Long-range and targeted ectopic recombination between the two homeologous chromosomes 11 and 12 in *Oryza* species

Research article

Jacquemin J*[1], Chaparro C[1], Laudié M[1], Berger A[2], Gavory F[2], Goicoechea JL[3], Wing RA[3], Cooke R[1] .

[1] Laboratoire 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, Perpignan, Cedex, France.

[2] Genoscope-Centre National de Séquençage, CP5706 91057 EVRY-CEDEX

[3] Arizona Genomics Institute, The University of Arizona, Tucson, AZ, USA.

*corresponding author: Julie Jacquemin

current address: Arizona Genomics Institute, The University of Arizona, Tucson, AZ, USA

email: juliej@cals.arizona.edu

Tel.: +1 (520) 626-9601

Fax: +1 (520) 232-4762

Running title: Recurrent gene conversion in rice

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

Abbreviations : MYA million years ago, WGD whole genome duplication, DSBR double strand break repair, HR homologous recombination, HJ Holliday junction, BIR break-induced replication, SDSA synthesis-dependent strand annealing, SIC simple indel coding, ML maximum likelihood, BI bayesian inference, Os *Oryza sativa*, Ob *Oryza brachyantha*, Og *Oryza glaberrima*

Abstract

Whole genome duplication (WGD) and subsequent evolution of gene pairs have been shown to have shaped the present day genomes 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 whole genome duplication ~50-70 million years ago (MYA) common to all cereals, and a segmental duplication between chromosomes 11 and 12 as recently as 5 MYA. 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 non-coding) sequences between two cultivated (*O. sativa, O. glaberrima*) and one wild (*O. brachyantha*) rice species using a novel approach of topological incongruency. The high levels of intra-species conservation of both gene and non-gene sequences, particularly in *O. brachyantha*, indicate long-range conversion events less than 4 MYA 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.

Introduction

The availability of genome sequences from closely-related species, such as yeasts (reviewed in Dujon 2010) or Drosophila (Hahn, Han MV and Han S 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 *O. sativa* ssp.

japonica c.v. Nipponbare 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 10 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 (van de Peer, Maere and Meyer 2009). Preliminary analysis of the rice RefSeq suggested a whole genome duplication, 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; The rice chromosomes 11 and 12 Sequencing Consortia 2005). More recent studies by ourselves (Jacquemin, Laudié and Cooke 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, Laudié and Cooke 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 sub-telomeric 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; Duret and Galtier 2009; De Muyt et al. 2009; Llorente, Smith and Symington 2008): 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 chromatids or homologous loci. As this process has mainly been described for multigene families and tandemly-duplicated genes (Gao and Innan 2004; Xu et al. 2008; Yang et al. 2009; Hogan and Bettencourt 2009; Ezawa et al. 2010), the long-term conservation of large genomic regions in rice and other cereals was unexpected and raises questions on the extent and pattern of gene conversion in plant genome evolution, as well as the recombination mechanisms involved.

Previous studies on the evolution of the region duplicated between chromosomes 11 and 12 were carried out either on the two very closely-related *O. sativa* subspecies (Wang et al. 2007) or widely-divergent species (rice, sorghum and *Brachypodium distachyon*), largely concentrating on protein coding sequences. We chose three species from the *Oryza* genus to carry out a deep comparative study of the duplication at the genome level. In addition to the RefSeq, we selected two annual African species, *O. glaberrima* S. (2n=24, AA genome) and *O. brachyantha* Chev. Et Roehr. (2n=24, FF genome). The former has the same genome type as Asian domesticated rice (Linares 2002), while the latter, which diverged from the AA lineage ~15 MYA (Tang et al. 2010), has the smallest genome in the genus (340 Mb) (Uozu et al. 1997) and may display a faster evolution rate (Zou et al. 2008). The divergence of *O.*

sativa and *O. glaberrima* genomes is estimated between 0.6 and 1 MYA (Ge, Guo and Zhu 2005; Zhu and Ge 2005; Roulin et al. 2010).

Here we studied the extent and pattern of paralogous conversion between chromosomes 11 and 12 since the *O. brachyantha*/genome AA and *O. glaberrima*/*O. sativa* divergences, focusing on a region from 1.5 to 2.5 Mb overlapping the colinearity breakpoint. We show recent long-range conversion, particularly in *O. brachyantha*, involving both coding and non-coding sequences. The breakpoint is located in syntenic positions in all three species and we discuss the mechanisms that could explain these observations.

Materials and Methods

A detailed version of all Methods is available in supplementary text S1.

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 1e-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, Higgins and Gibson 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). 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 non-synonymous (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, Thornton and Long 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 likelihood ratio test (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 Datamonkey webserver (Kosakovsky Pond and Frost 2005), with all sequences fitting the

Hasegawa-Kishono-Yano (HKY85) model. ML was implemented with PhyML 3.0 (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 BI. 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 2008). 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 to the paralogous pairs (figure 1). *O. sativa* and *O. glaberrima* orthologous contigs display weaker divergence than with the FF species. A complete list of large structural variants (> 5 kb) is presented in detail in supplementary table S1. Indels involve both genes and repeat elements. The most striking rearrangement is a specific inversion at the 5' end on *O. glaberrima* chromosome 12 covering at least 82 kb. Overall, contigs from *O. glaberrima* and *O. brachyantha* are shorter, compared with the RefSeq *O. sativa* chromosomes (table 2). The expansion of the chromosome 11 segment in *O. sativa* compared to *O. glaberrima* results from eight insertions/deletions (indels) for a total of 83.4 kb (table 2). Four events, three indels and one tandem duplication (~14 kb, 1832000-1846500 bp), contribute to the size expansion of both AA genomes compared to *O. brachyantha* (supplementary table S1 and table 2). *O. brachyantha* chromosome 11 also displays a specific inverted duplication of 42 kb.

The expansion of the chromosome 12 segment in *O. sativa* compared to *O. glaberrima* is explained by five insertions, but the Og12 region also displays two large insertions and a tandem duplication (table 2). The size difference is particularly striking for *O. brachyantha* chromosome 12 (628505 bp compared with 966580 bp for the RefSeq orthologous region). 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: 1925654-2117228), 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). 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). We observed four large events shared by paralogous chromosome pairs or in syntenic positions. *O. 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 outgroup (supplementary figure S1) 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 to *O. brachyantha*, the first varying from 10 to 38 kb, and the second covering approximately 29 kb (supplementary table S1).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 (sextets: see figure 1 and supplementary table S2). 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*, while 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.

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 their own clade, 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 Bayesian inference methods, 24 out of 50 sextets present topology 1 and 15 topology 1M (table 3 and supplementary table S2). For two 1M sextets (Os11g04200 and Os11g04500) 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 incongruent between Bayesian and Maximum likelihood 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.

Non-genic 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 non-gene regions. We first tested the frequently-used program GENECONV on the CDS sextet data set (see results in supplementary table S2). 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, Casalo and Hahn 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 bp 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 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). The distal region (zone 1), where more than 80% of the trees display topologies 1 (515) or 1M (136), extends to block 1047501-1048000 (total 800 blocks), corresponding to 2108257 bp and 2151747 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 (856501-857000), and only one (663001-663500) 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 (~106000 bp), 414 (135000 bp) and 415 (126000 bp) are located in intergenic regions for *O. sativa*, *O. glaberrima* and *O. brachyantha* respectively, whereas 315 (120000 bp) 238 (94000 bp) and 237 (93000 bp) overlap protein-coding sequences.

The proximal region (zone 2) extends from block 1182001-1182500 to the end and covers RefSeq chromosome 11 from 2195478 bp and chromosome 12 from 2214633 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 1552001_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 non-coding 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 nine 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 Bayesian method (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 2100000-2106000 bp on *O. sativa* chromosome 11, corresponding to 2120000-2128000 bp on *O. sativa* chromosome 12, in agreement with the topological analysis on sextets. It is at syntenic locations in *O. glaberrima*, between 591500-597000 bp and 599500-606000 bp on *O. glaberrima* contigs 11 and 12, respectively. The *O. brachyantha* breakpoint is slightly more proximal, between 518000-519000 bp and 374000-375000 bp on contigs 11 and 12 (2118000 and 2155000 bp on RefSeq chromosome 11 and 12 respectively). These breakpoints all map to the intermediate region 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 (figure 4). Distributions of distance values for zone 1 (figure 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). The second peak represents the distances between the sequences in the non-converted 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 MYA for *O. brachyantha* in the genus and ~0.8 MYA 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) using the formulas:

x(p11,p12)=(median(d(p11,p12))x0.8)/mean(median(d(Os11,Og11)),median(d(Os12,Og12)))

x(p11,p12)=(median(d(p11,p12))x15)/mean(median(d(Os11,Ob11)),median(d(Os12,Ob12)))

(where p11 and p12 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 2.5-4.0 MYA for the AA species and 1.5-3.5 MYA for *O. brachyantha*, much lower than previous estimations, from 5 to 21 MYA, given for the whole region in *O. sativa* but based only on coding sequences (Wang et al. 2005; The Rice Chromosomes 11 and 12 Consortia 2005; Goff et al. 2002; Salse et al. 2008). Using pairs from zone 2, we calculate 15-55 MYA for the AA species and 20-50 MYA for *O. brachyantha*. Age estimations for the WGD event are somewhat greater (50-90 MYA, 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 non-coding 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 towards 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, Yang and Gu 2005) or micro-arrays (Throude et al. 2009) did not detect significantly different expression patterns between gene pairs in the 11-12 duplication. However Yim, Lee and Jang (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 non-synonymous/synonymous ratios (ω) 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 ω 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 ω=1 (neutrality level), and one pair for each other species displayed ω>1 (indicator of positive selection). Under the likelihood ratio test (LRT), among 245 paralogous pairs, 112 showed an ω 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 ω 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 (d.f.=4, P=0.05), indicating significant variation in selective constraints among sites. For 22 sextets, both models M2 and M8, which allow the ω ratio to exceed 1, fit the data significantly better (d.f.=2, P=0.05) than models M1 and M7 (supplementary table S2). The number of sites with ω >1 varied from five 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, Ponger and Gaut 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 to 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 non-crossover 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, Smith and Symington 2008). It continues with a processive replication fork, and DNA synthesis proceeds to the end of the donor chromosome (Llorente, Smith and Symington 2008). BIR have been implicated in homogenization of subtelomeric regions in yeast (Bosco and Haber 1998) and their relative frequency increases towards telomeric regions, in which their consequences are less deleterious than in other regions of the chromosomes (Ricchetti, Dujon and Fairhead 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, Smith and Symington 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 non-homologous 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, Tang and Paterson 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 chromosome 11-12 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 inter-specific rearrangements larger than 10 kb with genes involved, we observe one insertion (5 genes) specific to the RefSeq, one expansion for the AA lineage compared to *O. brachyantha* (4 genes), and two tandem duplication, one for *O. brachyantha* (involving 10 genes) and one for the AA species (2 genes), all on the chromosome 11 1 Mb-segment. On chromosome 12, we observed one inversion (7 genes), one expansion (2 genes) specific to *O. glaberrima*, and one expansion (5 genes) on *O. sativa*. Contractions compared to 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 3-gene segment translocation in *O. coarctata*, a 3-gene segment insertion in *O. sativa*, and a single gene tandem duplication in *O. granulata*).

Wang, Tang and Paterson (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 to 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 inter-species 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, Myers and McVean 2007; Innan 2009; Innan and Kondrashov 2010) and the occurrence of conversion between homeologous genes during polyploid formation and divergence (Udall, Quijada and Osborn 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 is available on the MBE web site.

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.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403-410.

Anderson LK, Salameh N, Bass HW, Harper LC, Cande WZ, Weber G, Stack SM. 2004. Integrating Genetic Linkage Maps With Pachytene Chromosome Structure in Maize *Genetics*. **166**:1923-1933.

Benovoy D, Drouin G. 2009. Ectopic gene conversions in the human genome. *Genomics* **93**:27–32.

Betran E, Thornton K, Long M. 2002. Retroposed New Genes Out of the X in Drosophila. *Genome Res*. **12**:1854-1859.

Bosco G, Haber JE. 1998. Chromosome break-induced DNA replication leads to nonreciprocal translocations and telomere capture. *Genetics* **150**:1037-1047.

Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream M. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**:2672-2676.

Chaw SM, Chang CC, Chen HL, Li WH. 2004. Dating the monocot–dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J. Mol. Evol.* **58**:424–441.

Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**:762–775.

Darling ACE, Mau B, Blatter FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**:1394-1403.

De Muyt A, Mercier R, Mézard C, Grelon M. 2009. Meiotic recombination and crossovers in plants. *Genome Dyn.* **5**:14-25.

Dujon B. 2010. Yeast evolutionary genomics. *Nat. Rev. Genet.* **11:**512-524.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Ann. Rev Genomics Hum. Genet.* **10**:285-311.

Ezawa K, Ikeo K, Gojobori T, Saitou N. 2010. Evolutionary Pattern of Gene Homogenization between Primate-Specific Paralogs after Human and Macaque Speciation using the 4-2-4 method. *Mol. Biol. Evol.* **27**:2152-2171.

Fan C, Zhang Y, Yu Y, Rounsley S, Long M, Wing RA. 2008. The subtelomere of *Oryza sativa* chromosome 3 short arm as a hot bed of new gene origination in rice. *Mol. Plant* **1**:839-850.

Gao L, Innan H. 2004. Very Low Gene Duplication Rate in the Yeast Genome. *Science* **306**:1367-1370.

Gay J, Myers S, McVean G. 2007. Estimating Meiotic Gene Conversion Rates From Population Genetic Data. *Genetics* **177**:881-894.

Ge S, Guo Y, Zhu Q. 2005. Molecular phylogeny and divergence of the rice tribe Oryzeae, with special reference to the origin of the genus *Oryza*. In *Rice Is Life: Scientific Perspectives for the 21st Century* (ed K Toriyama, KL Heong, and B Hardy), pp40–44

International Rice Research Institute Publications.

Goff SA, Ricke D, Lan TH et al. (55 co-authors). 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**:92-100.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725-736.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**:696-704.

Hahn MW, Han MV, Han S. 2007. Gene Family Evolution across 12 Drosophila Genomes. *PLoS Genet.* **3**:e197.

Hogan CC, Bettencourt BR. 2009. Duplicate Gene Evolution Toward Multiple Fates at the Drosophila melanogaster HIP/HIP-Replacement Locus. *J. Mol. Evol.* **68**:337–350.

Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754-755.

Innan H. 2009. Population genetic models of duplicated genes. *Genetica* **137**:19–37.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**:97–108.

Jacquemin J, Laudié M, Cooke R. 2009. A recent duplication revisited: phylogenetic analysis reveals an ancestral duplication highly-conserved throughout the *Oryza* genus and beyond. *BMC Plant Biol.* **9**:146.

Kao FI, Cheng YY, Chow TY, Chen HH, Liu SM, Cheng CH, Chung MC. 2006. An integrated map of *Oryza sativa* L chromosome 5. *Theoret. Appl. Genet.* **112**:891–902.

Kijima TE, Innan H. 2010. On the estimation of the insertion time of LTR retrotransposable elements. *Mol. Biol. Evol.* **27**:896-904.

Kosakovsky Pond SL, Frost SD. 2005. DATAMONKEY: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21**:2531-2533.

Koszul R, Fischer G. 2009. A prominent role for segmental duplication in modeling Eukaryotic genomes. *C. R. Biologies* **332**:254266.

Kuang H, Woo SS, Meyers BC, Nevo E, Michelmore RW. 2004. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *The Plant Cell* **16**:2870-2894.

Lin YS, Byrnes JK, Hwang JK, Li WH. 2006. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. *Proc. Natl. Acad. Sci. U S A.* **103**:14412-14416.

Linares OF. 2002. African rice (*Oryza glaberrima*): History and future potential. *Proc. Natl. Acad. Sci. U S A.* **99**:16360-16365.

Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet.* **21**:602–607.

Llorente B, Smith CE, Symington LS. 2008. Break-induced replication. *Cell Cycle* **7**:859-864.

Lu F, Ammiraju JS, Sanyal A et al. (15 co-authors). 2009. Comparative sequence analysis of MONOCULM1 -orthologous regions in 14 *Oryza* genomes. *Proc. Natl. Acad. Sci. U S*

A. **106**:2071-2076.

Lydeard JR, Jain S, Yamaguchi M, Haber JE. 2007. Break-induced replication and telomerase-independent telomere maintenance require Pol32. *Nature* **448**:820–823.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454**:479–485.

Mansai SP, Innan H. 2010. The Power of the Methods for Detecting Interlocus Gene Conversion. *Genetics* **184**:512-527.

McGrath CL, Casola C, Hahn MW. 2009. Minimal Effect of Ectopic Gene Conversion Among Recent Duplicates in Four Mammalian Genomes. *Genetics* **182**:615-622.

Mézard C. 2006. Meiotic recombination hotspots in plants. *Biochem. Soc. Trans.* **34**:531–534.

Mondragon-Palomino M, Gaut BS. 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **22**:2444-2456.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929-936.

Ogden TH, Rosenberg MS. 2007. How should gaps be treated in parsimony? A comparison of approaches using simulation. *Mol. phylogenet. Evol.* **42**:817–826.

Paterson AH, Bowers JE, Bruggmann R et al. (45 co-authors). 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**:551–556.

R Development Core Team. 2008. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna Austria http://wwwR-project.org.

Ricchetti M, Dujon B, Fairhead C. 2003. Distance from the Chromosome End dertermine the efficiency of double strand break repair in subtelomeres of haploid yeast. *J. Mol. Biol.* **328**:847-862.

Rizzon C, Ponger L, Gaut BS. 2006. Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in Arabidopsis and Rice. *Plos Computational Biology* **9:**e115.

Roulin A, Chaparro C, Piégu B, Jackson S, Panaud O. 2010. Paleogenomic Analysis of the Short Arm of Chromosome 3 Reveals the History of the African and Asian Progenitors of Cultivated Rices. *Genome Biol. Evol.* **2010**:132-139.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944-5

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF. 2009. Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytol.* **186**:123-134.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. 2008. Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution. *Plant Cell* **20**:11-24.

Sanyal A, Jetty AS, Lu F et al. (14 co-authors). 2010. Orthologous comparisons of the Hd1 region across genera reveal Hd1 gene lability within diploid *Oryza* species and disruptions to microsynteny in sorghum. *Mol. Biol. Evol.* **27(11)**:2487-2506.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**:526-538.

Simmons MP, Ochoterena H. 2000. Gaps as characters in sequence-based phylogenetic analyses *Syst. Biol.* **49**:369–381.

Soderlund C, Nelson W, Shoemaker A, Paterson A. 2006. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res.***16**:1159-68.

Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**:GC1-10.

Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW. 1983. The double-strand-break repair model for recombination. *Cell* **33**:25–35.

Tang L, Zou X, Achoundong G, Potgieter C, Second G, Zhang D, Ge S. 2010. Phylogeny and biogeography of the rice tribe (Oryzeae): evidence from combined analysis of 20 chloroplast fragments. *Mol. Phylogenet. Evol.* **54**:266–277.

Teshima KM, Innan H. 2004. The effect of gene conversion on the divergence between duplicated genes. *Genetics* **166**:1553-1560.

The Rice Chromosomes 11 and 12 Sequencing Consortia (115 co-authors). 2005. The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biol.* **3**:20.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-4680.

Throude M, Bolot S, Bosio M et al. (14 co-authors). 2009. Structure and expression analysis of rice paleo duplications. *Nucleic Acids Res.* **37**:1248-1259.

Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, Jackson SA, Gaut BS, Ma J. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Research* **19:**2221-2230.

Udall JA, Quijada PA, Osborn TC. 2005. Detection of Chromosomal Rearrangements Derived From Homeologous Recombination in Four Mapping Populations of Brassica napus L. *Genetics* **169**:967-979.

Uozu S, Ikehashi H, Ohmido N, Ohtsubo H, Ohtsubo E, Fukui K. 1997. Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*. *Plant Mol. Biol.* **35**:791-799.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**:725–732.

Wang X, Shi X, Hao B, Ge S, Luo J. 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.***165**: 937–946.

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH. 2007. Extensive Concerted Evolution of Rice Paralogs and the Road to Regaining Independence. *Genetics* **177**:1753-1763.

Wang X, Tang H, Bowers JE, Paterson AH. 2009. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res.* **19**:1026-1032.

Wang X, Tang H, Paterson AH. 2011. Seventy Million Years of Concerted Evolution of a Homoeologous Chromosome Pair, in Parallel, in Major Poaceae Lineages. *Plant Cell*

23:27-37.

Wing RA, Ammiraju JS, Luo M et al. (17 co-authors). 2005. The *Oryza* Map Alignment Project: The Golden Path to Unlocking the Genetic Potential of Wild Rice Species. *Plant Mol. Biol.* **59**:53-62.

Wu J, Mizuno H, Hayashi-Tsugane M et al. (22 co-authors). 2003. Physical maps and recombination frequency of six rice chromosomes. *Plant J.* **36**:720-730.

Xu S, Clark T, Zheng H, Vang S, Li R, Wong GK, Wang J, Zheng X. 2008. Gene conversion in the rice genome. *BMC Genomics* **9**:93.

Yang Z. 1998. Likelihood Ratio Tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol*. **15(5)**:568-573.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**:1586-1591.

Yang Z, Gao Q, Sun C, Li W, Gu S, Xu C. 2009. Molecular evolution and functional divergence of HAK potassium transporter gene family in rice (*Oryza sativa* L). *J. Genet. Genomics* **36**:161–172.

Yim WC, Lee BM, Jang CS. 2009. Expression diversity and evolutionary dynamics of rice duplicate genes. *Mol. Genet. Genomics* **281**:483–493.

Yu J, Wang J, Lin W et al. (117 co-authors). 2005. The Genomes of *Oryza sativa*: A History of Duplications. *PLoS Biol.* **3**:e38.

Zhu Q, Ge S. 2005. Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. *New Phytol.* **167**:249–265.

Zou X, Zhang F, Zhang J, Zang L, Tang L, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol.* **9**:R49-R49.

Table 1.

General features of contigs of *O. glaberrima* and *O. brachyantha* and orthologous segments on the MSU Rice genome annotation v6.1 pseudomolecules of *O. sativa* ssp. *japonica* (RefSeq)

	RefSeq		*O. glaberrima*		*O. brachyantha*	
	Ch11	Ch12	Ch11	Ch12	Ch11	Ch12
Length (bp)	1090000	1200000	874636	971932	857170	628505
Genbank accession	/	/	FQ378034	FQ377974	FQ378032	FQ378033
Coordinates (kb)*	1.42-2.51	1.34-2.54	1.44-2.40	1.52-2.53	1.43-2.50	1.57-2.53
Number of genes	180	168	116	104	116	74
Density (genes/kb)	0.165	0.139	0.132	0.107	0.135	0.117
% GC	42.76	43.41	42.77	42.99	41.07	40.49
Coding %	37.4	32.4	33.6	29.8	37.1	30.4
TE %	15.3	33	10.2	8.5	1.8	3.7
Class I TE	18	47	18	18	6	4
Class II TE	43	36	32	24	3	5
MITES	153	136	115	101	74	117
Other	2	5	0	0	0	0

NOTE : Numbers of genes do not include alternative splicing forms and CDS with TE-related annotations.
*coordinates are relative to the RefSeq

Table 2.

Summary of expansion events between analyzed orthologous segments on chromosomes 11 and 12 in *O. sativa* (Os), *O. glaberrima* (Og) and *O. brachyantha* (Ob).

Segment size difference			Number of expansions	Size range	Total size
Os11/Og11	85.4	**Os11**	8	5.5-20.3	83.4
		Og11	1	9.6	9.6
Os11/Ob11	212.8	**Os11**	4	7.4-24.7	56.1
		Ob11	2	7.3-42.3	49.6
Os12/Og12	38.1	**Os12**	5	6.5-33.5	68.3
		Og12	3	11.6-15.5	39.7
Os12/Ob12	331.5	**Os12**	3	19-158	272.2
		Ob12	1	8.8	8.8

NOTE: Both indels and tandem duplications more than 5 kb long are considered. Sizes are indicated in kb.

Table 3.
Topology data for sextets of CDS and whole contig blocks (divided in three zones)

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

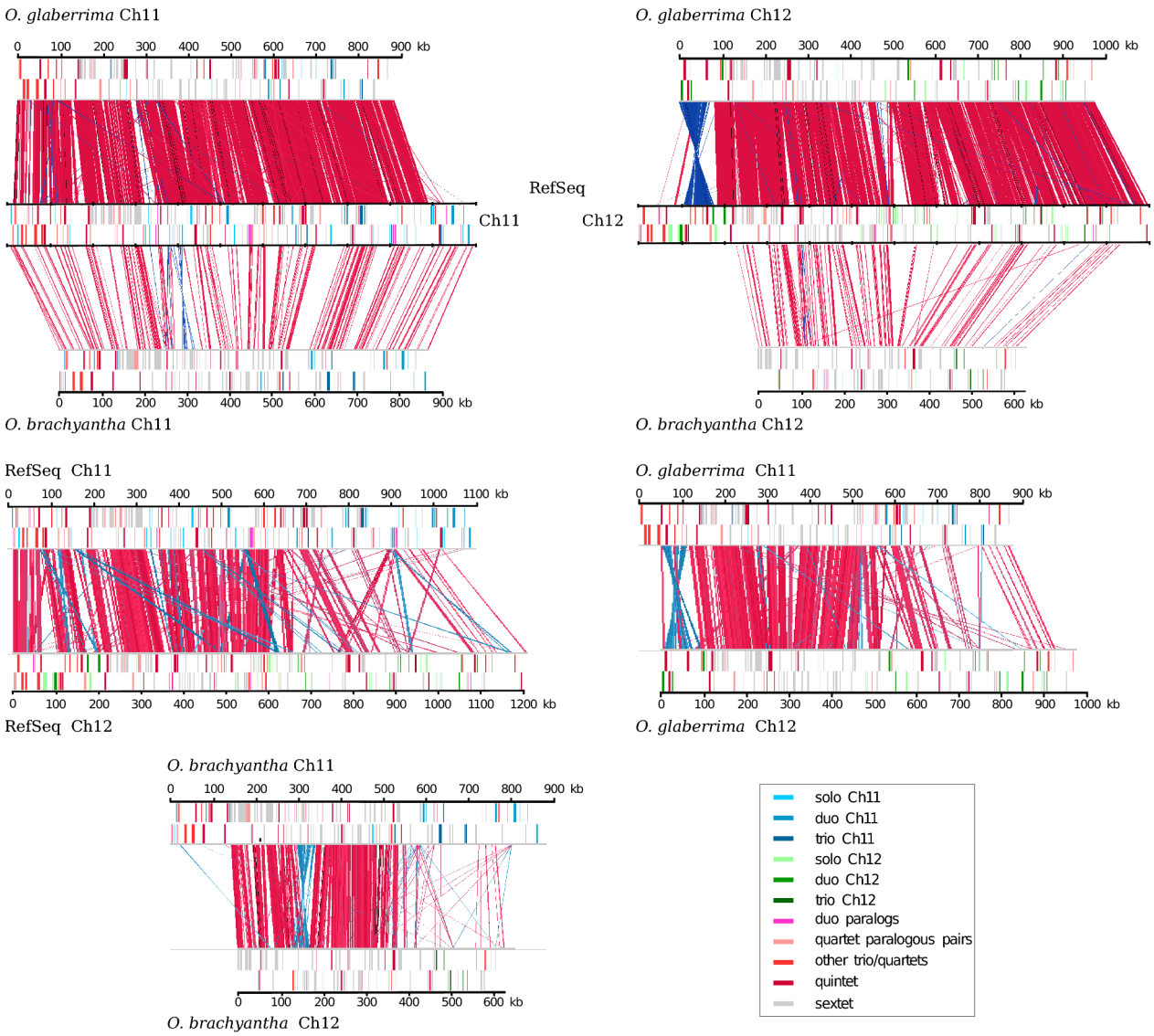

Figure 1. Graphical representation of synteny between the orthologous and paralogous 11 and 12 contigs in the RefSeq, *O. glaberrima* and *O. brachyantha*. Coordinates are indicated in kb. The segments for the RefSeq correspond to 1.42-2.51 Mb on chromosome 11 and 1.34-2.54 Mb on chromosome 12. Lines 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.

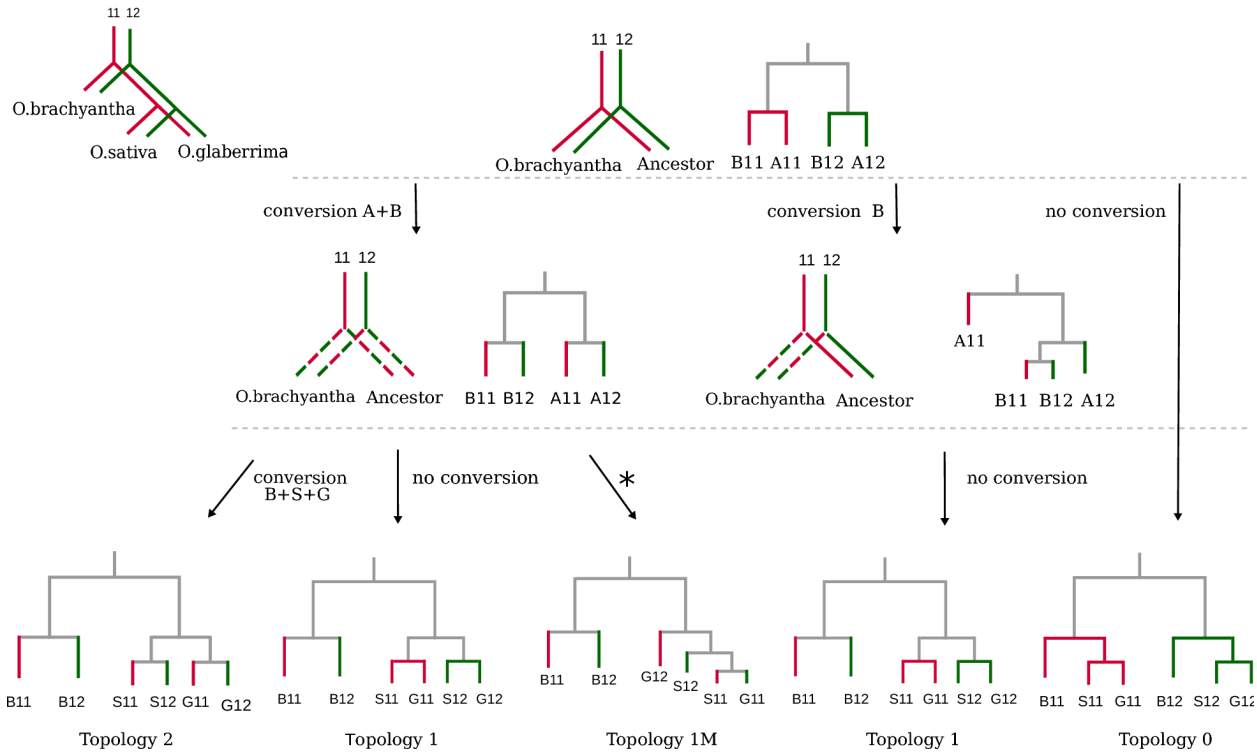

Figure 2. Evolutionary scheme of the 11-12 duplicated block in the *Oryza* genus, as a function of conversion events in the FF and AA lineages. A=Ancestor of AA lineage, B=*O. brachyantha*, G=*O. glaberrima*, S=*O. sativa*. Conversion is inferred based on topological incongruency with the topology 0. *Only one example of topology 1M is shown as we group several trees in this class: the first have only one orthologous pair, S11-G11 or S12-G12, clustered in a terminal branch, while the two remaining genes form intermediate branches between this cluster and the *O. brachyantha* node. The second have only one paralogous pair, S11-S12 or G11-G12, clustered in a terminal branch, while the two remaining genes form intermediate branches between this cluster and the *O. brachyantha* node. This topology is ambiguous as it could reveal (1) too weak divergence of the four AA genes to resolve their phylogenetic relationships, (2) the strong divergence of one of these genes blurring their true relationships, (3) conversion in one of the AA lineages after their divergence.

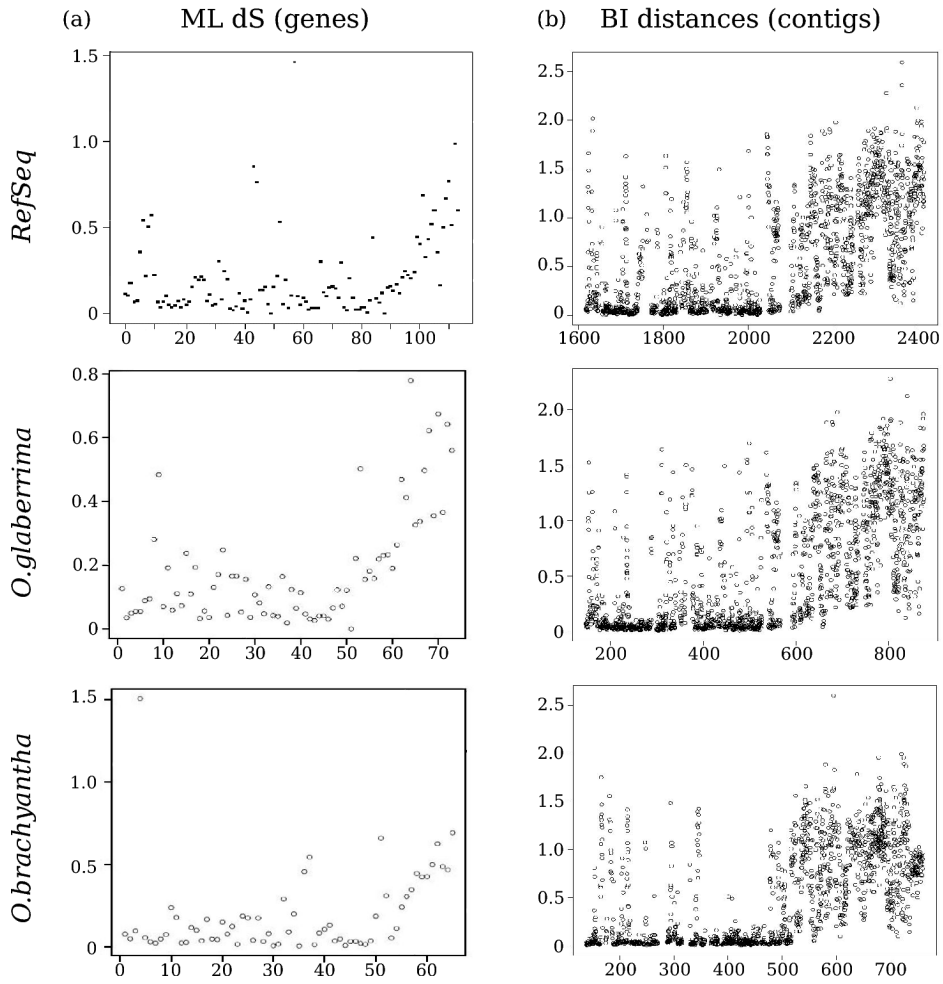

Figure 3. Spatial distribution of synonymous substitution rates (dS) between paralogous gene pairs computed with the basic ML codon model, plotted against the number of pairs (a) and BI nucleotide distances between paralogous 500 bp fragments of the whole contig alignment, plotted against the chromosome 11 coordinates (kb) for the three species (b).

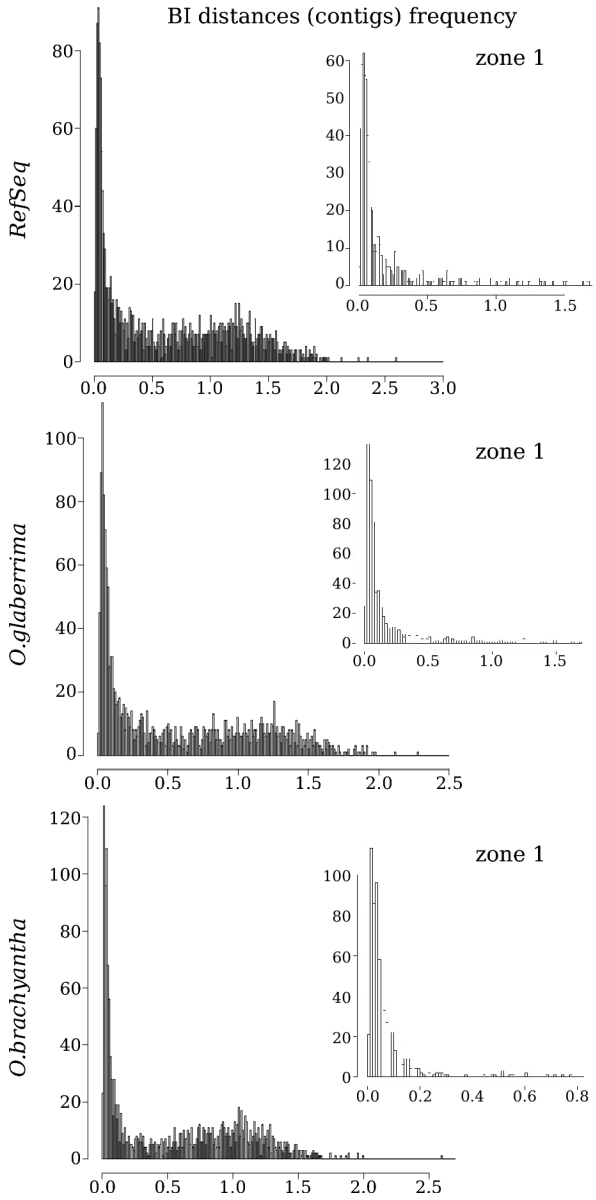

Figure 4. Frequency distribution of BI nucleotide distances between paralogous 500 bp fragments of the whole contig alignment. The insert histograms show distance distributions in converted zone 1 only.