gene pairs (85%) between IRD11 and IRD12 have probably diverged gene expression patterns because one copy of the duplicate gene pair has expression difference (absence/presence of EST) in one or more of the six tissues and nine developmental stages. Therefore, more than 80% of the gene pairs in the rice recently duplicated segments have diverged in expression even though their average $K_s$ is only 0.08. This result has to be taken with caution because the level of expression divergence was estimated only based on EST data, more functional data may result in different estimation of expression divergence between duplicate genes. It is also noteworthy that the $K_s$ value (0.08) was calculated using the 59 gene pairs that have only

<table>
<thead>
<tr>
<th>Gene number</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>F-test</th>
<th>P-value</th>
<th>T-test</th>
<th>P-value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>F-test</th>
<th>P-value</th>
<th>T-test</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-dup</td>
<td>1000</td>
<td>0.001</td>
<td>4.1 x 10^-6</td>
<td>0.017</td>
<td>0.018</td>
<td>0.005</td>
<td>0.003</td>
<td>2 x 10^-5</td>
<td>0.015</td>
<td>0.003</td>
<td>0.017</td>
<td>0.001</td>
</tr>
<tr>
<td>Sum of RD11,12</td>
<td>176</td>
<td>0.006</td>
<td>9 x 10^-4</td>
<td>0.001</td>
<td>0.011</td>
<td>5 x 10^-4</td>
<td>0.001</td>
<td>5 x 10^-4</td>
<td>0.018</td>
<td>0.005</td>
<td>0.018</td>
<td>0.019</td>
</tr>
<tr>
<td>RD11</td>
<td>85</td>
<td>0.004</td>
<td>5 x 10^-5</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>RD12</td>
<td>91</td>
<td>0.009</td>
<td>0.002</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

two copies in the whole rice genome as done by Yu et al. (2005).

**DISCUSSION**

$K_s$ is presumably a neutral rate and therefore could be used to date duplication events (Li, '97; Lynch and Conery, 2000). If we assume the synonymous substitution rate of rice genes is $6.5 \times 10^{-9}$ substitutions per silent site per year (Gaut et al., '96), based on the mean $K_s$ (0.08) between the RD11 and RD12 duplicated gene pairs in *indica*, the segmental duplication between chromosomes 11 and 12 occurred about 6.15 mya. Due to the observed accelerated evolution after segmental duplication, the real event could be more recent, which will make our conclusion more conservative. Compared to the age of the rice whole genome duplication (more than 50 mya, Yu et al., 2005), it is a very young duplication event, and therefore no wonder it could even be identified by traditional physical mapping approach (Wu et al., '98). According to the rice phylogeny, this duplication event happened accompanying with the divergence of C- and B-genome species (Ge et al., '99; Ge et al., 2004).

For old duplications, such as the duplicated regions resulting from the rice genome duplication, many duplicates have diverged so much that it is impossible to use them to investigate the early stage evolution after duplication (Vandepoele et al., 2003; Yu et al., 2005). This young segmental duplication between rice chromosomes 11 and 12 provides us an opportunity to examine the evolution patterns in duplicated segments. One prediction of the traditional gene duplication model (Ohno, '70) is that relaxation of selection should also occur at the duplicate segments. One way to evaluate this effect is to investigate the rearrangement rate within duplicate segments. Our results indeed demonstrate that the rearrangement rate is significantly higher at least in one of the duplicate segments than those in non-duplicated regions (Fig. 2). This rate is the fastest rearrangement rate in eukaryotic organisms reported so far. Coghlan and Wolfe (2002) estimated that the rearrangement rate in nematodes is 0.4–1.0 breakages/Mb per million years (myr), which is the previous fastest rate among eukaryotes. Because the divergence time of *indica* and *japonica* is about 0.4 mya (Ge et al., 2004; Ma and Bennetzen, 2004), the rearrangement rate in the rice duplications is about 6.5 breakages/Mb per myr (2.59 breakages/0.4 myr), which is about six times higher than that of nematodes. The estimated genome-wide rearrangement rate in rice (1.17/0.4 = 2.9 breakages/myr), however, is also higher than that of nematodes, suggesting that rice genomes likely have experienced faster rearrangement in evolution.

While our result of rapid nucleotide substitution in the duplicate segments (Table 2) unsurprisingly confirms previous conclusion that genes evolve faster after duplication (Kondrashov et al., 2002; Wagner, 2002), it also clearly shows that both copies, including the genes in the ancestral segment, experienced accelerated evolution after duplication, although this acceleration is asymmetrical between two segments (Table 2). Besides the relaxation of selection, we are more interested in observing emergence of new function after duplication. It has been shown that duplicate genes increase gene expression diversity within and between species (Gu et al., 2004), and changes in gene expression can lead to important phenotypic changes in evolution. (Carroll, 2000; Wray et al., 2003). Gu et al. (2002b) found that a large proportion of yeast duplicate genes has diverged rapidly in expression and about 40% of gene pairs show expression divergence when $K_s \leq 0.1$. Following their work, Makova and Li (2003) found that the expression of human duplicate genes diverged even faster than yeast duplicates. Their study shows that 73.3% gene pairs with an average $K_s$ of 0.064 have already evolved different expression in at least one human tissue. Our result using EST data further shows that more than 80% of the gene pairs in the rice recently duplicated segments could have acquired diverged expression even though their average $K_s$ is only 0.08. This conclusion is also supported by a previous report that shows diverged expression between some duplicated genes on the rice chromosomes 11 and 12 (Rice Chromosomes 11 and 12 Sequencing Consortia, 2005). Interestingly, rice chromosomes 11 and 12 are rich in disease-resistant genes (Zhou et al., 2004; Rice Chromosomes 11 and 12 Sequencing Consortia, 2005; Wu et al., 2005), such as WRKY which plays significant roles in response to biotic and abiotic stresses, and in developmental process (Eulgem et al., '99; Woese, 2002; Wu et al., 2005). In light of these observations, the rapid expression differentiation between genes in the duplicate segments might have an important impact on the phenotypic evolution of rice due to the large number of genes involved.

In summary, the results from the rice segmental duplication demonstrate that duplicate segments
experienced overall accelerated evolution at the levels of gene order, gene sequence, and gene expression. These observations suggest that relaxation of selection after duplication also holds true for duplicated segments. These large-scale changes might have played important roles in the evolution of rice. It is expected that large-scale duplications in other organisms might also have similar evolutionary patterns.

ACKNOWLEDGMENTS

The authors thank Haijing Yu for her critical reading on the manuscript and English editing.

LITERATURE CITED


