JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12548

Research Article

Genetic innovations: Transposable element recruitment and de novo formation lead to the birth of orphan genes in the rice genome

Gui-Hua Jin[1,2], Yan-Li Zhou[1], Hong Yang[1,2], Yan-Ting Hu[1,2], Yong Shi[1], Ling Li[1,2], Abu N. Siddique[1,3], Chang-Ning Liu[4], An-Dan Zhu[1], Cheng-Jun Zhang[1,5]*, and De-Zhu Li[1]*

[1]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Biotechnology, Bacha Khan University, Charsadda 24420, Pakistan
[4]Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Menglun 666303, Yunnan, China
[5]Haiyan Engineering and Technology Center, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
*Authors for correspondence. De-Zhu Li. E-mail: dzl@mail.kib.ac.cn; Cheng-Jun Zhang. E-mail: zhangchengjun@mail.kib.ac.cn

Received 27 February 2019; Accepted 3 November 2019; Article first published online 7 November 2019

Abstract Orphan genes are genetic innovations that lack homologs in other lineages. Orphan genes can rapidly originate and become substantially functional, yet the mechanisms underlying their origins are still largely unknown in plants. Here, we investigated the origin of orphan genes in the *Oryza sativa* ssp. *japonica* "Nipponbare" genome using genome-wide comparisons with 10 closely related *Oryza* species. We identified a total of 37 orphan genes in the Nipponbare genome that show short sequence lengths, elevated GC content, and absence of introns. Interestingly, half of the identified orphan genes originated by way of a distinctive mechanism that involved the generation of new coding sequences through independent and rapid divergence within the inserted transposable element. Our results provide valuable insight into genetic innovations in the model rice genome that formed on a very short timescale.

Key words: comparative genomics, origin, orphan gene, transposable element.

1 Introduction

Orphan genes are newly born genes, present in only one species or in a specific clade but are evidently absent in other closely related taxa (Fischer & Eisenberg, 1999; Boffelli et al., 2004). Compared to other protein-coding genes, orphan genes usually possess several atypical genetic characteristics, such as small coding regions (Lipman et al., 2002), fast evolutionary rates (Domazet-Loso & Tautz, 2003), uncharacterized functional domains (Daubin & Ochman, 2004), and variable guanine-cytosine (GC) content (Arendsee et al., 2014; Palmieri et al., 2014; Sun et al., 2015; Xu et al., 2015). Although the functions of the majority of orphan genes are largely unknown, these genes appear to play important, and sometimes even essential, roles in species-specific adaptation. For example, orphan genes can encode phylum-specific morphology in *Hydra* (Khalturin et al., 2009), confer *Fusarium* resistance in wheat (Perochon et al., 2015), participate in early human brain development (Zhang et al., 2011), and participate in primary metabolic pathways to allow *Arabidopsis thaliana (L.)* Heynh adapt to environmental changes (Jones et al., 2016).

The identification and classification of orphan genes largely depends on two factors. Both the evolutionary distance between the target taxa/clade and its nearest sequenced relatives, as well as the quality of genome sequence, are used (Tautz & Domazet-Loso, 2011). Due to limited data availability, previous studies have mainly focused on the identification of orphan genes specific to large taxonomic groups, which could span long evolutionary timescales. For example, 270 orphan genes were found in primates, based on comparisons with non-primate species (Toll-Riera et al., 2009). In plants, 1789 orphan genes were identified in Brassicaceae that lack homology in non-Brassicaceae plant species (Donoghue et al., 2011). However, having a long evolutionary history might actually reduce the power of analyses, because this could make it more difficult to identify orphan genes as the homology relationship might be masked by increased sequence divergence. Recently, with the increasing number of available genomes from closely related species, it has become feasible to identify orphan genes on small evolutionary timescales. These short timescale analyses can benefit our understanding of the origin and evolutionary processes of newly born genes.

Several excellent models have been proposed to explain the origin of orphan genes. For instance, a duplication-divergence mechanism is considered the main mechanism for the birth of orphan genes in zebrafish (Yang et al., 2013) and

Drosophila (Zhou et al., 2008). Transposable element (TE) exaptation is another vital mechanism to consider in the origin of orphan genes in primates (Toll-Riera et al., 2009) and likewise in silkworm (Sun et al., 2015). Other studies have indicated that the majority of orphan genes could arise de novo from noncoding sequences (Wissler et al., 2013; Zhang et al., 2019). Recently, Prabh & Rödelsperger (2019) also proposed a mixed origin mechanism, incorporating the notion that an orphan gene could generate through both duplication-divergence and the de novo mechanism. In plants, a previous study indicated the existence of four distinct origin mechanisms for the formation of orphan genes in *A. thaliana*, including overlap with conserved gene loci, duplication followed by sequence divergence, TE exaptation, and de novo formation (Donoghue et al., 2011). However, the origin and dynamics of orphan genes in other plants were less studied.

The Asian cultivated rice *Oryza sativa* ssp. *japonica* Shig.Kato is one of the youngest taxa (<0.55 Myr), and has a more comprehensively assembled and annotated genome. To date, at least 13 genomes of *Oryza* belonging to different lineages have been released, which provide an excellent opportunity to systematically identify orphan genes within these genomes (Stein et al., 2018). In this study, we attempted to find which origin mechanisms play an important role in the newly born genes. To answer this question, we investigated origin mechanisms for newly born orphan genes in the *Oryza sativa* ssp. *japonica* "Nipponbare" using genome-wide comparisons with 10 closely related *Oryza* species.

2 Material and Methods

2.1 Data sources

To investigate the origin of orphan genes, we selected the model rice plant *Oryza sativa* ssp. *japonica* as the focal taxon. Genomic data, including nucleotide and amino acid sequences of protein genes and the genome sequence of *O. sativa* ssp. *japonica* "Nipponbare" were downloaded from TIGR release 7 (Kawahara et al., 2013). We also downloaded the genomic data for *O. sativa* ssp. *indica* "Shuhui 498" (R498) from MBKbase (Du et al., 2017) and for nine other *Oryza* species (*O. barthii* A.Chev., *O. brachyantha* A.Chev. & Roehrich, *O. glaberrima* Steud., *O. glumaepatula Steud.*, *O. longistaminata* A.Chev. & Roehrich, *O. meridionalis* N.Q.Ng, *O. nivara* S.D.Sharma & Shastry, *O. punctata* Kotschy ex Steud, and *O. rufipogon* Griff.) from the Gramene database version 52 (http://www.gramene.org) to use as outgroups. The genomic data of *O. sativa* ssp. *japonica* "Nipponbare" were extensively compared with the data for three close relatives (*O. nivara*, *O. rufipogon*, and *O. sativa* ssp. *indica*). All of the genomic data used in this study are summarized in Table S1.

2.2 Identification of orphan genes

To reliably identify orphan genes in the *O. sativa* ssp. *japonica* "Nipponbare" (hereafter, the Nipponbare) genome, we undertook genome-wide pairwise comparisons between Nipponbare and 10 other *Oryza* genomes (outgroups). The identification of the orphan genes on long evolutionary timescales was mostly based on the use of amino acid alignments (Cai et al., 2006; Toll-Riera et al., 2009; Johnson & Tsutsui, 2011), which largely rely on genome annotation quality. Critically speaking, genome annotation artifacts within the outgroup could affect orphan gene identification. Therefore, we implemented strict procedures to reduce annotation artifacts from the outgroups (Fig. 1 for details).

Initially, all of the 55 986 coding sequences (CDS) annotated in the Nipponbare genome were used as queries to search against CDS in the outgroup using BLAT (at the threshold of 80% sequence coverage) (Kent, 2002) and to search against genomes in the outgroup with BLASTN (default task: Megablast, E value ≤ 0.01). Adopting this approach enabled us to filter any homologous genes that were perhaps misannotated in the outgroups. We also undertook a TBLASTN search in the NCBI NR database (up until 7 May 2018) using amino acid

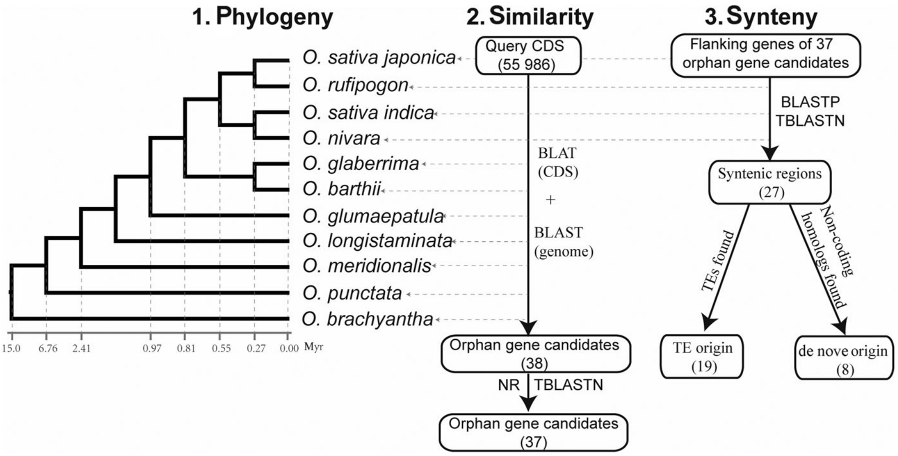

Fig. 1. Computational pipeline for the identification of orphan genes in *Oryza sativa* ssp. *japonica* "Nipponbare" based on similarity and synteny. The topology of the phylogeny referred to the published high-resolution tree in Stein et al. (2018). Orphan genes were identified in the Nipponbare genome using 10 other *Oryza* genomes as outgroups. The flanking genes of the Nipponbare orphan genes were used to detect syntenic regions in three close relatives, including *O. rufipogon*, *O. sativa* ssp. *indica* "Shuhui 498," and *O. nivara*. CDS, coding sequence; TE, transposable element.

sequences as queries to exclude any genes with homologs in other plants. Moreover, we implemented a similar procedure that has been described as the "flanking gene method" (Freeling et al., 2008). This procedure utilizes neighboring syntenic genes to identify the presence or absence of putative orphan genes in the three closest relatives. Five to 10 upstream and downstream genes from the orphan genes in Nipponbare were used for synteny analysis using BLASTP and TBLAST alignments (E value $\leq 10^{-5}$).

2.3 Characteristics of orphan genes

To better understand whether orphan genes possess special features, we analyzed three important genic characteristics of the orphan genes. Except for the identified orphan genes, all other protein-coding genes in the Nipponbare genome were considered non-orphan genes. Comparisons between orphan genes and non-orphan genes were carried out using custom Perl scripts, including calculations of the protein length, intron number, and GC content at the CDS level. The Mann–Whitney U test was also applied to assess the differences between orphan and non-orphan genes (Bauer, 1972).

2.4 Functional signatures

Expression signatures of orphan genes were first determined by searches against publicly available data on the website of http://rice.plantbiology.msu.edu, including full-length cDNA, expressed sequence tag (EST, Pontius et al., 2003), and microarray data. We also searched for all functional domains in orphan genes using the InterProScan programs at the European Bioinformatics Institute website (version 75.0) (http://www.ebi.ac.uk/Tools/InterProScan). Additionally, expression levels were examined on the basis of 284 high-quality RNA sequencing datasets, deposited in the Rice Expression Database (http://expression.ic4r.org/index). An orphan gene was considered to be expressed at a threshold of fragments per kilobase of transcript per million fragments mapped greater than 0 in at least two samples.

To verify orphan genes expression levels, reverse transcription–polymerase chain reaction (RT-PCR) experiments were carried out on eight randomly selected orphan genes. Primer information is provided in Table S2, and the housekeeping gene *Tubulin* was selected as an internal control. Total RNA was extracted from mixed rice tissue samples using an Eastep Super Total RNA Extraction Kit (Promega, Beijing, China). Based on concentration and quality assays, undertaken with a NanoDrop ND 1000, 2 μg total RNA was used to synthesize the first-strand cDNA with the GoScript Reverse Transcription System (Promega). Finally, high-quality cDNAs were used as templates in the RT-PCRs. The RT-PCR products were resolved on 3% agarose gels by electrophoresis at 120 V for 30 min.

2.5 Tracing the birth of orphan genes

Based on syntenic analysis of the flanking genes in Nipponbare and its close relatives, we compared multiple sequences in the syntenic regions of the genomes of Nipponbare and its three close relatives via the MUSCLE program (Edgar, 2004) using the codon alignment option with the Nipponbare CDS as the template. Any orphan genes that had partial homologs in the syntenic regions of its close relatives were considered as non-coding homologous sequences. If an orphan gene lacked non-coding homologous sequences in the close relatives, we searched for TEs using the CENSOR program against the RepBase database (Kohany et al., 2006) in the syntenic regions of Nipponbare and its close relatives.

3 Results

3.1 Identification of 37 functional orphan genes in the Nipponbare genome

We implemented a strict pipeline on the basis of both homology and synteny to search for newly born orphan genes (<0.55 Myr) in Nipponbare after it diverged from *Oryza rufipogon* (Fig. 1). We identified 38 genes without any homologs in the 10 *Oryza* outgroups through extensive comparisons. Among these genes, one gene (*LOC_Os04g34130*) was removed from further analysis because the sequence was identical to that of the *Escherichia coli ECs4062* gene, identified through the NCBI NR database (Fig. S1), implying possible sequence contamination in the Nipponbare genome. Therefore, after exclusion, a total of 37 orphan genes were identified in the Nipponbare genome (Table 1).

Syntenic information was further used to verify results of the gene homology searches. Due to the fact that orphan genes inherently lack phylogenetic conservation, the "flanking gene method" (Freeling et al., 2008) was used to deduce the syntenic locations of the orphan genes in the close relatives. Twenty-seven orphan genes were anchored to corresponding syntenic regions in at least one close relative (Table S3). In these cases, all the syntenic regions in the close relatives showed continuous sequences without any gaps, excluding the improbable orphan genes attributed to poor genome assemblies in the outgroup. Furthermore, no intact open reading frames were detected in the syntenic regions, suggesting these orphan genes were truly absent even in the close relatives.

Analysis of the functionality of orphan genes provides evidence for its authenticity. Functionality of all 37 identified orphan genes was supported by at least one piece of evidence, either expressed sequence tags, full-length cDNA sequences, microarray data, RNA sequencing experimental data, or InterPro protein domain alignments (Tables 1, S4). We found that 35 orphan genes contain intrinsically disordered protein domains by MobiDB-lite (Necci et al., 2017) analysis at the InterPro website (Table S5). We also successfully validated expression patterns of two orphan genes from among eight randomly selected orphan genes using RT-PCR experiments (Fig. S2).

3.2 Orphan genes showed genic features distinct from those of non-orphan genes

To investigate whether the identified orphan genes had distinct properties, we compared three genic features (protein length, intron number, and GC content) between the orphan genes and non-orphan genes. Peptide lengths of the orphan genes were significantly shorter than those of the non-orphan genes (Fig. 2A). The median length of the non-orphan proteins (336 amino acids) is approximately three times longer than that of the orphan proteins (105 amino acids) (Mann–Whitney U test, $P = 6.458e-13$). Orphan genes

Table 1 Information on orphan genes in the *Oryza sativa* ssp. *japonica* “Nipponbare” genome

Orphan gene	InterPro domain	FL-cDNA	ESTs	RNA-seq	Microarray	RT-PCR
LOC_Os01g50560	+	–	–	+	–	–
LOC_Os01g72920	+	–	–	–	–	–
LOC_Os03g17830	+	–	–	–	–	–
LOC_Os03g60419	+	+	+	+	–	–
LOC_Os04g11940	+	–	–	+	–	–
LOC_Os04g22510	+	–	–	+	+	+
LOC_Os05g42940	+	–	–	+	–	–
LOC_Os05g46650	+	–	–	+	–	–
LOC_Os05g48540	+	–	–	–	–	–
LOC_Os06g16530	+	–	–	+	–	–
LOC_Os06g19880	+	–	–	+	–	–
LOC_Os06g33910	+	–	–	–	–	–
LOC_Os06g38190	–	–	–	+	–	–
LOC_Os06g44390	+	–	–	+	–	–
LOC_Os06g51300	+	+	–	+	–	–
LOC_Os07g26240	+	–	–	+	–	–
LOC_Os07g26770	+	–	–	–	–	–
LOC_Os07g26890	+	–	–	+	–	–
LOC_Os08g05330	+	–	–	+	–	–
LOC_Os08g09680	+	–	–	–	–	–
LOC_Os08g26460	+	–	–	+	+	–
LOC_Os08g26960	+	–	–	–	–	–
LOC_Os08g36270	+	–	–	–	–	–
LOC_Os08g44980	+	–	–	+	–	–
LOC_Os09g13260	+	–	–	+	+	–
LOC_Os09g35640	+	–	–	+	–	–
LOC_Os10g09560	+	–	–	+	+	–
LOC_Os11g06950	+	–	–	+	–	–
LOC_Os11g30450	+	–	–	+	–	–
LOC_Os11g35600	+	–	–	+	–	–
LOC_Os11g44200	+	–	–	+	+	–
LOC_Os11g44270	+	–	–	+	+	–
LOC_Os11g44280	–	–	–	+	+	–
LOC_Os12g09880	+	–	–	+	+	–
LOC_Os12g11060	+	–	–	+	+	+
LOC_Os12g33250	+	–	–	+	–	–
LOC_Os12g43200	+	–	–	+	–	–

+, evidence present; –, evidence absent; EST, expressed sequence tag; FL-cDNA, full-length cDNA; RNA-seq, RNA sequencing; RT-PCR, reverse transcription-polymerase chain reaction.

contained fewer introns (with a median of 0) than the non-orphan genes (with a median of 2) (Mann–Whitney U test, *P* = 6.458e−13) (Fig. 2B). The GC content at the CDS level of the orphan genes (with a median of 74.23%) was notably higher than that of the non-orphan genes (with a median of 54.10%) (Mann–Whitney U test, *P* = 2.585e−14) (Fig. 2C).

3.3 Most orphan genes originated from a TE-mediated mechanism followed by rapid divergence

To obtain clues about the underlying processes involved in the birth of orphan genes, we compared multiple sequences in the syntenic regions of Nipponbare and of its three close relatives using the MUSCLE program (Edgar, 2004). We did not detect any homologous sequences in the close relatives for 19 Nipponbare orphan genes. In order to trace origin mechanisms for these 19 untraceable orphan genes, we searched for the TEs in the syntenic regions of the genomes of Nipponbare and of its close relatives. These searches were motivated by previous observations that suggested orphan genes are derived from TEs in primates (Toll-Riera et al., 2009) and in silkworms (Sun et al., 2015).

Determination of TE-derived sequences was the key evidence suggesting a TE-mediated origin of the orphan genes. We checked whether the TEs were intact with complete target site duplications (TSDs) in the syntenic regions. A typical example (*LOC_Os01g72920*) of how an orphan gene could be formed from TEs is shown in Fig. 3. *LOC_Os01g72920* is an orphan gene located on chromosome 1: 42288473–42289606, where it is embedded in an LTR-18C_OS-LTR (RepBase ID) retrotransposon. We further confirmed that this retrotransposon was recently inserted into the Nipponbare genome with “TTATG” as the TSD sequence, as it was not found in the syntenic regions of three close relatives (*O. rufipogon*, *O. sativa* ssp. *indica* “Shuhui 498,” and *O. nivara*). Sequences around the TE were

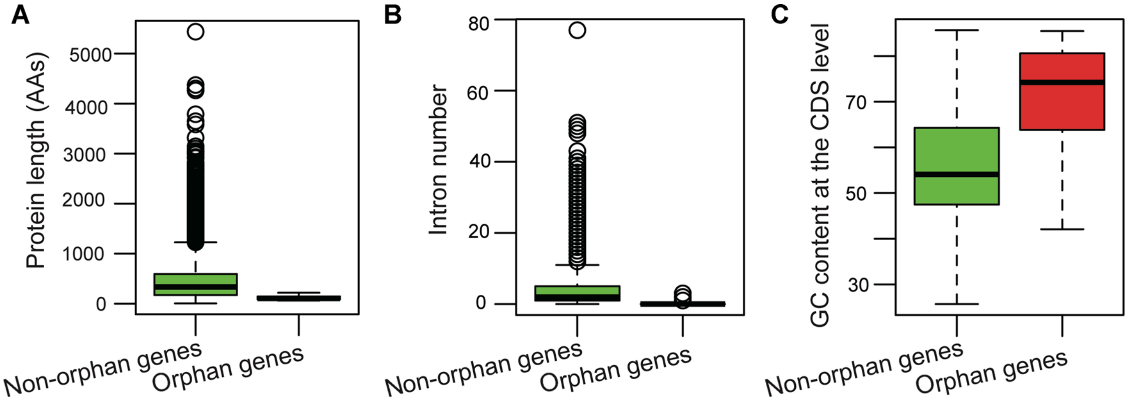

Fig. 2. Comparison of the genic features conserved between orphan genes and non-orphan genes in *Oryza sativa* ssp. *japonica* "Nipponbare" using a Mann–Whitney U test ($P < 0.05$). **A,** The protein-coding length of the orphan genes was much shorter than that of non-orphan genes. **B,** The intron number in orphan genes was significantly lower than that of non-orphan genes. **C,** The GC content of the orphan genes at the coding sequence (CDS) level was significantly higher than that of non-orphan genes. AA, amino acid.

conserved among Nipponbare and its close relatives. Therefore, specifically inserting LTR-18C_OS-LTR retrotransposon resulted in the formation of the orphan gene *LOC_Os01g72920*. The first exon of *LOC_Os01g72920* overlapped with the LTR-18C_OS-LTR retrotransposon, and the second exon combined the TSD sequence "TTATG" with the retrotransposon sequence.

In total, 19 of the identified Nipponbare orphan genes (51%) were derived from TE insertion events. The TE-mediated orphan genes can be sorted into two groups (Fig. 4A): 14 orphan genes were generated by recent TE insertion events that only occurred in the Nipponbare genome (e.g., *LOC_Os01g72920*) (Figs. 4A, S3A–S3I, S3P–S3R; Table 2), and five orphan genes were formed by ancient TE insertion events when TEs were inserted in the common ancestor of Nipponbare and its close relatives (Figs. 4A, S3J–S3M, S3O; Table 2).

Eleven different TEs participated in the origin of 19 orphan genes (Table 2). Transposable element expansion can duplicate orphan genes after its origin. If an orphan gene is duplicated through TEs, we hypothesized that we should be able to detect associated gene paralogs, and all those paralogs should be related to one TE type. We found nine orphan genes associated to three TEs that fit this hypothesis. In detail, five orphan genes (LOC_Os04g11940, LOC_Os05g48540, LOC_Os08g36270, LOC_Os08g44980, and LOC_Os11g30450) were paralogs (Fig. S4). All paralogs appear to be associated with ENSPM4 (Table 2), so we speculated that these five genes formed by ENSPM4 expansion, after one had initially originated. Similarly, *LOC_Os05g42940* and *LOC_Os06g16530* appear to have formed through SPMLIKE transposon expansion (Fig. S5), and *LOC_Os08g09680* and *LOC_Os11g35600* appear to have formed through SZ-67LTR expansion (Fig. S6). Although three orphan genes (*LOC_Os06g51300*, *LOC_Os08g26960*, and *LOC_Os07g26890*) appear to be related to SZ-67LTR insertion, they have completely different coding sequences. This suggests that these three orphan genes originated independently after TE insertion. In summary, 13 independent origin events and three TE expansion events account for 19 Nipponbare orphan genes (Table 2).

3.4 De novo origination served as another primary mechanism for orphan gene formation

The identification of non-coding orthologous sequences in the syntenic region of close relatives can provide strong evidence for de novo origination. The power of sequence similarity searches depends on the size of the query genes (Ruiz-Orera et al., 2018). As a result, orthologous sequences of short genes could be missed when using Megablast to rule out similar hits in the genomes of close relatives. By comparing multiple sequences of syntenic regions in Nipponbare and close relatives using the MUSCLE program, highly divergent or small homologous non-coding sequences can be identified in the close relatives.

In this study, eight orphan genes in Nipponbare (22%) were confidently defined as de novo origination events because non-coding orthologous sequences existed in the close relatives. During the process of de novo origination, it is crucial to obtain a start codon and to remove internal stop codons through mutations, such as indels (i.e., frameshift) and point mutations. For instance, non-coding sequences in Nipponbare were transformed into the orphan gene *LOC_Os04g22510* using three critical enabling mutations: (i) start codon acquisition; (ii) insertion of a single "T" at base pair position 46, which resulted in a frameshift that removed a stop codon at base pair position 50; and (iii) conversion of a premature "TAG" stop codon into "TCG" (which encodes serine) at base pair positions 293–295 by a point mutation (Fig. S7). In summary, of the eight de novo-formed orphan genes, four appear to have formed through start codon acquisition, and indel and point mutations; two appear to have formed by start codon acquisition and indels; one appears to have formed by start codon acquisition alone, and one by enabling indels alone (Table 3).

4 Discussion

In this study, we identified 37 recently originated orphan genes in the Nipponbare genome (Table 1). Orphan genes are

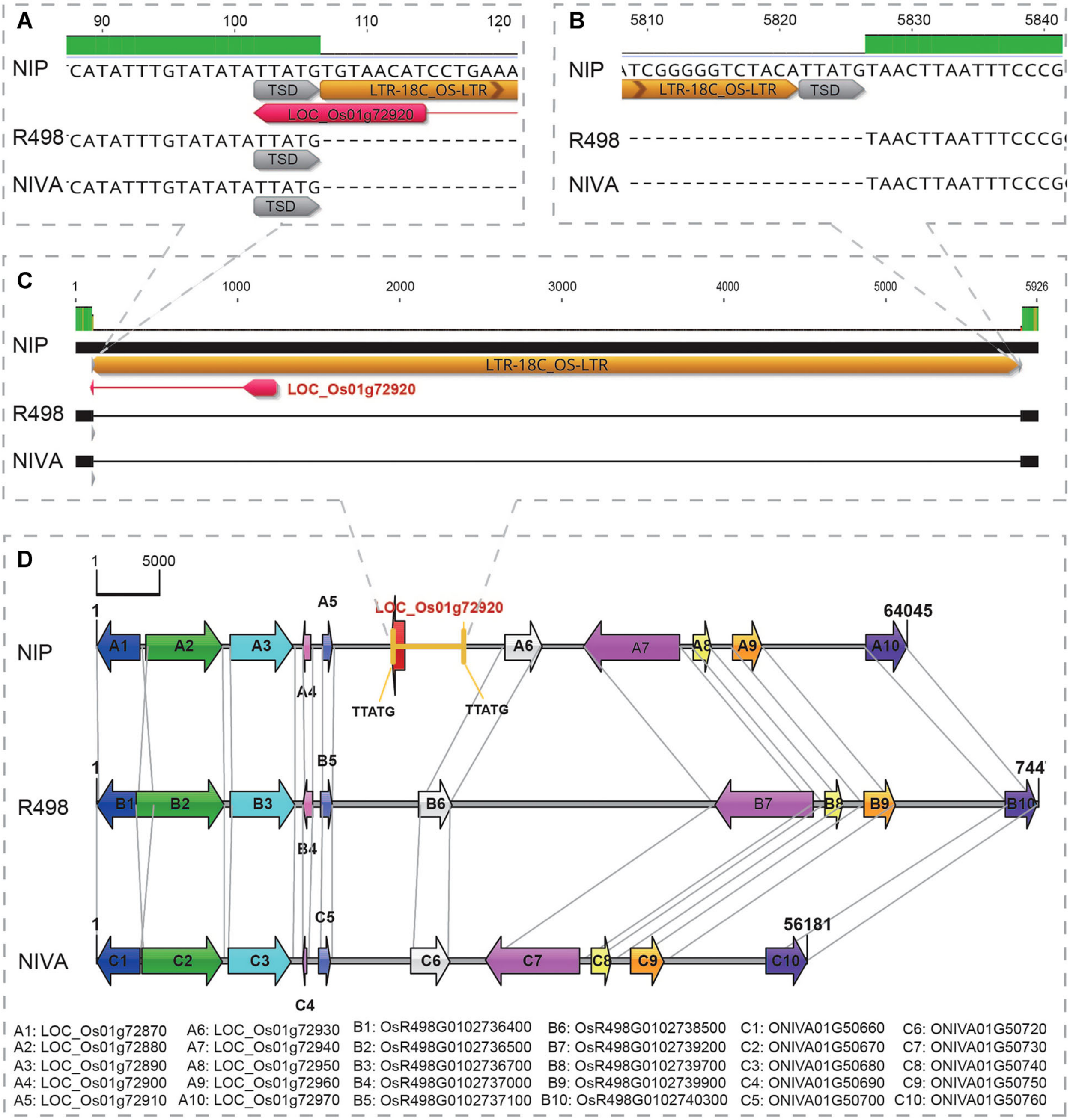

Fig. 3. Orphan gene *LOC_Os01g72920* originated through a transposable element (TE)-mediated mechanism. **A,** Enlarged view of the 5′-junction of the TE insertion. **B,** Enlarged view of the 3′-junction of the TE insertion. **C,** Illustration of the syntenic relationship between *Oryza sativa* ssp. *japonica* "Nipponbare" (NIP) and its close relatives. **D,** Gray lines represent syntenic chains among the four genomes. Boxes marked in the same color represent orthologous genes. Red box represents the orphan gene. NIVA, *O. nivara*; R498, *O. sativa* ssp. *indica* "Shuhui 498"; TSD, target site duplication.

sometimes considered as de novo genes originating from ancestral non-coding sequences (Chen et al., 1997; Knowles & McLysaght, 2009; Li et al., 2010; Murphy & McLysaght, 2012; Xie et al., 2012). In fact, de novo formation is just one way by which orphan genes can form, as many orphan genes are derived from other distinct evolutionary processes, such as the duplication-divergence mechanism (Schlotterer, 2015; Moyers & Zhang, 2016), TE exaptation (Toll-Riera et al., 2009), loss of homologous genes in related species (Zhao et al., 2015), repetition of low-complexity short peptides (Chen et al., 1997; Cheng & Chen, 1999), and horizontal gene transfer from fast-evolving donors (Keeling & Palmer, 2008; Husnik & McCutcheon, 2018). An orphan gene can also originate through a combination of several origin mechanisms, such as the mixed origin mechanism in nematodes (Prabh & Rödelsperger, 2019). Zhang et al. (2019) recently

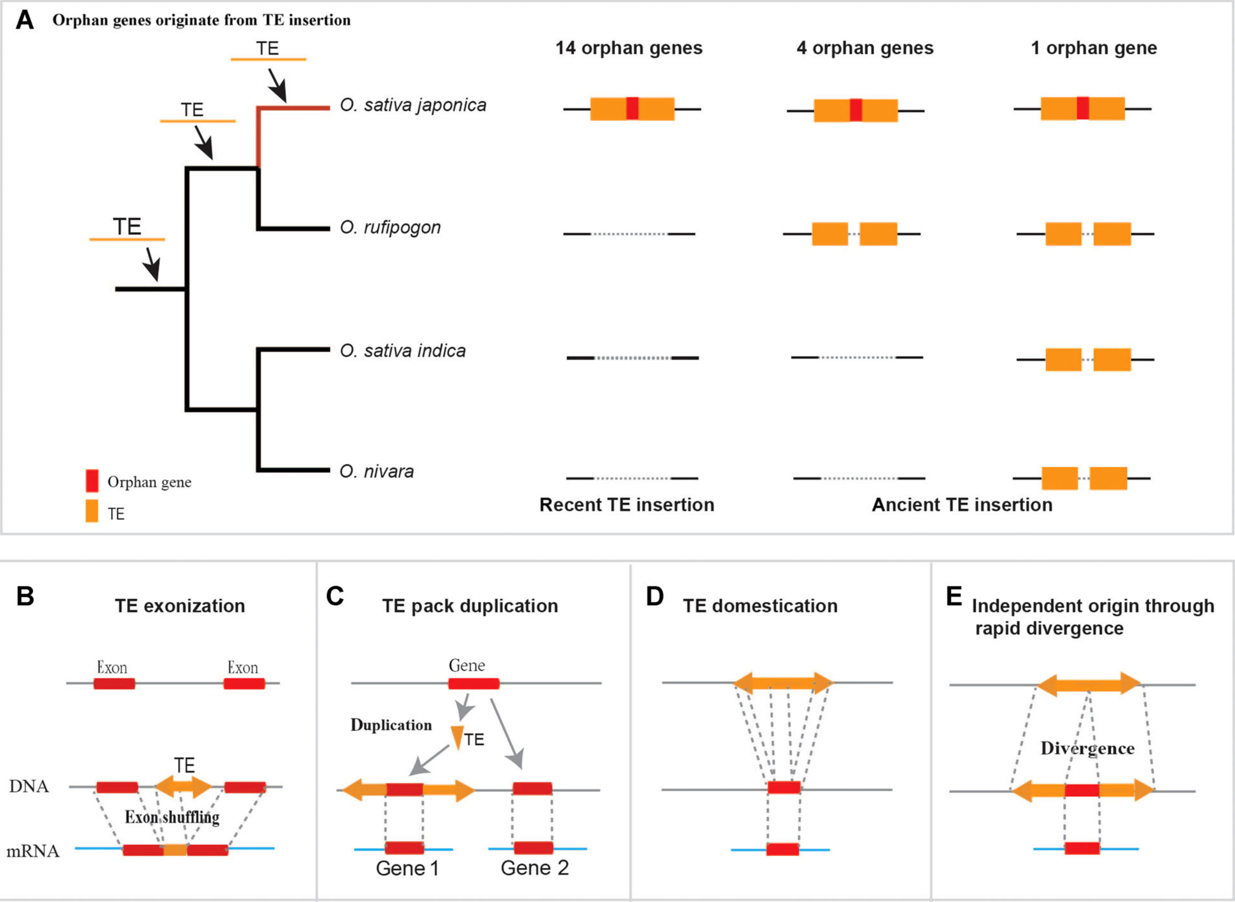

Fig. 4. Summary of new genes originated through the transposable element (TE)-mediated mechanism. **A,** Different TE (orange box) insertion events associated with orphan genes (red box/line) in *Oryza sativa* ssp. *japonica* "Nipponbare" are shown on the branches of the phylogenetic tree. **B,** TE exonization: a host gene recruits a segment of a TE sequence as a dependent exon or as part of an exon of the host gene, also called TE exaptation, referring to the description of Chen et al. (2013). **C,** Pack duplication model: genes or gene fragments are captured by TEs to generate duplications, referring to the pack-MULE model of Jiang et al. (2004). **D,** TE domestication: conserved TE sequences, such as those encoding transposases, can be domesticated through evolution to harbor specific traits, referring to the description of Jangam et al. (2017). **E,** Independent origin through rapid divergence: orphan genes originate independently through rapid divergence within inserted TEs. Dashed line, empty position of a closely related outgroup in the alignment.

reported that de novo genes contributed to the rapid evolution of *Oryza* protein diversity over 15 Myr. In order to provide more details about the origination of orphan genes, we investigated the origins of the youngest orphan genes in the model plant *O. sativa* ssp. *japonica* "Nipponbare" (<0.55 Myr).

Identification pipelines and genome annotation could affect the number of orphan genes identified. In this study, because we utilized a strict filtering procedure (excluding genes that have any MEGABLAST hits in the outgroup) to identify Nipponbare orphan genes, we exclude the de novo gene reported in Zhang et al. (2019), which highlighted highly similar in non-coding sequences across the relatives. The strict filtering procedure used in our study could lead to an underestimation in the number of de novo genes. Genome annotation of focal genome is another important factor which might affect the identification of orphan genes (Denton et al., 2014; Prabh & Rödelsperger, 2016). For instance, among 37 Nipponbare orphan genes in our analysis (MSU-RAP annotation), we found that only eight orphan genes (22%) were formed by de novo, but these genes were not annotated in the Zhang et al. (2019) study (OGE/IOMAP annotation).

In our study, the majority of the orphan genes (51%) were derived from TE-related sequences, indicating that TE could serve as the main genetic material for the origin of orphan genes in the Nipponbare genome. We identified a distinctive TE-mediated mechanism that differs from previously reported mechanisms involved in the birth of orphan genes. Previous studies indicated that TEs can contribute to novel gene formation through three distinct mechanisms. First, segments of TE sequences could be recruited as dependent exons or as parts of exons in host genes by a process, known as TE exonization or TE exaptation (Fig. 4B) (Nekrutenko & Li, 2001; Long et al., 2003, 2013; Shapiro, 2005; Chen et al., 2013). Second, several TEs, such as mutator-like TEs (MULEs),

Table 2 Origin of orphan genes in *Oryza sativa* ssp. *japonica* "Nipponbare" mediated by transposable element insertion

Orphan genes	TE class	TE IDs in RepBase	TE insertion time	Origin event
LOC_Os04g11940, LOC_Os05g48540, LOC_Os08g36270, LOC_Os08g44980, LOC_Os11g30450	DNA/CACTA	ENSPM4	Recent	1
LOC_Os05g42940, LOC_Os06g16530	DNA/CACTA	SPMLIKE	Recent	1
LOC_Os08g09680, LOC_Os11g35600	RNA/LTR	SZ-67LTR	Recent	1
LOC_Os06g19880	DNA/CACTA	SPMLIKE-B_OS	Recent	1
LOC_Os01g72920	RNA/LTR	LTR-18C_OS-LTR	Recent	1
LOC_Os05g46650	RNA/LTR	LTR-18_OS-LTR	Recent	1
LOC_Os06g44390	RNA/LTR	MuDR-N18C_OS	Recent	1
LOC_Os08g26460	RNA/LTR, LINE	COPIA1-LTR_OS, LINE1–11_OS	Recent	1
LOC_Os11g06950	RNA/LTR	LTR-18B_OS-LTR	Ancient	1
LOC_Os12g33250	RNA/LTR	LTR-18K_OS-LTR	Ancient	1
LOC_Os06g51300	RNA/LTR	SZ-67LTR	Ancient	1
LOC_Os07g26890	RNA/LTR	SZ-67LTR	Ancient	1
LOC_Os08g26960	RNA/LTR	SZ-67LTR	Ancient	1

TE, transposable element.

can capture host genes or gene fragments, such as the pack MULE model detailed in Jiang et al. (2004), referred to as "TE pack duplication" (Fig. 4C). This TE-mediated mechanism often generates duplicated genes, and some captured gene fragments which can become functional (Jiang et al. 2004; Flagel & Wendel, 2009; Panchy et al., 2016). Finally, conserved TE sequences, such as those encoding transposases, can be domesticated through evolution to harbor specific traits in a process called TE domestication (Fig. 4D) (Kapitonov & Jurka, 2005; Jangam et al., 2017). However, none of these TE-related mechanisms are absolutely consistent with our results.

In our analysis, orphan genes could originate independently through rapid divergence, after the TE insertions, a process that we consider independent origin within inserted TEs (Fig. 4E). Ten Nipponbare orphan genes had no identifiable paralogs in the genome and appear to be associated with different TEs, which indicates that these orphan genes formed independently after the TE insertions. But once an orphan gene is formed, it will be duplicated through TE expansion, for example, nine orphan genes were formed by three expanded TEs (ENSPM4, SPMLIKE, and SZ-67LTR), similar to TE pack duplication process (Fig. 4C). In these nine genes, it is difficult to distinguish which three are parent genes because they have no comparative homologs in the close relatives.

Table 3 Crucial mutations that transformed noncoding sequences into coding sequences in the de novo process

Orphan gene	Start codon acquisition	Enabling indels	Enabling point mutations
LOC_Os03g17830	+	+	+
LOC_Os03g60419	+	+	+
LOC_Os04g22510	+	+	+
LOC_Os06g33910	+	+	−
LOC_Os08g05330	+	−	−
LOC_Os09g35640	+	+	+
LOC_Os11g44200	+	+	−
LOC_Os12g43200	−	+	−

+, present; −, absent.

The identification of independent origin through rapid divergence within inserted TEs that explain Nipponbare orphan gene origination was based on a short timescale analysis of TE insertion events in the target taxon and its closely related species. Previous studies on orphan genes have mainly focused on detecting TE-related orphan gene sequences in target taxon/taxa by BLAST searches against a TE database (Toll-Riera et al., 2009; Donoghue et al., 2011; Wissler et al., 2013; Yang et al., 2013; Sun et al., 2015), without comparisons to its close relatives. Therefore, these analyses failed to reveal a detailed picture of the TE insertions or the emergence of TE-related orphan genes. In addition to searching for intact transposons in the syntenic regions of the focus genome and those of its close relatives, we also speculated on the relative times of TE insertion events and the origins of the orphan genes. This novel analysis led us to infer that the orphan genes can form independently through rapid divergence within inserted TEs. The origination of the orphan genes from independent divergence after TE insertion is unlikely to be Nipponbare-specific. This mechanism might also be present in other taxa that have not yet been characterized.

The emergence of orphan genes from TE-related sequences could be supported by the high TE mutation rates. Historically, TEs were considered "junk" sequences in genomes (Ohno, 1970). However, TEs are highly mutagenic due to their high genomic abundance, which can easily promote chromosomal rearrangements (Schrader & Schmitz, 2019). Transposable element insertions can drastically affect the evolution of the surrounding genes by altering their genetic structure and/or regulatory sequences. Most TEs remain silent, evolving in a neutral fashion, however, a proportion appear to gain adaptive roles (Arkhipova, 2018). Transposable element-derived proteins have been abundantly and recurrently domesticated, and they are now considered an important adaptive mechanism for evolutionary innovation (Jangam et al., 2017; Schrader & Schmitz, 2019). Transposable elements have very high mutation rates, a trait that has been confirmed in the model plant *A. thaliana* at the population level (Li et al., 2018), which is also consistent with the previous hypothesis that a high mutation rate will present along with an insertion (Tian et al., 2008)

The most prevalent TE-associated mechanism for orphan gene formation in Nipponbare (51%) is comparable to the prevalence of this mechanism in primates (53%) (Toll-Riera et al., 2009). Nevertheless, the frequency of the TE-associated mechanism for orphan gene formation in the Nipponbare genome is substantially higher than that in model plant *A. thaliana* (9.73%) (Donoghue et al., 2011). This difference could be due to the relative abundance of TE content in their genomes. The TE content in the *A. thaliana* genome is only 10% (Arabidopsis Genome Initiative, 2000) but reaches up to 48.6% in the Nipponbare genome (International Rice Genome Sequencing Project, 2005). The conclusion that a TE-mediated mechanism is the dominant driver of Nipponbare orphan gene formation is indeed inconsistent with the conclusions of previous studies on *Drosophila* (Zhou et al., 2008) and *A. thaliana* (Donoghue et al., 2011). These studies indicated that duplication-divergence was the primary mechanism for orphan gene formation based on comparisons with distant reference genomes. However, duplication requires much longer evolutionary timescales to accumulate sufficient nucleotide substitutions. Thus, this model was insufficient to explain the youngest genes formed on a short timescale that were identified in the Nipponbare genome (<0.55 Myr). The dominance of different mechanisms for orphan gene origination in various species suggests that different mechanisms might play different roles in each species.

Shorter protein length and fewer introns in the Nipponbare orphan genes were consistent with the findings of studies on orphan genes in both animals (Zhang et al., 2007; Murphy & McLysaght, 2012; Wissler et al., 2013; Yang et al., 2013; Palmieri et al., 2014; Mayer et al., 2015) and plants (Lin et al., 2010; Xu et al., 2015). This suggest that these two characteristics are conserved among the orphan genes in all eukaryotes. The elevated GC content of the Nipponbare orphan genes was consistent with previous reports in plants such as *A. thaliana* (Arendsee et al., 2014), *Citrus* × *sinensis* (L.) Osbeck (Xu et al., 2015), and Poaceae (Campbell et al., 2007). High GC content could cause intrinsically disordered proteins (Basile et al., 2017). In our study, we also found that almost all (95%) Nipponbare orphan genes have disordered protein properties, in agreement with previous results (Bornberg-Bauer et al., 2015; Wilson et al., 2017). Short protein lengths and high GC content made it difficult for us to design primers and to carry out RT-PCR experiments. Even though we obtained some expression evidence, it remains unclear which orphan gene began to function after origination because non-functional proteins tend to pervasively transcript and translate (Ruiz-Orera et al., 2018; Prabh & Rödelsperger, 2019). Although future studies will be required to determine which orphan genes are fixed in the population and associated with species-specific functions, our findings provide clear insight into genetic innovations in the rice genome.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 31571311) and the Pioneer Hundred Talents Program of the Chinese Academy of Sciences (Grant No. 29Y127E71). We are grateful to Dr. Wei-Shu Fan at the Kunming Institute of Botany for revising this manuscript.

References

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814): 796–815.

Arendsee ZW, Li L, Wurtele ES. 2014. Coming of age: Orphan genes in plants. *Trends in Plant Science* 19: 698–708.

Arkhipova IR. 2018. Neutral theory, transposable elements, and eukaryotic genome evolution. *Molecular Biology and Evolution* 35: 1332–1337.

Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan proteins to be intrinsically disordered. *PLoS Computational Biology* 13: e1005375.

Bauer DF. 1972. Constructing confidence sets using rank statistics. *Journal of the American Statistical Association* 67: 687–690.

Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nature Reviews Genetics* 5: 456–465.

Bornberg-Bauer E, Schmitz J, Heberlein M. 2015. Emergence of de novo proteins from 'dark genomic matter'by 'grow slow and moult'. *Biochemical Society Transactions* 43: 867–873.

Cai JJ, Woo PCY, Lau SKP, Smith DK, Yuen KY. 2006. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *Journal of Molecular Evolution* 63: 1–11.

Campbell MA, Zhu W, Jiang N, Lin H, Ouyang S, Childs KL, Haas BJ, Hamilton JP, Buell CR. 2007. Identification and characterization of lineage-specific genes within the Poaceae. *Plant Physiology* 145: 1311–1322.

Chen LB, DeVries AL, Cheng HC. 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences USA* 94: 3811–3816.

Chen SD, Krinsky BH, Long MY. 2013. New genes as drivers of phenotypic evolution. *Nature Reviews Genetics* 14: 645–660.

Cheng CHC, Chen LB. 1999. Evolution of an antifreeze glycoprotein. *Nature* 401: 443–444.

Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: The genealogy of ORFans in *E-coli*. *Genome Research* 14: 1036–1042.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Computational Biology* 10: e1003998.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Research* 13: 2213–2219.

Donoghue MTA, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evolutionary Biology* 11: 47–70.

Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, Ma B, Qi M, Li Y, Zhao X, Wang J, Liu K, Qin P, Yang X, Zhu L, Li S, Liang C. 2017. Sequencing and de novo assembly of a near complete indica rice genome. *Nature Communications* 8: 15324–15346.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* 183: 557–564.

Fischer D, Eisenberg D. 1999. Finding families for genomic ORFans. *Bioinformatics* 15: 759–762.

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Research* 18: 1924–1937.

Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology* 16: 67–79.

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

Jangam D, Feschotte C, Betran E. 2017. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends in Genetics* 33: 817–831.

Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR. 2004. Pack-mule transposable elements mediate gene evolution in plants. *Nature* 431: 569–573.

Johnson BR, Tsutsui ND. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genomics* 12: 164–174.

Jones DC, Zheng WG, Huang S, Du CL, Zhao XF, Yennamalli RM, Sen TZ, Nettleton D, Wurtele ES, Li L. 2016. A clade-specific *Arabidopsis* gene connects primary metabolism and senescence. *Frontiers in Plant Science* 7: 983–1001.

Kapitonov VV, Jurka J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biology* 3: 998–1011.

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu JZ, Zhou SG, Childs KL, Davidson RM, Lin HN, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, Buell CR, Matsumoto T. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6: 4–14.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9: 605–618.

Kent WJ. 2002. BLAT – the BLAST-like alignment tool. *Genome Research* 12: 656–664.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: Are taxonomically-restricted genes important in evolution? *Trends in Plant Science* 25: 404–413.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Research* 19: 1752–1759.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7: 474–484.

Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X, Du Q, Uhl GR, Liu QR, Wei L. 2010. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Computational Biology* 6: e1000734.

Li ZW, Hou XH, Chen JF, Xu YC, Wu Q, Gonzalez J, Guo YL. 2018. Transposable elements contribute to the adaptation of *Arabidopsis thaliana*. *Genome Biology and Evolution* 10: 2140–2150.

Lin HN, Moghe G, Ouyang S, Iezzoni A, Shiu SH, Gu X, Buell CR. 2010. Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana*. *BMC Evolutionary Biology* 10: 41–55.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. *BMC Evolutionary Biology* 2: 20–30.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nature Reviews Genetics* 4: 865–875.

Long MY, VanKuren NW, Chen SD, Vibranovski MD. 2013. New gene evolution: Little did we know. *Annual Review of Genetics* 47: 307–333.

Mayer MG, Rodelsperger C, Witte H, Riebesell M, Sommer RJ. 2015. The orphan gene *dauerless* regulates dauer development and intraspecific competition in nematodes by copy number variation. *PLoS Genetics* 11: e1005146.

Moyers BA, Zhang JZ. 2016. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Molecular Biology and Evolution* 33: 1245–1256.

Murphy DN, McLysaght A. 2012. De novo origin of protein-coding genes in murine rodents. *PLoS One* 7: e48650.

Necci M, Piovesan D, Dosztanyi Z, Tosatto SC. 2017. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* 33: 1402–1404.

Nekrutenko A, Li WHS. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends in Genetics* 17: 619–621.

Ohno S. 1970. *Evolution by gene duplication*. Berlin, Heidelberg: Springer-Verlag.

Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of *Drosophila* orphan genes. *Elife* 3: e01311.

Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of gene duplication in plants. *Plant Physiology* 171: 2294–2316.

Perochon A, Jia JG, Kahla A, Arunachalam C, Scofield SR, Bowden S, Wallington E, Doohan FM. 2015. TaFROG encodes a Pooideae orphan protein that interacts with SnRK1 and enhances resistance to the mycotoxigenic fungus *Fusarium graminearum*. *Plant Physiology* 169: 2895–2906.

Pontius JU, Wagner L, Schuler GD. 2003. UniGene: A unified view of the transcriptome. The NCBI Handbook, Bethesda (MD): National Library of Medicine.

Prabh N, Rödelsperger C. 2016. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics* 17: 226–239.

Prabh N, Rödelsperger C. 2019. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus* nematodes. *G3: Genes, Genomes, Genetics* 9: 2277–2286.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nature Ecology & Evolution* 2: 890–896.

Schlotterer C. 2015. Genes from scratch – the evolutionary fate of de novo genes. *Trends in Genetics* 31: 215–219.

Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution. *Molecular Ecology* 28: 1537–1549.

Shapiro JA. 2005. A 21st century view of evolution: Genome system architecture, repetitive DNA, and natural genetic engineering. *Gene* 345: 91–100.

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, Wei S, Wang J, Liao Y, Wang M, Jacquemin J, Becker C, Kudrna D, Zhang J, Londono CEM, Song X, Lee S, Sanchez P, Zuccolo A, Ammiraju JSS, Talag J, Danowitz A, Rivera LF, Gschwend AR, Noutsos C, Wu CC, Kao SM, Zeng JW, Wei FJ, Zhao Q, Feng Q, El Baidouri M, Carpentier MC, Lasserre E, Cooke R, Rosa Farias DD, da Maia LC, Dos Santos RS, Nyberg KG, McNally KL, Mauleon R, Alexandrov N, Schmutz J, Flowers D, Fan C, Weigel D, Jena KK, Wicker T, Chen M, Han B, Henry R, Hsing YC, Kurata N, de Oliveira AC, Panaud O, Jackson SA, Machado CA, Sanderson MJ, Long M, Ware D, Wing RA. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nature Genetics* 50: 285–296.

Sun W, Zhao XW, Zhang Z. 2015. Identification and evolution of the orphan genes in the domestic silkworm, *Bombyx mori*. *FEBS Letters* 589: 2731–2738.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* 12: 692–702.

Tian DC, Wang Q, Zhang PF, Araki H, Yang SH, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455: 105–108.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Alba MM. 2009. Origin of primate orphan genes: A comparative genomics approach. *Molecular Biology and Evolution* 26: 603–612.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecology & Evolution* 1: 0146–0165.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biology and Evolution* 5: 439–455.

Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, Zhang M, Zhang R, Wei L, Li CY. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genetics* 8: e1002942.

Xu YT, Wu GZ, Hao BH, Chen LL, Deng XX, Xu Q. 2015. Identification, characterization and expression analysis of lineage-specific genes within sweet orange (*Citrus sinensis*). *BMC Genomics* 16: 995–1005.

Yang LD, Zou M, Fu BD, He SP. 2013. Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. *BMC Genomics* 14: 65–80.

Zhang G, Wang H, Shi J, Wang X, Zheng H, Wong GK, Clark T, Wang W, Wang J, Kang L. 2007. Identification and characterization of insect-specific proteins by genome data analysis. *BMC Genomics* 8: 93–104.

Zhang L, Ren Y, Yang T, Li GW, Chen JH, Gschwend AR, Yu Y, Hou GX, Zi J, Zhou R, Wen B, Zhang JW, Chougule K, Wang MH, Copetti D, Peng ZY, Zhang CJ, Zhang Y, Ouyang YD, Wing RA, Liu SQ, Long MY. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nature Ecology & Evolution* 3: 679–690.

Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biology* 9: e1001179.

Zhao Y, Tang L, Li Z, Jin JP, Luo JC, Gao G. 2015. Identification and analysis of unitary loss of long-established protein-coding genes in Poaceae shows evidences for biased gene loss and putatively functional transcription of relics. *BMC Evolutionary Biology* 15: 66–76.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Research* 18: 1446–1455.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12548/suppinfo:

Table S1. Genomic information for the 11 selected *Oryza* genomes used in this study.

Table S2. Primer sequences used for the RT-PCR experiments.

Table S3. The flanking genes of the Nipponbare orphan genes used to detect syntenic regions in three close relatives, i.e., *O. rufipogon*, *O. sativa* ssp. *indica* 'Shuhui 498', and *O. nivara*. Minus and plus signs in the corresponding columns represent upstream and downstream flanking genes, respectively.

Table S4. RNA-seq expression values (FPKM) for Nipponbare orphan genes.

Table S5. Protein domain search for Nipponbare orphan genes by InterProscan program.

Fig. S1. Sequence alignment of *LOC_Os04g34130* from Nipponbare and the *Escherichia coli ECs4062* gene. The query sequence was *LOC_Os04g34130*, and the subject is represented as *E. coli*.

Fig. S2. Two orphan genes were verified via RT-PCR experiments from eight randomly selected orphan genes. The primer information is listed in Table S2. Amplification of the *Tubulin* gene was used as a loading control.

Fig. S3. Orphan genes originated from TE insertions. Red box with arrow: orphan gene and its orientation; orange box: TEs; gray line: gap in alignment; the green box is the ruler line: 100% identity in the alignment.

Fig. S4. Sequence alignments of five orphan genes origin from ENSPM4 expansion. The green box is the ruler line: 100% identity in the alignment.

Fig. S5. Sequence alignments of two orphan genes origin from SPMKIKE expansion. The green box is the ruler line: 100% identity in the alignment.

Fig. S6. Sequence alignment of two orphan genes origin from SZ-67LTR expansion. The green box is the ruler line: 100% identity in the alignment.

Fig. S7. The de novo formation of *LOC_Os04g22510*; The alignment is based on the coding sequence of the orphan gene and the orthologous noncoding sequences in the closely related outgroup. In the sequence alignments, the "-" symbol represents added empty positions in the alignment; the black box represents start codon creation in *LOC_Os04g22510*; the black star with a gray shadow represents the stop codon; the black arrows represents indel mutations; and the gray arrows represents point mutations.