

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. Marowsky², Manyuan Long¹, Rod A. Wing^{3,*}, and Chuanzhu Fan^{2,*}

¹Department of Ecology and Evolution, University of Chicago

²Department of Biological Sciences, Wayne State University

³Arizona Genomics Institute, School of Plant Sciences, University of Arizona

*Corresponding authors: E-mail: cfan@wayne.edu; rwing@ag.arizona.edu.

†These authors contributed equally to this work.

Accepted: May 1, 2013

Abstract

In an effort to identify newly evolved genes in rice, we searched the genomes of Asian-cultivated rice *Oryza sativa* ssp. *japonica* and its wild progenitors, looking for lineage-specific genes. Using genome pairwise comparison of approximately 20-Mb DNA sequences from the chromosome 3 short arm (Chr3s) in six rice species, *O. 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 1 Myr. We identified 28 functional *O. sativa* genes, which likely originated after *O. sativa* diverged from *O. glaberrima*. These genes account for around 1% (28/3,176) 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 that 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 K_a/K_s analysis and the presence of RNA-seq, cDNA, expressed sequence tag, massively parallel signature sequencing, 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

Key words: 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 (Long and Langley 1993; Zhang et al. 2002; Long et al. 2003; Jones et al. 2005; Des Marais and Rausher 2008; Fan, Emerson, et al. 2008; Zhou et al. 2008; Heinen et al. 2009; Parker et al. 2009; Chen et al. 2010; Ding et al. 2010; Potrzebowski et al. 2010; Charrier et al. 2012; Yeh et al. 2012). 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 (Long et al. 2003; Kaessmann et al. 2009; Cardoso-Moreira and Long 2012; 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 (Wang et al. 2000; Bachtrog and Charlesworth 2003; Jones and Begun 2005). 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). Because of

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 (Zhang et al. 2005; Wang et al. 2006; Fan et al. 2008; Zhu et al. 2009; Sakai et al. 2011).

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 nonmodel 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; Wang et al. 2004; Marques et al. 2005). This was not the case in plants, due to technical challenges, for example, 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 (Stein et al. 2003; Chimpanzee Sequencing and Analysis Consortium 2005; Marques et al. 2005; Yu et al. 2005; Clark et al. 2007; Jun et al. 2009; Liti et al. 2009; Marques-Bonet and Eichler 2009; Marques-Bonet et al. 2009; Green et al. 2010; Jensen and Bachtrog 2010; Gan et al. 2011; Hu et al. 2011; Kim et al. 2011; Locke et al. 2011; Zhang et al. 2011; Scally et al. 2012). 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 chromosome 3 short arm (Chr3s) using the bacterial artificial chromosome (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 that diverged over a relatively short time period approximately 15–20 Ma with broad diversification and largely solved phylogenetics (Ge et al. 1999; Zhu and Ge 2005; Ammiraju et al. 2008; Tang et al. 2010). *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 Ma (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 Ma (fig. 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 <http://www.iplantcollaborative.org/>. We performed genome pairwise comparisons between *O. sativa* ssp. *japonica* Chr3s coding sequences (CDSs) and other five species Chr3s genome sequences. The annotation and CDSs of *O. sativa* ssp. *japonica* were downloaded from Michigan State University (MSU) Rice Genome Annotation Project (RGAP, MSU V7) (<http://rice.plantbiology.msu.edu/downloads.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 30-kb DNA fragment

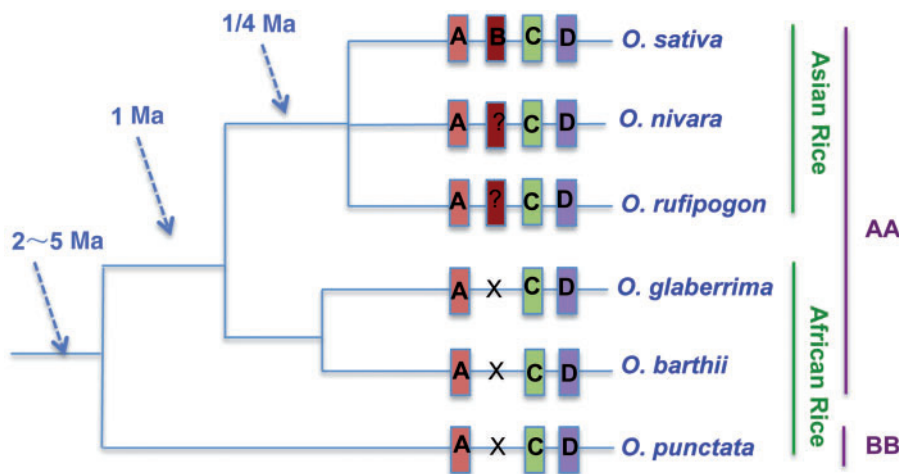


Fig. 1.—Phylogeny of six rice species showing the species divergence time and an illustration of new gene origination in *Oryza sativa*. Genes “A,” “C,” and “D” are orthologous in six species. Gene “B” is a new gene in *O. sativa* and/or Asian rice species. “AA” stands for the *Oryza* “A” genome type.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

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* (fig. 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 Ma. We further estimated the average rates of synonymous substitution (K_s) 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 CDSs of the candidate genes against all the CDSs of *O. sativa ssp. japonica* with the match length of the paralogous gene pair more than

100 bp and mismatch length/(mismatch length + match length) less than 0.1. We picked up only the paralogous gene pairs with K_s less than 0.0192, which is the average K_s of the orthologous gene pairs between *O. sativa ssp. japonica* and *O. glaberrima* corresponding to 1 Myr 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 whether these *O. sativa* lineage-specific new genes were ancient duplicate genes that lost in African *Oryza* species, we applied reciprocal BLASTP searches to identify whether 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.gemene.org/Multi/blatview>). 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 (K_a/K_s , denoted as “ ω ”) using maximum likelihood algorithm (codeml) implemented in the PAML package (Yang 2007). The significance of ω that deviated from neutrality ($\omega = 1$) 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 CDSs of the gene family were aligned using ClustalW (Larkin et al. 2007). The bootstrap analysis with 1,000 replicates was used to assess the robustness of the branches.

To address whether $\omega < 1$ 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. punctata* from <http://www.iplantcollaborative.org>. We identified the orthologous sequences of parental genes from the three outgroup species using ortholog search approach described earlier. 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), expressed sequence tag (EST; Pontius et al. 2003), RNA sequencing transcriptome data (RNA-seq) (He et al. 2010; Zemach et al. 2010; Davidson et al. 2012), 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/kilobase of transcript/million fragments mapped (FPKM) across 11 libraries including leaves—20 days, postemergence 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 S1, Supplementary Material online, 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 National Center for Biotechnology Information (NCBI) EST library collection of *O. 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 S2, Supplementary Material online). We used BLAT to identify the genes corresponding to the ESTs with Basic Local Alignment Search Tool (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, $\leq 1e-20$ 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 26,577 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 transcripts/million in 70 tissues (supplementary table S3, Supplementary Material online). Small RNA expression data were reported in the sum for the abundance of all the signatures in transcripts/quarter million in six tissues (stem, germinating seedlings, immature panicles, germinating seedling infected with *Magnaporthe grisea*, seedlings treated with Abscisic acid (ABA), and seedlings control for ABA treatment) (supplementary table S4, Supplementary Material online). 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 whether 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 polymerase chain reaction 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 32-bp 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 K_s between *O. sativa* ssp. *japonica*-specific gene and its paralog. On the basis of the average $K_s = 0.0192$ of 1,797 Chr3s orthologous genes between *O. sativa* and *O. glaberrima*, we inferred that the paralogous pairs with K_s less than 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 Ma. We further removed four 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 listed 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 30-kb telomeric region containing three functional new genes was generated through a segmental duplication of an unmapped annotated region in *O. sativa* genome (supplementary figs. S1 and S2A–C,

Supplementary Material online). Four adjacent annotated genes, *LOC_Os03g24960*, *LOC_Os03g24970*, *LOC_Os03g24980*, and *LOC_Os03g24990*, are located in the middle of Chr3s within a 13-kb 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 figs. S3 and S4, Supplementary Material online). 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 Ma). Both *LOC_Os03g24980* and *LOC_Os03g24990* that originated after the divergence of *O. sativa* and *O. glaberrima* (~0.5–1 Ma) 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 fig. S2W–X, Supplementary Material online).

For the remaining 23 new genes, 21 were apparently generated through the single-gene DNA level recombination-mechanism gene duplication (supplementary fig. S2, Supplementary Material online). 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 (fig. 2A), for example, *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 nonchimeric gene (fig. 2B), for example, *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 (fig. 2C), for example, *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 (fig. 2D), for example, *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 fig. S2N, Supplementary Material online). Similarly,

Table 1

The New Genes, Paralogs, and Creation Mechanisms

	New Gene	Annotation	Paralogs	Possible Formation Mechanisms
1	Os03g01008	Expressed protein	ChrSy.fgenes.h.mRNA.80	Segmental duplication
2	Os03g01014	Expressed protein	ChrSy.fgenes.h.mRNA.82	Segmental duplication
3	Os03g01020	Pectinesterase inhibitor domain containing protein	ChrSy.fgenes.h.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 gene duplication, 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

LOC_Os03g12580 was formed from shuffling the first exon of *LOC_Os06g01010* and its flanking sequences (supplementary fig. S2P, Supplementary Material online).

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 (fig. 3A). However, we did find one new gene, *LOC_Os03g09130*, which was developed from two genes and an insertion of a DNA fragment (fig. 3B). We further examined the transcription of chimerical CDS structure using the expression data. Using RNA-seq data, we found eight chimeric genes that contain RNA-seq reads covering all the breakpoints and three chimeric genes that have RNA-seq reads covering some breakpoints. Using EST data, we identified three chimeric genes that have EST sequences covering all the breakpoints and one chimeric gene that has EST sequence covering some breakpoints. Furthermore, five chimeric genes have FL-cDNA (supplementary table S5, Supplementary

Material online). 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 S6, Supplementary Material online). Because all new genes originated and evolved very recently (<1 Ma), we observed very low number and rates of both synonymous and nonsynonymous substitution (supplementary table S6, Supplementary Material online). Nineteen of the 28 paralogous pairs showed no synonymous substitution and/or nonsynonymous substitution. For the remaining nine paralogs, four of them had ω values less than neutrality, and five had ω values greater than 1 (supplementary table S6, Supplementary Material online). Furthermore, LRTs 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

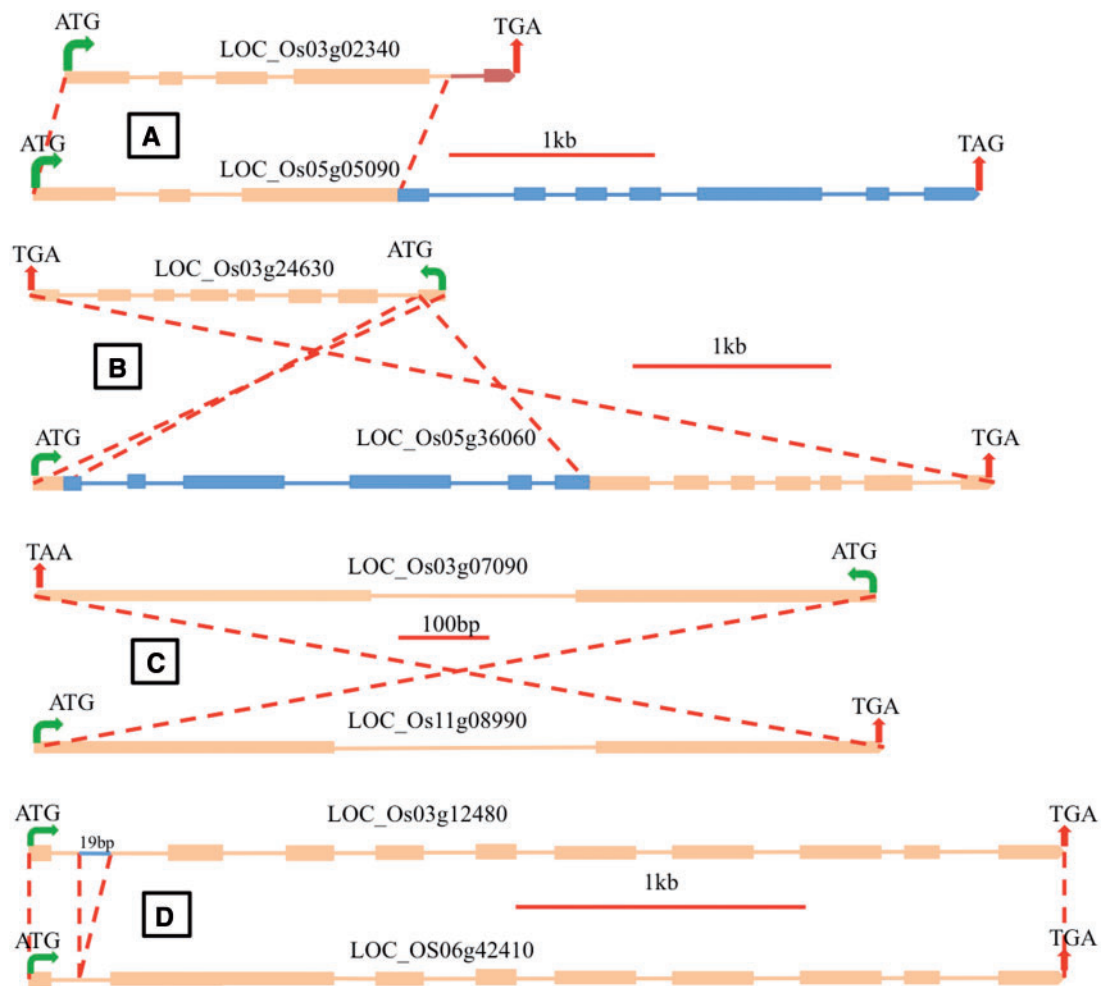


FIG. 2.—Illustration and example of four general patterns of new gene origination in *Oryza sativa* genome. The genes above are new genes and the genes below are parental genes. (A) New gene formed chimeric gene structure from partial parental gene sequence. (B) New gene formed intact and nonchimeric 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.

gene duplication that has not yet accumulated enough substitutions to give adequate statistical power.

Based on branch-specific ω analysis, six new genes have branch-specific ω less than 0.5. One new gene has branch-specific ω ranging between 0.5 and 1. Nine new genes have branch-specific ω more than 1 ranging from 1.92140 to 999.000. Moreover, LRTs showed that four new genes (*LOC_Os03g12480*, *LOC_Os03g21310*, *LOC_Os03g24990*, and *LOC_Os03g32526*) have branch-specific ω significantly smaller than 1 (supplementary table S7, Supplementary Material online).

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 S1, Supplementary Material online). Among them, the expression of the two genes including *LOC_Os03g01014* and *LOC_Os03g07270* was enriched in different tissues: *LOC_Os03g01014* was highly expressed in leaves. *LOC_Os03g07270* was mainly transcribed in preinflorescence, pistil, seed, and embryo (supplementary table S1, Supplementary Material online). Accumulation of mRNA from two 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 S2, Supplementary Material

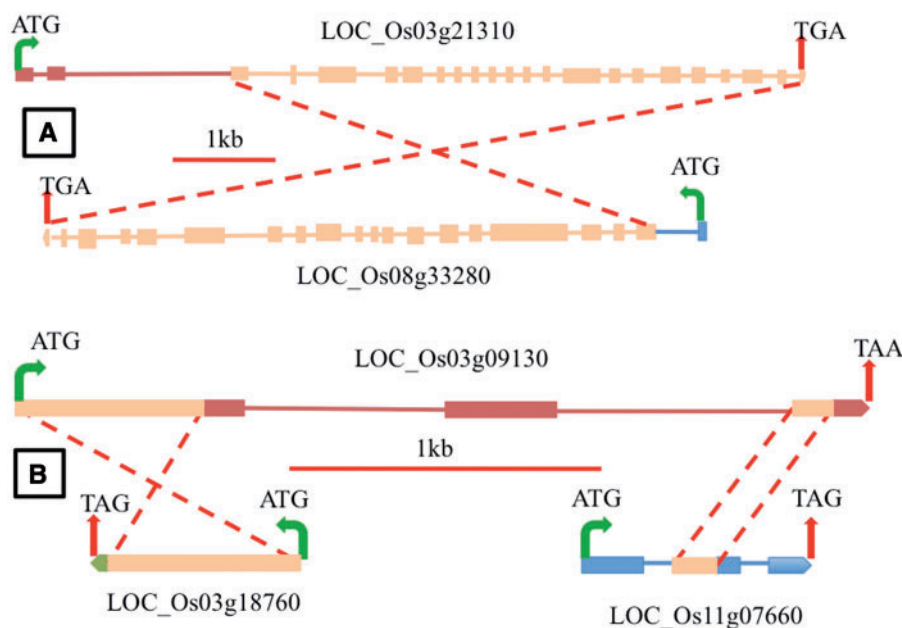


Fig. 3.—Illustration and example of chimeric new gene. (A) New gene formed from one parental gene. (B) New gene formed from two parental genes. Exon, filled box; intron, solid line; homologous region, dash line. The start and stop codons are marked for each gene.

online). Two genes, *LOC_Os03g01020* and *LOC_Os03g01490*, expressed substantial enrichments in MPSS (supplementary table S3, Supplementary Material online). 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 S4, Supplementary Material online). Moreover, these eight genes showed transcription of small RNA signatures in different tissues and developmental stages (supplementary table S4, Supplementary Material online). 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 approximately 3,100 annotated CDSs including hypothetical and transposable element (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 K_s 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 K_s 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 56,797 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 500–1,000 new genes (0.0088–0.017/gene/Myr), which evolved around 1 Ma after *O. sativa* ssp. *japonica* split from *O. glaberrima*. This new gene origination rate (per gene per Myr) in rice

Table 2Expression of New Genes in *Oryza 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.

genome was over 10-fold higher than in *Drosophila*, which was estimated at 5–11 genes/Myr (0.0004–0.00092/gene/Myr) 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. Because of 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 (Paulding et al. 2003; Sayah et al. 2004; Parker et al. 2009), to flies (Long and Langley 1993; Jones et al. 2005; Nozawa et al. 2005) and plants (Long et al. 1996; Wang et al. 2006; Fan et al. 2008). 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: $14 \times 20 = 280$ chimeric genes/Myr/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 (Wang et al. 2006; Fan et al. 2008). 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.

Supplementary Material

Supplementary tables S1–S7 and figures S1–S4 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgments

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

Literature Cited

- Achtul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
- Amiraju 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.
- 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.
- Charrier C, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* 149:923–935.
- Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. *Science* 330:1682–1685.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
- Clark AG, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218.
- Davidson RM, et al. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.* 71: 492–502.
- 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.
- Fan C, Emerson J, Long M. 2008. The origin of new genes. Sunderland (MA): 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*. *J Integr Plant Biol.* 49: 915–926.
- Fan C, et al. 2008. The subtelomere of *Oryza sativa* chromosome 3 short arm as a hot bed of new gene origination in rice. *Mol Plant.* 1: 839–850.
- Gan X, et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477:419–423.
- 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.
- He G, et al. 2010. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell* 22:17–33.
- 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.
- 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.
- 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.
- 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.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12: 656–664.
- Kim EB, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479:223–227.
- Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
- Liti G, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* 458:337–341.
- Locke DP, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* 469:529–533.
- 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.
- 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.
- Marques-Bonet T, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457: 877–881.
- Nakano M, et al. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res.* 34:D731–D735.
- Nobuta K, et al. 2007. An expression atlas of rice mRNAs and small RNAs. *Nat Biotechnol.* 25:473–477.
- 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. McEntyre J, Ostell J, editors. The NCBI handbook. Bethesda (MD): 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.
- 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* 34:477–483.
- 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.
- 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.
- Stein LD, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* 1:E45.
- Swofford D. 2002. PAUP, phylogenetic analysis using parsimony, version 4.0b10. Sunderland (MA): Sinauer Associates, Inc.
- Tang L, et al. 2010. Phylogeny and biogeography of the rice tribe (Oryzaceae): evidence from combined analysis of 20 chloroplast fragments. *Mol Phylogenet Evol.* 54:266–277.
- Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* 25:1105–1111.
- 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.
- 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.
- 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.
- Yang S, et al. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS Genet.* 4:e3.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
- 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.
- Yu J, et al. 2005. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* 3:e38.
- Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328: 916–919.
- Zhang C, Wang J, Long M, Fan C. 2013. gKaKs: the pipeline for genome level Ka/Ks calculation. *Bioinformatics* 29:645–646.
- 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.
- Zhang Y, Wu Y, Liu Y, Han B. 2005. Computational identification of 69 retrotransposons in *Arabidopsis*. *Plant Physiol.* 138:935–948.
- Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* 9:e1001179.
- Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays* 34:982–991.
- 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.

Associate editor: Esther Betran