the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2010) **63**, 430–442 doi: 10.1111/j.1365-313X.2010.04251.x

Spatio-temporal patterns of genome evolution in allotetraploid species of the genus *Oryza*

Jetty S. S. Ammiraju[1], Chuanzhu Fan[1], Yeisoo Yu[1], Xiang Song[1], Karen A. Cranston[2], Ana Clara Pontaroli[3], Fei Lu[4], Abhijit Sanyal[5], Ning Jiang[6], Teri Rambo[1], Jennifer Currie[1], Kristi Collura[1], Jayson Talag[1], Jeffrey L. Bennetzen[3], Mingsheng Chen[4], Scott Jackson[5] and Rod A. Wing[1,*]

[1]*Arizona Genomics Institute, School of Plant Sciences, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA,*
[2]*Biodiversity Synthesis Center, Field Museum of Natural History, 1400 South Lake Shore Drive, Chicago, IL 60605, USA,*
[3]*Department of Genetics, University of Georgia, Athens, GA 30602–7223, USA,*
[4]*State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China,*
[5]*Department of Agronomy, Purdue University, West Lafayette, IN 47907–2054, USA, and*
[6]*Department of Horticulture, Michigan State University, East Lansing, MI 48823, USA*

Received 10 March 2010; revised 26 April 2010; accepted 28 April 2010; published online 15 June 2010.
*For correspondence (fax +1 520 621 1259; e-mail rwing@ag.arizona.edu).

SUMMARY

Despite knowledge that polyploidy is widespread and a major evolutionary force in flowering plant diversification, detailed comparative molecular studies on polyploidy have been confined to only a few species and families. The genus *Oryza* is composed of 23 species that are classified into ten distinct 'genome types' (six diploid and four polyploid), and is emerging as a powerful new model system to study polyploidy. Here we report the identification, sequence and comprehensive comparative annotation of eight homoeologous genomes from a single orthologous region (*Adh1–Adh2*) from four allopolyploid species representing each of the known *Oryza* genome types (BC, CD, HJ and KL). Detailed comparative phylogenomic analyses of these regions within and across species and ploidy levels provided several insights into the spatio-temporal dynamics of genome organization and evolution of this region in 'natural' polyploids of *Oryza*. The major findings of this study are that: (i) homoeologous genomic regions within the same nucleus experience both independent and parallel evolution, (ii) differential lineage-specific selection pressures do not occur between polyploids and their diploid progenitors, (iii) there have been no dramatic structural changes relative to the diploid ancestors, (iv) a variation in the molecular evolutionary rate exists between the two genomes in the BC complex species even though the BC and CD polyploid species appear to have arisen <2 million years ago, and (v) there are no clear distinctions in the patterns of genome evolution in the diploid versus polyploid species.

Keywords: ***Oryza*, polyploidy, BACs, comparative genomics, genome evolution.**

INTRODUCTION

Polyploidy is a major evolutionary force for the diversification of flowering plants and their genomes (reviewed by Wendel, 2000; Leitch and Leitch, 2008; Doyle *et al.*, 2008; Jackson and Chen, 2009). Some patterns associated with the evolution of structural and functional plasticity of polyploid genomes, in particular those from recently evolved allopolyploids, have emerged from studies in wheat, *Brassica*, cotton, soybean and Arabidopsis; however, the nature, pace and magnitude of these patterns differ between various species (Cronn *et al.*, 1999; Comai, 2000, 2005; Comai *et al.*, 2000; Liu *et al.*, 2001; Osborn *et al.*, 2003; Pires *et al.*, 2004; Wang *et al.*, 2004, 2006; Salmon *et al.*, 2005; Udall *et al.*, 2005; Innes *et al.*, 2008; Ha *et al.*, 2009; Jackson and Chen, 2009; Ni *et al.*, 2009). To assess the relevance of these emerging patterns, additional studies are required across diverse plant families (Doyle *et al.*, 2008).

The genus *Oryza*, which includes rice and closely related wild relatives, has emerged as a powerful system to study the modes and mechanisms of genome evolution (Wing *et al.*, 2005; Ammiraju *et al.*, 2006, 2007, 2008; Piegu *et al.*, 2006; Ma *et al.*, 2007; Lu *et al.*, 2009). *Oryza* has undergone rapid diversification (Ge *et al.*, 1999; Vaughan *et al.*, 2003;

Zou *et al.*, 2008) within a short evolutionary time span of 15 million years (Ammiraju *et al.*, 2008; Lu *et al.*, 2009). The genus comprises approximately 23 species that have been grouped into six diploid and four allotetraploid genome types (BBCC, CCDD, HHJJ and KKLL) (Nayar, 1973; Aggarwal *et al.*, 1997; Ge *et al.*, 1999; Lu *et al.*, 2009). All *Oryza* polyploids are wild and contain important phenotypic traits that have the potential for use to improve cultivated rice (Brar and Khush, 1997). *Oryza* polyploids have been further broadly classified into two polyploid genome complexes, CC and HH. Six of the nine *Oryza* polyploid species belong to the CC polyploid complex (Vaughan *et al.*, 2003). Polyploid species containing the BB and CC genome types arose recently (Lu *et al.*, 2009; Tang *et al.*, 2009; Wang *et al.*, 2009) as a result of at least three independent polyploidization events (Vaughan *et al.*, 2003). In contrast, much of the diversity in the CCDD, HHJJ and KKLL genomes resulted from single polyploidization events (Ge *et al.*, 1999; Vaughan *et al.*, 2003; Bao and Ge, 2004; Guo and Ge, 2005). Among the species within the CC genome polyploid complex, both putative diploid donor species for the BBCC genome species [*O. punctata* (BB) and *O. officinalis* (CC)] (Ge *et al.*, 1999) and one presumptive donor for the CCDD genome species [*O. officinalis* (CC)] are extant. However, the diploid donor species for the DD, HH, JJ, KK and LL genomes are presumed to be extinct, and are currently known only as homoeologous genomes of polyploid *Oryza* species. Several studies have suggested that the EE genome is the potential diploid progenitor of the DD genome (Nayar, 1973; Fukui *et al.*, 1997; Ge *et al.*, 1999, 2001; Li *et al.*, 2001; Bao and Ge, 2004; Guo and Ge, 2005; Gong and Bao, 2008; Hirsch *et al.*, 2009), whereas the LL genome of *O. coarctata*, although related to the HH genome of *O. ridleyi*, was recently reclassified from the HH to the LL genome type (Lu *et al.*, 2009).

Because of the importance of rice as a major food crop and model organism, large-scale efforts are underway to understand the genome organization and evolution of all *Oryza* species, including the polyploids. For example, numerous public resources have been developed for the genus, including well-curated seed collections (http://beta.irri.org/seeds/; http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp; Jackson, 1997; Jackson *et al.*, 1999), inter-specific mapping populations (http://www.rgrc.dna.affrc.go.jp/stock.html and http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp), and a set of bacterial artificial chromosome (BAC)/end sequence physical maps (Ammiraju *et al.*, 2006; Kim *et al.*, 2008; http://www.omap.org) for 16 different species representing ten known genome types (diploid as well as polyploid) and various evolutionary time points within the 15 million years of *Oryza* radiation (Ammiraju *et al.*, 2008; Lu *et al.*, 2009).

Macro-level comparative analyses of these BAC-based physical maps and end sequences from the polyploid *Oryza* species [i.e. *O. minuta* (BBCC), *O. alta* (CCDD), *O. ridleyi* (HHJJ) and *O. coarctata* (LLKK)] with putative progenitor diploids [i.e. *O. officinalis* (CC) and *O. punctata* (BB)], as well as the cultivated rice genome (AA) revealed that: (i) homoeologous chromosomes of the polyploid species are highly divergent, and individual co-resident genome maps can be developed by BAC-based fingerprinting methods (Ammiraju *et al.*, 2006; Kim *et al.*, 2008), and (ii) part of the divergence is mediated by lineage-specific evolution of transposable elements in diploid and polyploid species (Ammiraju *et al.*, 2006, 2007; Zuccolo *et al.*, 2007, 2008; Kim *et al.*, 2008; Roulin *et al.*, 2008).

Comparative sequence analysis of large DNA segments, especially within well-defined phylogenetic frameworks, has already yielded important insight into the processes and mechanisms driving post-polyploidization genome restructuring and evolution (Cheung *et al.*, 2009; Chantret *et al.*, 2005; Dubcovsky and Dvorak, 2007; Fiebig *et al.*, 2004; Gu *et al.*, 2006; Grover *et al.*, 2004, 2007; Ha *et al.*, 2009; Rana *et al.*, 2004; Wicker *et al.*, 2003; Innes *et al.*, 2008). We recently reported genus-wide evolutionary comparisons of two orthologous genomic regions, *MOC1* and *Adh1* (Ammiraju *et al.*, 2008 and Lu *et al.*, 2009), providing the first glimpse into the nature, mechanistic basis, evolutionary origin and timing of various DNA rearrangements and their impact on *Oryza* genome diversity. These comparisons also revealed similarities and differences in the evolution of *Oryza* genomes in a local context. For example, 78% of the *Adh*1 region in *O. sativa* was found to be composed of tandemly arrayed gene families, whereas the *MOC1* region was predominantly composed of low-copy-number sequences. Analysis of the genomic structural stability across the *Oryza* phylogeny for these two regions revealed that the *Adh1* region was highly unstable (Ammiraju *et al.*, 2008), whereas the *MOC1* regions were conserved, thereby leading to an interesting correlation between regional genome stability and genic composition (Lu *et al.*, 2009). Although the *Adh1* regional analyses were confined to diploid lineages, the *MOC1* investigation included comparisons of both diploid and polyploid *Oryza* genomes, and remains the only systematic comparative genomics study that has addressed genome evolution in all *Oryza* polyploid genomes.

A comprehensive understanding of the molecular evolutionary consequences of polyploidization in *Oryza* cannot be revealed by studying a single locus or region alone. This was clearly demonstrated in allotetraploid cotton, in which the genome dynamics of two regions, *CesA* and *AdhA*, were compared (Grover *et al.*, 2004, 2007). The *AdhA* region was shown to have increased levels of illegitimate recombination and a higher frequency of small deletions, providing evidence for possible genome downsizing after polyploidization. Comparative analysis of the *Adh1* regions of polyploid *Oryza* provides a fitting contrast to the *MOC1* region based on differences in gene family content and

recombinational properties. Here, we report the isolation, sequencing and comparative phylogenomics analyses of eight homoeologous genomic regions from a single orthologous region (the *Adh1* locus) from four allopolyploid species representing each of the known *Oryza* genome types, namely *O. minuta* (BBCC), *O. alta* (CCDD), *O. ridleyi* (HHJJ) and *O. coarctata* (KKLL), in a well-defined phylogenetic context.

In addition to obtaining detailed information on genome microstructure and complexity for five extinct diploid genome types (DD, HH, JJ, KK and LL), this analysis, based on techniques of Bayesian relaxed-clock models, provides a robust temporal framework for the origin of polyploidy in *Oryza*. In addition, comparative analyses of these homoeologous regions, within and across species and ploidy levels, revealed several unique insights into the spatio-temporal dynamics of *Oryza* genome organization and evolution, and their similarities to and differences from known information from *Oryza* and other polyploid systems.

RESULTS

Identification, sequencing and annotation of eight homoeologous regions spanning the *Adh1* region of four allotetraploid *Oryza* species

We identified (Ammiraju *et al.*, 2006 and Figure S1) and sequenced a set of 11 BAC clones spanning the *Adh1* genomic region from the homoeologous genomes of four polyploid *Oryza* species (Table S1). These BACs were manually annotated, resulting in the identification of 87 intact genes and 16 apparent pseudogenes, plus a single case of a gene embedded in a transposable element (TE) (Table S2). For simplicity, when discussing genes and TEs from different species and genome types, the nomenclature shown in Table 1 was used. TE content and diversity are described in Table S3. TE content ranged from 26.4% (for L_{TC}) to 53.9% (for H_{TR}). No major deviations in GC content or exon and intron length were observed compared to the *Oryza* diploid genomes (Ammiraju *et al.*, 2008).

Table 1 Genome nomenclature used to distinguish individual genes, transposable elements and genomic regions from various species and ploidy backgrounds

Species	Genome type	Ploidy type	Genome nomenclature[a]
O. sativa ssp. *japonica*	AA	Diploid	J
O. punctata	BB	Diploid	B_d
O. officinalis	CC	Diploid	C_d
O. australiensis	EE	Diploid	E_d
O. minuta	BBCC	Polyploid	B_{TM}
			C_{TM}
O. alta	CCDD	Polyploid	C_{TA}
			D_{TA}
O. coarctata	KKLL	Polyploid	L_{TC}
			K_{TC}
O. ridleyi	HHJJ	Polyploid	H_{TR}
			J_{TR}

[a]Genes from each *Oryza* species are indicated by the first letter of each species, and the ploidy level is indicated by the subscript 'd' for diploid or 'T' for tetraploid.

Phylogenetic framework, molecular timing of polyploidization events and rate of sequence evolution

Previous sequence analysis of the *Adh1* region from the diploid *Oryza* phylogeny identified a core set of six conserved genes (6–1, 6–2, 7, 8, 9 and 10; Ammiraju *et al.*, 2008) that were used to deduce phylogenetic relationships and the timing of various speciation events. Protein coding sequences from the same core gene set from tetraploid *Oryza* species were similarly identified (with the exception of genes 7 and 9 in the K_{TC} genome) and used to: (i) test the independent phylogenetic evolution of homoeologous pairs of genes since polyploid formation, (ii) estimate the time of polyploid formation, (iii) estimate the effects of selection on homoeologous gene pairs, and (iv) test the equivalence of molecular evolutionary rates of two co-resident genomes relative to diploid progenitors.

Two phylogenetic approaches based on Bayesian and maximum-likelihood methods were used to infer the evolutionary relationships of diploid and polyploid *Oryza* lineages across the genus. Phylogenetic trees for the six core genes are shown in Figure S2(a–f) (Bayesian approach) and Figure S3(a–f) (maximum-likelihood approach). The resulting phylogenetic relationships for all 12 individual trees from both approaches mirrored the previously deduced evolutionary history of *Oryza* (Ge *et al.*, 1999, 2002; Guo and Ge, 2005). Importantly, homoeologous sequences from each polyploid species always maintained sister relationships with orthologous sequences from putative diploid progenitors (B_d–B_{TM} and C_d–C_{TM}–C_{TA}) or with orthologous sequences of a similar or closely related genome type from a different species background (D_{TA}–E_d and L_{TC}–H_{TR}). Combined trees of the six core genes also resulted in a similar genus topology (Figures S2g and S3g).

To estimate the timing of polyploid formation of the BBCC and CCDD species, we used a relaxed molecular clock approach in a Bayesian phylogenetic framework using the program BEAST (Drummond and Rambaut, 2007). We tested two models of rate variation against the molecular clock (the hypothesis that all rates are equal across the phylogeny): the uncorrelated log normal (UCLN) and the uncorrelated exponential (UCED) models. Both assume that rates are uncorrelated between adjoining branches. Our data strongly rejected the strict molecular clock in favor of models where rates varied across the phylogeny. The Bayes factors (BF), calculated against the molecular clock and either the UCLN or UCED models, were greater than 10 for all comparisons, with no significant difference between the models (BF <1). In addition, the BF tests allowed us to

determine that the 'general time reversible model' (GTR), with estimates for among-site rate heterogeneity (gamma distribution; GTR+G), was the best site-substitution model for the data. The results in Figure 1 are based on the UCLD relaxed-clock model with the GTR+G model of site substitution.

Results from these analyses indicated that the diploid and polyploid B genomes (B_d–B_{TM}) evolved independently for approximately 2.1 million years. Similarly the diploid and polyploid C genomes (C_d–C_{TM}–C_{TA}) radiated approximately 1.7 million years ago, and the C_d–C_{TM} genomes approximately 0.6 million years ago (Figure 1 and Table 2). Thus the polyploidization event leading to formation of *O. minuta* (BBCC) occurred approximately 0.6–2.1 million years ago, while that of *O. alta* (CCDD) occurred approximately 1.7 million years ago. The relaxed-clock models also indicated that the diploid B and C genomes last shared a common ancestor approximately 9.6 million years ago (Figure 1). Although we were not able to estimate the timing of the polyploidization events leading to formation of the HHJJ and KKLL genome species, we were able to estimate that the J_{TR}–L_{TC}–H_{TR} genomes last shared a common ancestor approximately 9.3 million years ago, and that the L_{TC}–H_{TR} genomes have evolved independently for approximately 8.9 million years (Figures 1 and S2g).

To test for different types of selection acting on the six core genes, we calculated the rates of synonymous (d_S) and non-synonymous (d_N) substitutions using a phylogenetic approach based on the maximum-likelihood algorithm in PAML (Yang, 2007). The ratio of d_N/d_S, referred to as 'omega' (ω), provides a stringent measure of the selective pressures acting on a protein coding gene (i.e. $\omega = 1$ indicates selectively neutral evolution; $\omega < 1$ indicates purifying selection; $\omega > 1$ indicates adaptive or diversifying selection (Yang, 2002). For all six core genes, and in all tested *Oryza* lineages (diploid and polyploid), the ω ratio was found to be <0.5, suggesting that these proteins are under strong purifying selection (Tables S4–S6).

To investigate whether the core gene sets in *O. minuta* (BBCC) and *O. alta* (CCDD) and their putative diploid progenitors *O. punctata* (BB) and *O. officinalis* (CC) are undergoing different rates of evolution, we calculated and compared their molecular evolution rates using BEAST (Drummond and Rambaut, 2007) to determine whether the mean rate was higher in the polyploid lineage and whether the 95% highest posterior density (HPD) interval overlapped between the polyploid and diploid lineages. Our analysis detected an elevated rate of molecular evolution in the B_{TM} genome relative to the B_d genome. In contrast, the C genomes from *O. alta* and *O. minuta* showed nearly

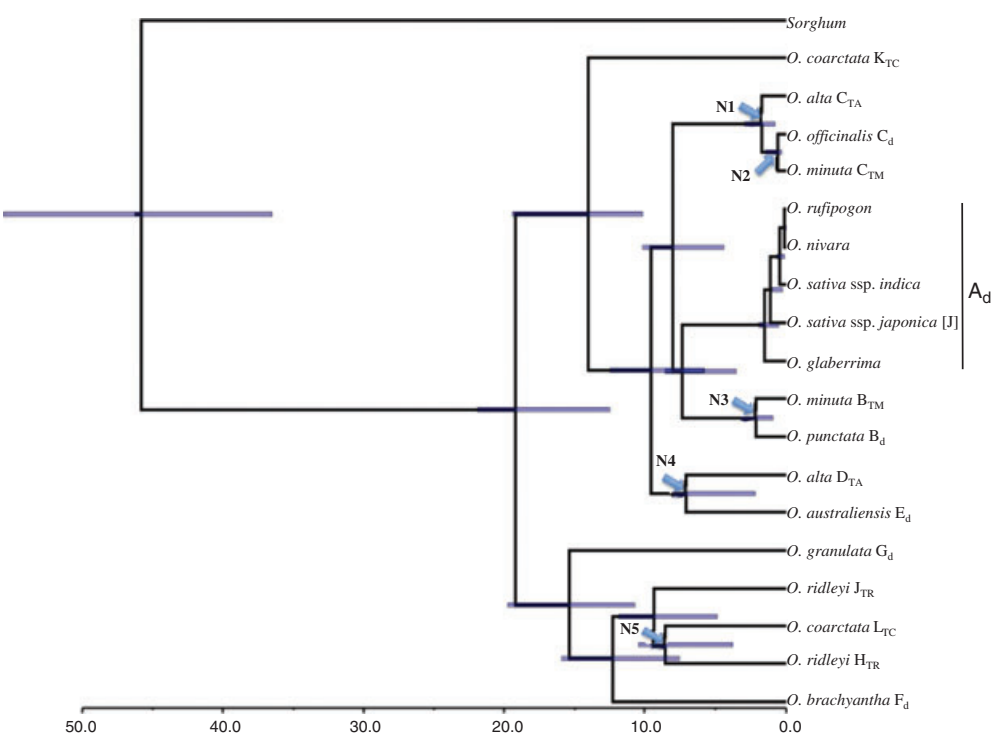

Figure 1. Paleogram of the genus *Oryza* obtained using a Bayesian relaxed-clock approach. Nodes leading to the formation of polyploid species are numbered. Blue boxes reflect the 95% highest posterior density (HPD) interval for the age of the respective nodes. The time scale below the figure indicates the number of million years. The nomenclature for genome type and ploidy of each species is indicated in Table 1. A_d, F_d and G_d indicate diploid *Oryza* species belonging to the genome types AA, FF and GG respectively.

Table 2 Time of origin for the CC genome complex polyploid species *O. minuta* (BC) and *O. alta* (CD), and the split of the diploid DD–EE and LL–HH genomes from their most recent common ancestor (MRCA)

Node in Figure 1	MRCA	Age (million years)	95% HPD interval
N1	All CC genomes	1.70	0.7636–2.9052
N1	*O. officinalis* C_d versus *O. alta* C_{TA}	1.70	0.7636–2.9052
N2	*O. officinalis* C_d versus *O. minuta* C_{TM}	0.60	0.3017–1.4674
N3	*O. punctata* B_d versus *O. minuta* B_{TM}	2.10	0.9283.3.1886
N4	*O. australiensis* E_d versus *O. alta* D_{TA}	7.08	2.1577–7.9699
N5	*O. coarctata* L_{TC} versus *O. ridleyi* H_{TR}	8.58	3.760–10.4406

Times in millions of years at 95% highest posterior density (HPD) intervals based on Bayesian relaxed-clock models are indicated.

identical rates of molecular evolution relative to *O. officinalis* (Table S7).

Evolution of intergenic space

The amount of intergenic space divergence between homoeologs should approximately equal that between the two orthologous regions of the two progenitor diploids, under a scenario of recent polyploidization and independence. This expectation was tested and confirmed using pairwise global sequence comparisons (Table S8). The amount of intergenic sequence conserved between the B_d and C_d genomes was similar to that between B_{TM} and C_{TM}. Similarly, the amount of sequence conservation between B_d and B_{TM} and between C_d and C_{TM} was high, in line with their recent origin (Table S8). A comparison of TE content indicated independent and non-uniform evolution of TEs as the basis for independent expansions and/or compressions in both diploid and polyploid genomes (Table S9). TEs were observed to make similar contributions to most of these genomes, varying from 49–61% (Table S3). The ratio of RNA to DNA TE contributions in the *Oryza* tetraploids varied from approximately 0.6 to 2.3, which is not unusual for plant species with small genomes (380–640 Mb) (Table S3). Larger genomes such as that of maize (Schnable *et al.*, 2009) have ratios of RNA to DNA TE genomic contributions of approximately 5. Individual element family representations, especially for the DNA elements, were found to vary in the studied orthologous regions, but this is likely to be a sampling outcome due to the small region investigated.

Variation in the rate of unequal recombination as gauged by LTR retrotransposon dynamics

A total of 13 intact LTR retrotransposons and 34 solo LTRs were identified from the polyploid comparative dataset (Table S10). Intact LTR retrotransposon content ranged from 0 in the L_{TC} genome to 3 in each of the D_{TA}, B_{TM} and C_{TM} genomes. Solo LTR content ranged from 1 in the L_{TC} and C_{TM} genomes to 9 in the D_{TA} genome. The estimated insertion times of intact elements ranged from 0.2–7 million years ago (Table S9). Pairwise global comparisons between various genomic counterparts across ploidy levels and genome types revealed that none of the intact LTR elements were shared, whereas four of 34 (11.7%) solo LTRs were shared (Table 3). However, for type II DNA TEs, a number of intact elements were shared in the CC genome complex species, with no such conservation among HH genome complex species. The ratio of intact LTR retrotransposons to solo LTRs is a commonly used method to determine rates of unequal/ectopic recombination (SanMiguel *et al.*, 1998). Variation in the rate of unequal recombination was observed in the compared diploid donor and polyploid counterparts. In particular, the B_d region has experienced fourfold higher unequal recombination than its B_{TM} counterpart, while the two homoeologous CC genomes of *O. alta* (CCDD) and *O. minuta* (BBCC) exhibited opposite patterns in the rate of unequal recombination, i.e. higher in C_{TA} and lower in C_{TM} relative to C_d (Table 3).

Synteny and DNA rearrangements

Three types of global comparisons were performed using windows of aligned *Adh1* regions: (i) comparisons of the same or closely related genome types across species and ploidy backgrounds (B_d–B_{TM}, C_d–C_{TM}, C_d–C_{TA}, L_{TC}–H_{TR} and D_{TA}–E_d), (ii) comparisons between the two homoeologous genomes from within each polyploid species (B_{TM}–C_{TM}, C_{TA}–D_{TA}, L_{TC}–K_{TC} and H_{TR}–J_{TR}), and (iii) comparisons of homoeologous polyploid genomes with the Os*Adh*1 reference sequence to obtain specific inferences about the structure or timing of a particular DNA arrangement. Here we present a summary of the various molecular and evolutionary mechanisms that underlie the maintenance and disruption of micro-synteny.

Evolution of paralogous gene families. Using a combination of synteny and phylogenetic analysis, the shared and unshared (lineage-specific) fraction of each gene family was identified for all polyploid species. When the BB and CC genomes (B_d–B_{TM}, C_d–C_{TM} and C_d–C_{TA}) were compared separately, gene content and order were highly conserved and the size of the single orthologous gene cluster (gene family 11) was stable. Therefore, the two homoeologous genomes of *O. minuta* (B_{TM} and C_d–C_{TM}) and the homoeologous genome of *O. alta* (C_{TA}) essentially mirrored the gene

Table 3 Unequal recombination rate variation in pairwise comparisons of the globally aligned orthologous *Adh1* regions

Selected global genome comparison	Genome specificity	Number of intact LTR retrotransposons	Number of solo LTRs	Age (million years ago)		Ratio of intact LTR retrotransposons
				Range	Mean	
B_d–B_{TM}	Present only in B_d	1	6	1.05	1.1	1:9
	Present only in B_{TM}	2	1	2.00–3.67	2.8	1:2
	Common to both	0	3	0	0	0
C_d–C_{TM}	Present only in C_d	0	1	0	0	0:1
	Present only in C_{TM}	3	0	0.57–2.03	1.3	3:0
	Common to both	0	0	0	0	0
C_d–C_{TA}	Present only in C_d	0	1	0	0	0:1
	Present only in C_{TA}	0	2	0	0	0:2
	Common to both	0	0	0	0	0
C_{TM}–C_{TA}	Present only in C_{TM}	3	1	0.57–2.03	1.3	3:1
	Present only in C_{TA}	0	2	0	0	0:2
	Common to both	0	0	0	0	0
	Common to all C genomes	0	0	0	0	0
D_{TA}–E_d	Present only in D_{TA}	3	8	0.44–7.03	2.7	1:3
	Present only in E_d	7	9	0.28–3.6	1.3	1:1.3
	Common to both	0	1	0	0	0
L_{TC}–H_{TR}	Present only in L_{TC}	0	0	0	0	0
	Present only in H_{TR}	0	3	0	0	0:3
	Common to both	0	0	0	0	0
B_{TM}–C_{TM}	Present only in B_{TM}	2	4	2.00–3.67	2.8	1:2
	Present only in C_{TM}	3	0	0.57–2.03	1.3	3:0
	Common to both	0	0	0	0	0
C_{TA}–D_{TA}	Present only in C_{TA}	0	3	0	0	0:3
	Present only in D_{TA}	3	9	0.44–7.03	2.7	1:3
	Common to both	0	0	0	0	0
L_{TC}–K_{TC}	Present only in L_{TC}	0	1	0	0	0:1
	Present only in K_{TC}	0	2	0	0	0:2
	Common to both	0	0	0	0	0
H_{TR}–J_{TR}	Present only in H_{TR}	0	1	0	0	0:2
	Present only in J_{TR}	2	1	3.04–3.06	3.1	1:1
	Common to both	0	1	0	0	0

space organization of their putative progenitors (Figure 2). However, a single apparent case of a lineage-specific insertional inactivation of a paralogous member by an LTR retrotransposon was identified in B_{TM} (Figure 2).

In contrast, comparisons between the L_{TC}–H_{TR} and D_{TA}–E_d genomes, and two homoeologous genomes from the polyploids *O. coarctata* (L_{TC}–K_{TC}) and *O. ridleyi* (H_{TR}–J_{TR}) revealed extensive variation in the size of almost every orthologous gene cluster (e.g. gene families 2, 5, 6, 11 and 13; Figures 2 and 3). Even among the shared members, several cases of apparent gene loss events were observed. Our analysis indicated that these two events (size change and silencing) were mediated by at least three major evolutionary forces, mostly in a lineage-specific fashion. Pertinent examples are discussed below.

(a) Gain of new paralogous family members through tandem duplication. For example, a fourth *Adh* family member (K_{TC}6–4; Figure 2) located approximately 65 kb upstream of the *Adh* cluster, and in an opposite transcriptional orientation, was identified on the K genome of *O. coarctata*. We determined that its origin was recent and lineage-specific (K_{TC}) by sequence divergence estimates of *Adh* family members.

(b) Gene loss by either (i) frameshift mutations or mutational decay, e.g. K_{TC}6–3 and gene family 11 members in a number of compared lineages (Figures 2 and 3), insertion of transposons, e.g. several gene family 11 members in D_{TA} and one member in H_{TR}, or complete deletion of a member or a complete tandem array, e.g. the entire NBS–LRR family 2 in L_{TC} and loss of an entire protein kinase gene cluster 5 in J_{TR}.

(c) Sequence divergence, e.g. several members of gene families 2, 5, 11 and 13 (Figures 2 and 3).

Single-copy gene deletions. In addition to the rapid birth and death pattern of gene family evolution seen among gene clusters, several cases of single-copy gene loss events were also discovered. An interesting case was the complete

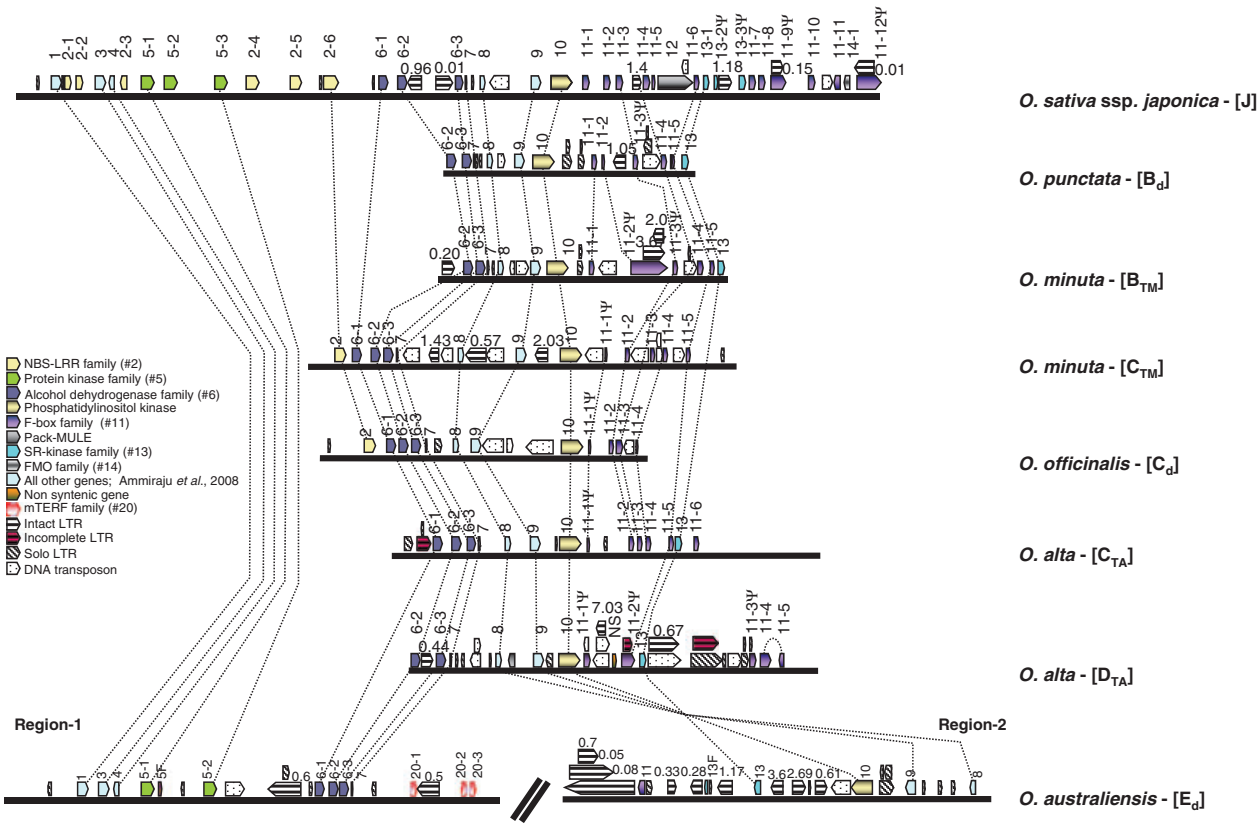

Figure 2. Phylogenomic view of the orthologous *Adh*1 regions from CC genome complex polyploid species [*O. minuta* (BBCC) and *O. alta* (CCDD)] and their putative diploid progenitors *O. punctata* (BB) and *O. officinalis* (CC) relative to the closely related species *O. australiensis* (EE) and *O. sativa* (AA). The nomenclature for genome types is indicated in Table 1. Each gene family is color-coded, and the various types of TEs are indicated. Insertion time estimates for intact LTR retrotransposons (in millions of years) are indicated above each element.

absence of genes 7 and 9 in the K_{TC} genome. As these genes are conserved in all other *Oryza* genome types (diploid and polyploid) at the *Adh1* region, these events probably represent deletions that arose as a consequence of polyploidization specifically in the K_{TC} genome.

Gene transposition. Another mechanism that contributed to deviation from strict micro-synteny was gene transposition. Two non co-linear genes were identified in the D_{TA} genome that were absent at the corresponding genomic location in all other genomes analyzed. The first was a gene fragment that was found embedded in a Pack-MULE (Os0874 family; Jiang *et al.*, 2004) with intact structural features, and was located in the gene interval D_{TA}8–9 (Figure 2). Homology searches provided evidence that this gene fragment belonged to a USP family protein (LOC_Os01g57450) located on chromosome 1 of cultivated rice, suggesting movement to its current location in D_{TA} with the aid of a MULE element. The second non-syntenic gene (D_{TA}NS) was a member of NBS–LRR family 2. The most closely related gene was found on chromosome 8 (LOC_Os08g10430.1) of cultivated rice (International Rice Genome Sequencing Project, 2005). This gene was not embedded in any known TE, suggesting gene movement by an unknown mechanism.

DISCUSSION

The main objectives of this study were to apply comparative phylogenomics approaches to: (i) understand the genomic consequences of natural polyploidization in *Oryza* through detailed sequence-level characterization of an orthologous set of genomic regions across the *Oryza* phylogeny, and (ii) determine the molecular evolutionary forces and events that shaped the eight genomes of polyploid *Oryza* in light of known patterns of genome evolution in diploid *Oryza*, as well as known patterns in other polyploid systems. The analysis yielded several insights into the tempo of genome evolution in natural polyploids of *Oryza*, and these are discussed below.

Timing of *Oryza* polyploid formation

The origin and distribution of the BBCC and CCDD *Oryza* allopolyploids have been widely debated (Ge *et al.*, 1999; Vaughan *et al.*, 2003; Second and Rouhan, 2008). Here we used a relaxed-clock approach (Sanderson, 2002;

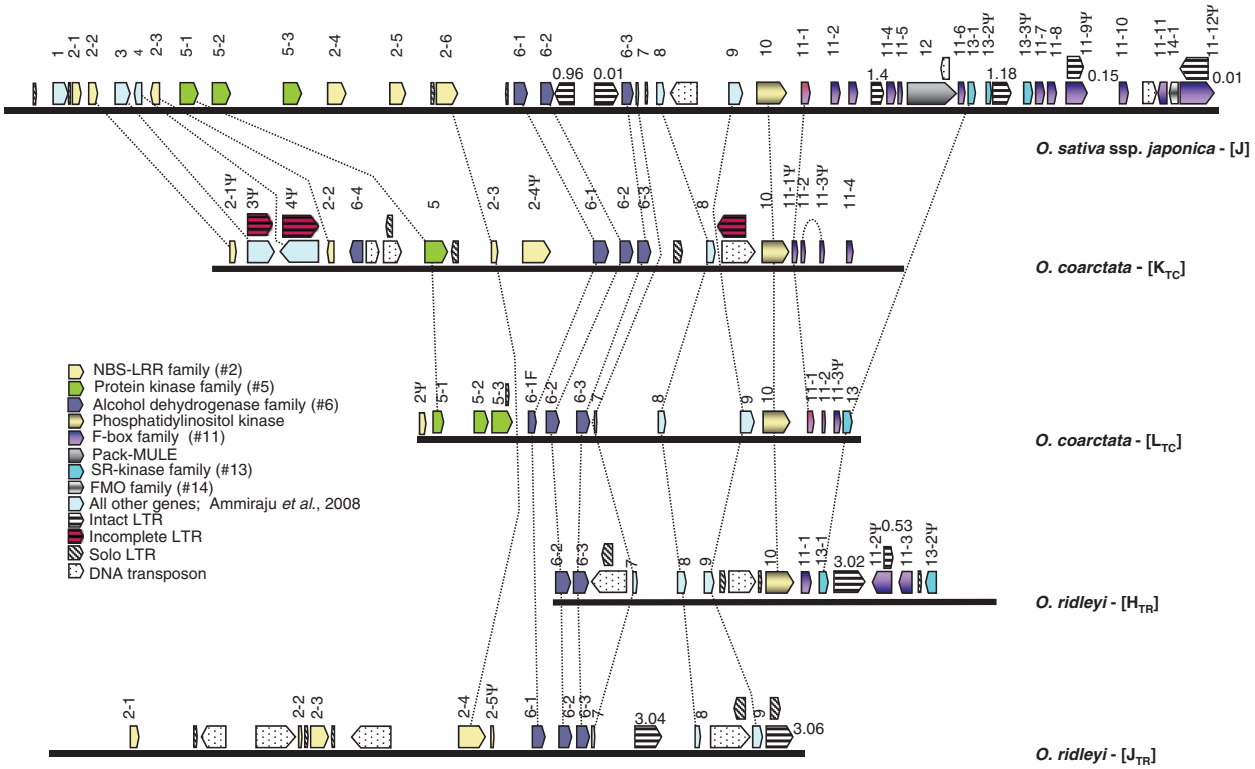

Figure 3. Phylogenomic view of orthologous *Adh*1 regions from the HH genome complex polyploid species *O. coarctata* (KKLL) and *O. ridleyi* (HHJJ) relative to *O. sativa* (AA). The nomenclature for genome types is indicated in Table 1. Each gene family is color-coded, and the various types of TEs are indicated. Insertion time estimates for intact LTR retrotransposons (in millions of years) are indicated above each element.

Drummond *et al.*, 2006; Drummond and Rambaut, 2007) to estimate the times of origin of polyploidy in *Oryza*. The main advantages of this method are that (i) it accounts for rate heterogeneity in genes or lineages, (ii) there is no need for application of a local clock based on an externally calibrated rate (e.g. universal nucleotide substitution rate), and (iii) it provides robust molecular dates with plausible ranges of uncertainty (i.e. 95% highest posterior density interval). The results from these analyses broadly agreed with our previous estimates using a local clock, and those of another study that used relaxed-clock methods (Tang *et al.*, 2009). Additionally, the molecular clock findings reported for the *MOC1* locus (Lu *et al.*, 2009), and here for the *Adh1* locus, dated the BC and CD polyploidization events to approximately 2 million years ago. Thus, our results exclude the possibility of ancient polyploidization events occurring approximately at the time of maize–rice divergence and continental drift (Chang, 1976, 2003; Khush, 1997). Therefore, the present-day distribution of diploid and polyploid species on different continents, and the molecular clock results obtained here, can only be explained by a 'long-distance dispersal' hypothesis (Vaughan *et al.*, 2005, 2008; Zhang and Ge, 2007; Wang *et al.*, 2009).

The D and E genomes of *Oryza*

As the diploid DD genome species is presumed extinct, and is only present in three American CCDD polyploid species, various hypotheses have been proposed to explain its origin and relationship to other *Oryza* genome types. Several studies have suggested that the EE genome is a strong candidate as the potential donor for the DD genome (Gong and Bao, 2008; Guo and Ge, 2005; Bao and Ge, 2004; Fukui *et al.*, 1997; Ge *et al.*, 1999, 2001; Li *et al.*, 2001; Nayar, 1973). If this is true, and given the recent origin of the CCDD and BBCC genomes (<2 million years ago), it is expected that the DD and EE genomes would have a divergence time that is comparable to that for all other donor and recipient genomes involved in BBCC and CCDD polyploidization events. The phylogeny and molecular clock data obtained here indicate that the D_{TA} and E_d genomes have experienced independent evolution for a long time (>7 million years; Figure 1) compared to B_d–B_{TM} (approximately 2.1 million years) and C_d–C_{TM} (approximately 0.6 million years). This long evolutionary divergence between the DD and EE genomes could have resulted from an accelerated rate of sequence substitutions in one of these genomes. Our analyses uncovered no such phenomenon at the *Adh1* region

between the DD and EE genomes (Table S7). In addition, the sequence homology in the intergenic regions was lower and the number of structural rearrangements was higher between the D_{TA} and E_d genomes, compared to those among the BB and CC genomes in different ploidy backgrounds, further supporting the above hypothesis. Although the majority of E_d genome-specific structural variation is of recent origin (Piegu *et al.*, 2006; Ammiraju *et al.*, 2007, 2008), the current degree of divergence observed in this region nevertheless supports the existing independent genome type designation for the D genome of *O. alta*.

Parallel and ongoing evolution of genes and genome micro-structure of co-habitating *Oryza* homoeologous genomes at the *Adh1* region

Genes duplicated by polyploidization usually undergo one or more of three known evolutionary fates (Wendel, 2000): (i) functional divergence due to relaxed selection on one copy (Adams and Wendel, 2005; Lu *et al.*, 2009; Stephens, 1951; Wendel, 2000; Zhang, 2003), (ii) retention of both copies, and their original or similar function (Wendel, 2000; Zhang, 2003; Freeling, 2008; Veitia *et al.*, 2008; Edger and Pires, 2009), and (iii) gene loss by pseudogenization or complete elimination (Wendel, 2000; Doyle *et al.*, 2008; Otto, 2008). Analysis of orthologous *Adh1–Adh2* regions from *Oryza* polyploid species revealed that, barring the few exceptions described below, most of the low-copy-number homoeologous duplicated gene pairs were retained and were under strong functional constraints. A classic example of functional retention of duplicated genes is the *Adh* family studied in this region. Duplication of *Adh* pre-dates the divergence of grasses (Gaut *et al.*, 1999; Ammiraju *et al.*, 2008), and all paralogs in cultivated rice are still functionally intact and under strong purifying selection, but show organ-specific expression patterns (Xie and Wu, 1989; Terada *et al.*, 2007), indicating sub-functionalization. These results are in contrast with previous findings indicating that more than half of the homoeologous gene pairs in the *MOC1* region of *O. minuta* (BBCC) experienced accelerated rates of amino acid substitutions, suggesting functional diversification (Lu *et al.*, 2009). However, current findings from the *Adh1* region are consistent with recent findings showing that retention of duplicate genes plays a role in increasing expression diversity (Ha *et al.*, 2009).

Orthologous *Adh1* regions of diploid *Oryza* are predominantly composed of multi-gene families that are as old as or older than the genus itself (Ammiraju *et al.*, 2008). Their dynamic evolution, through lineage-specific birth and death processes, was shown to be a frequent cause of synteny disruption. Combined phylogenetic and synteny analyses revealed that most paralogous gene sequences form sister clades with respective sequences from putative diploid donors or closely related genomes (Figures S4–S7), or form genome-specific groups that arose independently after polyploidization. The absence of phylogenetic violations further supports the hypothesis that each of the two co-habitating genomes in each *Oryza* polyploid species is undergoing independent evolution. This analysis indicated that the two homoeologous C genomes of *O. minuta* and *O. alta* have been evolving independently since their respective polyploidization events. This time frame is close to that of the divergence time of all AA genomes (<2 million years) (this study; Zhu and Ge, 2005). In the *Adh1* region, a high degree of gene content conservation was observed between the homoeologous regions of *O. minuta* and *O. alta* and their respective diploid progenitors (Table S11). This is in contrast to the approximately 2–8% gene flux (percentage of unshared genes due to independent lineage-specific gain or loss) observed among the AA genomes of *Oryza* at the same orthologous region (Ammiraju *et al.*, 2008). Similarly, an approximate 22% gene flux was observed between the D_{TA} and E_d genomes and between the two L_{TC}–H_{TR} genomes (Table S11). This extent of gene content variation is not surprising based on a comparison of diploid *Oryza* lineages with similar temporal resolutions (*O. sativa* and *O. punctata*, divergence 7.4 million years ago, gene content variation 28%; *O. sativa* and *O. australiensis*, divergence 9.6 million years ago, gene content variation 37%). Much of the observed gene flux was caused by rapid evolution of multi-gene families through independent genome-specific duplications and deletions. These observations are consistent with the dynamics of F–box and NBS–LRR gene families observed in other plant genomes (Freeling, 2008). The extent of intergenic space conservation in the B and C genomes in various ploidy backgrounds was comparable to that observed among the AA genomes, suggesting that the observed differences in intergenic sequence evolved independently in a clock-like manner, rather than occurring as a result of rapid induction of instability conditioned by *de novo* polyploidy. Independent and genome-specific insertions/deletions of TEs were the major underlying causes for the observed intergenic space divergence. By measuring the age of intact LTR retrotransposons, it was shown that most of these changes occurred independently of and subsequent to respective polyploidization events. This means that some of the 'non-additive' changes observed in the *Oryza* polyploids are derived states that occurred subsequent to polyploidization.

Some studies have indicated that extensive genome modifications, such as non-additive genetic change (rapid DNA loss, homoeologous recombination, gene conversion, ectopic recombination) (Song *et al.*, 1995; Feldman *et al.*, 1997; Liu *et al.*, 1998a,b; Ozkan *et al.*, 2001; Osborn *et al.*, 2003; Pires *et al.*, 2004; Udall *et al.*, 2005) and epigenetic modifications (TE suppression/release, methylation, histone modifications and gene expression changes) (Comai, 2000, 2005; Shaked *et al.*, 2001; Kashkush *et al.*, 2002, 2003; Adams *et al.*, 2003; He *et al.*, 2003; Wang *et al.*, 2004, 2006; Ha *et al.*, 2009; Ni *et al.*, 2009), are major mechanisms of genome

stabilization in nascent polyploids. Several intergenomic synthetic amphiploids have been described previously for the genus *Oryza* to clarify the relationships between various genome types (Katayama, 1977 and references therein; Katayama *et al.*, 1977; Katayama and Onizuka, 1978; Katayama, 1982). However, no information is available on the patterns of genome evolution in these synthetic polyploids. The natural and young *Oryza* polyploid species *O. minuta* (BBCC) and *O. alta* (CCDD) studied here did not exhibit dramatic genomic modifications in relation to their extant diploid parents at this genomic location. Further supporting these results is the macro-level observation that, in *O. minuta* (BBCC), the only polyploid species with two presumed diploid progenitors that currently survive, the genome size (1124 Mb) is the sum of those of its parents [425 Mb for *O. punctata* (BB) and 651 Mb for *O. officinalis* (CC)] (Ammiraju *et al.*, 2006). The minor micro-structural differences observed at this region were probably due to ongoing independent evolution of polyploid genomes subsequent to polyploidization events and/or haplotypic differences in the parental lines used as progenitors for genome comparisons. Similar levels of genome micro-structure stability between homoeologous genomes in other recently evolved natural polyploids have also been observed using comparative genomics approaches (Cheung *et al.*, 2009; Chantret *et al.*, 2005; Dubcovsky and Dvorak, 2007; Fiebig *et al.*, 2004; Gu *et al.*, 2006; Grover *et al.*, 2004, 2007; Ha *et al.*, 2009; Rana *et al.*, 2004; Wicker *et al.*, 2003; Innes *et al.*, 2008).

The degree of micro-level structural variation appears to depend on the genetic divergence between the parental genomes involved in polyploid formation and/or the age of the polyploid. Genome restructuring is less extensive in young CC polyploid complex species than in older HH genome complex polyploid species. An intriguing case is that of *O. coarctata*, which has the smallest genome size of the *Oryza* polyploids (771 Mb; Kim *et al.*, 2008). The L_{TC} genome of *O. coarctata* has the lowest TE content among all polyploid genomes: almost half of the total TE content of the closely related H_{TR} genome from *O. ridleyi*. Most TEs observed in the L_{TC} genome of *O. coarctata* were fragmented, and no recent TE insertions were found relative to the H_{TR} genome of *O. ridleyi*. As a whole, *O. coarctata* has an unusually low repeat content and a small percentage of common repeats shared with rest of the *Oryza* species (Zuccolo *et al.*, 2007). Taken together, it is tempting to speculate that the small genome size in *O. coarctata* is a possible example of 'genomic downsizing', which is a widespread biological phenomenon observed after polyploid formation (Leitch and Bennett, 2004).

A surprising finding was the apparent variation in the rates of sequence evolution between the B and C genomes of *O. minuta* at the *Adh1* region, with the B genome exhibiting a moderately accelerated rate of evolution relative to its C genome counterpart.

To date, the wheat genome has been the best model for understanding the consequences of natural polyploidy in grasses. The timing of polyploidization events in wheat was estimated to be as recent as that for the CC genome complex polyploids of *Oryza* (0.5–3 million years ago for AABB allopolyploid formation and 7000–9000 years for AABBDD hexaploid formation from the A, B and D progenitors that diverged 2.5–6 million years ago). Analysis of a number of genetic loci across various ploidy levels in wheat has indicated that, despite low intergenic space conservation, gene content and micro-colinearity are largely preserved, with an approximate 10–20% gene flux in the homoeologous regions (Wicker *et al.*, 2003; Chantret *et al.*, 2005; Gu *et al.*, 2006; Dubcovsky and Dvorak, 2007). Despite being a highly recombinogenic region, analysis of *Oryza* CC genome complex polyploids uncovered a low level of gene flux (approximately 4% in the *O. minuta* B genome; Table S11), suggesting similar patterns of polyploid genome evolution in *Oryza* and wheat.

In conclusion, analysis of the tetraploid *Oryza* genomes indicates that this genus provides an excellent and opportunistic model system to study the consequences of natural polyploidy. Future investigations will focus on expanding the comparative phylogenomic analyses to whole chromosome arms and eventually full genome and epigenome sequences to understand not only the rates of genetic and genomic changes but also ecological and physiological effects of polyploidy on phenotype (Wang *et al.*, 2006; Ni *et al.*, 2009) and the historical consequences of multiple origins at different evolutionary times.

EXPERIMENTAL PROCEDURES

Sequencing and annotation

Sequencing, annotation and phylogenetic analysis were performed as previously described (International Rice Genome Sequencing Project, 2005; Ammiraju *et al.*, 2008). LTR retrotransposon insertion dates were calculated as described by SanMiguel *et al.* (1998), using a rate of 1.3×10^{-8} mutations per site per year (Ma and Bennetzen, 2004). Quantitative estimates of sequence conservation in selected global sequence comparisons were obtained using MLAGAN (Brudno *et al.*, 2003).

Divergence time estimation

Divergence times for the *Oryza* phylogeny were estimated using a relaxed-clock approach, implemented in BEAST version 1.5.2 (Drummond and Rambaut, 2007). This approach permits simultaneous estimation of tree topology, model parameters and both the rate and time components on branch lengths, thereby allowing ages to be assigned to each node in the phylogeny. This analysis requires external (prior) information on the dates of one or more nodes in the tree. Ideally, these dates would come directly from fossil information, but, given a lack of fossil data for all *Oryza* species, dates from two previous molecular studies in the tribe Oryzeae and all grasses were used. A prior distribution on the age of the rice and Sorghum divergence (normal distribution, mean 50 million years, standard deviation 5 million years; Vicentini *et al.*, 2008) and the base of the *Oryza* clade (normal distribution, mean

15 million years, standard deviation 3 million years; Tang *et al.*, 2009) was specified. A number of different nucleotide and relaxed-clock models using Bayes factors (Suchard *et al.*, 2001) were tested to determine which model provided the best fit for the data. These included Hasegawa, Kishino and Yano model with Gamma distribution (HKY+G) and GTR+G model for nucleotide substitutions, and a strict molecular clock (one rate) or relaxed-clock models with log normal and exponential distributions. For all analyses, 25 million Markov chain Monte Carlo (MCMC) generations were run, with sampling every 1000 generations. TRACER version 1.4.1 (http://beast.bio.edu.ac.uk/tracer) was used to check convergence and to calculate Bayes factors. The sampled trees were summarized using TREEANNOTATOR version 1.5.2 (http://beast.bio.ed.ac.uk) to determine the maximum clade credibility tree and estimates for node ages and branch rates. Trees were visualized using FIGTREE version 1.2.3 (http://tree.bio.ed.ac.uk/software/figtree).

Tests for selection

Two d_N/d_S tests were performed using PAML and the likelihood ratio test. First, protein sequences of the homologous gene pairs were aligned using MUSCLE (http://www.drive5.com/muscle/), and then the protein alignment was converted into a codon-based nucleotide alignment using Pal2nal (Suyama *et al.*, 2006). Pairwise d_N/d_S (ω) ratios of orthologous genes were calculated using the maximum likelihood algorithm in PAML (Yang, 2007). The significance of ω values that deviated from neutrality ($\omega = 1.0$) was tested using the likelihood ratio test. Second, to test the d_N/d_S ratios along branches using a tree-based branch model, a phylogenetic tree for each core gene was constructed using the maximum likelihood algorithm and appropriate models estimated using ModelTEST (Posada and Crandall, 1998) implemented in PAUP*4.0b10 (Swofford, 2002). A tree-based d_N and d_S analysis was performed using the codon model (codeml) (Goldman and Yang, 1994; Yang, 1998) in PAML4 (Yang, 2007). For the initial codeml analyses, two models, using either one ratio (ω) for all branches or a free ratio (ω) for each branch, were employed to determine whether the d_N/d_S ratios were indeed different among lineages. If they were, subsequent tests with multiple-ratio branch models were performed. d_N/d_S ratio differences among branches were evaluated using the likelihood ratio test. Log likelihoods of the defined models were compared using a χ^2 distribution with degrees of freedom equal to the difference in the number of variable parameters between the nested models. The numbers of synonymous and non-synonymous substitutions along each branch were calculated based on branch length (t) and the d_N/d_S ratios (ω), together with the estimated transition/transversion ratio (κ) under the free-ratio model.

Accession numbers

BAC clone addresses and species names are listed in Table S1. All sequences are deposited in Genbank (GQ203296–GQ203303).

ACKNOWLEDGEMENTS

This work was supported by the US National Science Foundation (grant number DBI-0321678 to R.A.W. and S.J.), the Bud Antle Endowed Chair (to R.A.W.), and the Georgia Research Alliance (to J.L.B.). We thank Professor Mike Sanderson (Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ), for suggestions on the molecular clock analysis. We thank members of the Arizona Genomics Institute's BAC/EST Resource, Sequencing, Finishing and Bioinformatics groups for generating and processing the high-quality data used in this analysis.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Identification of homoeologous BAC clones from each tetraploid *Oryza* species using genome-specific CAP and STS markers.

Figure S2. Chronograms of the genus *Oryza* based on six core genes obtained using a Bayesian relaxed-clock approach.

Figure S3. Phylogenetic trees inferred from six core genes using maximum-likelihood approaches.

Figure S4. Unrooted phylogenetic tree demonstrating the evolutionary origin and diversification of F–box or F–box-like genes in polyploid *Oryza* species and their proposed diploid progenitors.

Figure S5. Unrooted phylogenetic tree demonstrating the evolutionary origin and diversification of gene family 2 in polyploid *Oryza* species and their proposed diploid progenitors.

Figure S6. Unrooted phylogenetic tree demonstrating the evolutionary origin and diversification of gene family 5 in polyploid *Oryza* species and their proposed siploid progenitors.

Figure S7. Unrooted phylogenetic tree demonstrating the evolutionary origin and diversification of gene family 13 in polyploid *Oryza* species and their proposed diploid progenitors.

Table S1. Comparative sequence dataset of sequenced BAC clones spanning eight homoeologous genomes from four allotraploid species of *Oryza*.

Table S2. Genes identified in the comparative sequence dataset.

Table S3. Compositional diversity and nucleotide contribution of various classes of transposable elements in the homoeologous regions of each polyploid *Oryza* species.

Table S4. Model comparison for detecting lineage-specific selection using the likelihood ratio test.

Table S5. Likelihood values and parameters estimated using various maximum-likelihood models.

Table S6. Purifying selection revealed by pairwise comparisons of six core genes across species/genome/ploidy backgrounds.

Table S7. Variation of the molecular evolutionary rate in the diploid and polyploid lineages of *Oryza*.

Table S8. Selected global sequence comparisons between species and ploidy backgrounds, showing the extent of sequence conservation in the intergenic regions.

Table S9. Independent evolution of various transposable element classes in the aligned regions.

Table S10. Intact LTR retrotransposons and solo LTRs identified in the polyploid dataset, and their structural features.

Table S11. Percentage shared and unshared gene content in the aligned windows of various *Oryza* genome type comparisons.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

Adams, K. and Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141.

Adams, K.L., Cronn, R., Percifield, R. and Wendel, J.F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl Acad. Sci. USA*, **100**, 4649–4654.

Aggarwal, R.K., Brar, D.S. and Khush, G.S. (1997) Two new genomes in the *Oryza* complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. *Mol. Gen. Genet.* **254**, 1–12.

Ammiraju, J.S., Luo, M., Goicoechea, J.L. *et al.* (2006) The *Oryza* bacterial artificial chromosome library resource, construction and analysis of 12

deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. *Genome Res.* **16**, 140–147.

Ammiraju, J., Zuccolo, A., Yu, Y. *et al.* (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *Plant J.* **52**, 342–351.

Ammiraju, J., Lu, F., Sanyal, A. *et al.* (2008) Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell*, **20**, 3191–3209.

Bao, Y. and Ge, S. (2004) Origin and phylogeny of *Oryza* species with the CD genome based on multiple-gene sequence data. *Plant Syst. Evol.* **249**, 55–66.

Brar, D.S. and Khush, G.S. (1997) Alien introgression in rice. *Plant Mol. Biol.* **35**, 35–47.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., Batzoglou, S. and Progra, N.C.S. (2003) LAGAN and multi-LAGAN, efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731.

Chang, T.T. (1976) The origin, evolution, cultivation, dissemination and diversification of Asian and African rices. *Euphytica*, **25**, 425–441.

Chang, T.T. (2003) Origin, domestication and diversification. In *Rice: Origin, History, Technology, and Production* (Wayne, S.C. and Dilday, R.H., eds). Chichester: John Wiley & Sons, pp. 3–25.

Chantret, N., Salseb, J., Sabot, F. *et al.* (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell*, **17**, 1033–1045.

Cheung, F., Trick, M., Drou, N. *et al.* (2009) Comparative analysis between homoeologous genome segments of *Brassica napus* and its progenitor species reveals extensive sequence-level divergence. *Plant Cell*, **21**, 1912–1928.

Comai, L. (2000) Genetic and epigenetic interactions in allopolyploid plants. *Plant Mol. Biol.* **43**, 387–399.

Comai, L. (2005) The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846.

Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y. and Byers, B. (2000) Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell*, **12**, 1551–1568.

Cronn, R.C., Small, R.L. and Wendel, J.F. (1999) Duplicated genes evolve independently after polyploid formation in cotton. *Proc. Natl Acad. Sci. USA*, **96**, 14406–14411.

Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S. and Wendel, J.F. (2008) Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461.

Drummond, A.J. and Rambaut, A. (2007) BEAST, Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. and Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, 699–710.

Dubcovsky, J. and Dvorak, J. (2007) Genome plasticity, a key factor in the success of polyploid wheat under domestication. *Science*, **316**, 1862–1866.

Edger, P.P. and Pires, J.C. (2009) Gene and genome duplications, the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17**, 699–717.

Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, A.A. and Vega, J.M. (1997) Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics* **147**, 1381–1387.

Fiebig, A., Kimport, R. and Preuss, D. (2004) Comparisons of pollen coat genes across Brassicaceae species reveal rapid evolution by repeat expansion and diversification. *Proc. Natl Acad. Sci. USA*, **101**, 3286–3291.

Freeling, M. (2008) The evolutionary position of subfunctionalization, downgraded. *Genome Dyn.* **4**, 25–40.

Fukui, K., Shishido, R. and Kinoshita, T. (1997) Identification of the rice D–genome chromosomes by genomic *in situ* hybridization. *Theor. Appl. Genet.* **95**, 1239–1245.

Gaut, B.S., Peek, A.S., Morton, B.R. and Clegg, M.T. (1999) Patterns of genetic diversification within the *Adh* gene family in the grasses (*Poaceae*). *Mol. Biol. Evol.* **16**, 1086–1097.

Ge, S., Sang, T., Lu, B.R. and Hong, D.Y. (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Natl Acad. Sci. USA*, **96**, 14400–14405.

Ge, S., Sang, T., Lu, B.R. and Hong, D.Y. (2001) Rapid and reliable identification of rice genomes by RFLP analysis of PCR-amplified *Adh* genes. *Genome*, **44**, 1136–1142.

Ge, S., Li, A., Lu, B., Zhang, S. and Hong, D. (2002) A phylogeny of the rice tribe *Oryzeae* (Poaceae) based on matK sequence data. *Am. J. Bot.* **89**, 1967–1972.

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

Gong, P.C. and Bao, Y. (2008) Nucleotide diversity of the homoeologous *adh1* loci in the American allotetraploid *Oryza* species. *Plant Syst. Evol.* **276**, 243–253.

Grover, C.E., Kim, H., Wing, R.A., Paterson, A.H. and Wendel, J.F. (2004) Incongruent patterns of local and global genome size evolution in cotton. *Genome Res.* **14**, 1474–1482.

Grover, C.E., Kim, H., Wing, R.A., Paterson, A.H. and Wendel, J.F. (2007) Microcolinearity and genome evolution in the *AdhA* region of diploid and polyploid cotton (*Gossypium*). *Plant J.* **50**, 995–1006.

Gu, Y.Q., Salse, J., Coleman-Derr, D. *et al.* (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics*, **174**, 1493–1504.

Guo, Y.L. and Ge, S. (2005) Molecular phylogeny of *Oryzeae* (Poaceae) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. *Am. J. Bot.* **92**, 1548–1558.

Ha, M., Kim, E.D. and Chen, Z.J. (2009) Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc. Natl Acad. Sci. USA*, **106**, 2295–2300.

He, P., Friebe, B.R., Gill, B.S. and Zhou, J.M. (2003) Allopolyploidy alters gene expression in the highly stable hexaploid wheat. *Plant Mol. Biol.* **52**, 401–414.

Hirsch, C.D., Wu, Y., Yan, H. and Jiang, J. (2009) Lineage-specific adaptive evolution of the centromeric protein CENH3 in diploid and allotetraploid Oryza species. *Mol. Biol. Evol.* **26**, 2877–2885.

Innes, R.W., Ameline-Torregrosa, C., Ashfield, T. *et al.* (2008) Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiol.* **148**, 1740–1759.

International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Jackson, M.T. (1997) Conservation of rice genetic resources – the role of the International Rice Genebank at IRRI. *Plant Mol. Biol.* **35**, 61–67.

Jackson, S. and Chen, Z.J. (2009) Genomic and expression plasticity of polyploidy. *Curr. Opin. Plant Biol.* **13**, 1–7.

Jackson, M.T., Pham, J.L., Newbury, H.J., Ford-Lloyd, B.V. and Virk, P.S. (1999) A core collection for rice – needs, opportunities and constraints. In *Core Collections for Today and Tomorrow* (Johnson, R.C. and Hodgkin, T., eds). Rome, Italy: International Plant Genetic Resources Institute, pp. 18–27.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.

Kashkush, K., Feldman, M. and Levy, A.A. (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics*, **160**, 1651–1659.

Kashkush, K., Feldman, M. and Levy, A.A. (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* **33**, 102–106.

Katayama, T. (1977) Cytogenetical studies on genus *Oryza*. 9. F1 hybrids from synthesized amphiploid X species with BBCC genomes. *Jpn. J. Genet.* **52**, 301–307.

Katayama, T. (1982) Cytogenetical studies on the genus *Oryza*. 13. Relationship between the genome–E and genome–D. *Jpn. J. Genet.* **57**, 613–621.

Katayama, T. and Onizuka, W. (1978) Dihaploid-BC plants produced in F_1 plants between synthesized amphiploid (BBCC) and diploid (BB) or tetraploid (BBCC)-*Oryza-punctata*-kotschy-ex-steud. *Jpn. J. Genet.* **53**, 67–70.

Katayama, T., Shin, Y.B. and Onizuka, W. (1977) Cytogenetical studies on genus *Oryza*. 10. Cytogenetics of tetraploid F1 plant between amphiploid punctata-eichingeri and BBCC genome species. *J. Fac. Agric. Kyushu Univ.* **22**, 99–105.

Khush, G.S. (1997) Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35**, 25–34.

Kim, H., Hurwitz, B., Yu, Y. *et al.* (2008) Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. *Genome Biol.* **9**, R45.

Leitch, I.J. and Bennett, M.D. (2004) Genome downsizing in polyploid plants. *Biol. J. Linn. Soc.* **82**, 651–663.

Leitch, A.R. and Leitch, I.J. (2008) Genomic plasticity and the diversity of polyploid plants. *Science*, **320**, 481–483.

Li, C., Zhang, D., Ge, S., Lu, B. and Hong, D. (2001) Differentiation and intergenomic relationships among C, E and D genomes in the *Oryza officinalis*

complex (Poaceae) as revealed by multicolor genomic in situ hybridization. *Theor. Appl. Genet.* **103**, 197–203.

Liu, B., Vega, J.M., Segal, G. *et al.* (1998a) Rapid genomic changes in newly synthesized amphiploids of *Triticum* and *Aegilops*. I. Changes in low-copy noncoding DNA sequences. *Genome*, **41**, 272–277.

Liu, B., Vega, J.M. and Feldman, M. (1998b) Rapid genomic changes in newly synthetized amphiploids of *Triticum* and *Aegilops*. II. Changes in low-copy coding DNA sequences. *Genome*, **41**, 535–542.

Liu, B., Brubaker, G., Cronn, R.C. and Wendel, J.F. (2001) Polyploid formation in cotton is not accompanied by rapid genomic changes. *Genome*, **44**, 321–330.

Lu, F., Ammiraju, J.S.S., Sanyal, A. *et al.* (2009) Comparative sequence analysis of MONOCULM1-orthologous regions in 14 *Oryza* genomes. *Proc. Natl Acad. Sci. USA*, **106**, 2071–2076.

Ma, J. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA*, **101**, 12404–12410.

Ma, J.X., Wing, R.A., Bennetzen, J.L. and Jackson, S.A. (2007) Evolutionary history and positional shift of a rice centromere. *Genetics*, **17**, 1217–1220.

Nayar, N.M. (1973) Origin and cytogenetics of rice. *Adv. Genet.* **17**, 153–292.

Ni, Z., Kim, E.D., Ha, M., Lackey, E., Liu, J., Zhang, Y., Sun, Q. and Chen, Z.J. (2009) Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature*, **457**, 327–331.

Osborn, T.C., Butrulle, D.V., Sharpe, A.G., Pickering, K.J., Parkin, I.A., Parker, J.S. and Lydiate, D.J. (2003) Detection and effects of a homoeologous reciprocal transposition in *Brassica napus*. *Genetics*, **165**, 1569–1577.

Otto, S.P. (2008) Evolutionary consequences of polyploidy. *Cell*, **131**, 452–462.

Ozkan, H., Levy, A.A. and Feldman, M. (2001) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops–Triticum*) group. *Plant Cell*, **13**, 1735–1747.

Piegu, B., Guyot, R., Picault, N. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269.

Pires, J.C., Zhao, J.W., Schranz, M.E., Leon, E.J., Quijada, P.A., Lukens, L.N. and Osborn, T.C. (2004) Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (*Brassicaceae*). *Biol. J. Linn. Soc.* **82**, 675–688.

Posada, D. and Crandall, K.A. (1998) MODELTEST, testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Rana, D., van den Boogaart, T., O'Neill, C.M., Hynes, L., Bent, E., Macpherson, L., Park, J.Y., Lim, Y.P. and Bancroft, I. (2004) Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. *Plant J.* **40**, 725–733.

Roulin, A., Piegu, B., Wing, R.A. and Panaud, O. (2008) Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon *RIRE1* within the genus *Oryza*. *Plant J.* **53**, 950–959.

Salmon, A., Ainouche, M.L. and Wendel, J.F. (2005) Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (*Poaceae*). *Mol. Ecol.* **14**, 1163–1175.

Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45.

Schnable, P.S., Ware, D., Fulton, R. *et al.* (2009) The B73 maize genome, complexity, diversity and dynamics. *Science*, **326**, 1112–1115.

Second, G. and Rouhan, G. (2008) Human-mediated emergence as a weed and invasive radiation in the wild of the CD genome allotetraploid rice species (*Oryza*, *Poaceae*) in the neotropics. *PLoS ONE*, **3**, e2613.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M. and Levy, A.A. (2001) Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell*, **13**, 1749–1759.

Song, K., Lu, P., Tang, K. and Osborn, T. (1995) Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proc. Natl Acad. Sci. USA*, **92**, 7719–7723.

Stephens, S.G. (1951) Possible significance of duplication in evolution. *Adv. Genet.* **4**, 247–265.

Suchard, M.A., Weiss, R.E. and Sinsheimer, J.S. (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18**, 1001–1013.

Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL, robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612.

Swofford, D. (2002) *PAUP, Phylogenetic Analysis Using Parsimony, version 4.0b10*. Sunderland, MA: Sinauer Associates.

Tang, L., Zou, X., Achoundong, G., Potgieter, C., Second, G., Zhang, D. and Ge, S. (2009) Phylogeny and biogeography of the rice tribe (Oryzeae). Evidence from combined analysis of 20 chloroplast fragments. *Mol. Phylogenet. Evol.* **54**, 266–277.

Terada, R., Johzuka-Hisatomi, Y., Saitoh, M., Asao, H. and Iida, S. (2007) Gene targeting by homologous recombination as a biotechnological tool for rice functional genomics. *Plant Physiol.* **144**, 846–856.

Udall, J.A., Quijada, P.A. and Osborn, T.C. (2005) Detection of chromosomal rearrangements derived from homologous recombination in four mapping populations of *Brassica napus* L. *Genetics*, **169**, 967–979.

Vaughan, D.A., Morishima, H. and Kadowaki, K. (2003) Diversity in the *Oryza* genus. *Curr. Opin. Plant Biol.* **6**, 139–146.

Vaughan, D.A., Kadowaki, K., Kaga, A. and Tomooka, N. (2005) On the phylogeny and biogeography of the genus *Oryza*. *Breed. Sci.* **55**, 113–122.

Vaughan, D.A., Lu, B.R. and Tomooka, N. (2008) The evolving story of rice evolution. *Plant Sci.* **174**, 394–408.

Veitia, R.A., Bottani, S. and Birchler, J.A. (2008) Cellular reactions to gene dosage imbalance, genomic, transcriptomic and proteomic effects. *Trends Genet.* **24**, 390–397.

Vicentini, A., Barber, J.C., Aliscioni, S.S., Giussani, L.M. and Kellogg, E.A. (2008) The age of the grasses and clusters of origins of C4 photosynthesis. *Glob. Chang. Biol.* **14**, 2963–2977.

Wang, J., Tian, L., Madlung, A., Lee, H.S., Chen, M., Lee, J.J., Watson, B., Kagochi, T., Comai, L. and Chen, Z.J. (2004) Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. *Genetics*, **167**, 1961–1973.

Wang, J., Tian, L., Lee, H.S. and Chen, Z.J. (2006) Nonadditive regulation of FRI and FLC loci mediates flowering-time variation in Arabidopsis allopolyploids. *Genetics*, **173**, 965–974.

Wang, B., Ding, Z., Liu, W., Pan, J., Li, C., Ge, S. and Zhang, D. (2009) Polyploid evolution in *Oryza* officinalis complex of the genus *Oryza*. *BMC Evol. Biol.* **14**, 250.

Wendel, J.F. (2000) Genome evolution in polyploids. *Plant Mol. Biol.* **42**, 225–249.

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E. and Liu, Z.D. (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. *Plant Cell*, **15**, 1187–1197.

Wing, R.A., Ammiraju, J.S., Luo, M. *et al.* (2005) The *Oryza* map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol. Biol.* **59**, 53–62.

Xie, Y. and Wu, R. (1989) Rice alcohol dehydrogenase genes, anaerobic induction, organ specific expression and characterization of cDNA clones. *Plant Mol. Biol.* **13**, 53–68.

Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573.

Yang, Z. (2002) Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12**, 688–694.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Zhang, J. (2003) Evolution by gene duplication, an update. *Trends Ecol. Evol.* **18**, 292–298.

Zhang, L.B. and Ge, S. (2007) Multilocus analysis of nucleotide variation and speciation in *Oryza officinalis* and its close relatives. *Mol. Biol. Evol.* **24**, 769–783.

Zhu, Q.H. and 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.H., Zhang, F.M., Zhang, J.G., Zang, L.L., Tang, L., Wang, J., Sang, T. and Ge, S. (2008) Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol.* **9**, R49.

Zuccolo, A., Sebastian, A., Talag, J., Yu, Y., Kim, H., Collura, K., Kudrna, D. and Wing, R.A. (2007) Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evol. Biol.* **7**, 152.

Zuccolo, A., Ammiraju, J.S., Kim, H., Sanyal, A., Jackson, S. and Wing, R.A. (2008) Rapid and differential proliferation of the Ty3–Gypsy LTR retrotransposon Atlantys in the genus *Oryza*. *Rice*, **1**, 85–99.

