Genome Biology and Evolution Advance Access published May 7, 2013 doi:10.1093/gbe/evt071

High Occurrence of Functional New Chimeric Genes in Survey of Rice Chromosome 3 Short Arm Genome Sequences

Chengjun Zhang[1†], Jun Wang[2†], Nicholas C. Marowsky[2], Manyuan Long[1], Rod A. Wing[3*], and Chuanzhu Fan[2*]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

[2]Department of Biological Sciences, Wayne State University, Detroit, MI 48202, USA

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

†These authors contributed equally to this work.

*Author for Correspondence:

Chuanzhu Fan, Department of Biological Sciences, Wayne State University, 5047 Gullen Mall, Detroit, MI 48202; Phone: 313-577-6451; Fax: 313-577-6891; email: cfan@wayne.edu.

Rod A. Wing, Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721,USA. Phone: 520-626-9595; Fax: 520-621-1259; email: rwing@ag.arizona.edu.

Running title: New chimeric genes in rice genome

Abstract

In an effort to identify newly evolved genes in rice, we searched the genomes of Asian cultivated rice *O. sativa* ssp. *japonica* and its wild progenitors, looking for lineage specific genes. Using genome pairwise comparison of ~20Mb DNA sequences from the chromosome 3 short arm (Chr3s) in six rice species, *Oryza sativa, O. nivara, O. rufipogon*, *O. glaberrima*, *O. barthii*, and *O. punctata*, combined with synonymous substitution rate tests and other evidence, we were able to identify potential recently duplicated genes which evolved within the last one million years. We identified 28 functional *O. sativa* genes which likely originated after *O. sativa* diverged from *O. glaberrima*. These genes account for around 1% (28/3176) of all annotated genes on *O. sativa*'s Chr3s. Among the 28 new genes, two recently duplicated segments contained eight genes. Fourteen of the 28 new genes consist of chimeric gene structure derived from one or multiple parental genes and flanking targeting sequences. Although the majority of these 28 new genes were formed by single or segmental DNA-based gene duplication and recombination, we found two genes which were likely originated partially through exon shuffling. Sequence divergence tests between new genes and their putative progenitors indicated that new genes were most likely evolving under natural selection. We showed all 28 new genes appeared to be functional, as suggested by Ka/Ks analysis and the presence of RNA-seq, cDNA, EST, MPSS, and/or small RNA data. The high rate of new gene origination and of chimeric gene formation in rice may demonstrate rice's broad diversification, domestication, its environmental adaptation, and the role of new genes in rice speciation

Keywords: chimera, comparative genomics, gene duplication, new gene, *Oryza*

Introduction

The genetic fundamental of organismal biodiversity is considerably relied on origination of new genetic elements. Myriad examples have provided evidence supporting newly evolved gene involvement in adaptive changes (Charrier et al. 2012; Chen et al. 2010; Des Marais and Rausher 2008; Ding et al. 2010; Fan et al. 2008a; Heinen et al. 2009; Jones et al. 2005; Long et al. 2003; Long and Langley 1993; Parker et al. 2009; Potrzebowski et al. 2010; Yeh et al. 2012; Zhang et al. 2002; Zhou et al. 2008). Understanding the molecular mechanisms involved in the formation of new genes is progressing rapidly, although many details of these mechanisms and their interactions await further investigation. As reviewed previously (Cardoso-Moreira and Long 2012; Kaessmann et al. 2009; Long et al. 2003; Ranz and Parsch 2012), the major mechanisms of new gene origination include, but not limited to: tandem gene duplication, exon shuffling, retroposition, mobile elements, horizontal gene transfer, gene fusion/fission, de novo origination, or a combination of two or more of the mechanisms (Bachtrog and Charlesworth 2003; Jones and Begun 2005; Wang et al. 2000). Systematical comparative genomic analysis using *Drosophila* genomes revealed that DNA-based gene duplication and retroposition played major roles in the formation of new genes (Yang et al. 2008; Zhou et al. 2008). Due to the limitation of genome sequence data and genetic resources, we do not know yet about the prospect of new gene formation in the plant kingdom as much as in animals, though a few recent studies have demonstrated that many similarities exist between plants and animals (Fan et al. 2008b; Sakai et al. 2011; Wang et al. 2006; Zhang et al. 2005; Zhu et al. 2009).

In order to understand the molecular processes and mechanisms governing the evolution of new genes and their functions, we must search for genes that originated recently and study their origination patterns and functions. The methods of detecting new genes have evolved

dramatically with the advancement of experimental and computational technology and massive DNA sequence data generated in both model and non-model organisms. Early discoveries of new genes were largely based on the detection of a single gene by chance. Phylogenetic comparisons of genetic signals (e.g. fluorescence in situ hybridization and genomic southern blotting) have also been used as an efficient and reliable way to identify new protein-coding genes in *Drosophila* and mammals at a larger scale (Betrán et al. 2002; Marques et al. 2005; Wang et al. 2004). This was not the case in plants, due to technical challenges, e.g. difficulty of cytogenetic analysis for plant chromosomes and low efficiency and high false positive rate of genomic southern blotting analysis to identify gene duplication events. In plants, previous works have also provided a tool that used array-based comparative genomic hybridization to identify potential new genes in the closely related *Arabidopsis* species (Fan et al. 2007). However, the most effective technique for finding duplications and further identifying new genes would be a genomic sequence comparison based on the availability of genome sequences. Similar efforts have been applied in the analysis of several other genomes, and yielded a fair amount of information contributing to our understanding of the evolution of genes and genomes (Chimpanzee Sequencing and Analysis Consortium 2005; Clark et al. 2007; Gan et al. 2011; Green et al. 2010; Hu et al. 2011; Jensen and Bachtrog 2010; Jun et al. 2009; Kim et al. 2011; Liti et al. 2009; Locke et al. 2011; Marques et al. 2005; Marques-Bonet and Eichler 2009; Marques-Bonet et al. 2009; Scally et al. 2012; Stein et al. 2003; Yu et al. 2005; Zhang et al. 2011). Moreover, comparing closely related species, as demonstrated in the *Drosophila melanogaster* subgroup (Yang et al. 2008; Zhou et al. 2008), provided more powerful strategy for identifying gene duplication events across the entire genome and for revealing the extent and pattern of new gene originations.

As part of an international effort to characterize the functions of all rice genes (Zhang et al. 2008), sequences of Chr3s using the BAC-based physical maps to select minimum tilling paths of BAC clones in most *Oryza* species have been finished and are publically available. Therefore, these genomic sequence data provide an opportunity to decipher gene and genome evolution at the phylogenetic level within a single genus using comparative genomics approaches. The genus *Oryza* is composed of 23 species which diverged over a relatively short time period ~15-20 million years ago (MYA) with broad diversification and largely solved phylogenetics (Ammiraju et al. 2008; Ge et al. 1999; Tang et al. 2010; Zhu and Ge 2005). *Oryza sativa* ssp. *japonica* and *O. glaberrima* are Asian and African cultivated rice species, respectively. Phylogenetically, *O. sativa* ssp. *japonica* and *O. glaberrima* belong to the AA genome type in the genus *Oryza*, which diverged roughly from 0.5 to 1 MYA (Ammiraju et al. 2008; Tang et al. 2010). Species of *O. punctata* belongs to the BB genome type and is used as outgroup of the AA genome *Oryza* species for phylogenetic analysis. AA and BB genome type species diverged at around 2~5 MYA (Figure 1)(Ammiraju et al. 2008; Tang et al. 2010). Through the genome sequence comparisons between Asian rice species (including *O. sativa, O. nivara*, and *O. rufipogon*) and African rice species (including *O. glaberrima*, *O. barthii*, and *O. punctata*), this study aimed to identify Chr3s potential new genes which recently originated in *O. sativa* and/or its wild species progenitors, *O. nivara* and *O. rufipogon*.

Materials and Methods

Searching *O. sativa* ssp. *japonica* specific new genes by comparative genome analysis

Sequence data of Chr3s in *O. glaberrima, O. punctata, and O. barthii, O. nivara and O. rufipogon* were downloaded from Gramene (http://www.gramene.org/). Chr3s sequences of *O. sativa* ssp. *indica* were downloaded from 2003/10/7 BGI version (ftp://ftp.genomics.org.cn/pub/ricedb/rice_update_data/genome/9311). The whole genome sequences of *O. glaberrima* were downloaded from ftp://ftp.gramene.org/pub/gramene/weix/oge/oge-toplevel-seq.tar.bz2. We performed genome pairwise comparisons between *O. sativa* ssp. *japonica* Chr3s coding sequences (CDS) and other five species Chr3s genome sequences. The annotation and CDSs of *O. sativa* ssp. *japonica* were downloaded from MSU Rice Genome Annotation Project (RGAP, MSU V7) (http://rice.plantbiology.msu.edu/data_download.shtml). To search for the *O. sativa* specific new genes, the first step was to identify the Chr3s orthologous genes among six species. We used two criteria to define the orthologous genes. First, we conducted a BLAT (Kent 2002) search for Chr3s orthologous genes by aligning genome sequences of *O. glaberrima, O. sativa* ssp. *indica*, *O. barthii, O. punctata, O. nivara*, and *O. rufipogon* against the CDSs of *O. sativa* ssp. *japonica.* We had two requirements: the alignment of the orthologous sequence needed to cover over 95% of the length of the *O. sativa* ssp. *japonica* CDSs, and must be located in the synteny region of all the genomes. Whether an *O. sativa* ssp. *japonica* gene was considered in the synteny region was defined by the presence of at least two flanking genes in the 30Kb DNA fragment containing the gene hit in other genomes. Second, the orthologous sequences were defined as two sequences with reciprocal best hits of each other. We conducted the reciprocal searches using BLAT and defined a pair of sequences from two genomes having the best hit against each other as "reciprocal" best hits. We descendingly sorted the hits according to the BLAT alignment score and then BLAT identity score (http://genome.ucsc.edu/FAQ/FAQblat.html#blat4 for methods to

compute these two scores). We then defined the ones ranking in the first as the "best" hits. After we identified the orthologous genes, we filtered them out and picked the remaining annotated genes which are only present in *O. sativa* ssp. *japonica* and/or the other three Asian rice species (*O. sativa* ssp. *indica, O. rufipogon, O. nivara*), but are absent in all the African rice species *O. glaberrima, O. barthii and O. punctata* (Figure 1). We further BLAT CDSs of *O. sativa* ssp. *japonica* specific genes to the entire O. *glaberrima* genome and identified their homologous regions in *O. glaberrima*. The results were then BLAT back to all CDSs of *O. sativa* ssp. *japonica*. We only selected *O. sativa* ssp. *japonica* genes which did not have reciprocal BLAT best hits in *O. glaberrima* genome as *O. sativa* ssp. *japonica* new gene candidates. These genes likely originated after the divergence between Asian rice species and African rice species about 1 MYA. We further estimated the average rates of synonymous substitution (Ks) using gKaKs pipeline with Yn00 method for all Chr3s orthologous genes earlier identified between *O. sativa* ssp. *japonica* and *O. glaberrima* (Zhang et al. 2013).

To determine the origination pattern of these recently evolved new genes in *O. sativa* ssp. *japonica*, we searched for their paralogs in the *O. sativa* ssp. *japonica* genome. To identify paralogous gene pairs, we BLAT the CDS sequences of the candidate genes against all the CDS sequences of *O. sativa* ssp. *japonica* with the match length of the paralogous gene pair >100bp and mismatch length/(mismatch length+match length) < 0.1. We picked up only the paralogous gene pairs with Ks less than 0.0192, which is the average Ks of the orthologous gene pairs between *O. sativa* ssp. *japonica* and *O. glaberrima* corresponding to 1 million years divergence time. We further removed the genes with "retrotransposon protein" and "transposon protein" terminology in their annotations to define the list of *O. sativa* ssp. *japonica* new gene candidates. Next, to test if these *O. sativa* lineage specific new genes were ancient duplicate genes that lost

in African *Oryza* species, we applied reciprocal blastp searches to identify if these new gene candidates contain orthologous copies in other distantly related species. We BLASTP protein sequences of these new gene candidates to all proteins in Uniprot (http://www.uniprot.org/) which includes SwissProt and TrEMBL data. If a new gene candidate had hits in other species, we BLASTP these hits back to all *O. sativa* ssp. *japonica* proteins (http://www.gramene.org/Multi/blastview). If this best hit from blastp search was the new gene, we deleted this new gene candidate. We also used Repeatmasker (RepeatMasker libraries version: rm-20120418) to scan the transposons existing in CDSs of new gene candidates.

Sequence divergence and phylogenetic analysis

We calculated the ratio of nonsynonymous substitution and synonymous substitution rates (Ka/Ks, donated as 'ω') using maximum likelihood algorithm (codeml) implemented in the PAML package (Yang 2007). The significance of ω that deviated from neutrality (ω =1) was tested using the likelihood ratio test (LRT). We aligned the sequences of paralogous/orthologous gene pairs using bl2seq (Altschul et al. 1997). We used codeml to calculate the ω value between the two sequences (Yang and Nielsen 2000). We then used codeml with two models (ω fixed at 1 and ω varying freely) to test whether any of the identified new genes were statistically under natural selection (Yang 2007). Phylogenetic analysis of the gene tree was performed using Neighbor Joining algorithm implemented in PAUP (Swofford 2002). The CDS sequences of the gene family were aligned using ClustalW (Larkin et al. 2007). The bootstrap analysis with 1000 replicates was used to assess the robustness of the branches.

To address if $\omega < 1$ is due to that the parental gene is under strong purifying selection and the new gene is a pseudogene evolving neutrally, we applied PAML branch model to calculate ω

values for the branch leading to new genes. We first downloaded the recently completed whole genome sequences of *O. glaberrima*, *O. barthii*, and *O. punctate* from http://www.iplantcollaborative.org. We identified the orthologous sequences of parental genes from the three outgroup species using ortholog search approach described above. We aligned only homologous region for all sequences using MAFFT (Katoh et al. 2005) and Perl scripts. We estimated ω for the foreground branch leading to the *O. sativa* ssp. *japonica* lineage specific new gene and for background branches leading to the parental genes and their orthologous genes in outgroup species (*O. glaberrima*, *O. barthii*, and *O. punctata*). We used a two-ratio model allowing different ω in foreground and background branches with PAML codeml. The significant level of foreground branch ω was tested using LRT compared with the null hypothesis of a model where foreground ω fixed to 1 and background ω varied freely (Yang 2007).

Expression analysis

The expression of identified new genes was determined by the presence of full-length cDNA (FL-cDNA), EST (Pontius et al. 2003), RNA sequencing transcriptome data (RNA-seq) (Davidson et al. 2012; He et al. 2010; Zemach et al. 2010), Massively Parallel Signature Sequencing (MPSS) (Nakano et al. 2006) and small RNA sequencing signatures (Nobuta et al. 2007). RNA-seq data, which were processed by RGAP, were downloaded from http://rice.plantbiology.msu.edu/expression.shtml. The transcription abundance was reported in fragments per kilobase of transcript per million fragments mapped (FPKM) across 11 libraries including leaves-20 days, post-emergence inflorescence, pre-emergence inflorescence, anther, pistil, seed-5 DAP, embryo-25 DAP, endosperm-25 DAP, seed-10 DAP, shoots, and seedling

four-leaf stage (Supplementary table 1, DAP = Days After Pollination). RGAP used Tophat v1.2.0 to map the sequence reads to the version 7 pseudomolecules in RGAP (Trapnell et al. 2009) and used Cufflinks v0.9.3 to calculate the expression abundances for RNA-seq libraries (Trapnell et al. 2010).

The NCBI EST library collection of *Oryza sativa* ssp. *japonica* was downloaded from http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=4530&CUTOFF=0, which contained 1,047,507 ESTs from 259 EST libraries expressed in 12 tissues (Supplementary table 2). We used BLAT to identify the genes corresponding to the ESTs with blast tabular format as output (the blat option –out=blast8). The criteria to define the corresponding gene of an EST were as follows: 1) the CDS of the gene was the first best hit of the EST; 2) the alignment of the EST and the best hit gene had an at least 95% identity, ≤ 1e-20 E value, and at least 100 blast score; and 3) the blast score of the first best gene hit was at least 5 points higher than that of the second gene hit (Wang et al. 2012). Thus, the corresponding relationships between ESTs and 26577 current annotated genes were constructed. We then collected the EST information for all *O. sativa* new genes.

MPSS and small RNA expression data were obtained from http://mpss.udel.edu/rice/mpss_index.php. MPSS expression data were reported in the sum for the abundance of unique signatures in TPM (transcripts per million) in 70 tissues (Supplementary table 3). Small RNA expression data were reported in the sum for the abundance of all the signatures in TPQ (transcripts per quarter million) in 6 tissues (stem, germinating seedlings, immature panicles, germinating seedling infected with *M. grisea*, seedlings treated with ABA, and seedlings control for ABA treatment) (Supplementary table 4). Because small RNAs can be biologically active in more than one sequence that they match, sequence matches

for small RNA were not required to be a unique signature.

Identification of new chimeric genes

After we compared the new genes with their paralogs, we detected that many new genes have formed chimerical gene structures with flanking sequences or other gene sequences. If the flanking or other gene sequences that a new gene recruited in the CDS are larger than 30 bp, we considered it as a new chimeric gene. To identify if a new chimeric genes has transcription evidence for the chimerical CDS structure, we mapped EST, full-length cDNA and RNA-seq sequences to the junctions of chimera. We obtained RNA-seq raw data from NCBI Sequence Read Archive (SRA: SRR352184.sra, SRR352187.sra, SRR352189.sra, SRR352190.sra, SRR352192.sra, SRR352194.sra, SRR352204.sra, SRR352206.sra, SRR352207.sra, SRR352209.sra, SRR352211.sra, SRR042529.sra, SRR034580.sra, SRR034581.sra, SRR034582.sra,SRR034583.sra) from http://sra.dnanexus.com/dispatch_many. We preprocessed the RNA-seq data with quality-control using trim_galore (Version 0.2.5) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) before mapping. We removed duplications existing in aligned reads due to PCR using picard-tools-1.79 (http://picard.sourceforge.net/) after mapping. Given the length of the RNA-seq reads ranging from 35 to 40 bases, we extracted 32bp DNA sequences of upstream and downstream flanking regions center at the breakpoints of a chimeric gene. We then mapped the RNA-seq reads to the extracted flanking DNA sequences with Tophat v2.0.7 (Trapnell et al. 2009). Finally, we checked whether any RNA-seq reads aligned on these flanking sequences and crossed the chimerical breakpoints. We applied similar approach to map the EST sequence data to the extracted chimera breakpoint flanking DNA sequences with BLAT. We also checked whether

these chimeric genes have FL-cDNA through browsing http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/.

Results

Identification of potential new gene candidates for *O. sativa* ssp. *japonica*

Three steps were carried out to detect the potential new genes that recently originated in *O. sativa* ssp. *japonica* and its wild progenitors. First, the comparative genomic analysis of Chr3s pseudomolecules among six species identified 862 annotated genes only present in *O. sativa* ssp. *japonica* and/or its progenitors. Second, from the 862 gene candidates, we filtered out the gene candidates which had reciprocal best hits in the *O. glaberrima* whole genome sequence. This yielded 753 *O. sativa* ssp. *japonica* specific gene candidates. Third, we BLAT these 753 candidates to all the CDSs of *O. sativa* ssp. *japonica* to find the best-hit paralogs and then calculated the Ks between *O. sativa* ssp. *japonica* specific gene and its paralog. Based on the average Ks=0.0192 of 1797 Chr3s orthologous genes between *O. sativa* and *O. glaberrima*, we inferred that the paralogous pairs with Ks < 0.0192 were potential new genes that likely originated after the divergence of *O. sativa* ssp. *japonica* and *O. glaberrima* from their common ancestor around 0.5 ~1 MYA. We further removed 4 new gene candidates, which have orthologs in other plant species presented in Uniprot database by reciprocal blastp approach. These four genes likely were old duplicate genes that later lost in *O. glaberrima*. Overall, we identified 28 new genes in *O. sativa* as shown in Table 1.

Origination pattern of *O. sativa* ssp. *japonica* new genes

The origination patterns of these new genes were revealed by the location, gene structure and sequences comparison between new genes and their paralogous progenitors in *O. sativa* ssp. *japonica*. A 30Kb-telomeric region containing 3 functional new genes was generated through a segmental duplication of an unmapped annotated region in *O. sativa* genome (Supplementary Figures 1, 2A-C). Four adjacent annotated genes, *LOC_Os03g24960*, *LOC_Os03g24970*, *LOC_Os03g24980*, and *LOC_Os03g24990*, are located in the middle of Chr3s within a13Kb fragment which is unique to AA genome rice species. By identifying the paralogs of these four genes, we concluded that these genes originated through segmental gene duplication followed by tandem duplication. *LOC_Os04g30860* and *LOC_Os04g30870* appeared to be the most closely related parental genes given their structure, sequence similarity and phylogenetic analysis (Supplementary Figures 3, 4). A partial segment of the region between these two genes was involved in a segmental duplication which possibly gave rise to *LOC_Os03g24960* and *LOC_Os03g24970* after the divergence of *O. sativa* and *O. punctata* (~2-5 MYA). Both *LOC_Os03g24980* and *LOC_Os03g24990* that originated after the divergence of *O. sativa* and *O. glaberrima* (~0.5-1 MYA) appeared to be chimeric. *LOC_Os03g24990* was possibly generated by DNA-level recombination of *LOC_Os03g24960* and its target flanking sequence. *LOC_Os03g24980* recruited exons of *LOC_Os03g24970* and local sequences as its intron (Supplementary Figure 2W-X).

For the remaining 23 new genes, 21 were apparently generated through the single gene DNA level recombination-mechanism gene duplication (Supplementary Figure 2). Comparing gene DNA sequences and exon-intron structure between new genes and parental genes, we observed four general patterns of DNA-based recombination and duplications for new gene origination in *O. sativa* Chr3s: (1) The new gene recruited partial parental gene sequences to

form a new chimerical gene structure (Figure 2A), e.g. *LOC_Os03g01490, LOC_Os03g02340, LOC_Os03g07270, LOC_Os03g09130, LOC_Os03g11860, LOC_Os03g15110, LOC_Os03g18650, LOC_Os03g21310, LOC_Os03g25950,* and *LOC_Os03g29140*. (2) The new gene recruited partial parental gene sequences formed an intact non-chimeric gene (Figure 2B), e.g. *LOC_Os03g02130, LOC_Os03g03050, LOC_Os03g04760, LOC_Os03g07690, LOC_Os03g15060*, and *LOC_Os03g24630*. (3) The new gene adopted the entire parental gene sequences and both genes shared the same exon-intron gene structure (Figure 2C), e.g. *LOC_Os03g07090,LOC_Os03g32526*, and *LOC_Os03g33920*. (4) The new gene recruited the entire parental gene sequences but formed a different exon-intron gene structure (Figure 2D), e.g. *LOC_Os03g12480* and *LOC_Os03g16320*.

Though DNA-based gene duplication seems to be the major mechanism generating new genes in rice, we also found two genes generated through exon duplication and shuffling. *LOC_Os03g10840* was originated from the last exon of *LOC_Os03g11130* and formed a chimeric gene by recruiting the flanking region of its insertion site (Supplementary Figure 2N). Similarly, *LOC_Os03g12580* was formed from shuffling the first exon of *LOC_Os06g01010* and its flanking sequences (Supplementary Figure 2P).

Chimeric gene formation appears to be very common in new rice genes. Among 28 *O. sativa* new genes that we observed, 14 new genes are chimerical. The chimerical CDS structure of a new gene is mostly formed by recruiting entire or partial parental gene sequences and DNA sequences from the insertion site (Figure 3A). However, we did find one new gene, *Os03g09130*, which was developed from two genes and an insertion of a DNA fragment (Figure 3B). We further examined the transcription of chimerical CDS structure using the expression data. Using RNA-seq data, we found 8 chimeric genes that contain RNA-seq reads covering all the

breakpoints and 3 chimeric genes that have RNA-seq reads covering some breakpoints. Using EST data, we identified 3 chimeric genes that have EST sequences covering all the breakpoints and 1 chimeric gene that has EST sequence covering some breakpoints. Furthermore, five chimeric genes have FL-cDNA (Supplementary table 5). In summary, the chimerical CDS structure for all 14 chimeric genes was confirmed by RNA-seq, EST and/or FL-cDNA sequence data.

Evolution pattern of *O. sativa* ssp. *japonica* new genes

We calculated ω values to gain insight into the evolution of *O. sativa* ssp. *japonica* new genes (Supplementary table 6). Since all new genes originated and evolved very recently (< 1 MYA), we observed very low number and rates of both synonymous and non-synonymous substitution (Supplementary table 6). Nineteen of the 28 paralogous pairs showed no synonymous substitution and/or non-synonymous substitution. For the remaining 9 paralogs, four of them had ω values less than neutrality, and five had ω values greater than 1 (Supplementary table 6). Furthermore, LRT tests for the sequence divergence of the majority of 32 paralogous genes did not show significant deviation from neutrality. This was likely due to the recent gene duplication which has not yet accumulated enough substitutions to give adequate statistical power.

Based on branch specific ω analysis, six new genes have branch specific $\omega < 0.5$. One new genes have branch specific ω ranging between 0.5 and 1. Nine new genes have branch specific $\omega > 1$ ranging from 1.92140~999.000. Moreover, LTR tests showed that 4 new genes (*LOC_Os03g12480, LOC_Os03g21310, LOC_Os03g24990,* and *LOC_Os03g32526*) have branch specific ω significantly smaller than 1 (Supplementary table 7).

Expression of new genes in *O. sativa* ssp. *japonica*

All 28 new *O. sativa* genes appeared to be transcribed, as evidenced by the presence of RNA-seq, EST and/or FL-cDNA sequence, and/or small RNA/MPSS sequencing signature (Table 2). Sixteen of the 28 new genes had at least two evidences of expression (Table 2). Three genes, *LOC_Os03g01014, LOC_Os03g01490,* and *LOC_Os03g07270* had high mRNA enrichment in RNA-seq data (Supplementary table 1). Among them, the expression of the two genes including *LOC_Os03g01014 and LOC_Os03g07270* were enriched in different tissues: *LOC_Os03g01014* was highly expressed in leaves. *LOC_Os03g07270* was mainly transcribed in Pre-inflorescence，pistil, seed and embryo (Supplementary table 1). Accumulation of mRNA from three genes (*LOC_Os03g01020,* and *LOC_Os03g01490*) appeared to be fairly high in vivo, as revealed by the presence of 9 and 40 independent EST sequences in GenBank, respectively (Supplementary table 2). Two genes, *LOC_Os03g01020* and *LOC_Os03g01490*, expressed substantial enrichments in MPSS sequencing signature (Supplementary table 3). Eight genes, *LOC_Os03g11860, LOC_Os03g29140, LOC_Os03g12850, LOC_Os03g25950, LOC_Os03g02340, LOC_Os03g02130, LOC_Os03g24630* and *LOC_Os03g24980*, appeared to be enriched in small RNA sequencing signatures (Supplementary table 4). Moreover, these eight genes showed transcription of small RNA signatures in different tissues and developmental stages (Supplementary table 4). To compare the general pattern of small RNA expression signatures between new genes and regular functional genes, we randomly picked up 500 functional genes and found that 82.2% of the 500 genes show small RNA expression signature, thus the small RNA signature was higher in regular functional genes than in the new genes.

Discussion

High rate of new gene origination in rice genome

Oryza sativa ssp. *japonica* Chr3s contains ~3100 annotated CDS sequences including hypothetical and TE-related genes. In our effort to systematically search for potential new genes which recently evolved in *O. sativa* ssp. *japonica*, we were able to identify 28 new genes, which account for 1 % of total genes on Chr3s. However, it is likely that we underestimated or possibly overestimated the true number of new genes in *O. sativa* ssp. *japonica*. These values may be underestimates of the true number of new genes considering two reasons. First, we filtered out all TE-related genes ("retrotransposon protein" and "transposon protein") after the unique *O. sativa* ssp. *japonica* genes were found. Second, we used the average Ks value of orthologous genes between *O. sativa* ssp. *japonica* and *O. glaberrima* as a cutoff value to define the age of the paralogous duplication event. It is likely that some new genes evolved quickly and that the substitution rate may be elevated. These criteria could possibly ignore some new genes based on their high synonymous substitution rate. Meanwhile, the number of new genes that we identified might be overestimates of the true number of new genes given two possibilities. First, although *O. sativa* ssp. *japonica* new genes do not have orthologs in *O. glaberrima*, it is possible to have orthologs present outside of Chr3s in other rice species due to chromosomal rearrangement (e.g. segmental duplication and transposition). Second, the low Ks values, which can be resulted from gene conversion and locally reduced mutation rate, may not truly reflect the age of duplications. Therefore, considering both situations, we estimated that *O. sativa* ssp. *japonica* specific new genes would account for 0.8-2% of total annotated genes in the entire rice genome. RGAP annotated a total of 56797 genes including putative, expressed, hypothetical, and TE-related genes (http://rice.plantbiology.msu.edu/riceInfo/info.shtml#Genes). Therefore we deduced that

the rice genome (a total of ~57,000 genes) might have 500-1000 new genes (0.0088~0.017/gene/million years) which evolved around 1 MYA after *O. sativa* ssp. *japonica* split from *O. glaberrima*. This new gene origination rate (per gene per million years) in rice genome was over ten-fold higher than in *Drosophila*, which was estimated at 5 to 11 genes per million years (0.0004~0.00092/gene/million years) for the *D. melanogaster* subgroup genomes (a total of 12,000 genes) (Zhou et al. 2008). A caveat in this estimate was our assumption that the new gene distributions on the sequenced Chr3s were representative of the whole rice genome. However, this pilot analysis already revealed the high rate of new gene origination in the recent evolution of these species. One major force was likely responsible for the rapid occurrence of new genes in rice genome. Though genus *Oryza* stands as a small group in the plant kingdom containing only 23 species, the diversity and ecological adaptability of rice, which is found in a wide range of habitats from forest, savanna, and mountainsides to river and lakes, is remarkable and could drive the rapid occurrence of new genes in rice genome (Ge et al. 1999; Vaughan et al. 2003).

New gene originated as chimera in rice genome

Chimeric genes represent a class of genes that originated from multiple parental sources in coding and/or noncoding (regulatory site) sequences. Due to their unique origination, chimeric genes are unlikely to retain their parental characteristics and thus evolve novel functions. By surveying previous new genes detected in other organisms, it can be concluded that chimeric new genes account for a high percentage of total new genes identified in a variety of organisms ranging from mammals (Parker et al. 2009; Paulding et al. 2003; Sayah et al. 2004), to flies (Jones et al. 2005; Long and Langley 1993; Nozawa et al. 2005) and plants (Fan et al. 2008b;

Long et al. 1996; Wang et al. 2006). A recent investigation systematically searched through new genes using the *Drosophila* genome comparisons and found 30% of the new genes in the *D. melanogaster* species complex recruited various genomic sequences and formed chimeric gene structures. These findings suggest structure innovation is important to the generation of new genes (Zhou et al. 2008). This is similar to what was reported previously in the genomic analysis of *O. sativa* ssp. *indica* (Wang et al. 2006). A previous study reported that cultivated rice (*O. sativa* ssp. *indica*) genome encodes 898 functional retroposed genes, of which 380 were predicted to have chimerical protein sequence structures (Wang et al. 2006). Because the most recent divergent time can better record the recent evolutionary events, our observation provided additional solid evidence for the high rate of new gene origination. Consistent with previous finding, we annotated a total of 28 new genes on *O. sativa* ssp. *japonica* Chr3s, 14 (50%) of which appeared to be chimeric genes generated by segmental duplication and DNA-level recombination. Our current study revealed a high rate of chimeric gene origination as: 14 x 20 =280 chimeric genes / million years / genome. The higher rates of chimeric gene formation and the generation of a large number of functional genes in rice again demonstrated the broad diversification and adaptation of the grass species. Both our previous and current studies all demonstrated that rice genomes displayed an accelerated gene origination rate and generated a high number of chimeric gene structures that held potential to evolve novel functions (Fan et al. 2008b; Wang et al. 2006). However, these findings are in contrast to the recently reported lower gene origination rate, which may result from extremely conservative genome annotation (Sakai et al. 2011). Conservative annotation is an approach that has been widely used in functional genomics and molecular functional analysis but may not fit the need for evolutionary genomic

study. In practice, new evolutionary changes, including new genes, are seriously underestimated by this approach (Zhang et al. 2012).

Previous studies in *Drosophila* have demonstrated that repetitive elements could facilitate recombination to generate high occurrences of chimeric genes (Yang et al. 2008). In rice, the abundance of Pack-MULEs could capture fragment(s) of genomic DNA sequence while also rearranging and fusing with target sequence to generate a large amount of new reading frame and chimerical transcripts (Jiang et al. 2004). Therefore, mechanisms such as these could be responsible for the chimeric gene formation in rice genome.

Acknowledgements

C.F. was supported by start-up fund from Wayne State University. M.L. and R.A.W were supported by National Sciences Foundation Grant MCB1026200. We are also grateful for the grid computing service from Computing & Information Technology of Wayne State University. We thank three anonymous reviewers for valuable comments and suggestions.

References

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Ammiraju JS, 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. doi: 10.1105/tpc.108.063727

Bachtrog D, Charlesworth B 2003. On the genomic location of the exuperantia1 gene in Drosophila miranda: the limits of in situ hybridization experiments. Genetics 164: 1237-1240.

Betrán E, Thornton K, Long M 2002. Retroposed new genes out of the X in Drosophila. Genome Res 12: 1854-1859.

Cardoso-Moreira M, Long M 2012. The origin and evolution of new genes. Methods Mol Biol 856: 161-186. doi: 10.1007/978-1-61779-585-5_7

Charrier C, et al. 2012. Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation. Cell 149: 10.1016/j.cell.2012.1003.1034. doi: 10.1016/j.cell.2012.03.034

Chen S, Zhang YE, Long M 2010. New genes in Drosophila quickly become essential. Science 330: 1682-1685. doi: 10.1126/science.1196380

Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87. doi: 10.1038/nature04072

Clark AG, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-218. doi: 10.1038/nature06341

Davidson RM, et al. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J 71: 492-502. doi: 10.1111/j.1365-313X.2012.05005.x

Des Marais D, Rausher M 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454: 762-765.

Ding Y, et al. 2010. A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. PLoS Genet 6: e1001255. doi: 10.1371/journal.pgen.1001255

Fan C, Emerson J, Long M 2008a. *The Origin of New Genes*. Sunderland, Massachusett 01375: Sinauer Associates, Inc.

Fan C, Vibranovski M, Chen Y, Long M 2007. A Microarray Based Genomic Hybridization Method for Identification of New Genes in Plants: Case Analyses of Arabidopsis and Oryza. Journal of Integrative Plant Biology 49: 915-926.

Fan C, et al. 2008b. The Subtelomere of Oryza sativa Chromosome 3 Short Arm as a Hot Bed of New Gene Origination in Rice. Molecular Plant 1: 839-850.

Gan X, et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477: 419-423. doi: 10.1038/nature10414

Ge S, Sang T, Lu B, Hong D 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc Natl Acad Sci U S A 96: 14400-14405.

Green RE, et al. 2010. A draft sequence of the Neandertal genome. Science 328: 710-722. doi: 10.1126/science.1188021

He G, et al. 2010. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. Plant Cell 22: 17-33. doi: 10.1105/tpc.109.072041

Heinen T, Staubach F, Häming D, Tautz D 2009. Emergence of a new gene from an intergenic region. Curr Biol 19: 1527-1531.

Hu TT, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet 43: 476-481. doi: 10.1038/ng.807

Jensen JD, Bachtrog D 2010. Characterizing recurrent positive selection at fast-evolving genes in Drosophila miranda and Drosophila pseudoobscura. Genome Biol Evol 2: 371-378. doi: 10.1093/gbe/evq028

Jiang N, et al. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569-573.

Jones C, Begun D 2005. Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci U S A 102: 11373-11378.

Jones C, Custer A, Begun D 2005. Origin and evolution of a chimeric fusion gene in Drosophila subobscura, D. madeirensis and D. guanche. Genetics 170: 207-219.

Jun J, Ryvkin P, Hemphill E, Nelson C 2009. Duplication mechanism and disruptions in flanking regions determine the fate of Mammalian gene duplicates. J Comput Biol 16: 1253-1266. doi: 10.1089/cmb.2009.0074

Kaessmann H, Vinckenbosch N, Long M 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10: 19-31.

Katoh K, Kuma K, Toh H, Miyata T 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518. doi: 10.1093/nar/gki198

Kent WJ 2002. BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664. doi: 10.1101/gr.229202. Article published online before March 2002

Kim EB, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479: 223-227. doi: 10.1038/nature10533

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948. doi: 10.1093/bioinformatics/btm404

Liti G, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458: 337-341. doi: 10.1038/nature07743

Locke DP, et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469: 529-533. doi: 10.1038/nature09687

Long M, Betrán E, Thornton K, Wang W 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet 4: 865-875.

Long M, de Souza S, Rosenberg C, Gilbert W 1996. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. Proc Natl Acad Sci U S A 93: 7727-7731.

Long M, Langley C 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260: 91-95.

Marques AC, et al. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 3: e357. doi: 10.1371/journal.pbio.0030357

Marques-Bonet T, Eichler EE 2009. The evolution of human segmental duplications and the core duplicon hypothesis. Cold Spring Harb Symp Quant Biol 74: 355-362. doi: 10.1101/sqb.2009.74.011

Marques-Bonet T, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457: 877-881. doi: 10.1038/nature07744

Nakano M, et al. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 34: D731-735. doi: 10.1093/nar/gkj077

Nobuta K, et al. 2007. An expression atlas of rice mRNAs and small RNAs. Nat Biotechnol 25: 473-477. doi: 10.1038/nbt1291

Nozawa M, Aotsuka T, Tamura K 2005. A novel chimeric gene, siren, with retroposed promoter sequence in the Drosophila bipectinata complex. Genetics 171: 1719-1727.

Parker H, et al. 2009. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325: 995-998.

Paulding C, Ruvolo M, Haber D 2003. The Tre2 (USP6) oncogene is a hominoid-specific gene. Proc Natl Acad Sci U S A 100: 2507-2511.

Pontius JU, Wagner L, Schuler GD 2003. UniGene: a unified view of the transcriptome. In: *The NCBI Handbook*, The NCBI Handbook. Bethesda, MD, USA: National Center for Biotechnology Information.

Potrzebowski L, Vinckenbosch N, Kaessmann H 2010. The emergence of new genes on the young therian X. Trends Genet 26: 1-4. doi: 10.1016/j.tig.2009.11.001

Ranz JM, Parsch J 2012. Newly evolved genes: Moving from comparative genomics to functional studies in model systems: How important is genetic novelty for species adaptation and diversification? Bioessays. doi: 10.1002/bies.201100177

Sakai H, et al. 2011. Retrogenes in rice (Oryza sativa L. ssp. japonica) exhibit correlated expression with their source genes. Genome Biol Evol 3: 1357-1368. doi: 10.1093/gbe/evr111

Sayah D, Sokolskaja E, Berthoux L, Luban J 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature 430: 569-573.

Scally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-175. doi: 10.1038/nature10842

Stein LD, et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biol 1: E45. doi: 10.1371/journal.pbio.0000045

Swofford D 2002. *PAUP, Phylogenetic Analysis Uisng Parsimony, Version 4.0b10*. Sunderland, Massachusett 01375: Sinauer Associates, Inc.

Tang L, et al. 2010. Phylogeny and biogeography of the rice tribe (Oryzeae): evidence from combined analysis of 20 chloroplast fragments. Mol Phylogenet Evol 54: 266-277. doi: 10.1016/j.ympev.2009.08.007

Trapnell C, Pachter L, Salzberg SL 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111. doi: 10.1093/bioinformatics/btp120

Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515. doi: 10.1038/nbt.1621

Vaughan D, Morishima H, Kadowaki K 2003. Diversity in the Oryza genus. Curr Opin Plant Biol 6: 139-146.

Wang J, Long M, Vibranovski MD 2012. Retrogenes moved out of the z chromosome in the silkworm. J Mol Evol 74: 113-126. doi: 10.1007/s00239-012-9499-y

Wang W, Yu H, Long M 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. Nat Genet 36: 523-527.

Wang W, et al. 2000. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in Drosophila melanogaster. Mol Biol Evol 17: 1294-1301.

Wang W, et al. 2006. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791-1802. doi: 10.1105/tpc.106.041905

Yang S, et al. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. PLoS Genet 4: e3. doi: 10.1371/journal.pgen.0040003

Yang Z 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591. doi: 10.1093/molbev/msm088

Yang Z, Nielsen R 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32-43.

Yeh SD, et al. 2012. Functional evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition. Proc Natl Acad Sci U S A 109: 2043-2048. doi: 10.1073/pnas.1121327109

Yu J, et al. 2005. The Genomes of Oryza sativa: a history of duplications. PLoS Biol 3: e38. doi: 10.1371/journal.pbio.0030038

Zemach A, McDaniel IE, Silva P, Zilberman D 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328: 916-919. doi: 10.1126/science.1186366

Zhang C, Wang J, Long M, Fan C 2013. gKaKs: The pipeline for genome level Ka/Ks calculation. Bioinformatics. doi: 10.1093/bioinformatics/btt009

Zhang J, Zhang Y, Rosenberg H 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet 30: 411-415.

Zhang Q, et al. 2008. Rice 2020: a call for an international coordinated effort in rice functional genomics. Mol Plant 1: 715-719. doi: 10.1093/mp/ssn043

Zhang Y, Wu Y, Liu Y, Han B 2005. Computational identification of 69 retroposons in Arabidopsis. Plant Physiol 138: 935-948. doi: 10.1104/pp.105.060244

Zhang YE, Landback P, Vibranovski M, Long M 2012. New genes expressed in human brains: Implications for annotating evolving genomes. Bioessays 34: 982-991. doi: 10.1002/bies.201200008

Zhang YE, Landback P, Vibranovski MD, Long M 2011. Accelerated recruitment of new brain development genes into the human genome. PLoS Biol 9: e1001179. doi: 10.1371/journal.pbio.1001179

Zhou Q, et al. 2008. On the origin of new genes in Drosophila. Genome Res 18: 1446-1455.

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

Zhu Z, Zhang Y, Long M 2009. Extensive structural renovation of retrogenes in the evolution of the Populus genome. Plant Physiol 151: 1943-1951. doi: 10.1104/pp.109.142984

Figure legend

Figure 1. Phylogeny of six rice species showing the species divergence time and an illustration of new gene origination in *O. sativa*. Gene 'A', 'C', and 'D' are orthologous in six species. Gene 'B' is a new gene in *O. sativa* and/or Asian rice species. 'AA' stands for the *Oryza* 'A' genome type. 'BB' stands for *Oryza* 'B' genome type.

Figure 2. Illustration and example of four general patterns of new gene origination in *O. sativa* genome. The genes above are new genes and the genes below are parental genes. A: new gene formed chimeric gene structure from partial parental gene sequence. B: new gene formed intact and non-chimeric structure from partial parental gene. C: new gene formed from entire parental gene and shared same exon-intron gene structure. D: new gene formed from entire parental gene but with different exon-intron gene structure. Exon: filled box; Intron: solid line; Homologous region: dash line. The start and stop codons are marked for each gene.

Figure 3. Illustration and example of chimeric new gene: A: new gene formed from one parental gene. B: new gene formed from two parental genes. Exon: filled box; Intron: solid line; Homologous region: dash line. The start and stop codons are marked for each gene.

Table 1. The new genes, paralogs and creation mechanisms

	New gene	Annotation	Paralogs	Possible formation mechanisms
1	Os03g01008	Expressed protein	ChrSy.fgenesh.mRNA.80	Segmental duplication
2	Os03g01014	expressed protein	ChrSy.fgenesh.mRNA.82	Segmental duplication
3	Os03g01020	pectinesterase inhibitor domain containing protein	ChrSy.fgenesh.mRNA.85	Segmental duplication
4	Os03g01490	expressed protein	Os03g01420	Tandem duplication, chimera
5	Os03g02130	hypothetical protein	Os01g63170	Gene duplication
6	Os03g02340	expressed protein	Os05g05090	Gene duplication, chimera
7	Os03g03050	expressed protein	Os07g20240	Gene duplication
8	Os03g04760	expressed protein	Os05g11820	Gene duplication
9	Os03g07090	expressed protein	Os11g08990	Gene duplication
10	Os03g07270	glycine-rich cell wall protein	Os01g57250	Gene duplication, chimera
11	Os03g07690	expressed protein	Os01g22910	Gene duplication
12	Os03g09130	expressed protein	Os03g18760/ Os11g07660	Gene duplication, chimera
13	Os03g10840	expressed protein	Os03g11130	Exon shuffling, chimera
14	Os03g11860	expressed protein	Os01g09060	Gene duplication, chimera
15	Os03g12480	expressed protein	Os06g42410	Gene duplication
16	Os03g12580	expressed protein	Os06g01010	Exon shuffling, chimera
17	Os03g15060	expressed protein	Os01g19250	Gene duplication, chimera
18	Os03g15110	expressed protein	Os03g46230	Gene duplication, chimera
19	Os03g16320	expressed protein	Os04g50840	Gene duplication
20	Os03g18650	hypothetical protein	Os05g38540	Gene duplication
21	Os03g21310	ulp1 protease family	Os08g33280	Gene duplication, chimera
22	Os03g24630	hypothetical protein	Os05g36060	Gene duplication
23	Os03g24980	SWIM zinc finger family protein	Os03g24970	Tandem geneduplication,chimera
24	Os03g24990	ulp1 protease family	Os03g24960	Tandem gene duplication, chimera

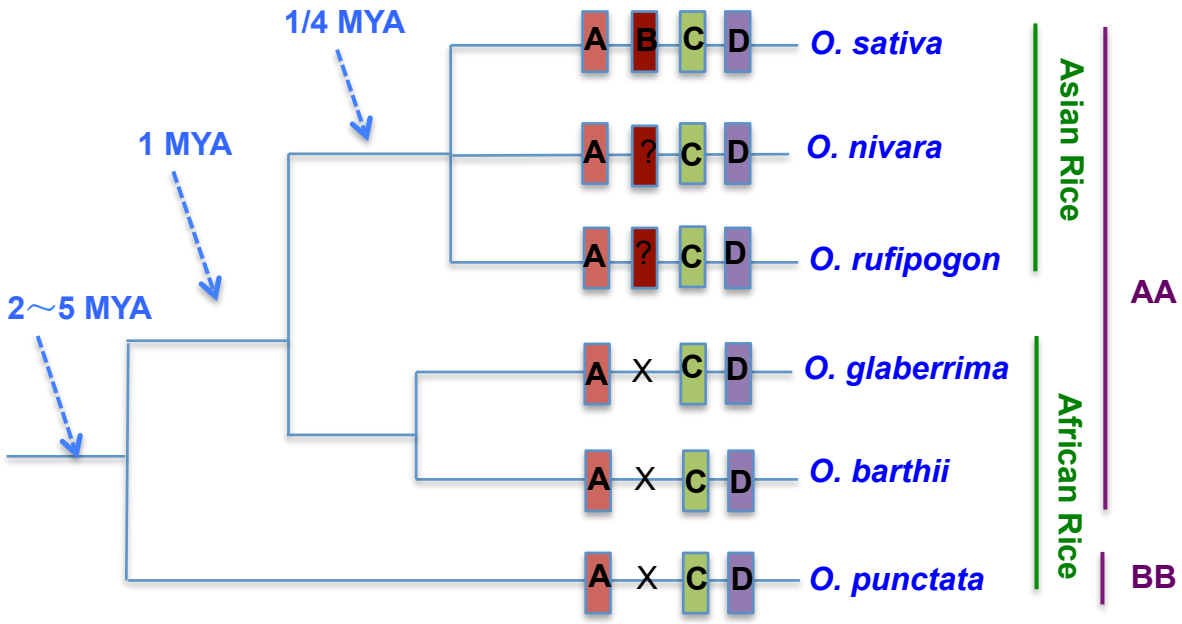
25	Os03g25950	expressed protein	Os12g32810	Gene duplication, chimera
26	Os03g29140	expressed protein	Os01g09060	Gene duplication, chimera
27	Os03g32526	tRNA-splicing endonuclease positive effector-related	Os06g20500	Gene duplication
28	Os03g33920	conserved hypothetical protein	Os06g36630	Gene duplication

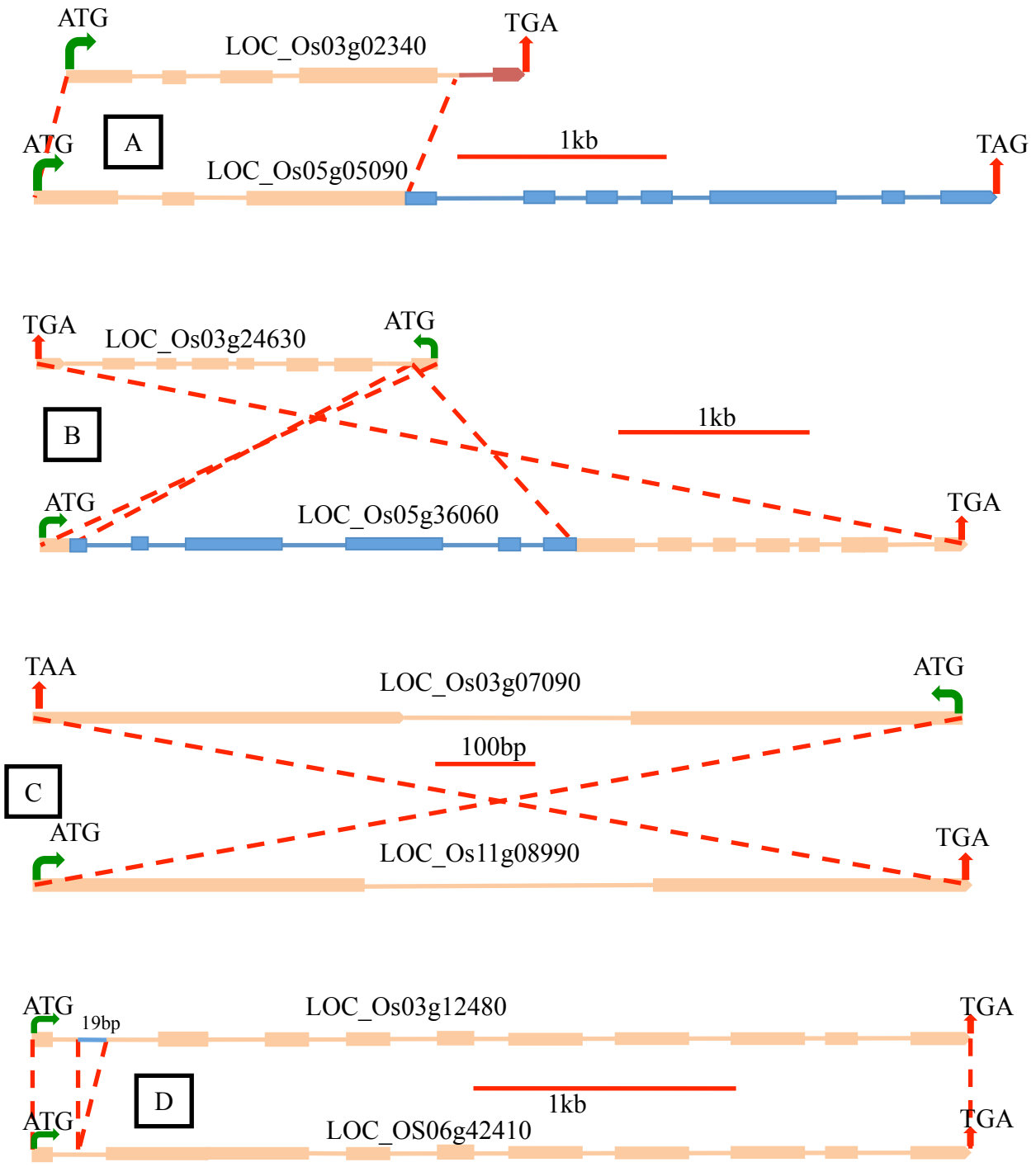
Table 2. Expression of new genes in *O. sativa*

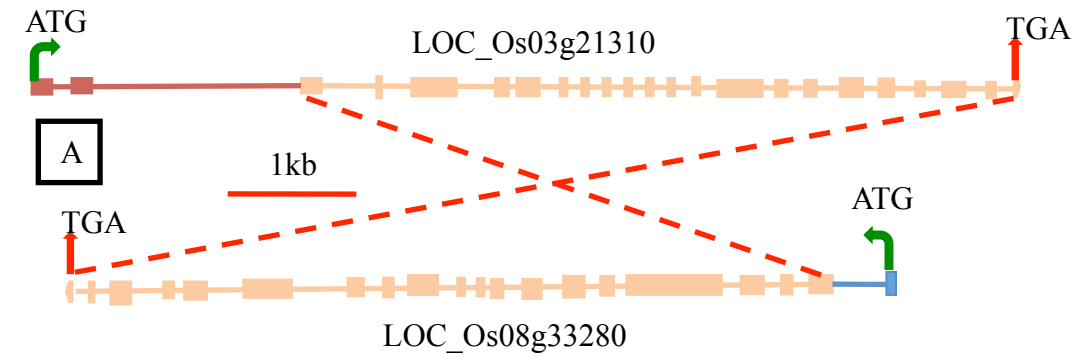
Locus	RNA seq data	EST	MPSS	Small RNA
Os03g01008	+	-	-	-
Os03g01014	+	-	-	-
Os03g01020	+	+	+	+
Os03g01490	+	+	+	+
Os03g02130	-	-	-	+
Os03g02340	+	-	-	+
Os03g03050	+	+	-	+
Os03g04760	+	-	-	+
Os03g07090	-	-	-	+
Os03g07270	+	+	-	+
Os03g07690	+	-	-	+
Os03g09130	-	-	-	+
Os03g10840	+	-	-	+
Os03g11860	-	-	+	+
Os03g12480	-	-	-	+
Os03g12580	-	-	-	+
Os03g15060	+	-	-	+
Os03g15110	+	+	-	+
Os03g16320	+	-	-	+
Os03g18650	-	-	-	+
Os03g21310	+	+	+	+

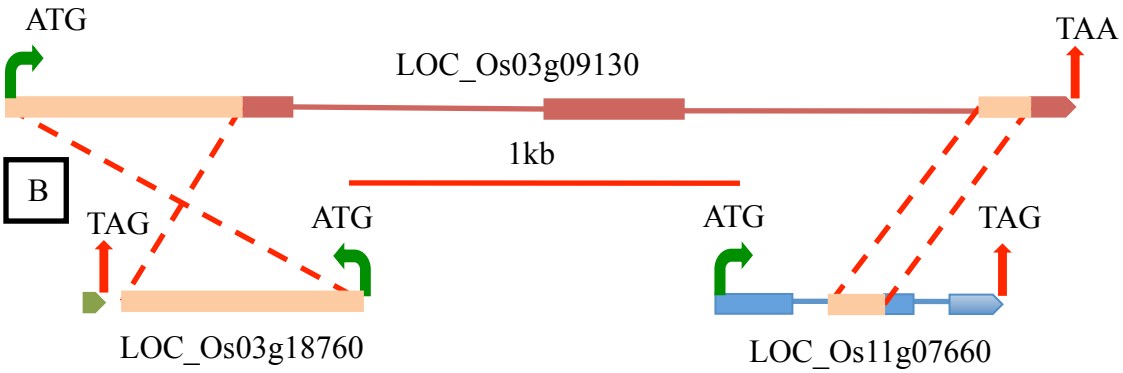
Os03g24630	-	-	-	+
Os03g24980	-	-	-	+
Os03g24990	-	-	-	+
Os03g25950	+	-	-	+
Os03g29140	+	-	-	+
Os03g32526	+	+	-	+
Os03g33920	-	-	-	+

Note: +: present; -: absent.


1/4 MYA
1 MYA
2～5 MYA
A B C D
O. sativa
A ? C D
O. nivara
A ? C D
O. rufipogon
A X C D
O. glaberrima
A X C D
O. barthii
A X C D
O. punctata
Asian Rice
African Rice
AA
BB


ATG
LOC_Os03g02340
TGA
ATG
A
1kb
TAG
LOC_Os05g05090
TGA
LOC_Os03g24630
ATG
B
1kb
ATG
LOC_Os05g36060
TGA
TAA
LOC_Os03g07090
ATG
100bp
C
ATG
TGA
LOC_Os11g08990
ATG
19bp
LOC_Os03g12480
TGA
D
1kb
ATG
LOC_OS06g42410
TGA


ATG
LOC_Os03g21310
TGA
A
1kb
TGA
ATG
LOC_Os08g33280


ATG
LOC_Os03g09130
TAA
1kb
B
TAG
ATG
ATG
TAG
LOC_Os03g18760
LOC_Os11g07660