Molecular Plant Advance Access published November 8, 2013

15 MYA of evolution in the Oryza genus shows extensive gene family expansion

Julie Jacquemin[1], Jetty S.S. Ammiraju[1], Georg Haberer[2], Dean D. Billheimer[3], Yeisoo Yu[1], Liana C. Liu[1], Luis F. Rivera[1], Klaus Mayer[2], Mingsheng Chen[4], Rod A. Wing[1,5].

[1]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

[2]MIPS/IBIS, Helmholtz Center Munich, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany

[3]Arizona Statistics Consulting Laboratory, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

[4] State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

[5] Senior Corresponding Author

Corresponding author contact informations
Rod A. Wing
email. rwing@ag.arizona.edu
Tel. (+1) 520.626.9595
Fax. (+1) 520.621.1259

Running title: Gene family expansion within the *Oryza* genus

Summary: The evolutionary dynamics of seven gene families in three rice species were characterized by a comparative genomic analysis. After large size expansion were observed between these closely-related species, the molecular and adaptive mechanisms contributing to these size variation were investigated.

Abstract

In analyzing gene families in the whole genome sequences available for *O. sativa* (AA), *O. glaberrima* (AA) and *O. brachyantha* (FF), we observed large size expansion in the AA genomes compared to FF genomes for the super-families F-box and NB-ARC, and 5 additional families: the Aspartic proteases, BTB/POZ proteins, Glutaredoxins, Trypsin α-amylase inhibitor proteins, and Zf-Dof proteins. Their evolutionary dynamic was investigated to understand how and why such important size variations are observed between these closely-related species.

We show that expansions resulted from both amplification, largely by tandem duplications, and contraction by genes losses. For the F-box and NB-ARC gene families, the genes conserved in all species were under strong purifying selection while expanded orthologous genes were under more relaxed purifying selection. In F-box, NB-ARC and BTB, the expanded groups were enriched in genes with little evidence of expression, in comparison with conserved groups. We also detected 87 loci under positive selection in the expanded groups.

These results show that most of the duplicated copies in the expanded groups evolve neutrally after duplication because of functional redundancy but a fraction of these genes were preserved following neofunctionalization. Hence the lineage-specific expansions observed between *Oryza* species were partly driven by directional selection.

Keywords: gene family expansion, *Oryza*, tandem duplication, duplicated gene evolution

Introduction

A major goal in the field of evolutionary biology is to understand how genetic diversity affects phenotypic differences between species. Functional evolutionary innovations frequently result from gene duplications and subsequent divergence or reciprocal gene losses events, which can lead to reproductive isolation and speciation (Dobzhansky-Muller incompatibilities, Orr and Presgraves, 2000). Thus, large changes in gene family size, as revealed by comparative genomic studies (Lynch and Conery, 2003; Hua et al., 2011), have been hypothesized to serve as major contributors in the evolution of complex traits, diversification, and adaptation (Tähtiharju et al., 2012; Chang and Duda, 2012). For example, the dramatic expansion of the Receptor-Like Kinase/Pelle gene family, involved in biotic stress responses, coincided with the establishment of land plants. It expanded at higher rates than other kinase families as a consequence of adaptation to fast-evolving pathogens (Lehti-Shiu et al., 2009). The evolution of bilateral floral symmetry in the Lamiales, an order of Asteridae, was shown to result from duplication followed by functional divergence in the TCP gene family (Reeves and Olmstead, 2003). Studies in *Arabidopsis* have shown that K-box-containing MADS-box proteins form complexes with each other, and these complexes have different DNA-binding affinities, targetting different genes with roles in transcriptional activation and floral organ identity. Hence the amplification of K-box-containing MADS-box proteins has allowed the diversification of these genes' function in plant development through new protein-protein interactions (Hofer and Ellis, 2002). Theoretical models proposed for gene family evolution combine neutral processes like the stochastic birth and death model (BD), which predicts that gene families continuously undergo random gain and loss events, and directional processes, related to the functional fates of gene duplicates (Zimmer et al., 1980; Reed and Hughes, 2004; Hahn et al., 2005). Indeed, according to evolutionary models proposed for the fate of duplicated genes, new duplicated copies are randomly fixed by genetic drift and most of them are then randomly lost through recombination-dependent deletion, or the accumulation of loss-of-function mutations (pseudogenization). However, new copies can be fixed and subsequently preserved by selection for novel functions (neofunctionalization) or partitioning of ancestral functions (subfunctionalization) (Conan and Wolfe, 2008; Innan and Kondrashov, 2010), in which case the random BD model is violated and lineage-specific expansion events can be observed (Hahn et al., 2005).

Comparing closely-related lineages is crucial to better understand the evolutionary dynamic of gene families and its consequences, although most studies on gene duplication and gene family evolution

have focused on distantly related species, e.g. (Hanada et al., 2008; Volokita et al., 2011; Xu et al., 2012). Hahn *et al.* (2007) showed that 40% of all gene families within the *Drosophila* genus displayed size variation between 12 *Drosophila* species (~60 million years [MY] of divergence). The authors also identified newly evolved lineage-specific gene families. Even gene families with an apparent stasis in copy number were shown to have experienced rapid turnover (gain and loss) of individual genes. In cone snails, Chang and Duda Jr. (Chang and Duda, 2012) showed that high rates of gene duplication and rapid turnover led to dramatically divergent arrangements of A-superfamily conotoxin genes among four closely-related *Conus* species which diverged ~11 million years ago (MYA).

Because of its large number of species, well-characterized phylogeny, and numerous genetic resources available for rice (*Oryza sativa*), the *Oryza* genus is an exceptional model system to study short-term evolutionary dynamics in the plant kingdom. In addition to Asian rice, the genus includes African domesticated rice (i.e. *O. glaberrima*) as well as 22 wild relatives, that have diversified across a broad ecological range (Vaughan et al., 2003) within a relatively short time scale (~15 MY). A considerable amount of useful genetic resources is preserved in the wild species and can be utilized to improve cultivated rice, the developing world's most important food crop. A large array of molecular resources has been developed for the 10 distinct genome types (x = 12) in the genus (Wing et al., 2005), including genome sequences of several species, as part of the International *Oryza* Map Alignment Project (I-OMAP) (Jacquemin et al., 2013; http://www.gramene.org; http://www.genome.arizona.edu). In a previous comparative analysis of an orthologous genomic region (~120-300kb) encompassing the *Adh*1-*Adh*2 loci across nine *Oryza* species, dynamic patterns of gene gain and loss were observed for several gene families (Ammiraju et al., 2008). The F-box family in particular displayed significant copy number variation throughout the evolution of the *Oryza* genus, with a tendency to expand in the more recently diverged AA genomes. Only two F-box genes were identified in the *Adh*1-*Adh*2 region of *O. brachyantha*, a wild relative of rice harboring the FF genome type, while 12 family members were identified in the orthologous region of *O. sativa* ssp. *japonica* cv. Nipponbare (AA genome).

In the present study, we extend our previous analysis of gene family evolution based on a single orthologous genomic region across the *Oryza* phylogeny, to a genome-wide investigation of gene families across three *Oryza* species for which sequences are available: *O. sativa* subspecies *japonica* (International Rice Genome Sequencing Project, 2005) and *indica* (Yu et al., 2002), *O. glaberrima S.* (AA, unpublished), and the wild annual African species *O. brachyantha* Chev. et Roehr (FF, Chen et al., 2013) (Figure 1). Initially, 32 families (Supplementary Table 4 online) were surveyed for significant

copy number variation between these species. The majority of these families showed very little copy number variation and thus appeared to be conserved. However 7 families listed in Table 1 exhibited large copy number variations between the FF genome and the two domesticated AA genome species and were selected for further investigation. These 7 gene families: i.e. F-box proteins, Aspartic proteases (Asp), BTB/POZ proteins (BTB), Glutaredoxins (GRX), NB-ARC proteins, Trypsin α-amylase inhibitor proteins (Tryp-α-amyl) and Zf-Dof proteins; have all been shown to play critical roles in plant stress responses or development (see Table 2).

We investigated the molecular mechanisms (either expansion by gene gain or contraction by gene loss) contributing to the dramatic size variation observed for these seven gene families between the AA and FF *Oryza* genome types. Here we focused primarily on tandem duplication which has been shown to be a primary mechanism for new gene formation and have contributed significantly to the expansion of plant gene families (Zhang and Gaut, 2003; Rizzon et al., 2006). In both *O. sativa* and *Arabidopsis thaliana,* 29% and 18% of genes have closely related paralogs generated from recent tandem duplications respectively (International Rice Genome Sequencing Project, 2005; Lockton and Gaut, 2005). We then investigated the mechanisms involved in the evolution of the lineage-specific gene family members, by searching for the presence of signatures of selective pressure on the genes that contributed to the observed expansions. To support the results of these analyses, and to determine if there is a functional bias for expanded genes *vs*. conserved genes, differences in the levels of transcription between these groups were also tested.

Results

Size variation in seven Oryza gene families

The size of the seven selected gene families, Asp, BTB/POZ, F-box, GRX, NB-ARC, Tryp-α-amyl and Zf-Dof, identified in four *Oryza* genome sequences (*Oryza sativa* ssp. *japonica* (Oj), *O. sativa* ssp. *indica* (Oi), *O. glaberrima* (Og) and *O. brachyantha* (Ob)) are shown in Table 1. Analysis of the F-box gene family revealed a major size variation between the FF genome (249 copies) and the AA genomes (649-767). This trend was observed for all other gene families with the exception of the NB-ARC family which showed a significant difference between the Og genome (398 members) relative to the Oj (610) and Oi (687) genomes. These large size variations happened in a relatively short evolutionary time frame of ~15 MY (Tang et al., 2010), stressing out the importance of analyzing closely-related species to capture this dynamic. A phylogenetic analysis was performed to cluster the families into sub-

families and test if one or few sub-families within each of the seven gene families were responsible for major size variation. We observed that a majority of the sub-families contributed to the global size variation (Supplementary Table 1 online), and not just one or few specific clusters. Only two sub-families, F-box-9 and NB-ARC-4, contained more Ob members than Oj members. The BTB-3 sub-family was AA genome-specific.

Orthologous relationships were established with OrthoMCL (Table 1), and the orthologous groups were then sorted with respect to major expansion events (Oj-Oi-Og and Oj-Oi), and groups with exactly one ortholog in each species (non-expanded). This last category represented up to 51% of the total number of groups for GRX, but the two largest families, F-box and NB-ARC, displayed smaller percentages (22 and 26% respectively) (Figure 2). The number of Oj-Oi-Og groups correlated with size variations. The large percentage (19%) of Oj-Oi groups (with orthologs present in *indica* and *japonica* genomes and not in Og and Ob genomes) found for the NB-ARC family correlate with the apparent size expansion observed for *Oryza sativa* genome. Thirty-two percent of the NB-ARC groups were found to belong to different categories, like Oj-Oi-Ob (65), Oi-Ob (36), Oj-Og (29) or Oj-Ob (17) (Figure 2), thereby suggesting that the NB-ARC family experienced high levels of gene birth and death over time, with many recent gene duplication events and differential gene losses. Inparalogs, paralogous genes resulting from post-speciation duplications, were also identified with OrthoMCL. The higher number of inparalogs found in the NB-ARC gene family (from 118 to 275) compared to the F-box family (from 14 to 52) (Table 1) also confirmed its highly dynamic evolution. Although the sizes of the seven gene families appear to be reduced in the FF genome, a fraction of Ob genes (15-35%) did not pair with an ortholog in the rice sequence (Oj) and remained lineage-specific in our analysis (Supplementary Table 2 online). This showed that expansion also occurred, at a smaller scale, in Ob when compared to the AA genomes.

Collinearity disruption can be ascribed not only to rearrangement events like inter or intra-chromosomal sequence translocations, but also to the misidentification of highly conserved paralogs in place of true orthologs. Such misidentifications could result from differential gene loss subsequent to gene duplications, or rapid divergence of the orthologous genes after speciation; both eventualities are expected to be considerable in the case of large gene family size variation. However, taking in account only the genes belonging to the 7 gene families under investigation, we observed that 85.7 to 100% of the Oj-Oi, Oj-Og and Oj-Ob orthologous pairs (OPs) were collinear (Supplementary Figure 1 online). Only 112 non-collinear and 122 non-syntenic OPs (out of a total of 3421) were counted for all families

between Oj and the remaining genomes (Supplementary Table 3 online), among which 44% and 38% belonged to the large F-box and NB-ARC families respectively, and 60% were Oj-Oi pairs. The surprisingly large number of non-syntenic Oj-Oi pairs observed for F-box (36) and NB-ARC (37) families could be partly explained by the large rearrangements (several deletions, inversions, translocations) detected in one region of Oi chromosome 11, compared with its homologous region in Oj (21.9-27.5 Mb) (Supplementary Table 3 online), resulting in the identification of paralogous relationships between genes on Oj chromosome 11 and genes located on different Oi chromosomes. Overall, collinearity within each family was highly preserved, showing that differential gene loss events are not frequent, and large family size variation is likely the result of specific duplications in the AA lineage, and specific losses in the FF genome.

The program CAFE (De Bie et al., 2006) was used to infer the direction of change in gene family size, i.e. lineage-specific duplication in species with expanded families or gene loss in species with smaller families. First, the global random birth and death rate (λ) of gene families in *Oryza* was estimated using our initial dataset of 32 families, with *Brachypodium distachyon* as outgroup (Supplementary Table 4 online). The BD rates estimated for the branches leading to and within the AA group (6.99x10-6) and for the Ob branch (6.67x10-6) were three times higher than the rate estimated for the outgroup branch (2.09x10-6), showing higher evolutionary dynamics of these families within the *Oryza* genus (Figure 1, Supplementary Figure 2 online). CAFE then tested if the seven selected families were evolving according to the estimated BD model or not, and determined the lineages along which the model has been violated for a specific gene family (and hence where large changes have taken place). All families except for the Zf-Dof family deviated significantly from the stochastic model ($p<0.05$). For the F-box family, both a significantly large contraction along the Ob branch and a significantly large expansion at node n2 ($p<0.05$) were observed. Asp, GRX, Tryp-α-amyl and Zf-Dof families showed a similar trend, but were not supported by P-values. For the BTB and NB-ARC gene families, the Og and Ob branches displayed significantly large contractions, while the Oi branch and the AA divergence nodes n1 and n2 displayed significantly large expansions ($p<0.05$). Globally these results showed that size variation in the seven gene families were the consequence of both gene losses in Ob, or Ob and Og in the case of the NB-ARC family, and lineage-specific duplications in all of the AA lineages.

Amplification mechanisms

The importance of tandem and segmental duplications in the expansion of the seven gene families were

investigated. For the F-box and BTB families, the numbers of tandem duplication arrays and tandemly duplicated genes were considerably different and correlated to family size between the AA and FF genomes (Figure 3 A and B). For the F-box family, 142, 144, and 109 genes were organized in 37, 38 and 30 tandem arrays for Oj, Oi and Og respectively, and only 2 arrays were detected for Ob on chromosomes 7 and 8, with a total of 8 tandemly duplicated genes. For the BTB family, 65, 84 and 50 tandem genes were identified in 10, 15 and 8 arrays, respectively for Oj, Oi and Og, but only 12 genes in 3 arrays for Ob. The differences resulted from the detection of very large AA genome-specific BTB arrays on chromosome 10 (around 12-15 Mb) (Supplementary Table 5 online), as well as additional arrays on chromosome 8. The BTB family also displayed the highest percentages of genes organized in tandem arrays for the AA genomes (Figure 3 C). For the NB-ARC family, a large number of tandem arrays and duplicated genes were observed in both *O. sativa* subspecies (34/138 in Oj and 19/88 in Oi), and smaller values for Og (6/30), which is again correlated with family sizes in each species (Figure 3 A and B). However, more NB-ARC genes were organized in tandem arrays in Ob (17/64) as compared with Og, even though the NB-ARC gene family size was similar in both species. Additional arrays in Ob were located on chromosomes 1, 4, 8, and 11. For the Tryp-α-amyl gene family, three tandem arrays for each of the AA genomes were observed, with 10 to 12 duplicated genes, but none were found for Ob. For the Asp, GRX and Zf-Dof gene families, no large variations were observed as the number of arrays were low or null. Tandem genes represented only 21% of the total number of Oj, Oi and Og genes in all Oj-Oi-Og groups for the F-box and Tryp-α-amyl gene families, and 22% of the total number of Oj and Oi genes in all Oj-Oi groups for the NB-ARC family, showing that tandem duplication is not the sole mechanism that can explain these expansions. This percentage was higher for the BTB family (56% for Oj-Oi-Og groups), pointing out again the larger importance of tandem duplications events for this family. To test the impact of local recombination rates on gene duplication and family size, the correlation between the number of genes and the mean recombination rate (cM/Mb) for 1 Mb intervals for the rice (Oj) was investigated using MareyMap (Rezvoy et al., 2007) (Supplementary Table 6 online). A significant correlation ($p<0.05$) was observed for the Asp, F-box, GRX and Tryp-α-amyl gene families. Correlations between recombination rate and the occurrence of tandem array genes were previously detected in rice (Rizzon et al., 2006) and *Arabidopsis* (Zhang and Gaut, 2003).

For Oj, Oi, Og and Ob, 39, 21, 26 and 23 segmental duplication blocks were detected, respectively (Supplementary Table 7 online). Relatively few genes in each family belonged to those blocks, even for

the largest F-box family (Figure 3 D).

Selective forces and expression analyses

To analyze the evolution of duplicated gene family members, and determine if genes in expanded and conserved groups are under different selective pressure and present functional biases, we tested two hypotheses. First, for the subset of genes in non-expanded groups we expected to observe signatures of purifying selection ($\omega<1$), and higher expression levels relative to expanded groups as they are conserved between all four *Oryza* genomes and may have essential functions. On the other hand, a part of the expanded groups (i.e. Oj-Oi-Og and Oj-Oi) are predicted to be the result of lineage-specific duplications (in an AA ancestor, or the Oj-Oi ancestor). Many of these new copies may have accumulated mutations neutrally because functional redundancy suppresses purifying selection pressure, and may be on their way to pseudogenization. Hence, expanded groups are predicted to contain coding sequences with no selection pressure, or relaxed purifying selection ($\omega\approx1$), and be enriched in loci with little or no evidence of expression. To test this hypothesis, the ω ratio (nonsynonymous/synonymous substitution ratio) was calculated pairwise between Oj, Oi, Og and Ob orthologs in expanded and non-expanded groups, and expression levels were analyzed *in silico*. Secondly, in the large expanded groups, we expected that a proportion of genes have been maintained after duplication because their functional traits were under adaptive selection during lineage divergence. This would apply whether an "expansion" is the result of lineage-specific duplication or gene loss. New gene copies may increase an organism's fitness by evolving into a beneficial function (neofunctionalization), in which case the duplicated genes are under positive selection in the preservation phase (Figure 4). To test for the presence of positive selection signatures, selective pressure was measured at each individual site for orthologs in the expanded groups. Other widely described models about the fixation and long-term maintenance of new copies in the genome include subfunctionalization (Force et al., 1999) and positive dosage (Ohno, 1970), but they were not tested in this study (see discussion).

Pairwise ω analysis between orthologs

The ω ratio was calculated pairwise between Oj, Oi, Og and Ob orthologs in OGs with a basic codon substitution model. The ω frequency distribution between all pairs in non-expanded groups had a peak between 0.1-0.3 for the Asp, BTB, F-box (Supplementary Figure 3 online), Tryp-α-amyl, and Zf-Dof

gene families, and between 0-0.05 for GRX. For the NB-ARC family, the observed peak was between 0.2 and 0.4 (Supplementary Figure 3 online). The infinite values (dS=0, ω=99) were displayed in a separate category in the frequency distribution, and were excluded for computing mean values, as an infinite ω value may indicate positive selection only if the dN (rate of nonsynonymous substitutions) value is high. The average ω values for all the orthologous pairs in non-expanded groups were less than 0.5, except for the NB-ARC Oj-Oi pairs (0.52±0.04), Zf-Dof Oj-Og pairs (0.58±0.25), and Zf-Dof Oi-Og pairs (0.82±0.31) (Figure 5). The ω frequency distributions between all pairs in Oj-Oi-Og groups had peak values between 0-0.05 for Asp, GRX, Tryp-α-amyl, and Zf-Dof; 0.1-0.2 for BTB; 0.2-0.4 for F-box; and 0.5-0.7 for NB-ARC. For the F-box and NB-ARC gene families, all the mean ω values for the different pairs of species were higher than 0.5 (Figure 5). For Oj-Oi groups, the ω frequency distribution between all pairs had peak values between 0.4-0.5 for F-box and 0.5-0.6 for NB-ARC. The low number of comparisons did not allow us to construct histograms for the smaller families. The mean ω values for all Oj-Oi pairs were higher than 0.5 for F-box and NB-ARC. Considering the distributions, mean and deviation of ω values and the number of comparisons for each OG class, the expanded groups appeared to be under more relaxed purifying selection compared to non-expanded groups for the F-box and NB-ARC families, whereas major differences for the remaining families were not observed.

Expression analysis

We tested the hypothesis that non-expanded groups may be enriched in genes with higher levels of expression compared with expanded groups by looking at differences in global expression levels using Genevestigator (Hruz et al., 2008). Distributions of mean levels of expression were computed for Oj genes over 1275 samples/conditions from whole-genome array experiments. The mean and median signal intensity values of the distributions were larger for non-expanded groups in all families except for the median value of the GRX and Tryp-α-amyl families (Figure 6 and Supplementary Figure 4 online). Nine genes in all non-expanded groups had mean signal intensities over 50,000, but only 1 for all Oj-Oi-Og, and 1 for all Oj-Oi groups. However, the larger mean values observed for non-expanded groups were not explained only by the presence of a few genes with extreme values. Looking directly at the distributions (Supplementary Figure 4 online), 34% of signal intensity values were >5000 for all the non-expanded groups, while only 8.6% and 10% were >5000 for all Oj-Oi and Oj-Oi-Og groups respectively. Signal intensity distributions of non-expanded and Oj-Oi groups differed significantly (Mann-Whitney-U, $p<0.05$ two-tailed) for the BTB, F-box, NB-ARC, Zf-Dof gene families.

Distributions of non-expanded and Oj-Oi-Og groups differed significantly for the Asp, BTB, F-box, NB-ARC families (Supplementary Figure 4 online). No significant difference was observed between the Oj-Oi and Oj-Oi-Og distributions. This preliminary *in silico* analysis of gene expression confirmed our hypothesis that expanded groups are enriched in loci with little or no evidence of expression relative to the non-expanded groups, for the BTB, F-box and NB-ARC families.

Site specific selective pressure between orthologs and paralogs

A branch-site codon substitution model (Yang and Nielsen, 2002) was applied to measure selection pressure at each individual site for orthologs in the expanded groups Oj-Oi-Og and Oj-Oi (foreground branches) compared to their closest paralogs in each family, considered as their putative ancestor (background branches). For each expanded group, the proportions of codon sites belonging to each ω category p0, p1, p2a, p2b (Table 3) in the foreground branches were analyzed. The p2b category was considered equivalent to p1 if ω2b =1 (with ω1=1). The sites of category p2a with ω2a=1 (ω0=0) were classified in a different category, p2a-1, to distinguish between neutral evolution and positive selection. In 60 to 100% of expanded OGs in the Asp, BTB, GRX, Tryp-α-amyl and Zf-Dof families, the orthologous sequences were composed of more than 50% of purifying selection sites (i.e. p0), and the remaining sites behaved in a neutral manor (i.e. p1 and p2a_1 sites). For 9 Oj-Oi-Og groups (1, 3, 3 and 2 groups for the Asp, BTB, GRX, and Tryp-α-amyl families, respectively), orthologous sequences were composed of more than 50% positive selection sites (i.e. p2a_x+p2b). For the F-box and NB-ARC expanded groups, the proportion of OGs with foreground branches composed of more than 50% purifying selection sites was smaller (33.3 to 55%), and we observed more OGs for which neutral sites represented more than 50% of the orthologous sequences (26.7 to 42.8%), and more OGs for which positive selection sites represented more than 50% of the orthologous sequences (9 and 20 for F-box Oj-Oi and Oj-Oi-Og, 5 and 6 for NB-ARC Oj-Oi and Oj-Oi-Og) compared to other families. The branch-site test for positive selection (Zhang et al., 2005) was applied to isolate individual sites under significant positive selection in expanded groups. A total of 105 loci under positive selection were detected, but 142 sites overall as the identified mutation is occasionally present on 2 or 3 of the orthologous genes in a particular group. The presence of these polymorphisms in at least one resequenced population allowed us to confirm 82% (117/142) of these sites (see Methods). They were distributed among 87 loci distributed in 68 genes (29 Oj, 27 Oi, 12 Og) and 42 different OGs; 2 Oj-Oi-Og in Asp, 10 Oj-Oi-Og in BTB, 1 Oj-Oi and 6 Oj-Oi-Og in F-box, 20 Oj-Oi and 3 Oj-Oi-Og in NB-ARC (Supplementary Table 8 online). Our results suggest that the expanded groups in the smaller gene

families globally evolve under purifying selection with very few sequences under positive selection, while the two largest families, F-box and NB-ARC, display more sequences with no selection pressure and positive selection

Discussion

Size variations observed for 6 gene families between *O. brachyantha* and the two domesticated species *O. sativa* and *O. glaberrima* resulted from both expansions in the AA and contractions in the FF genome lineages. Numerous duplication events occurred during the divergence of the ancestral AA genome, and then in *O. sativa* after its divergence from *O. glaberrima*, which confirms the ongoing trend for the steady increase of genes *via* duplication in the AA genome (Ammiraju et al., 2008). For the NB-ARC family however, *O. sativa* ssp. *japonica* contained 610 members, while *O. glaberrima* and *O. brachyantha* had 398 and 373 members respectively. The large difference in gene family content appears to have resulted from multiple gene duplication events in *O. sativa*, and gene loss events in *O. glaberrima* and *O. brachyantha*. Although the overall gene family content in *O. brachyantha* is much smaller, relative to the AA genome lineages, lineage-specific duplications also occurred in the FF genome. Considering the relatively similar estimated whole genome size of these three species (*O. brachyantha*: 362 Mb (Ammiraju et al., 2006), *O. glaberrima*: ~357 Mb (Martinez et al., 1994), *O. sativa:* ~389 Mb (International Rice Genome Sequencing Project, 2005)), in comparison to the 3-fold variation observed across the entire *Oryza* genus, there do not seem to be a relationship between genome size and copy number variation for the 7 gene families studied. The sub-species *indica* underwent large expansions of the BTB, F-box, GRX, and NB-ARC families (from 11 to 77 additional genes) compared to *japonica*. Also, the majority of non-collinear and non-syntenic pairs were observed in the F-box and NB-ARC families between *japonica* and *indica*, although these genomes diverged only ~0.4 MYA (Zhu and Ge, 2005; Ge et al., 2005). However, the *O. glaberrima*, *O. brachyantha*, and *O. sativa* ssp. *indica* genome, sequenced with a whole-genome shotgun approach, are likely of lesser quality and less complete as the IRGSP gold standard sequence of *O. sativa* ssp. *japonica*, which was sequenced and finished using a clone-by-clone approach. Our observations therefore must be considered cautiously as the gene family content in each species could be influenced by the integrity of the genome assemblies (see Supplementary Table 9 online for the length of the pseudomolecules used in this study). In this era of high-throughput next-generation sequencing, it is important to remember that sequencing, assembly and annotation accuracy is critical to correctly

perform comparative genomic analysis and draw conclusions about genome evolution (Alkan et al., 2011).

For the F-box and NB-ARC gene families, the genes conserved in all species (non-expanded groups) showed evidence of strong purifying selection, while expanded OGs showed overall more relaxed purifying selection. Our preliminary *in silico* analysis of gene expression of the BTB, F-box and NB-ARC gene families showed that the non-expanded groups were substantially enriched in loci with higher average levels of transcription relative to the expanded groups, while the expanded groups were enriched in loci with little or no evidence of transcription. Thus, our results for the largest families, i.e. the BTB, F-box and NB-ARC gene families, supported our first hypothesis about the fate of the lineage-specific genes in the expanded groups: i.e. duplicated copies in the expanded groups Oj-Oi-Og and Oj-Oi mainly evolve neutrally after duplication because of functional redundancy. We can postulate that a large portion of these lineage-specific genes are on a path towards pseudogenization or may already be pseudogenes, thus they are not transcribed or their proteins expressed anymore. These genes under neutral evolution could serve as reservoir for upcoming genetic novelties as well. On the contrary, genes found in conserved clusters are fixed because they encode ancient proteins that perform essential tasks. Similar results on selection pressure and gene expression were obtained by Hua *et al.* (Hua et al., 2011) for F-box genes in 18 plant species divided into highly conserved and lineage-specific subsets, using EST data to estimate transcriptional activity.

Our second hypothesis was that a fraction of the lineage-specific expanded genes were preserved due to neofunctionalization. This model assumes that a duplication confers no selective advantage or disadvantage when it appears in the population, so that fixation of the duplicated copy is a neutral process. As the new gene copy is redundant, it will be relieved from negative selection while the original copy maintains its function. During this brief period of relaxed selection new copies can acquire new functions that will be consequently maintained by positive selection (Lynch and Conery, 2003; Innan and Kondrashov, 2010). We confirmed the presence of 87 loci under positive selection, distributed among 68 genes, most of which were members of the largest gene families (i.e. BTB, F-box and NB-ARC). BTB and F-box genes with sites under positive selection belonged mostly to the Oj-Oi-Og groups, while NB-ARC genes were in Oj-Oi groups, which correlates with the expansion patterns of these families. This indicate that the lineage-specific expansions observed between *Oryza* species, at least for the largest families under investigation, were not dependent only on the random birth and death rates of the gene families but were partly driven by ecological adaptation.

The number of genes with sites under positive selection was low compared to the overall gene family content, which supports the contemporary models of gene duplication evolution. Because most *de novo* mutations are likely to be neutral or deleterious, it is widely accepted that pseudogenization is the most probable fate for redundant genes (Kimura and King, 1979), and the creation of new gene function by neofunctionalization is rare. However, we may have missed some sites under positive selection, as the branch-site model applied is very conservative and assumes that all orthologs in expanded OGs are under the same selective constraints, which is unrealistic, as the progenitors of *O. glaberrima* and *O. sativa* evolved independently since their divergence ~500,000 YA. According to Conant and Wolfe (Connan and Wolfe, 2008), most duplicated genes that arise when the duplication itself is not advantageous are actually preserved in the genome by a subfunctionalization mechanism (Duplication-degeneration-complementation model, Force et al., 1999). This model assumes that degenerate mutations are fixed in both original and new copies by drift, with subsequent partition of the original function in terms of expression or protein function, so both copies must be maintained to perform the original function. As this mechanism may occur with or without the involvement of positive selection in the final preservation phase, it was not possible to detect it with our analysis (Innan and Kondrashov, 2010). Moreover the division of function can be due to mutations in the regulatory regions of the gene copies, and not only in the genes encoding the proteins. Hence a fraction of loci in expanded groups may have developed lineage-specific sub-functions that we could not detect. The positive dosage model can also explain the preservation of duplicated genes in the long-term, and consequently the expansion of gene families. This model assumes that the duplication itself is advantageous, because higher gene expression levels resulting from an additional copy is beneficial. Throughout the preservation phase, the selection pressure does not vary and both copies will be under negative or slightly relaxed purifying selection. Thus this model can not be tested by looking at signature of selection pressure and we cannot exclude that a fraction of the expanded genes in the seven gene families have been directly fixed by selection for increased gene dosage.

In 2007, Jain *et al.* showed that the F-box gene family in the rice genome (i.e. IRGSP Oj reference sequence) expanded in size *via* localized tandem duplication events. Our results showed that this same mechanism was a predominant force in the expansion of this family and 6 additional families throughout the *Oryza* genus in the last ~15 MY. Rapid gene gain through tandem duplication could be a major advantage for plant stress responses, favoring a directional evolution hypothesis. The Receptor-like Kinase/Pelle gene family, associated with biotic stress responses, was shown to have undergone a

dramatic expansion of specific sub-families in the plant lineage by means of tandem duplication (Lehti-Shiu et al., 2009). The authors suggested that this expansion was related to the importance of these sub-families in plant defense. Also Hanada *et al.* (2008), studying the functional bias of retained duplicate genes during vascular plant evolution, showed that genes in orthologous groups that expanded *via* tandem duplication tended to be involved in responses to environmental stimuli, and consequently an organism's adaptation, while those that expanded *via* non-tandem mechanisms tended to serve in intracellular regulatory roles.

An AA genome specific cluster of BTB gene family members (BTB-3) was observed. Seventy-two percent of the proteins belonging to BTB-3 in all 3 AA genomes contained a protein binding domain MATH (Meprin And TRAF Homology) in addition to the BTB/POZ domain, and 90% of BTB-3 Oj genes were annotated as expressed. BTB/POZ-MATH domain containing proteins have been shown to assemble with cullin proteins into functional E3 ligases, conferring substrate specificity to the complex in the same way as F-box proteins (Weber et al., 2005). F-box protein-coding genes are key regulators in many pathways including cell signaling, transcription, and the cell cycle, as their C-terminal domain is highly divergent and confers the crucial role of conferring specificity in the SCF complex. It is one of the largest and most polymorphic families in the plant kingdom, showing the importance of the SCF complex in protein regulation (Xu et al., 2009; Hua et al., 2011). Because of F-box and BTB proteins' substrate recognition properties the amplification and diversification of these families would be directly advantageous, as it will enable the recognition of variable and expanding protein targets for degradation. According to Xu *et al.* (2009), F-box expansion in *Arabidopsis*, poplar, and rice, since their divergence from a common ancestor, is due to sub-families which tend to be involved in specialized processes (e.g. pollen recognition or pathogen responses). Plant R genes, a majority of which bear a NB-ARC site along with leucine-rich repeats (NB-LRR), are involved in plant defense. A large number (20) of NB-ARC *O. sativa*-specific genes under positive selection were identified. The family displays high birth and death rates among the *Oryza* species, as shown by the diverse composition of orthologous groups, the large inparalog repertoire, and low collinearity values. This is consistent with previous observations of the highly dynamic nature of NB-LRR genes. Between *A. thaliana* and its close relative *A. lyrata* (~10 MY of divergence) only a little more than half of all NB-LRR genes are orthologous (Beilstein et al., 2010; Guo et al., 2011). R genes evolve rapidly due to selection pressures related to pathogens and disease defense (McHale et al., 2006). We can postulate that the fixation of new genes and subsequent gene family expansions in pathogen (NB-ARC) and

substrate (F-box and BTB) recognition families have been selected for by the need to diversify their reactive sites and recognize ever-evolving proteins. With respect to the smaller Tryp-α-amyl, Glutaredoxins, Zf-Dof, and Asp gene families, our results did not allow to draw any significant conclusions.

Since Ohno's proposition in 1970 on the importance of gene duplications on the origin of ecological adaptations, many studies have demonstrated that role. For example, a recent study by Tähtiharju *et al*. (2012) showed that the expansion of the CYC/TB1 gene family (a class of plant-specific TCP domain transcription factors) in Asteraceae is connected with the evolution of the increased complexity of the inflorescence architecture in this highly successful plant family. The large expansions of the F-box, BTB and NB-ARC gene families are certainly selected for because of their role in protein substrate recognition or biotic stress defense. The three *Oryza* species investigated in this study have faced different biotic and abiotic stresses because of their specific geographic range and local environments. *O. sativa* was domesticated in Asia and its wild ancestor *O. rufipogon* grows in Southeast Asia, while *O. glaberrima* was domesticated and is cultivated in West Africa. The two cultigens underwent human artificial selection pressure throughout the domestication process, unlike *O. brachyantha,* a wild species distributed in West and East Africa, and this may be a key factor to explain the large expansion of the families in Asian and African domesticated rice. This study has opened up several prospects for testing more amplification mechanisms (retroposition, types of recombination) and hypothesis about the fate of duplicated copies (pseudogenization, sub-functionalization, gene dosage), starting with the analysis of gene structure and orientation, pseudogene content, variations in the regulatory regions and transcriptomic changes. As reference genomes will be completed in the near-future for 17 of the 23 *Oryza* species, we will be able to extend our investigation of gene family copy number variation to the entire *Oryza* genus. This will allow us to better define the relative age of these amplifications/contractions events, and determine if they are specific or not to the AA and FF genomes. As mRNA sequences are produced currently by our group, we will be able to confirm the results obtained with our preliminary *in silico* analysis of gene expression levels in *O. sativa* ssp. *japonica* with transcriptome data for all the *Oryza* reference species. In-depth functional genomic studies to determine the lineage-specific function of each gene family copy would also help us understand how the expansions of gene families are associated with phenotypic and life-history trait divergence within the *Oryza* genus.

Methods

Identification of gene families, orthology relationships and collinearity

We initially targeted 32 gene families (Supplementary Table 4 online) across the four *Oryza* species for which genome sequences were available. The HMM-profile of each family domain (see Supplementary Table 4 online for Pfam identifiers), downloaded from Pfam 26.0 (http://www.pfam.sanger.ac.uk; Punta et al., 2012), was used as query to search the annotated proteomes of *Oryza sativa* ssp. *japonica* and *indica*, *O. glaberrima* and *O. brachyantha* (E-value≤1.0) with HMMER 3.0 (http://www.hmmer.janelia.org). Information on genome sequences and annotation data for the four *Oryza* genomes are displayed in Supplementary Table 9 online. Alternative transcripts and TE related proteins were filtered. For the large F-box family, to identify additional genes missed in the annotation, a similarity-based re-annotation was performed (Supplementary Method 1 online). Among the 32 families, 7 families were selected (Supplementary Table 10 online) with the largest size differences between *O. brachyantha* and the *Oryza sativa* ssp. *japonica* reference sequence *(*Ob/Oj < 0.70 and at least 10 more genes in Oj compared to Ob). The families were divided into sub-families using an initial neighbor-joining tree constructed with Clustal 2.0 (Larkin et al., 2007). To identify domain composition of the AA-specific sub-family BTB-3, we searched the PfamA database using HMMER3 (E-value≤ 1.0).

OrthoMCL (Li et al., 2003) was used to cluster the proteins into orthologous groups (OGs) (Supplementary Method 1). Once the final OGs were obtained, the genes without chromosome coordinates were excluded, and syntenic (genes located on homologous chromosomes), collinear (syntenic genes in corresponding orders) and non-syntenic orthologous genes were defined for each pair of species (see more details in Supplementary Method 1 online). Circos was used to visualize the gene distribution and collinearity (Krzywinski et al., 2009).

Evolution of gene family sizes

CAFE (Computational Analysis of gene Family Evolution) was runned to compute the global birth and death rate (λ, probability of both gene gain and loss per gene per MY, assuming that they are equally probable) of gene families in the *Oryza* genus and calculates the most likely ancestral family sizes (internal nodes), from which we can infer if the variation is the consequence of expansion or contraction (De Bie et al., 2006). The dataset for 32 families (Supplementary Table 4 online) and the species tree displayed in Figure 1 were used in the analyses. The number of random samples for the

Monte Carlo re-sampling procedure was 1000 (see Supplementary Method 1 online for more details).

Composition of tandem duplication and presence in segmental duplications

The tandem arrays were identified with DAGchainer (Haas et al., 2004). To create a list of putative paralogous pairs, an all-versus-all BLAST search was run on the protein dataset for each family and each species (E-value<1e-30). Arrays with at least 2 copies (-A 1) and a maximum distance between two matches of 20 kb were marked at putative tandem arrays. To determine if any of the genes in our gene families were part of segmental duplications, segmental duplication blocks were constructed for the four genomes using a homology search at the DNA level. To define putative paralogous pairs, an all-versus-all BLAST search was run on the whole CDS dataset for each species (E-value<1e-10). The results were sorted according to the CIP (cumulative identity percentage) and CALP (cumulative alignment length percentage) parameters of Salse *et al.* (2008), both defined at 70%. DAGchainer was run with a minimum of 5 aligned pairs in one block, and a maximum of 1 Mb allowed between two matches. If two blocks on the same chromosomes were less than 1 Mb apart (on both chromosomes), they were concatenated. The lists of genes in our families were then compared to the list of collinear genes in each segmental duplication block. Local recombination rates (cM/Mb) were recovered every 100 kb along each Oj chromosome using MareyMap (Rezvoy et al., 2007), by comparing rice genetic (Muyle et al., 2011; http://www.rgp.dna.affrc.go.jp) and physical maps (Marey's map). Cubic splines were chosen as the interpolation method. The mean recombination rates for 1 Mb-intervals were computed and compared with the number of *japonica* genes from each family in the same interval. The correlations were then tested using a Poisson regression analysis (Supplementary Table 6 online).

Selective pressure analysis

Codon-based CDS alignments for each individual OG were created from protein alignments (MUSCLE) with the BioPerl module bp_mrtrans (Jason Stajich). ML trees were constructed with RAxML using the GTRGAMMA model and default settings. The basic codon substitution model of Goldman and Yang (1994) in Codeml (Yang, 2007) was used to compute pairwise ω ratios between orthologs. The branch-site codon substitution model (Yang and Nielsen, 2002) and the branch-site likelihood ratio test of positive selection (Zhang et al., 2005) were used to look at the selective pressure on the orthologs in the Oj-Oi-Og or Oj-Oi groups compared to their paralogs. The branch-site likelihood ratio test (LRT) was applied for alignments with p2a and/or p2b> 0 and at least 1 site under

positive selection in one of the foreground branches. The BEB (Bayes Empirical Bayes) method (Yang et al., 2005) gave posterior probabilities that each site was from a particular class. For alignments with a significant test in favor of model A, the sites with high posterior probabilities to belong to the 2a or 2b classes (P > 90%) were selected only (see Supplementary Method 1 online for more details). Finally, each putatively positively selected site was manually checked at both the protein and DNA levels to remove false positives due to misalignments.

All sites that appeared to be under positive selection in the three AA genome species were validated using Illumina HiSeq resequencing data generated at AGI (3 *O. sativa* ssp. *japonica*, 3 *O. sativa* ssp. *indica*, 5 *O. glaberrima* accessions: Wing et al., unpublished), or obtained from Dr. Wen Wang (2 accessions each of the 2 *O. sativa* subspecies (Xu et al., 2012) (Supplementary Table 11 online). BWA (Li and Durbin, 2009) and SAMtools (Li et al., 2009) were used for read alignments which were visualized using Tablet (Milne et al., 2010). A polymorphism was considered validated if 5 or more reads from a single accession confirmed the presence of the positively selected site.

Expression analysis

Genevestigator (Hruz et al., 2008) is a gene expression database in which public microarray data are collected, manually curated and normalized. At the time of this analysis, it contained data for 607 samples (1275 including replicates) from 63 whole genome 51K array (Affymetrix GeneChips) experiments for the IRGSP rice reference sequence (*Oryza sativa* ssp. *japonica* cv. Nipponbare). This database was queried for gene expression levels of Oj genes in expanded and non-expanded OGs, across the entire set of experimental treatments. After identifying the probe set ID corresponding to the specified genes, Genevestigator displayed signal intensities from Affymetrix probe sets. Only probes that targeted a single gene were selected. If a gene was targeted by several probe sets, only one was selected. Signal intensities were averaged over 1275 samples for each gene.

Supplementary data

The following supplementary data are available with the online version of this paper.

Supplementary Table 1. Sub-family copy number.

Supplementary Table 2. Number and percentage of orthologous groups and orthologous genes for each gene family.

Supplementary Table 3. Non-syntenic orthologous pairs for all gene families between *Oryza sativa*

ssp. *japonica* (Oj) and *O. sativa* ssp. *indica* (Oi), *O.glaberrima* (Og), *O.brachyantha* (Ob). Gene IDs are simplified compared to the initial annotations so only the species (Oj, Oi, Og or Ob) and the unique number are indicated.

Supplementary Table 4. Size of 32 families in the four *Oryza* genomes and *Brachypodium distachyon*.

Supplementary Table 5. DAGchainer's tandem arrays specific to each gene family. Gene IDs are simplified compared to the initial annotations so only the species (Oj, Oi, Og or Ob) and the unique number are indicated.

Supplementary Table 6. Results of the Poisson regression analysis for the relationship between recombination rate and gene family number.

Supplementary Table 7. Segmental duplication blocks in the current assemblies of four *Oryza* genomes with gene family members included in these blocks. Gene IDs are simplified compared to the initial annotations so only the species (Oj, Oi, Og or Ob) and the unique number are indicated.

Supplementary Table 8. List of loci with signature of positive selection. Gene IDs are simplified compared to the initial annotations so only the species (Oj, Oi, Og or Ob) and the unique number are indicated.

Supplementary Table 9. Genomic sequences and annotation information.

Supplementary Table 10. Genomic coordinates for members of 7 gene families across 3 *Oryza* species (*Oryza sativa* ssp. *japonica* and *indica*, *O. glaberrima*, and *O. brachyantha*). Gene IDs are simplified compared to the initial annotations so only the species (Oj, Oi, Og or Ob) and the unique number are indicated.

Supplementary Table 11. Informations about the population resequencing data used for selection pressure validation.

Supplementary Table 12. Correspondence between new IDs for F-box genes and simplified IDs of the initial annotations (Beijing Genomics Institute annotations for *Oryza sativa* ssp. *indica,* MSU v.6.1 for *Oryza sativa* ssp. *japonica*, Arizona Genomics Institute annotations for *Oryza glaberrima*, and Chinese Academy of sciences-IDGB annotations for *Oryza brachyantha*). The columns “status” indicate if the genes in the new annotations are either newly predicted genes (NEW), existed in a different form in the initial annotation (DIFF), or are the exact equivalent of the corresponding genes in the initial annotation (EQ).

Supplementary Figure 1. (a) Relationships of orthologous pairs between *Oryza sativa* ssp. *japonica* and both *O. glaberrima* and *O. brachyantha* for each family, (b) Number and percentage of syntenic,

collinear, and non-syntenic orthologous pairs (OP) for each pair of *Oryza* species.

Supplementary Figure 2. Results of CAFE analysis.

Supplementary Figure 3. Frequency distribution of ω ratios of all pairwise comparisons for the different types of orthologous groups for the F-box and NB-ARC gene families.

Supplementary Figure 4. Results of the global gene level expression analysis.

Supplementary Method 1. More detailed description of the methods applied in the study.

Acknowledgements

Authors kindly acknowledge Liu Hui and Dr. Wen Wang for providing resequencing data for accessions of the two *O. sativa* subspecies. This material is based upon work supported by the National Science Foundations under Grant #NSF1026200. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Figure legends

Figure 1. Phylogenetic relationships and divergence times between three *Oryza* species and *Brachypodium distachyon.* Rice and *Brachypodium* belong to the BEP clade (Bambusoideae-Ehrhartoideae-Pooideae) that diverged ~46 MYA (Sanderson, 1997). The AA and FF progenitors diverged ~15 MYA (Tang et al., 2010). Progenitors of *O. sativa* and *O. glaberrima* diverged ~0.7 MYA (Ge et al., 2005; Ma and Bennetzen, 2004), while the progenitors of the two *O. sativa* sub-species diverged ~0.4 MYA (Zhu and Ge, 2005; Ge et al., 2005).

Figure 2. Orthologous group composition for each gene family. The total number of groups is indicated above the histograms.

Figure 3. Number of tandemly and segmentally duplicated genes identified for each of the seven gene families.

Figure 4. Selection pressure expected for genes in the Oj-Oi-Og and non-expanded orthologous groups. The model is valid for Oj-Oi groups as well. Here we propose that the expansion group originated from a specific duplication in the AA ancestor, but it could also be derived from the loss of a corresponding ortholog in *O. brachyantha*. Orange genes are orthologs in Oj-Oi-Og groups while blue genes represent ancestral paralogs which are still conserved among the four *Oryza* species. In the case of neofunctionnalization for the expanded genes different hypotheses are possible. If, for example, a new function was fixed in the AA ancestor and subsequently conserved across the three AA species we should observe $\omega<1$ when we compare the orthologous copies. However, if the new copy evolved independently under positive selection in one of the AA species, we would observe $\omega>1$ for the ortholog comparison as well.

Figure 5. Pairwise ω values for the different classes of orthologous groups. Mean ω ratio and standard error are displayed for all pairwise comparisons in the different classes of orthologous groups for each family. The number of comparisons are displayed below the histogram.

Figure 6. Comparison of expression levels between the different classes of orthologous groups. Average values of the distribution of signal intensity values (expression unit) of Oj genes belonging to each type of OG (green bar: conserved groups, red bar: Oj-Oi-Og, blue bar: Oj-Oi) and each gene family.

Figure 7. **Percentages of orthologous groups with different proportions of site classes in each expansion group**. Below is an example for Oj-Oi-Og groups in the F-box gene family. p0: $\omega<1$ for background (BB) and foreground branches (FB), p1: $\omega=1$ and identical for BB and FB, p2a_1: $\omega<1$ for

BB and ω=1 for FB, p2a_x: ω<1 for BB and ω>1 for FB, p2b: ω=1 for BB and ω>1 for FB.

Tables

Table 1. Gene family composition. Pfam identifiers are indicated in brackets. The number of genes in OrthoMCL orthologous groups (OG) and the number of inparalogs are indicated for each species.

	japonica	*indica*	*O. glaberrima*	*O. brachyantha*	*B. distachyon*
Total	1680	1750	1120	972	485
Asp (PF00026)	**111**	**110**	**98**	**69**	**70**
Genes in OGs	101	98	87	55	/
Inparalogs	3	3	11	4	/
BTB (PF00651)	**166**	**200**	**134**	**83**	**110**
Genes in OGs	153	150	122	66	/
Inparalogs	14	22	15	15	/
F-box (PF00646)	**735**	**767**	**649**	**249**	**535**
Genes in OGs	703	701	594	204	/
Inparalogs	33	32	52	14	/
GRX (PF00462)	**60**	**71**	**61**	**40**	**57**
Genes in OGs	56	57	55	36	/
Inparalogs	3	3	6	1	/
NB-ARC (PF00931)	**610**	**687**	**398**	**373**	**364**
Genes in OGs	509	506	344	289	/
Inparalogs	254	275	118	135	/
Tryp-α-amyl (PF00234)	**138**	**126**	**113**	**84**	**93**
Genes in OGs	131	118	110	70	/
Inparalogs	3	4	3	8	/
Zf-Dof (PF02701)	**32**	**31**	**28**	**12**	**28**
Genes in OGs	30	28	26	10	/
Inparalogs	0	1	1	0	/

Table 2. Putative functions of the seven gene families under investigation, as described in the literature.

Family	Role	Reference
F-box	Subunit of Skp1-cullin-F-box (SCF) proteins, a major class of E3 ligases, which are components of the ubiquitin mediated protein degradation pathway. Role in embryogenesis, hormonal responses, seedling development, floral organogenesis, senescence, pathogen resistance, self incompatibility.	Kipreos and Pagano, 2000 Jain et al., 2007
Aspartic proteases	Proteolytic enzymes Role in various processes like processing of seed storage proteins, mobilization of nitrogen resources during seed germination, organ senescence, pollen/pistil interaction or defense responses against microbial pathogens and insects	Schaller, 2004 Xia et al., 2001
BTB/POZ	Domain BTB (BR-C, ttk, bab) or POZ (Pox virus, Zinc finger) involved in homophilic and heterophilic interactions, or the ubiquitin proteasome pathway like F-box proteins. Role in transcription regulation	Weber et al., 2005
Glutaredoxins	Glutathione-dependent redox enzymes Involved in oxidative and osmotic stress responses, by reducing and regulating toxic reactive oxygen species that accumulate in cells during abiotic stress and damage macromolecule and cell structures Also potentially involved in DNA synthesis, petal development, pathogen responses and iron sulfur cluster assembly	Rouhier et al., 2008 Wu et al., 2012
NB-ARC	Nucleotide binding ARC domain Plant resistance proteins (R genes) involved in pathogen recognition and activation of immune responses	Van Ooijen et al., 2008
Trypsin α-amylase inhibitor	Involved in plant defense as they impede insect digestion through bifunctional inhibition of their digestive α-amylases and proteinases	Franco et al., 2002
Zf-Dof	Highly conserved DNA-binding domain with a single zing finger Plant specific transcriptional activators or repressors involved in plant growth and development, stress, light and hormone-responses, phytochrome signaling, seed germination, and regulation of genes involved in storage or carbon metabolism	Gualberti et al., 2002 Yanagisawa, 2004

Table 3. Branch-site model A of codon evolution parameters (Zhang et al., 2005; Yang et al., 2005)

Site class	Proportion	Background	Foreground
0	p0	$0<\omega 0<1$	$0<\omega 0<1$
1	p1	$\omega 1=1$	$\omega 1=1$
2a	p2a	$0<\omega 0<1$	$\omega 2>1$
2b	p2b	$\omega 1=1$	$\omega 2>1$

Abbreviations

BD: birth and death; MY: million years; MYA: million years ago; IOMAP: International *Oryza* Map Alignment Project; Asp: Aspartic proteases; BTB: BTB/POZ proteins; GRX: Glutaredoxins proteins, Tryp-α-amyl: Trypsin α-amylase inhibitor proteins; Oj: *Oryza sativa* ssp. *japonica*; Oi: *Oryza sativa* ssp. *indica*; Og: *Oryza glaberrima*; Ob: *Oryza brachyantha*; OP: orthologous pair; OG: orthologous group; ML: maximum likelihood.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JJ participated in the design of the study, carried out the comparative genomic and bioinformatic analysis, performed the statistical analysis and wrote the manuscript. JSSA conceived the study and participated in its design. DDB carried out the Poisson regression analysis. GH, LCL, LFR, YY, KM and MC participated in the acquisition of the data and helped with the comparative genomic and bioinformatic analyses. RAW participated in the coordination of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

References

Alkan, C., Sajjadian, S. and Eichler, E.E. (2011). Limitations of next-generation genome sequence assembly. Nat Methods. **8(1)**, 61-65.

Ammiraju, J.S.S., Luo, M., Goicoechea, J.L., Wang, W., Kurdna, D., Mueller, C., Talag, J., Kim, H., Sisneros, N.B., Blackmon, B., Fang, E., Tomkins, J.B., Brar, D., MacKill, D., McCouch, S., Kurata, N., Lambert, G., Galbraith, D.W., Arumuganathan, K., Rao, K., Walling, J.G., Gill, N., Yu, Y., SanMiguel, P., Soderlund, C., Jackson, S., and Wing, R.A. (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.S.S., Lu, F., Sanyal, A., Yu, Y., Song, X., Jiang, N., Pontaroli, A.C., Rambo, T., Currie, J., Collura, K., Talag, J., Fan, C., Goicoechea, J.L., Zuccolo, A., Chen, J., Bennetzen, J.L., Chen, M., Jackson, S., and Wing, R.A. (2008). Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. Plant Cell. **20**, 3191-3209.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana.* Proc. Natl. Acad. Sci. USA. **107 (43),** 18724-18728.

Chang, D., and Duda, Jr. T.F. (2012). Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. Mol. Biol. Evol. **29(8),** 2019-2029.

Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., Li, B., Bai, Z., Goicoechea, J.L., Liang, C., Chen, C., Zhang, W., Sun, S., Liao, Y., Zhang, X., Yang, L., Song, C., Wang, M., Shi, J., Liu, G., Liu, J., Zhou, H., Zhou, W., Yu, Q., An, N., Chen, Y., Cai, Q., Wang, B., Liu, B., Min, J., Huang, Y., Wu, H., Li, Z., Zhang, Y., Yin, Y., Song, W., Jiang, J., Jackson, S.A., Wing, R.A., Wang, J., and Chen, M. (2013). Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. Nat. Communications. **4,** 1595.

Conan, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. Nat. Rev. Genet. **9,** 938-950.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. **27,** 1164-1165.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for

the study of gene family evolution. Bioinformatics. **22,** 1269-1271.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acid Res. **32(5),** 1792-1797.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics. **151,** 1531-1545.

Franco, O.L., Rigden, D.J., Melo, F.R., and Grossi-de-Sa, M.F. (2002). Plant α-amylase inhibitors and their interaction with insect α-amylases. Eur. J. Biochem. **269,** 397-412.

Ge, S., Guo, Y., and 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*. Edited by Toriyama K, Heong KL, Hardy B. Los Banos, Philippines: International Rice Research Institute Publications, 40-44.

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

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. Inform Software Tech. **47(15),** 965-978.

Gualberti, G., Papi, M., Bellucci, L., Ricci, I., Bouchez, D., Camilleri, C., Costantino, P., and Vittorioso, P. (2002). Mutations in the Dof Zinc Finger Genes DAG2 and DAG1 Influence with Opposite Effects the Germination of *Arabidopsis* Seeds. Plant Cell. **14,** 1253-1263.

Guo, Y., Fitz, J., Schneeberger, K., Ossowski, S., Cao, J., and Weigel, D. (2011). Genome-Wide Comparison of Nucleotide-Binding Site-Leucine-Rich Repeat-Encoding Genes in *Arabidopsis*. Plant Physiol. **157,** 757-769.

Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics. **20(18),** 3643-6.

Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C., and Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. **15,** 1153-1160.

Hahn, M.W., Han, M.V., and Han, S. (2007). Gene Family Evolution across 12 *Drosophila* Genomes. PLoS Genet. **3(11),** e197.

Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K., and Shiu, S. (2008). Importance of Lineage-

Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. Plant Physiol. **148,** 993-1003.

Hofer, J., and Ellis, N. (2002). Conservation and diversification of gene function in plant development. Curr. Opin. Plant Biol. **5**, 56-61.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Adv. Bioinformatics. **2008**, 420747.

Hua, Z., Zou, C., Shiu, S., and Vierstra, R.D. (2011). Phylogenetic comparison of F-box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift. PLoS one. **6(1),** e16219.

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

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

Jacquemin, J., Bhatia, D., Singh, K., and Wing, R.A. (2013). The International *Oryza* Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. Curr. Opin. Plant Biol. **16,** 1-10.

Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A.K., and Khurana, J.P. (2007). F-box Proteins in Rice. Genome-Wide Analysis, Classification, Temporal and Spatial Gene Expression during Panicle and Seed Development, and Regulation by Light and Abiotic Stress. Plant Physiol. **143,** 1467-1483.

Kimura, M., and King, J.L. (1979). Fixation of a deleterious allele at one of two duplicate loci by mutation pressure and random drift. Proc. Natl. Acad. Sci. USA. **76,** 2858-2861.

Kipreos, E.T., and Pagano, M. (2000). The F-box protein family. Genome Biol. **1(5),** 3002.1-3002.7.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an Information Aesthetic for Comparative Genomics. Genome Res. **19,** 1639-1645.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics. **23,** 2947-2948.

Lehti-Shiu, M.D., Zou, C., Hanada, K., and Shiu, S. (2009). Evolutionary History and Stress Regulation of Plant Receptor-Like Kinase/Pelle Genes. Plant Physiol. **150,** 12-26.

Li, L., Stoeckert, C.J. Jr., and Roos, D.S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. **13,** 2178-2189.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. Bioinformatics. **25,** 2078-9.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **25,** 1754-60.

Lockton, S., and Gaut, B.S. (2005). Plant conserved non-coding sequences and paralogue evolution. Trends Genet. **21,** 60-65.

Lynch, M., and Conery, J.S. (2003). The evolutionary demography of duplicate genes. J. Struct. Funct. Genomics. **3,** 35-44.

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

Martinez, C.P., Arumuganathan, K., Kikuchi, H., and Earle, E.D. (1994). Nuclear DNA content of ten rice species as determined by flow cytometry. Jpn. J. Genet. **69,** 513-523.

McHale, L., Tan, X., Koehl, P., and Michelmore, R.W. (2006). Plant NBS-LRR proteins: adaptable guards. Genome Biol. **7(4),** 212.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet - next generation sequence assembly visualization. Bioinformatics. **26(3),** 401-402.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glemin, S. (2011). GC-Biased Gene conversion and Selection affect GC content in the *Oryza* genus (rice). Mol. Biol. Evol. **28(9),** 2695-2706.

Ohno, S. (1970). Evolution by gene duplication. New-York: Springer.

Orr, H.A., and Presgraves, D.C. (2000). Speciation by postzygotic isolation: forces, genes and molecules. Bioessays. **22,** 1085-1094.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., and Finn, R.D. (2012). The Pfam protein families database. Nucleic Acid Res. **40,** D290-D301.

Reed, W.J., and Hughes, B.D. (2004). A model explaining the size distribution of gene and protein families. Math Biosci. **189,** 97-102.

Reeves, P.A., and Olmstead, R.G. (2003). Evolution of the TCP gene family in Asteridae: cladistic and network approaches to understanding regulatory gene family diversification and its impact on morphological evolution. Mol. Biol. Evol. **20(12),** 1997-2009.

Rezvoy, C., Charif, D., Gueguen, L., and Marais, G.A. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics. **23,** 2188-2189.

Rizzon, C., Ponger, L., and Gaut, B.S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol. **2,** e115.

Rouhier, N., Lemaire, S.D., and Jacquot, J. (2008). The Role of Glutathione in Photosynthetic Organisms: Emerging Functions for Glutaredoxins and Glutathionylation. Annu Rev. Plant Biol. **59,** 143-166.

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U.M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C. (2008). Identification and characterization of shared duplications between rice and wheat provide new insights into grass genome evolution. Plant Cell. **20,** 11-24.

Sanderson, M.J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. **14,** 1218-1231.

Schaller, A. (2004). A cut above the rest: the regulatory function of plant proteases. Planta. **220,** 183-197.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics. **22(21),** 2688–2690.

Storey, J.D., Taylor, J.E., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc.

Series B 66, 187-205.

Tähtiharju, S., Rijpkema, A.S., Vetterli, A., Albert, V.A., Teeri, T.H., and Elomaa, P. (2012). Evolution and Diversification of the CYC/TB1 Gene Family in Asteraceae-A Comparative Study in Gerbera (Mutisieae) and Sunflower (Heliantheae). Mol. Biol. Evol. **29(4),** 1155-1166.

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

The International Brachypodium Initiative. (2010). Genome sequencing and Analysis of the model grass Brachypodium distachyon. Nature. **463,** 763-8.

Van Ooijen, G., Mayr, G., Kasiem, M.M.A., Albrecht, M., Cornelissen, B.J.C., and Takken, F.L.W. (2008). Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. J. Exp. Bot. **59(6),** 1383-1397.

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

Volokita, M., Rosilio-Brami, T., Rivkin, N., and Zik, M. (2011). Combining Comparative Sequence and Genomic Data to ascertain Phylogenetic Relationships and Explore the Evolution of the large GDSL-Lipase Family in Land Plants. Mol. Biol. Evol. **28(1),** 551-565.

Weber, H., Bernhardt, A., Dieterle, M., Hano, P., Mutlu, A., Estelle, M., Genschik, P., and Hellmann, H. (2005). *Arabidopsis* AtCUL3a and AtCUL3b Form Complexes with Members of the BTB/POZ-MATH Protein Family. Plant Physiol. **137**:83-93.

Wing, R.A., Ammiraju, J.S.S., Luo, M., Kim, H., Yu, Y., Kudrna, D., Goicoechea, J.L., Wang, W., Nelson, W., Rao, K., Brar, D., Mackill, D.J., Han, B., Soderlund, C., Stein, L., SanMiguel, P., and Jackson, S. (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, Q., Lin, J., Liu, J., Wang, X., Lim, W., Oh, M., Park, J., Rajashekar, C.B., Whitham, S.A., Cheng, N., Hirschi, K.D., and Park, S. (2012). Ectopic expression of *Arabidopsis* glutaredoxin AtGRXS17 enhances thermotolerance in tomato. Plant Biotechnol. J. **10(8),** 945-955.

Xia, Y., Suzuki, H., Borevitz, J., Blount, J., Guo, Z., Patel, K., Dixon, R.A., and Lamb, C. (2001).

An extracellular aspartic protease functions in *Arabidopsis* disease resistance signaling. EMBO J. **23,** 980-988.

Xu, G., Ma, H., Nei, M., and Kong, H. (2009). Evolution of F-box genes in plants: Different modes of sequence divergence and their relationships with functional diversification. Proc. Natl. Acad. Sci. USA. **106(3),** 835-840.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., Li, J., He, W., Zhang, G., Zheng, X., Zhang, F., Li, Y., Yu, C., Kristiansen, K., Zhang, X., Wang, J., Wright, M., McCouch, S., Nielsen, R., Wang, J., and Wang, W. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. **30(1),** 105-114.

Xu, J., Bennetzen, J.L., and Messing, J. (2012). Dynamic Gene Copy Number Variation in Collinear Regions of Grass Genomes. Mol. Biol. Evol. **29(2),** 861-871.

Yanagisawa, S. (2004). Dof Domain Proteins: Plant-Specific Transcription Factors Associated with Diverse Phenomena Unique to Plants. Plant Cell Physiol. **45(4),** 386-391.

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19,** 908-917.

Yang, Z., Wong, W.S.W., and Nielsen, R. (2005). Bayes Empirical Bayes Inference of Amino Acids Sites Under Positive selection. Mol. Biol. Evol. **22(4),** 1107-1118.

Yang, Z. (2007). PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24,** 1586-1591.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L. *et al.* (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296(5565),** 79-92.

Zhang, L., and Gaut, B.S. (2003). Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? Genome Res. **13,** 2533-2540.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular level. Mol. Biol. Evol. **22(12),** 2472-2479.

Zhu, Q., 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.

Zimmer, E.A., Martin, S.L., Beverley, S.M., Kan, Y.W., and Wilson, A.C. (1980). Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc. Natl. Acad. Sci. USA. **77(4),** 2158-2162.

Figure-1

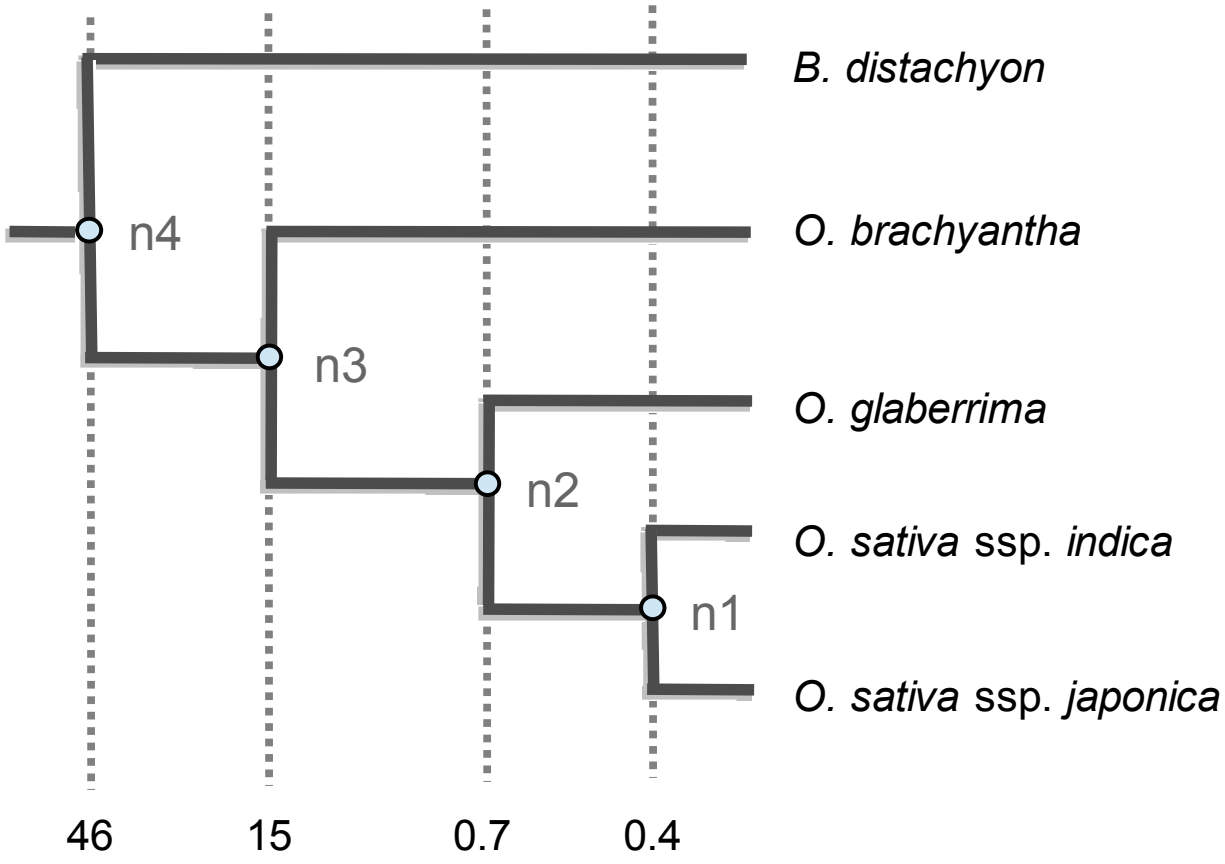


Figure-2

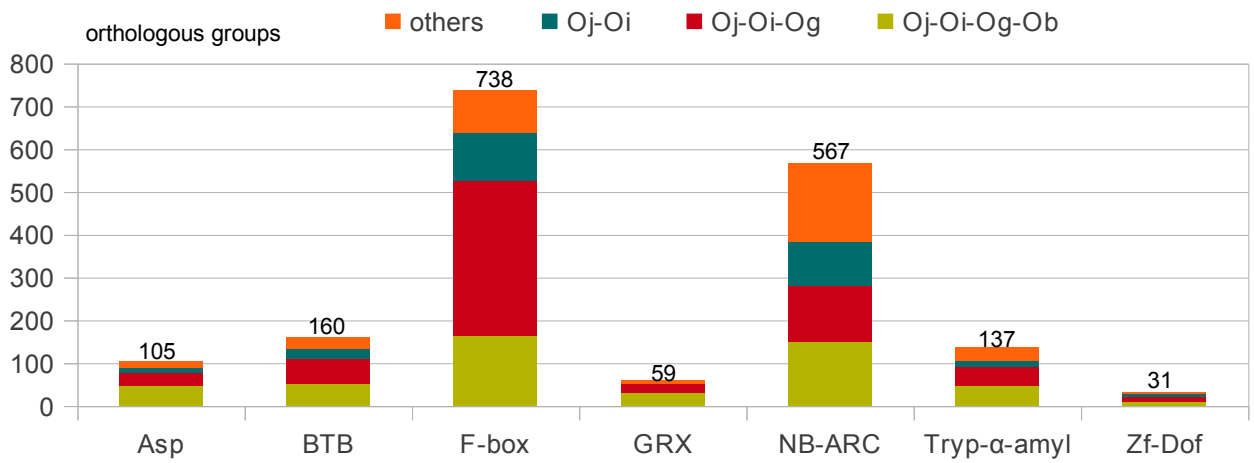


Figure-3

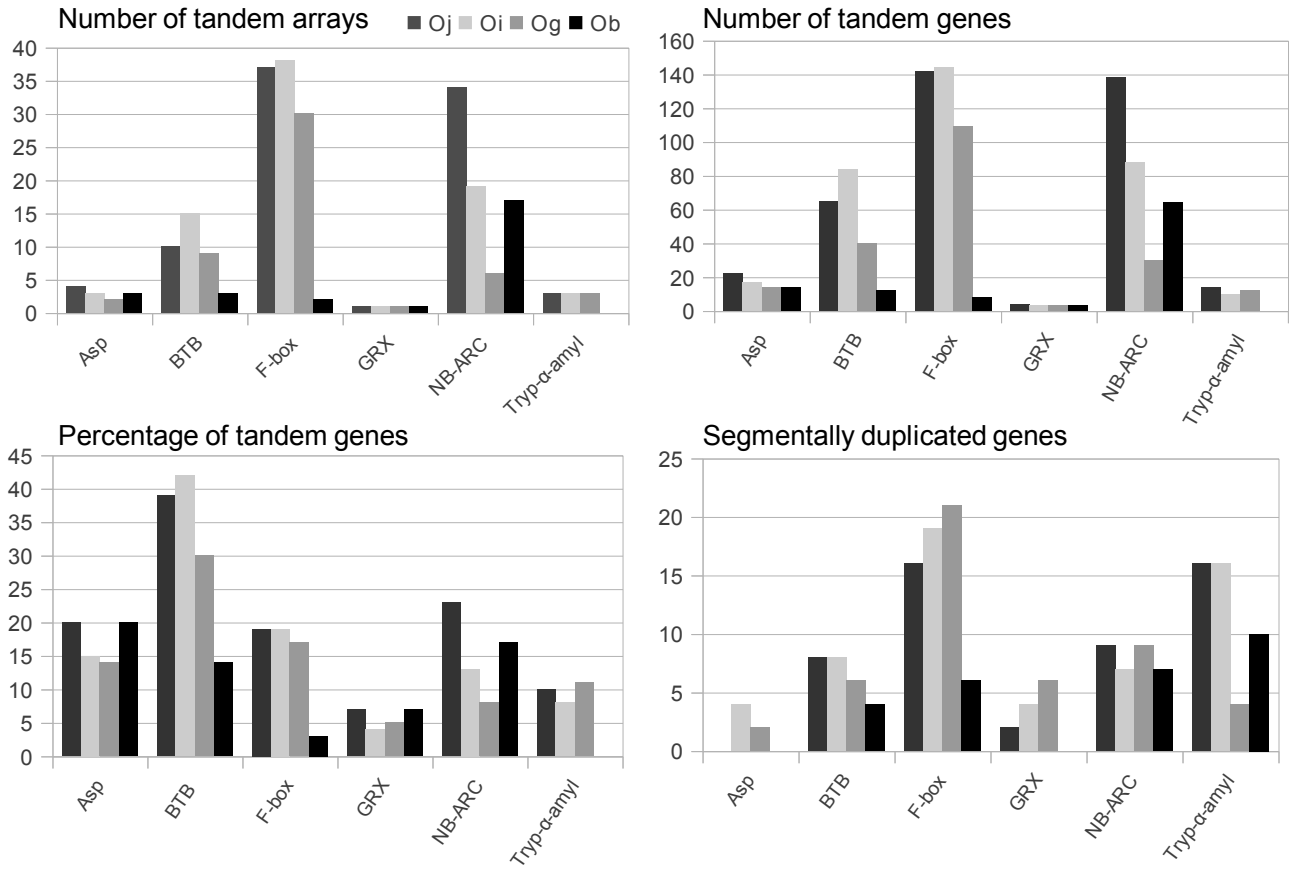


Figure-4

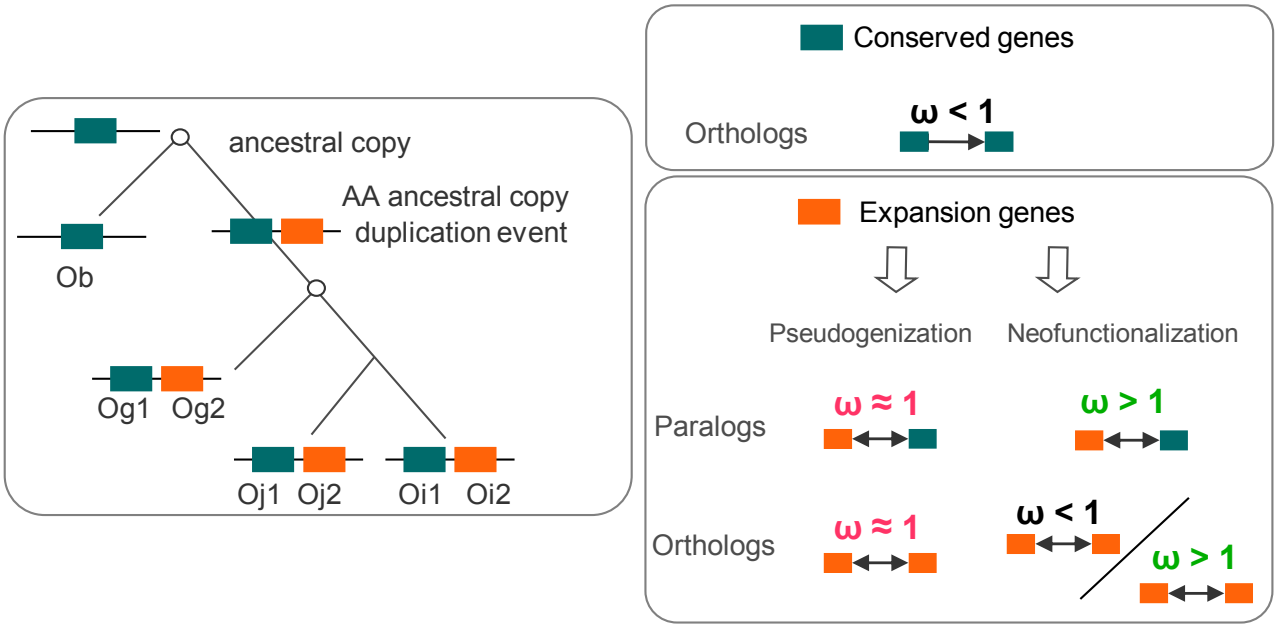
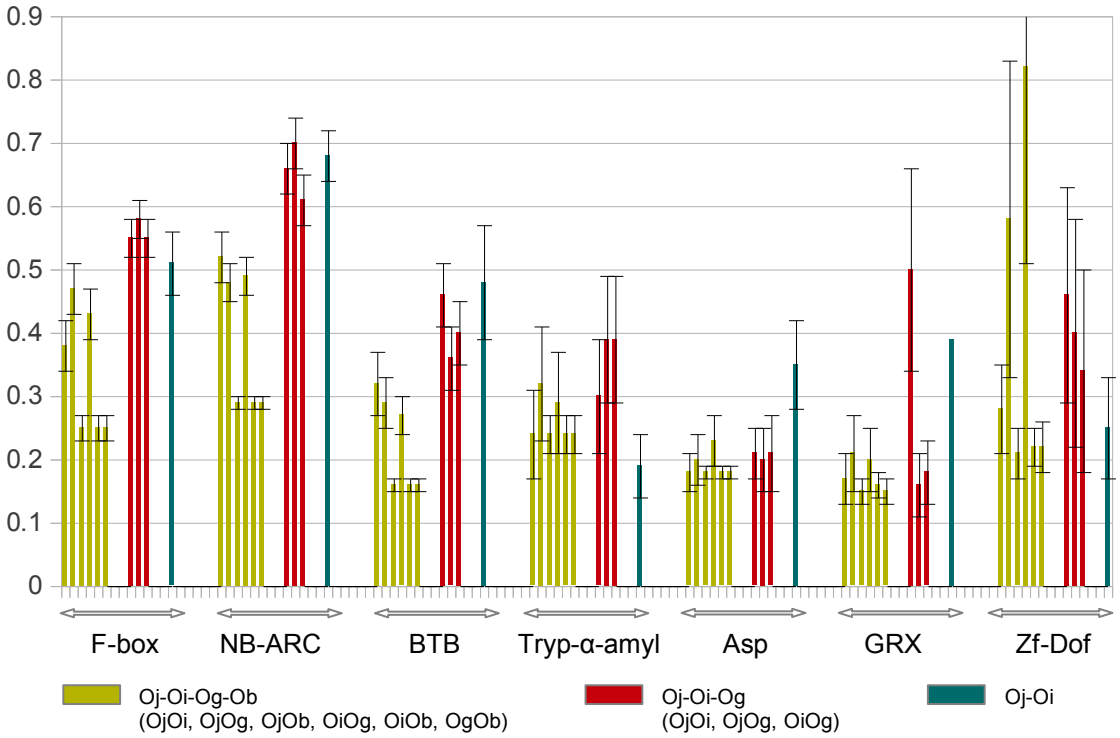


Figure-5



OGs	OPs	Fbox	NB-ARC	BTB	Tryp-α	Asp	GRX	Zf-Dof
Oj-Oi-Og-Ob	OjOi	126	138	44	35	42	26	8
	OjOg	134	128	42	35	43	23	7
	OjOb	99	139	50	38	43	29	9
	OiOg	140	129	46	39	43	25	7
	OiOb	98	139	50	38	43	29	9
	OgOb	99	128	50	38	42	29	9
Oj-Oi-Og	OjOi	295	112	53	39	27	18	13
	OjOg	319	95	49	38	27	17	7
	OiOg	313	94	47	40	27	18	10
Oj-Oi	OjOi	92	99	25	13	8	1	5

Figure-6

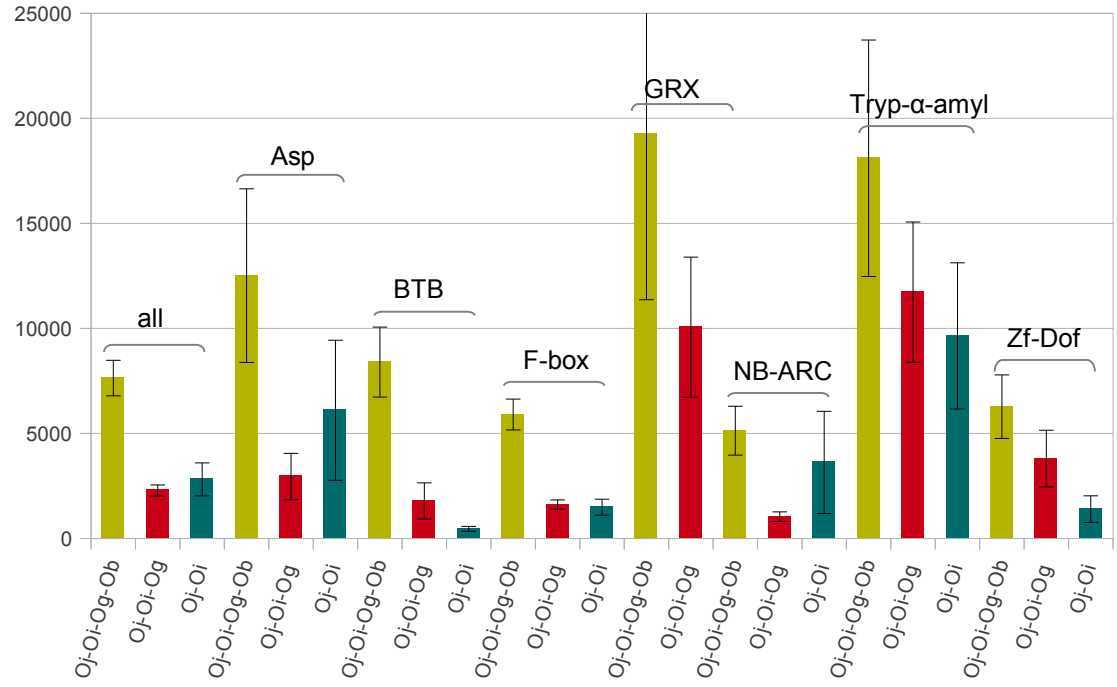


Figure-7

(%)	Total	po>50%	p1+p2a_1>50%	p2a_x + p2b>50%
Asp Oj-Oi	3	100	0	0
Asp Oj-Oi-Og	21	67.7	24	4.8
BTB Oj-Oi	18	77.8	16.7	0
BTB Oj-Oi-Og	45	73.3	11.1	6.7
F-box Oj-Oi	57	33.3	42	15.8
F-box Oj-Oi-Og	173	41.6	42.8	11.6
GRX Oj-Oi-Og	10	60	0	30
NB-ARC Oj-Oi	60	55	26.7	8.3
NB-ARC Oj-Oi-Og	82	50	29.3	7.3
Tryp-α-amyl Oj-Oi	6	66.6	16.7	16.7
Tryp-α-amyl Oj-Oi-Og	18	83.3	5.5	11.1
Zf-Dof Oj-Oi	1	100	0	0