nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-019-0822-5

Rapid evolution of protein diversity by de novo origination in *Oryza*

Li Zhang[1,7], Yan Ren[2,7], Tao Yang[3], Guangwei Li[3], Jianhai Chen[1], Andrea R. Gschwend[1], Yeisoo Yu[4], Guixue Hou[3], Jin Zi[2], Ruo Zhou[2], Bo Wen[2], Jianwei Zhang[4], Kapeel Chougule[4], Muhua Wang[4], Dario Copetti[4], Zhiyu Peng[2], Chengjun Zhang[1,5], Yong Zhang[6], Yidan Ouyang[3], Rod A. Wing[4]*, Siqi Liu[2]* and Manyuan Long[1]*

New protein-coding genes that arise de novo from non-coding DNA sequences contribute to protein diversity. However, de novo gene origination is challenging to study as it requires high-quality reference genomes for closely related species, evidence for ancestral non-coding sequences, and transcription and translation of the new genes. High-quality genomes of 13 closely related *Oryza* species provide unprecedented opportunities to understand de novo origination events. Here, we identify a large number of young de novo genes with discernible recent ancestral non-coding sequences and evidence of translation. Using pipelines examining the synteny relationship between genomes and reciprocal-best whole-genome alignments, we detected at least 175 de novo open reading frames in the focal species *O. sativa* subspecies *japonica*, which were all detected in RNA sequencing-based transcriptomes. Mass spectrometry-based targeted proteomics and ribosomal profiling show translational evidence for 57% of the de novo genes. In recent divergence of *Oryza*, an average of 51.5 de novo genes per million years were generated and retained. We observed evolutionary patterns in which excess indels and early transcription were favoured in origination with a stepwise formation of gene structure. These data reveal that de novo genes contribute to the rapid evolution of protein diversity under positive selection.

New genes that arise de novo from non-coding sequences[1,2] can dramatically enhance protein diversity above and beyond the well-known mechanisms of duplication/divergence and recombination of existing genic components[3–5]. It has been difficult to experimentally validate de novo origination, leading to early predictions against its validity[4,6]. This dissent arose due to the belief that, while recombination and duplication of pre-existing genes (or their parts that encode protein domains) can create new proteins, random non-coding sequences would be unlikely to generate novel functional domains[6–9], as they would not be seen as a unit of evolution. Despite such doubts, a number of de novo genes have been reported in an increasing number of organisms, from yeast[10,11] to plants[12–15] to *Drosophila*[2,16–21] to vertebrates[1,22–29]. However, reinspection of most reported de novo genes showed, with a few exceptions[1,23,26–28], that they can be defined just as orphan genes (that is, lineage-specific genes with no discernible homologous sequences in distantly related species[30,31]). Orphan genes can be derived from several distinct molecular evolutionary processes, such as extensive divergence of an ancestral orthologous gene[31,32], loss of homologous genes in related species[33], repeating of short peptides of low complexity[1,34], lateral gene transfer from fast-evolving donors (for example, from viruses and bacteria)[31,35] and de novo origination directly from ancestral non-coding sequences[1,2,36].

Major stumbling blocks for the routine identification of de novo-originated genes are twofold. First, there is a need for a set of high-quality reference genomes from closely related species that can be used to detect ancestral sequences before such sequences are scrambled beyond recognition as a consequence of genome turnover. Next, once a de novo gene is detected, there needs to be evidence for both transcription and translation[37], indicating that the new gene has the functional elements required to produce both messenger RNA (mRNA) and protein. Using *Oryza* genome datasets from 13 closely related species that were recently sequenced[38], we detected high abundance of young de novo candidate genes with clear evidence for ancestral non-coding sequences and translation.

Results

To detect and study the formation of de novo genes, we scrutinized the 13-genome data package of *Oryza* Genome Evolution and the International *Oryza* Map Alignment Project (OGE/IOMAP) by searching for and examining annotated homologous genes[38]. This resource consisted of high-quality genome assemblies and multitissue transcriptome data from 10 members of the genus *Oryza* that diverged within the most recent 12.1 Myr, including *O. sativa* subspecies *japonica* and *indica*, *O. rufipogon*, *O. nivara*, *O. glaberrima*, *O. barthii*, *O. glumaepatula*, *O. meridionalis*, *O. punctata*, *O. brachyantha*, as well as the outgroup *Leersia perrieri* from a grass genus diverged around 14.3 million years ago (Ma) (Fig. 1)[39]. All species were annotated using a common MAKER-P-based software pipeline[38]. RNA sequencing (RNA-Seq)-based evidence from multiple tissues and species, homology among species, ab initio prediction with various gene models, and repeat masking of transposable

[1]Department of Ecology and Evolution, The University of Chicago, Chicago, IL, USA. [2]BGI-Shenzhen, Shenzhen, China. [3]National Key Laboratory of Crop Genetic Improvement and National Centre of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan, China. [4]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ, USA. [5]Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China. [6]Key Laboratory of Zoological Systematics and Evolution & State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [7]These authors contributed equally: Li Zhang, Yan Ren. *e-mail: rwing@mail.arizona.edu; siqiliu@genomics.cn; mlong@uchicago.edu

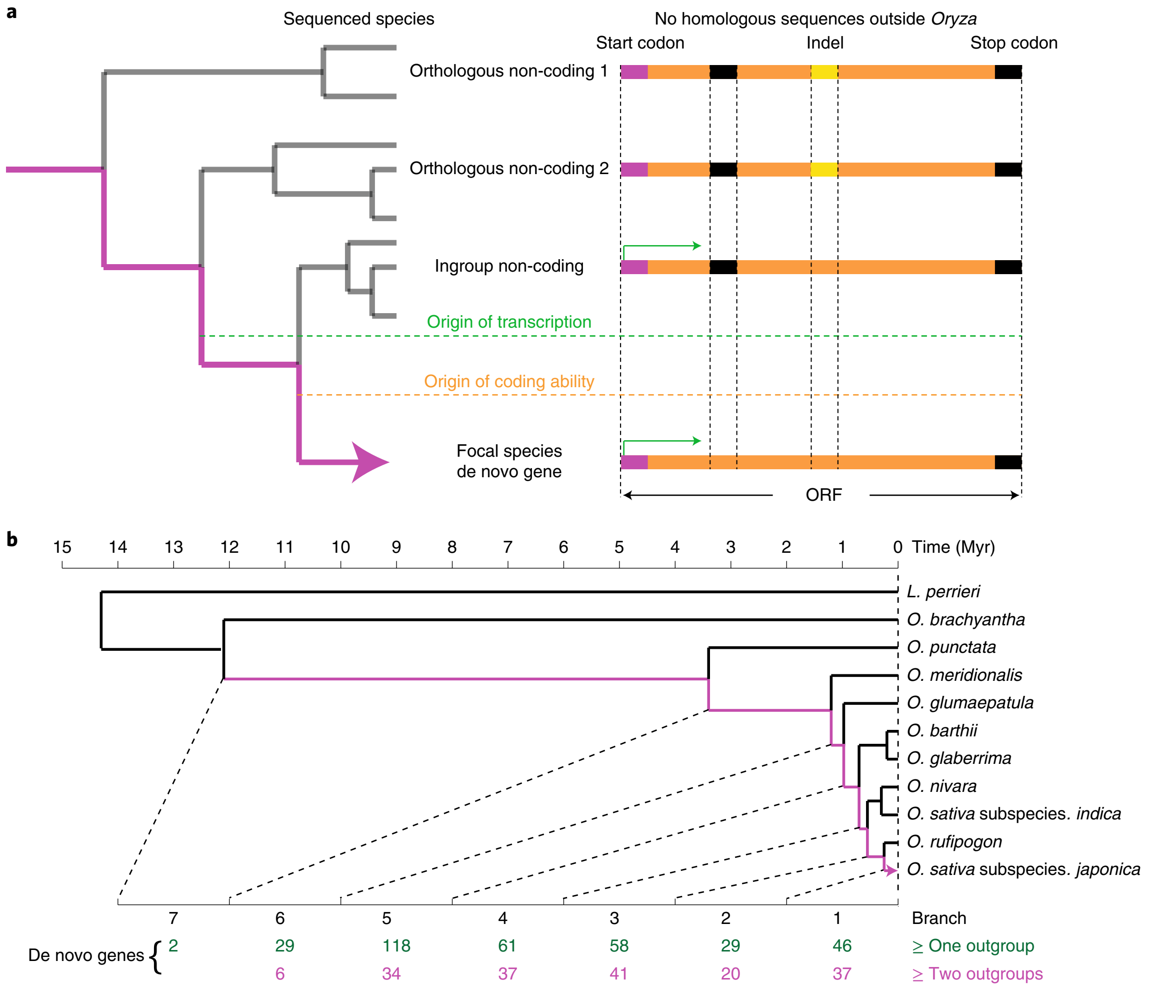

Fig. 1 | Identification of de novo genes that originated recently during *Oryza* diversification. a, Over evolutionary time, a non-coding sequence became a coding ORF due to an indel event (yellow → orange), followed by transcription being enabled (green dashed line) and the removal of a premature stop codon (black → orange). The general process of de novo origination was searched, using the *O. sativa* subspecies *japonica* ORF DNA sequences (orange block), starting with a regular translational start codon (ATG, purple) and ending with a termination codon (TAA, TAG or TGA, black). **b**, Evolutionary distribution of origination events for the identified de novo genes in the lineage towards *O. sativa* subspecies *japonica*, with the presence of at least one (green) and two outgroup non-coding sequences (purple). The divergence time was retrieved from the TimeTree database[39].

elements were incorporated by the MAKER pipeline to generate gene annotation.

We further compared the OGE/IOMAO assembly and annotation quality with two other databases of *Oryza* genomes—the MSU Rice Genome Annotation Project (MSU-RAP)[40] and the Rice Annotation Project Database (RAP-DB)[41]—using the evaluation method of evolutionarily considered BUSCO[42] (Methods). We found that the OGE/IOMAP showed very high completeness (detected as 96.1% complete genes from the total genes sampled to measure the completeness of assembly and annotation), revealing a very low proportion of fragmented and missing genes in the database. Furthermore, we observed that the high completeness of OGE/IOMAP was repeatable in MSU-RAP and RAP-DB, which showed high BUSCO complete metrics of 95.7 and 87.0%, respectively (Supplementary Fig. 1). These comparisons suggest high genome quality of the OGE assembly-annotated database for evolutionary analyses of de novo genes.

High rates of de novo gene origination. To exclude cases of functionalization of previously pseudogenized ancestral protein-coding genes, we started by identifying orphan genes in *O. sativa* subspecies *japonica* (the focal species for analysis) that have no homologous genes in the outgroup species *O. brachyantha* and *L. perrieri*. We searched the homologous sequences by examining the synteny relationship between genomes using reciprocal-best whole-genome alignments. We then sought evidence for orthologous non-coding sequences within the *Oryza* genus to identify recently evolved de novo genes (Fig. 1a). Before more evidence for their functionality was known, we initially denoted them as open reading frames (ORFs) instead of genes. We then searched for evidence of transcription in these ORFs, without considering intronic emergence due to possible expression in its non-coding ancestral state.

Based on these considerations, we developed a set of search pipelines and algorithms that can accurately extract orthologous non-coding sequences in outgroup species (Supplementary Fig. 2) to scan for de novo ORFs from the *O. sativa* subspecies *japonica* reference genome (hereafter, called the Nipponbare RefSeq). In this way, we identified 230 de novo ORFs (set 1) out of 38,757 OGE/IOMAP-annotated ORFs (Supplementary Table 1 and Supplementary File 1). Only ORFs with orthologous non-coding sequences in at least two outgroups (that is, three or more groups most recently diverged outside the ORF-containing ingroup that have non-coding orthologous sequences) were classified as de novo ORFs, as opposed to ORFs originating from gene loss in outgroup species, following the parsimony method that is well applied to the identification of new genes[43]. This approach yielded a 0.9% (8/929) false positive rate based on internal gene loss event estimates. Some 455 de novo ORFs (set 2) were also identified when we required orthologous non-coding sequences in one or more outgroups (that is, two or more species most recently diverged outside the ORF-containing ingroup, with an increased false positive rate of 7.6% (71/929)) (Supplementary Table 1).

We then looked for transcriptional evidence from the OGE/IOMAP-derived transcriptome data and Nipponbare RefSeq gene annotations[41]. The transcriptome data have, on average, 800× depth per sample, which is sufficient in most cases to detect lowly expressed genes. Meanwhile, the Nipponbare RefSeq annotations are expected to cover most rice genes since they were generated based on the most current available supporting data, including full-length complementary DNA (cDNA), expressed sequence tag (EST) and proteomic data. Of the 230 candidate de novo ORFs, we confirmed that 201 have transcriptional evidence that can be detected in at least one *Oryza* species in OGE/IOMAP or Nipponbare RefSeq datasets (Supplementary Table 2). These 201 candidate de novo genes are statistically much greater in number than expected de novo ORFs based on the random shuffling of nucleotides of intergenic regions in the focal species *O. sativa* subspecies *japonica* ($P = 1.032 \times 10^{-7}$; Poisson with $\lambda = 1$; Methods).

All except one of the candidates were supported by OGE/IOMAP transcriptome data. However, the Nipponbare RefSeq-derived evidence covered only 25 candidate de novo genes. The other 176 candidate de novo genes may have a reduced expression level and therefore went undetected in this dataset, whereas the high-coverage OGE/IOMAP transcriptome dataset has captured transcription not detected in the currently available gene expression datasets. This observation is consistent with the relatively lower expression patterns detected from previously identified new genes[44,45]. Of the 25 genes with Nipponbare RefSeq-derived evidence, 96% (24/25) can be found in transcriptome data, suggesting that the OGE/IOMAP-derived evidence is well supported by standard Nipponbare RefSeq gene annotations. Among the 200 candidate de novo genes with OGE/IOMAP transcriptome support, 74.5% (149/200) were detected in ≥3 species.

Because the sequence that matches with distant species may be shaped by horizontal transfer or evolution from ancestral sequences, we purged de novo gene candidates with even lower levels of homologous matches (BLASTP e value: 0.01 against the non-redundant database) in the species outside the *Oryza* genus to avoid any potential cases of pseudogenization from ancient genes (Supplementary Table 3). In total, by employing a rigorous requirement for the absence of de novo genes in two or more outgroup species, we identified the presence of 175 *O. sativa* subspecies *japonica* de novo genes (set 1) that emerged after the divergence of *O. punctata* ~3.4 Ma[39], suggesting that on average 51.5 de novo genes per million years were generated and retained in the genomes of extant species. Using the same criteria above, we identified 343 candidate de novo genes (set 2) supported by OGE/IOMAP transcriptome data (their phylogenetic distribution is shown in green in Fig. 1b and Supplementary Table 4), which resulted in a much higher rate of 100 de novo genes per million years since the divergence from the *O. punctata* lineage. Set 2 contains many more de novo genes in every branch compared with set 1 (Fig. 1b). In the following sections, we will present results mainly from the analyses of set 1 data (relevant analyses on the larger set 2 dataset are reported in Supplementary Tables 1–4).

Stepwise origination process of de novo genes. Indels and substitutions are two ORF triggers (also called enablers[26,27]) that can transform non-coding sequences into protein-coding genes. Indels create frameshifts that can quickly restructure a non-coding sequence into a proper ORF. Substitutions can cause point mutations that may lead to the creation of start codons or exon splicing sites, or the removal of premature stop codons. Analysing the stepwise origination processes of the hundreds of identified de novo genes provides insight into the roles of these evolutionary ORF triggers in the formation of de novo genes. We reconstructed the stepwise evolutionary processes for each of these de novo genes by comparing the sequence of the novel ORF with its closest outgroup non-coding and ingroup potential coding sequences. These analyses detected crucial changes leading to the eventual transformation from non-coding to coding sequences (identified mutations are summarized in Supplementary Table 4 and can be found in sequence alignments in Supplementary File 1).

The *Osjap05g30030* and *Osjap06g21910* genes are two examples showing a stepwise process to form de novo genes. *Osjap05g30030* (Fig. 2) was found to be transformed de novo gradually in six steps, becoming a new gene as the result of one frameshift mutation and one substitution inside a premature stop codon (Supplementary Fig. 3). In contrast, *Osjap06g21910* (Fig. 3) represents an older de novo gene than *Osjap05g30030*, which can be traced back to an ancestor predating *L. perrieri* (Fig. 3 and Supplementary Fig. 3). It transformed through a distinct evolutionary process that comprised four steps. As opposed to *Osjap05g30030*, which acquired its transcription ability after the origin of the de novo ORF, *Osjap06g21910* acquired expression at a significantly earlier stage, predating *L. perrieri*.

Summarizing the ORF triggers identified in the 175 de novo genes, in total, we identified 251 ORF triggers, including 200 frameshifts (indels) and 51 ORF triggers that dispensed premature stop codons by substitution, which were crucial for the transformation of non-coding to coding sequences in these genes (Supplementary Table 4). Of the 175 *O. sativa* subspecies *japonica* de novo genes, 18 were formed by both indel and substitution mutations, 123 by indels only and 30 by substitutions only, with 4 genes undetermined. Indels are usually one order of magnitude less common than substitutions in the standing variation of genomic changes in populations of metazoans[46], Asian cultivars[47] *O. sativa*, their wild direct ancestors *O. rufipogon*, the African cultivar *O. glaberrima* and its ancestor *O. barthii*[48].

To detect a pattern in the distribution of indels and substitutions, we compared the evolutionary divergence with the within-species variation of indels and substitutions using a statistical test similar to the McDonald–Kreitman test[49,50]. The genomic differences between indels and substitutions in a population, such as *O. sativa*[47] and the wild species *O. barthii*[51], after normalization by sample size (that is, numbers of accessions) as θ_w (which is defined in Supplementary Table 5), reflect the differences among the mutation rates of indels and substitutions[49,52]. The null hypothesis of neutrality predicts that the rates of mutation and evolution are equal[50], thus predicting a similar pattern of substitution/indel ratios in both the polymorphic stage and fixed stage. The standing variation in *O. sativa* and *O. barthii* genomes showed 11.91 (14.54 for intergenic regions) and 12.47 times more substitutions, respectively, than indels (Supplementary Table 5). Therefore, the neutrality hypothesis predicts that the vast majority of the 251 ORF triggers should be the substitution type. This prediction is the opposite of the observed pattern: the vast majority (200/251 = 79.68%) are indel ORF triggers in the de novo genes (*G*-test of independence: $G = 850.34$; $P = 2.2 \times 10^{-16}$; Supplementary Table 5 and Methods). This result suggests powerful positive selection for the use of indel mutations in the formation of de novo genes.

Patterns of de novo gene origination. Generally, de novo origination of a new gene can evolve in three different ways with regards to the temporal order in the appearance of transcription and coding

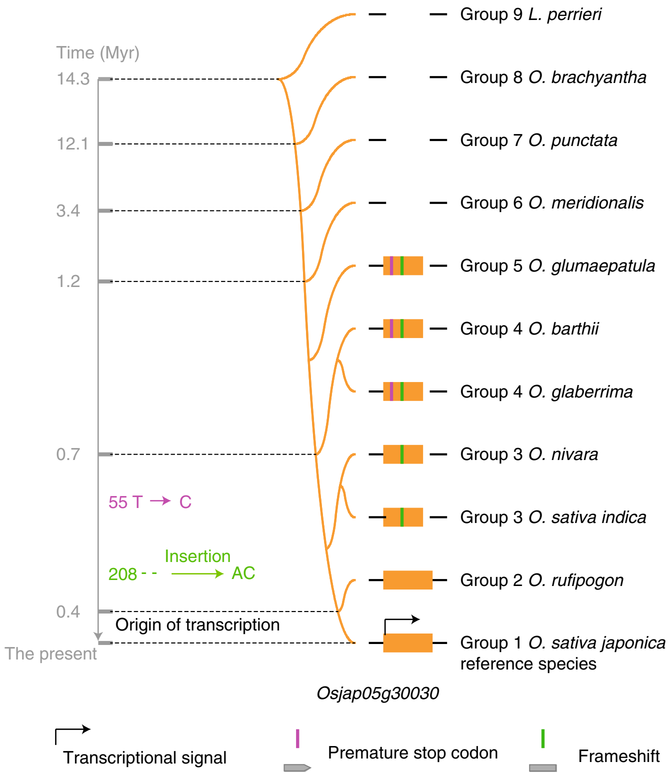

Fig. 2 | Stepwise origination processes for the de novo gene *Osjap05g30030*. The black arrow over the reference species gene indicates that a gene expression pattern was detected. Purple and green bars show a premature stop codon and frameshift from the ORF in the de novo gene, respectively. Base pair substitutions converted a premature stop codon into an amino acid-encoding codon, whereas an indel created a new ORF. The most distant orthologous non-coding sequence of *Osjap05g30030* is present in *O. glumaepatula*, which shares one premature stop codon with the orthologous non-coding sequences in *O. glaberrima* and *O. barthii* and one 2-bp frameshift with orthologous sequences in *O. glaberrima*, *O. barthii*, *O. sativa* subspecies *indica* and *O. nivara*. Its expression is detected in *O. sativa* subspecies *japonica*, suggesting a recent origination of gene expression even after the formation of the ORF. It appears that *Osjap05g30030* was generated de novo by experiencing six evolutionary stages, as follows. (1) An orthologous non-coding sequence probably appeared initially in the most recent common ancestor (MRCA) of groups 1, 2, 3, 4 and 5, with a later-used start codon, one premature stop codon in the first exon, one frameshift in the second exon and one 78-bp fragment missing at the 3′ end, compared with the functional de novo gene. (2) After *O. glumaepatula* and *O. barthii*, along with its African-derived cultivar *O. glaberrima*, diverged independently, both the frameshift and premature stop codon persisted. (3) After the two species of African *Oryza* diverged, the premature stop codon was restored by the aforementioned T → C substitution. (4) After the divergence of *O. sativa* subspecies *indica* and *O. nivara*, the frameshift was resolved by the 2-nucleotide insertion discussed above in the MRCA of groups 1 and 2. (5) During the period in which the 2-nucleotide insertion occurred, a fragment of 78 nucleotides was inserted at the 3′ end in the MRCA of groups 1 and 2, providing a translational termination codon. These molecular changes resulted in the generation of a new ORF in the MRCA of groups 1 and 2, but this was transcriptionally silenced. (6) Finally, the MRCA of *O. sativa* subspecies *japonica* acquired a distinct expression pattern and turned *Osjap05g30030* into an active de novo gene.

ability, as described in the early ORF–late transcription model[11,21], late ORF–early transcription model[10,19,20,28] and simultaneous ORF transcription model) (Fig. 4a). Since both the coding and transcription ability of each de novo gene in each species were known by inference from the 13-genome OGE/IOMAP datasets, we compared the coding and transcription states in each phylogenetic group and found that transcription ability emerged earlier than coding ability in most cases. Of 175 de novo genes, 159 were initially transcribed as non-coding transcripts, 10 first appeared as de novo ORFs but were not transcribed, and 6 obtained their coding and transcription ability in the same species or branch group (in the case of sister species) (Fig. 4a and Supplementary Table 4). In other words, we observed that 90.9% of de novo genes were derived from non-coding transcripts, which is statistically dominant compared with the other two models ($\chi^2 = 122$; $P < 0.0001$). These data suggest that non-coding transcripts, instead of pre-existing ORFs, are the major source of de novo genes. This late ORF–early transcription observation is consistent with the reported pervasive transcription in genomes[53], suggesting that non-coding RNAs or short peptides may serve as intermediate targets of adaptive selection[28,54–56].

To study the expression pattern of de novo genes, we focused on the 175 de novo genes identified above and a set of 4,965 old (predating *L. perrieri*) singleton genes expressed in *O. sativa* subspecies *japonica* leaf, panicle and root tissues. Here, old singleton genes are defined as *O. sativa* subspecies *japonica* genes that have only one orthologous sequence copy in each species, including the 10 *Oryza* species and *L. perrieri*. We analysed fragments per kilobase of transcript per million mapped reads (FPKM) values for 112 de novo genes and 4,955 old singleton genes that were identified in at least 1 of the 3 tissues (Supplementary Table 6). Figure 4b shows that de novo genes have lower expression patterns in all 3 tissues compared with old singleton genes ($P < 0.0001$ for all 3 between-tissue comparisons, Wilcoxon rank-sum test with continuity correction), although a few de novo genes have become highly expressed. Then, we examined the expression specificity of de novo genes using a specificity score defined by Yanai et al.[57]. Compared with old singleton genes, de novo gene expression is highly tissue specific, even with only 3 tissues, as in this analysis (Fig. 4c; $P < 0.0001$, Wilcoxon rank-sum test with continuity correction). In particular, we identified many de novo genes that are expressed in the root and leaf, adding to past studies that detected narrower expression in grass and *Arabidopsis*[13,14].

New orphan genes have been shown to be short and to contain mostly single exons, based on a limited dataset of a few probable authentic de novo genes in mammals[23,26,27]. Our large dataset of de novo genes in *Oryza* should therefore reveal more repeatable patterns associated with the structure of de novo genes. To accomplish this task, we compared the average ORF length, isoform number, exon number, exon length and gene length (including exons and introns) of the 175 de novo genes with those of the 4,955 old singleton genes and all 38,757 OGE/IOMAP annotated gene datasets (Supplementary Table 7 and Fig. 4d,e). This comparison yielded numerous statistically significant findings (Supplementary Table 7). First, de novo genes have shorter ORFs than old singleton genes. Second, de novo genes have fewer isoforms than old singleton genes. Third, de novo genes have fewer exons than old singleton genes (two- and three-exon de novo genes, instead of single-exon de novo genes, were frequent, possibly facilitating gene recombination for a greater protein diversity in *Oryza* than other organisms[58]). Fourth, de novo genes were found to have shorter exons than old singleton genes. Finally, de novo genes have longer gene lengths than the genome average at a marginal level ($P = 0.1084$), but are similar to old singleton genes. These observations reveal the stepwise evolution of de novo gene structures: de novo genes gradually recruited more exons, expanded their exon lengths and derived more isoforms.

To further understand the evolution of complex gene structure that was impacted by the evolutionary dynamics of exons and

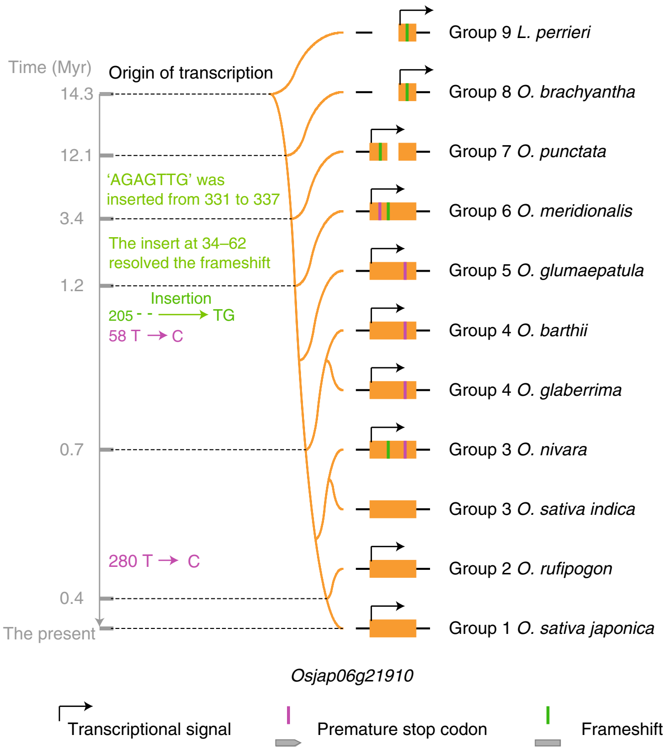

Fig. 3 | Stepwise origination process for the de novo gene *Osjap06g21910*. Black arrows indicate that a gene expression pattern was detected. Purple and green bars show a premature stop codon and frameshift from the ORF in the de novo gene, respectively. Differing from *Osjap05g30030*, the non-coding ancestral sequences of *Osjap06g21910* acquired an expression pattern at an early stage via a four-step process, as follows. (1) One insert resolved the frameshift at the 3′ end in *L. perrieri* and *O. brachyantha*, while initial sequence expansion with accumulated indels created a new frameshift at the 5′ end in *O. punctata*. (2) Continuous sequence expansion with accumulated indels resolved the frameshift in *O. punctata* but created one new frameshift and one premature stop codon in *O. meridionalis*. (3) One 'TG' insert resolved the frameshift, and one 'T → C' substitution resolved the premature stop codon in *O. meridionalis*. However, another 'C → T' substitution created a new premature stop codon in *O. glumaepatula*. (4) The premature stop codon in *O. glumaepatula* was preserved in African rice and resolved in Asian rice by one 'T → C' substitution. The evolutionary history of the Asian rice group is complex, since *O. nivara* retained the premature stop codon and one additional derived frameshift, while *O. sativa* subspecies *japonica*, *O. rufipogon* and *O. sativa* subspecies *indica* possess the de novo ORF. Considering the frequent introgression among Asian cultivated rice[106], a possible scenario could be the introgression of this de novo ORF into *O. sativa* subspecies *indica*. Sequence alignments can be found in Supplementary Fig. 2.

introns[58], we examined the intron phases of de novo genes by taking advantage of the fact that many de novo genes contain one or more introns. We calculated the intron phase (that is, the position of an intron between the codons or within a codon after the first and second nucleotides, respectively[59]) for each of 362 introns in the 175 de novo genes. We found that the distribution of phase 0, 1 and 2 introns follows a ratio of 175:85:102 (Supplementary Table 8). This distribution differs significantly from an equal-probable distribution of intron phases comparing the frequency of the between-codon position (phase 0) with the within-codon after the first nucleotide position (phase 1) and the within-codon after the second nucleotide position (phase 2) ($P = 0.0007$, Kolmogorov–Smirnov test). Our observation suggests that phase 0 introns are most common, even for very recently derived de novo genes, similar to intron phase distribution in metazoans[58]. Interestingly, these introns, as exons, were derived from non-coding ancestral sequences as well.

Evidence for the functionality of de novo genes. Although the de novo gene candidates were identified from the strictly annotated genes in the 13 *Oryza* genomes using a uniform annotation pipeline that minimized potential methodological artefacts[38], we further examined these gene candidates for their functionality and, especially, their translational evidence. We examined the potential functionality of these de novo gene candidates with several lines of evidence by characterizing their structure, expression and evolutionary constraints. First, all candidates have intact ORFs that are, on average, 137 amino acids long, with GC content typical of protein-coding genes (59.1% compared with a 56.8% genome average; Supplementary Table 4). Second, every de novo gene candidate has evolved a tissue-biased or -specific expression pattern, just like the other functional genes in the genome (Supplementary Table 6). Third, all candidates have intact gene structures.

We analysed the sequence evolution to detect substitution signals and determine the functionality of de novo gene sequences. We applied the branch model in PAML[60] to 236 candidate de novo genes that have ≥ 3 orthologous sequences, with the aim of identifying genes that showed a signal of natural selection as detected in their sequence substitutions at synonymous sites (d_S), non-synonymous sites (d_N) and $\omega = d_N/d_S$. The likelihood ratio test and Akaike information criterion (AIC) identified 28 candidate de novo genes that are incompatible with the model of neutrality, with ω either significantly lower than 1 (22 genes) or higher than 1 (6 genes), suggesting that they may undergo negative or positive selection (Supplementary Table 9). In the 45 candidate de novo genes that have only 2 orthologues, we detected 2 genes with ω significantly lower or higher than 1, whereas most of them had ω ratios lower than 1. Together, we detected 30 candidate de novo genes with substitution signals of negative or positive selection. The remaining genes had lower statistical power due to small numbers of substitutions (Supplementary Table 9). These results support the coding potential of the candidate de novo genes, prompting us to further explore experimental evidence for their translation.

We experimentally verified the protein products of candidate de novo genes. Considering the potential presence status of de novo proteins in tissue-dependent mode[13–15], which the orphan genes were more often observed to express in pollens and anthers, we adopted a powerful and sensitive proteomics method based on mass spectrometry with multiple-reaction monitoring (MRM-MS[61,62]; see Supplementary Materials), including protein targeting and synthesized control peptide matching, to capture the peptides derived from de novo genes. We tested three tissues dissected in the plants that were grown in the rice field research station in Hainan, including anthers, pistils and glumes, and flag leaves in the focal species *O. sativa* subspecies *japonica*. We used the MRM-MS method to identify the candidate de novo genes that were expressed at the protein level. Figure 5 provides an example of the identification of two peptides ('TFFDVGSATGGGVPR' and 'FTLILLNGAPR') in the candidate de novo gene *Osjap05g20760* from the three tissues in *japonica*, confirmed by the synthesis peptide, as shown by the overlays of MRM-MS signals. *Osjap06g21910* (Fig. 3) is another example with two peptides ('EDEGDKPEVEVK' and 'VGGSSILAYNALANNSGE') detected (Supplementary Table 10).

Using the targeting MRM-MS method, we successfully identified 109 unique peptides from the 167 peptides detected over the 3 tissues for 36.6% (64/175) of set 2 de novo gene candidates and 30.3% (104/343) of set 1 de novo gene candidates. The sizes of these peptides ranged from 8–23 amino acids (Supplementary Table 10). The vast majority of these genes (79 genes) were detected at least twice by different peptides or in different tissues, with the peptides

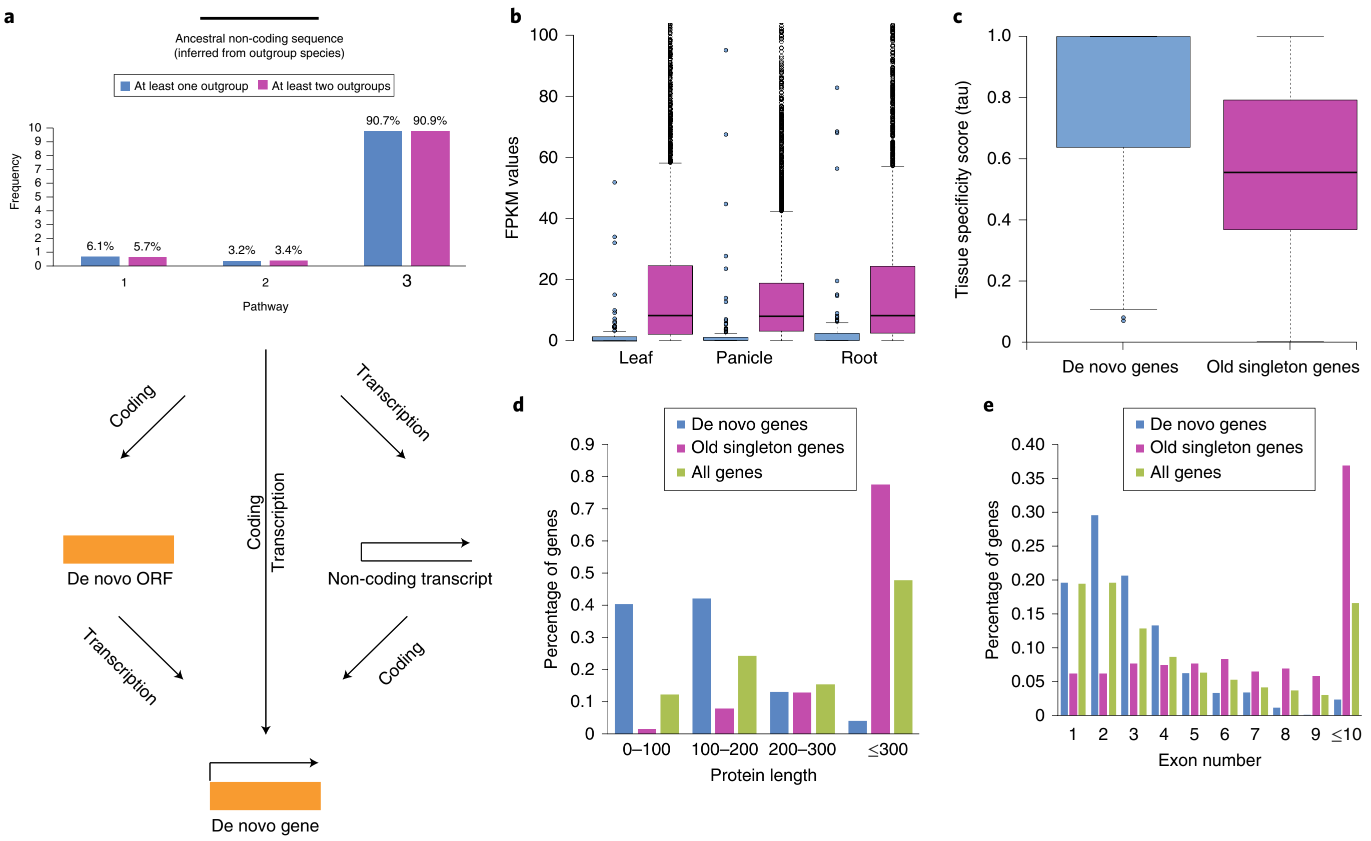

Fig. 4 | Patterns of de novo origination in evolution, expression and gene structures. a, Observed frequencies of three pathways of de novo origination from inferred ancestral non-coding sequences in 175 de novo genes: (1) the early ORF–late transcription model[11,21] (mutations transformed a non-coding sequence into a de novo ORF in the first step, after which newly recruited regulatory elements activated the de novo ORF into a de novo gene in the second step); (2) the simultaneous ORF transcription model; and (3) the late ORF–early transcription model. **b**,**c**, Expression patterns: de novo genes (blue) have reduced expression (**b**) and are more tissue specific (**c**) compared with old singleton genes (purple). A Wilcoxon rank-sum test with continuity correction was used to determine significance (leaf: $P<0.0001$; panicle: $P<0.0001$; root: $P<0.0001$ in **b**; $P<0.0001$ in **c**). **d**, Compared with old singleton genes or all genes combines, de novo genes are shorter. **e**, Although de novo genes tend to have fewer exons, the two-exon structure is most abundant.

of 25 genes being identified once, 53 genes 2–4 times and 26 genes 5–10 times (Fig. 6a). Figure 5b shows the distribution of protein expression of the 104 candidate de novo genes in the tissues of *O. sativa* subspecies *japonica*. Notably, there were excess genes (24 in *O. sativa* subspecies *japonica*) with tissue-specific protein expression in the male reproductive tissue (anther), while few genes were found that express specifically in the leaf or female tissues (Fig. 6b). For example, the candidate de novo genes *Osjap02g40470* and *Osjap01g09610* were identified to generate 4 and 3 peptides in 3 tissue samples, respectively (Fig. 6c). We further artificially synthesized 44 peptides (40.4% of the total 109 peptides), and 37 of them provided positive support to their corresponding peptides in samples, with a retention time shift of less than ±3 min in a 3 h gradient (<2.5%) and a transition abundance change of <20% in individual transitions (as required by the Commission Decision 2002/657/EEC[63]). Using these synthesized peptides, we confirmed that 84.1% of detected peptides were true peptides (Fig. 5 and Supplementary Fig. 4). These analyses suggest that the vast majority of the 109 peptides detected in the MRM-MS experiments exist in the tissues (Supplementary Figs. 4 and 5).

We also retrieved eight nanoscale liquid chromatography tandem mass spectrometry (MS/MS) datasets from ProteomeXchange, prepared from seeds, pistils and pollens, ovaries and seedlings/flowers in *O. sativa* subspecies *japonica* (Supplementary Table 11), and analysed previously unanalysed spectra with IPeak[64], to find possible matches to our candidate de novo gene set. We detected an additional 6 peptide signatures with lengths of 9–14 amino acids matching to 6 candidate de novo genes in 4 spectra datasets. *Osjap02g25860*—1 of these 6 genes—is the same as 1 of the 104 genes identified experimentally above, but identified by a different unique peptide (Supplementary Table 12).

We examined additional evidence for the translational potential of the candidate de novo genes by analysing published ribosomal profiling data[65] from *O. sativa* complementary to the direct proteomic evidence above. The well-developed method 'translating ribosome affinity purification followed by mRNA sequencing' (TRAP-Seq) from the p35S:HF-OsRPL18 transgene strain provided the inferred translatomes in calluses, panicles and seedlings[65]. We pooled the three samples together and assembled them de novo into strand-specific transcripts for TRAP-Seq with Trinity[66] (Methods). We identified 130 de novo genes that were associated with ribosomes in the test of TRAP-Seq. A further comparison of the 130 de novo genes with the gene list identified by the MRM-MS analysis revealed that 45 were on the lists of both ribosomal profiling and proteomics (Fig. 5b, Supplementary Table 13 and Supplementary File 2). To understand the regulation of translation of de novo genes, we investigated adaptation of the codons of the de novo genes to the transfer RNA (tRNA) pool of *O. sativa*. A computation of tRNA adaptation index (tAI) values[67] (Methods and Supplementary Table 14) revealed a slightly lower but statistically significant mean value for de novo genes compared with old genes (0.3653431 for 175 de novo genes (182 isoforms) and 0.3756544 for 4,965 single-copy

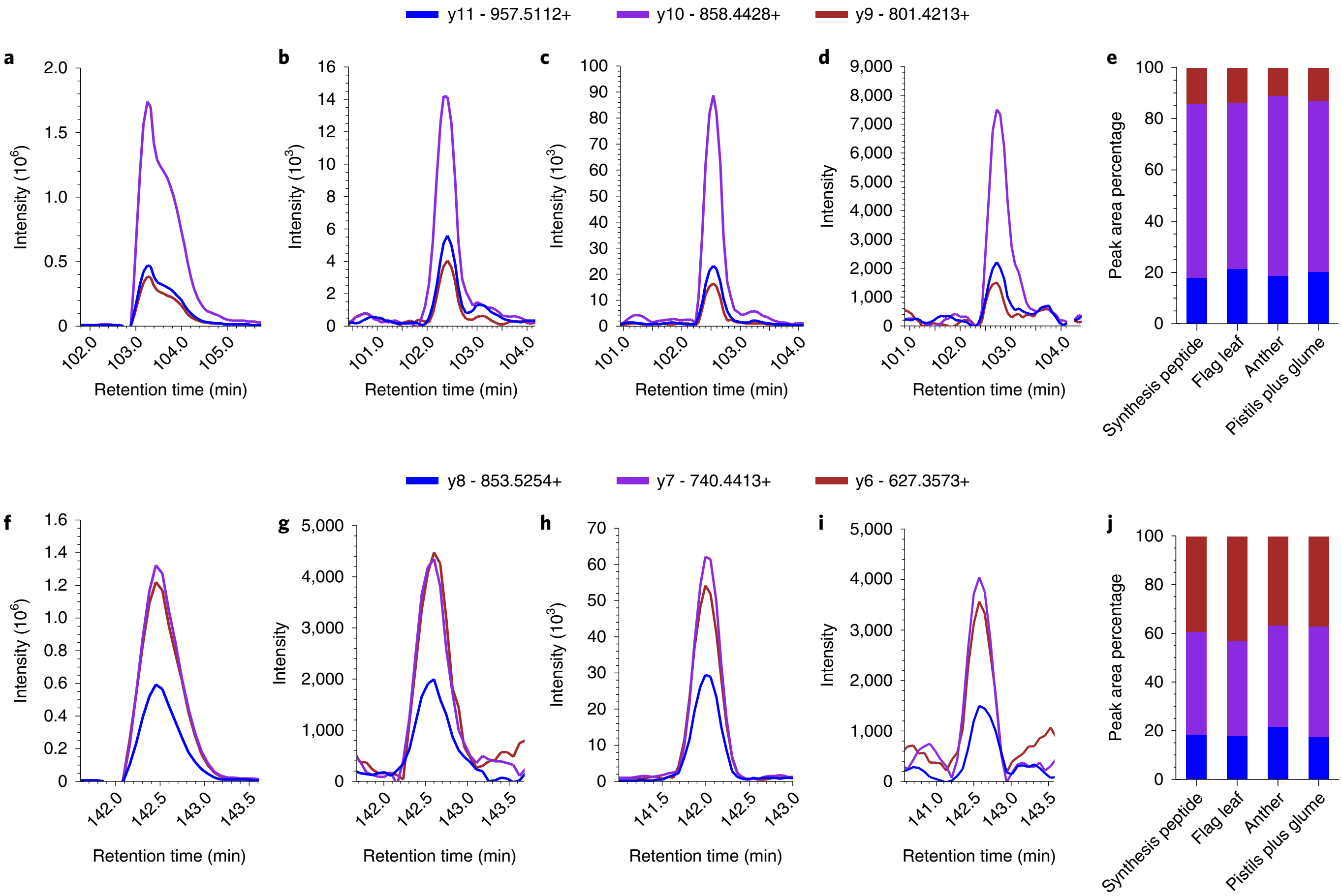

Fig. 5 | Example of the verification of protein products translated from a candidate de novo gene, *Osjap05g20760*. MRM-MS was used to verify protein products in three tissue types from *O. sativa* subspecies *japonica*. **a**–**j**, Overlays of MRM signals for 'TFFDVGSATGGGVPR' (**a**–**d**) and 'FTLILLNGAPR' (**g**–**i**) acquired from synthesized peptides (**a** and **f**), flag leaves (**b** and **g**), anthers (**c** and **h**) and pistils and glumes (**d** and **i**), and the distribution of the transition ion intensities for each peptide in the different samples (**e** and **j**, respectively). The parameters, including *m/z* ratios for the mass spectrometry spectra, are given in the legends.

genes (7,044 isoforms); ($P<0.0001$)). This suggests that de novo genes have lower adaptation to the tRNA pool that control translation, in accordance with their young ages.

In summary, we found evidence that 194 of the de novo genes (194/343 = 56.6%) were translated, suggesting great potential for the use of these approaches in the identification of novel proteins encoded by de novo genes in plants, in addition to the identification of proteins translated from a few de novo genes in yeast[10], humans[23,26,28] and rodents[27].

Discussion

This analysis of de novo origination in the genus *Oryza* detected critical evidence in both ancestral non-coding origination and translational products of de novo genes. We identified at least 175 candidate de novo genes with significantly similar ancestral non-coding sequences, implicating an unexpectedly high rate (51.5 de novo genes per Myr) of generation and retention of de novo genes in the genomes of young *Oryza* species compared with reported lower numbers of de novo genes in metazoans[1,2,16–29] and plants[12–15]. These observations show that de novo genes are important in the evolution of protein diversity in *Oryza*. These data reveal the power of evolution in natural populations in the generation of de novo genes that arise from non-coding sequences, probably via various possible molecular processes over time[54,68], compared with artificial synthesis and selection[69,70]. We also note that the observed high rate of de novo gene generation and retention in the recent evolution of *Oryza* might have been in equilibrium with the loss of unrelated genes in genome evolution, as was previously detected in cereal species[33]. We anticipate that, whether or not such dynamics of gene evolution can be extrapolated beyond the *Oryza* genus backwards to farther ancestral lineages of Poaceae, monocots and angiosperms will be an important problem to explore in the future.

At first glimpse, the rapid evolution of new genes in *Oryza* is associated with a remarkable feature of *Oryza* species: they experience a strikingly high level of diversity in ecological environments where they preside[71], although this level is still not adequate to interpret the high rate of generation and retention detected in the recent evolution of *Oryza*. Recent genomic analyses of the history of evolution of plants, especially *O. sativa* and other grass species, have revealed much higher adaptive plasticity than in animal genomes, in support of the theoretical predictions of plant plasticity[72,73]. These data suggest that positive selection for the highly excess indel ORF triggers, as our McDonald–Kreitman test showed, may have prepared a more gene-like structure in the first place, as observed similarly in mice and yeast[74]. We detected a small proportion of genes that present significant d_N/d_S signals of natural selection. Over their short lives, these de novo genes experienced a limited number of substitution events for a powerful test of selection or, alternatively, many of them are still driven by genetic drift in neutral evolution.

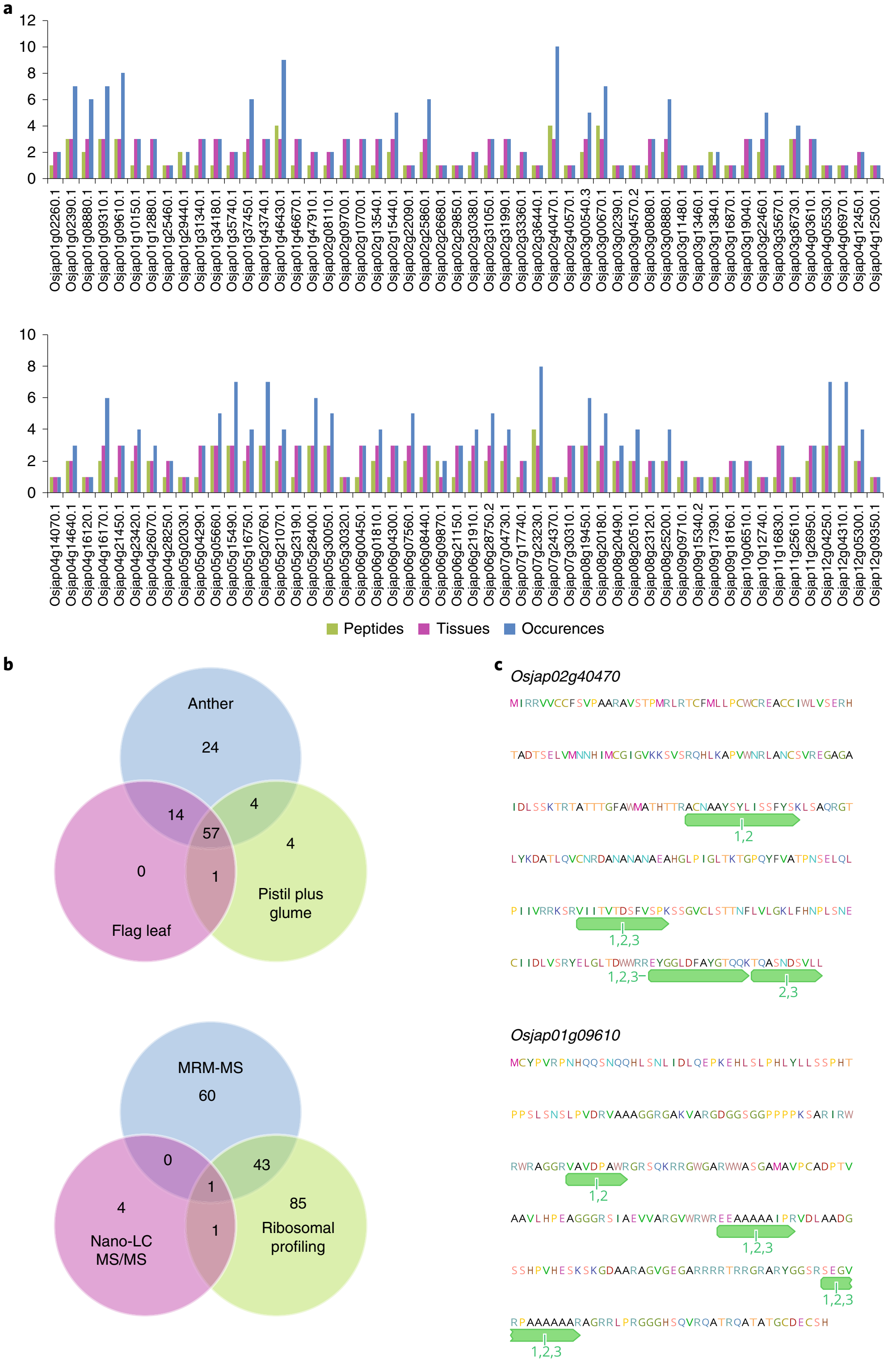

Fig. 6 | Summary of the protein products translated from candidate de novo genes in *O. sativa* subspecies *japonica*, as detected by experimental proteomics and ribosomal profiling analyses. **a**, Frequency distribution of the peptides identified using the MRM-MS-based technique. Numbers detected by different peptides and in different tissues are shown, along with total occurrences. **b**, Venn diagrams of the distribution of candidate de novo genes identified using proteomics and ribosomal profiling analyses. Top, distribution of protein expression for the 104 candidate de novo genes in anthers, pistils and glumes, and flag leaves, as detected by MRM-MS. Bottom, distribution of protein expression for the 194 candidate de novo genes detected using MRM-MS, nanoscale liquid chromatography MS/MS (nano-LC MS/MS) and ribosomal profiling. **c**, Examples of two candidate de novo genes, *Osjap02g40470* and *Osjap01g09610*, that translate to proteins. The peptides (green arrows) detected by MRM-MS in the anthers (1), flag leaves (2) and pistils and glumes (3) are indicated.

Overall, this study detected an unexpectedly important role of de novo origination as a mechanism responsible for protein diversity, adding fresh evidence of ancestral origination and translation to previous observations of (and discussions about) de novo genes[30–32,75].

Methods

Species groups, genome assemblies and annotations. Based on the *Oryza* phylogenetic tree, the 11 species were assigned to 9 branch groups (Fig. 2): *O. sativa* subspecies *japonica* (group 1); *O. rufipogon* (group 2); *O. sativa* subspecies *indica* and *O. nivara* (group 3); *O. glaberrima* and *O. barthii* (group 4); *O. glumaepatula* (group 5); *O. meridionalis* (group 6); *O. punctata* (group 7); *O. brachyantha* (group 8); and *L. perrieri* (group 9).

Genome assemblies and annotations were retrieved from the OGE/IOMAP 13-genome package[38]. The expression evidence from multiple tissues and multiple species, homology among species, ab initio prediction with various gene models, and repeat masking of transposable elements were incorporated by MAKER to generate gene annotation. IRGSP gene annotations supported by full-length cDNA, EST or protein data were downloaded from RAP-DB[41]. The minimum and maximum ORF lengths were 32 and 5,436 amino acids, respectively, with an average of 294 amino acids and a median of 291 amino acids. The independent evidence-supported shortest intron length was 16 nucleotides, and the maximum intron length was 268,114 nucleotides, with an average of 615 nucleotides and a median of 196 nucleotides.

To avoid potential bias in assembly and annotation against evolutionary novelties, as was previously observed in major public gene content databases[45], we further evaluated the completeness and repeatability of three *Oryza* genome databases created and maintained by independent groups: OGE/IOMAP[38], MSU-RAP[40] and RAP-DB[41]. We used BUSCO (version 3.0.1)—a widely used method based on phylogenetically distributed single-copy orthologues—to evaluate assembly annotation quality[42]. Relevant to the OGE analysis, this method was designed with consideration for evolutionary expectations.

Detection of de novo ORFs. First, we identified a large number of new genes in the genus *Oryza* using *O. punctata* and *O. brachyantha*, which diverged around 3.4 and 12.1 Ma, respectively, as outgroup species[39], and *L. perrieri* from a grass genus that diverged around 14.3 Ma as an outgroup outside the *Oryza* genus[39]. Given the young ages of these new genes (mostly <3.4 Ma), we inferred that if there are de novo genes among them, we might be able to detect their ancestral non-coding sequences from which de novo genes were explicitly derived. We designed a set of pipelines and algorithms to detect de novo ORFs whose origination processes occurred after the *Oryza* lineage divergence, meaning de novo genes arose and were shaped into intact and functional genes within 15 Myr. To facilitate comparison with other organisms, we chose the most widely used TimeTree Database (http://www.timetree.org) for the estimated divergence times of involved species. The major innovation of our algorithm is that we considered the coding ability of orthologue sequences outside the reference species (*O. sativa* subspecies *japonica*), which were directly treated as coding genes in previous pipelines[76,77]. An overview of the de novo ORF discovery pipelines is provided in Supplementary Fig. 2. We note that the in-exon disruptive mutation in Supplementary Fig. 2 is defined as the mutation in non-focal species that disrupts an exon in the start codon, stop codon or splicing sites. In-exon mutations that create premature stop codon(s) and/or frameshift(s) are considered as evidence of orthologous non-coding sequences.

We compared synteny relationships among genomes to identify orthologues using the reciprocal-best-whole-genome alignment method. We started with 38,757 predicted annotated ORFs in *O. sativa* subspecies *japonica* and used protein-to-protein BLAT[78] to scan the annotated gene sequences for all 10 of the other OGE/IOMAP species. If exact matches covered at least 20% of the amino acids corresponding to the *O. sativa* subspecies *japonica* ORF, 1 effective hit was accepted. Only *O. sativa* subspecies *japonica* ORFs with no effective hits in *O. brachyantha* and *L. perrieri* (the outgroup species), and no more than one effective hit in the remaining eight species (to avoid ambiguous cases, such as gene fusion and fission), were identified as potential orphan ORFs. This search resulted in 9,094 *O. sativa* subspecies *japonica* orphan ORF candidates. Next, we used nucleotide-to-nucleotide BLAT to align these 9,094 *O. sativa* subspecies *japonica* candidate orphan ORFs to the genome sequences of the other 10 species. If exact nucleotide matches covered at least 20% of the corresponding *O. sativa* subspecies *japonica* ORF, 1 effective hit was accepted. Only *O. sativa* subspecies *japonica* ORFs that had no more than one effective hit in each species were retained for subsequent analyses. This process narrowed down 930 *O. sativa* subspecies *japonica* orphan ORF candidates. Orthologous sequences for these orphan genes (the longest isoform), including the orthologous non-coding sequences in the species outside the orphan gene sequences, were then extracted from whole-genome reciprocal-best alignments. Moreover, we further used BLAT to align these orthologous sequences to *O. sativa* subspecies *japonica* ORFs, to retrieve highly similar orthologous sequences.

In total, 929 orthologous clusters were retrieved. One candidate (*Osjap04g11920*) was ruled out because no orthologous sequence outside *O. sativa* subspecies *japonica* was found. Each of the 929 orthologous sequence clusters was aligned based on codon alignment with the template *O. sativa* subspecies *japonica* ORF sequences using the MACSE programme (-seq Osjap.fa -seq_lr others.fa -fs 100 -stop 100 -fs_lr 20 -stop_lr 10)[79] (Supplementary File 1). Each 'N' in the MASCE sequence alignments meant a frameshift caused by one nucleotide deletion. Some 230 ORFs were identified as de novo ORFs that had orthologous non-coding sequences in at least 2 outgroup species.

To ensure the quality of the alignments generated by the procedure above, for each MASCE alignment, we manually adjusted sequence alignments in Geneious R9 (ref. [80]), checked in-exon disruptive mutations and labelled the status. '0' represents a full ORF, '1' represents one premature stop codon, '10' represents one frameshift, 'S' represents a partial sequence (including large in-frame indels and incomplete gene structure (for example, an undetected start or stop codon or undetected splicing sites)) without in-exon disruptive mutation, and 'NA' represents no orthologous sequence detected (Supplementary Table 1). We then skipped undetermined states (for example, 'S' and 'NA') and mapped protein-coding and non-coding states on the *Oryza* phylogenetic tree.

The ancestral/derived states of a gene were assigned following the parsimony rule. If one sibling species was undetermined, its status was thought to be the same as another one. If one sibling species was protein coding and another was non-coding, the group was thought to be protein coding to make sure our estimation was conservative. If both sibling species were non-coding, only one outgroup non-coding sequence or one gene loss event was considered. Out of 929 orphan gene candidates, 1% (8) had 3 or more internal gene-loss events. Based on these observations, we required that each de novo ORF should have at least two outgroup orthologous non-coding sequences and one ingroup orthologous non-coding sequence, which was equal to a 1% false positive rate under stringent criteria compared with previous systematic studies[76,77]. Alternatively, 455 de novo ORF candidates were identified, which corresponded to a 7.6% false positive rate (Supplementary Table 1) if at least one outgroup species was demanded.

After considering transcriptional evidence and orthologous sequence numbers, we collected 391 candidate de novo genes with at least 1 outgroup species. The protein sequences of 391 candidate de novo genes were then aligned to the National Center for Biotechnology Information non-redundant database version 20150219 (*e* value: 10×10^{-3})[81] using BLASTP. In total, 48 candidate de novo genes with potential protein matches outside *Oryza* species were excluded (Supplementary Table 3). The final numbers of *Oryza* candidate de novo genes were 343 (at least 1 outgroup) and 175 (at least 2 outgroups), as shown in Fig. 1b.

Transcription evidence. The OGE/IOMAP-derived transcriptome data package includes RNA-Seq data from three tissues (leaf, panicle and root) for ten species (not including *O. sativa* subspecies *indica*). These data have an average depth of 800× and should therefore be able to detect rare transcripts, but may lose tissue-specific transcripts. We collected IRGSP genes supported by widely available full-length cDNA, EST and protein data as a reference for comparison. Raw reads for all three tissues were pooled together, followed by de novo transcriptome assembly with Trinity version 20140717 (ref. [66]). Next, de novo ORFs and their orthologous sequences were aligned against the de novo-assembled transcripts from each corresponding species and IRGSP-annotated *O. sativa* subspecies *japonica* genes using BLAT. If exact matches covered at least 20% of a corresponding sequence, one effective hit was accepted. As long as one orthologous sequence of a de novo ORF was detected, it was recognized as a de novo gene. The *O. sativa* subspecies *japonica* tissue transcriptome data were analysed with tophat2 version 2.0.12 and cufflinks version 2.2.1 (ref. [82]). Tissue specificity scores were calculated as defined[57].

ORF triggers. A mutation that transforms a region with one or more nonsense codons in an ancestral reading frame into an ORF region can be called an ORF trigger or enabler[26,27]. Indels and substitutions are two types of ORF trigger. Mutations that created an ORF from a non-coding sequence that contained one or more premature stop codons were considered to be ORF triggers. Among two types of ORF trigger, substitutions can turn premature stop codons into regular codons, whereas indels can cause frameshifts, erase premature stop codons and even change gene structures.

Simulation of de novo ORF by chance. Before performing more in-depth evolutionary analyses of these 201 candidate de novo ORFs, we ran a sequence simulation test with 1,000,000 randomizations of intergenic regions on the focal species, *O. sativa* subspecies *japonica*, to estimate the probability that the candidate ORFs detected in this study were derived by random shuffling of nucleotides in non-coding regions. We collected 35,602 intergenic sequences from the *O. sativa* subspecies *japonica* genome. On average, they were 6,688 base pairs (bp) long with a GC content of A:T:C:G = 0.29:0.28:0.21:0.22. We randomly generated 1,000,000 sequences with the same length and GC content and predicted ORFs from these sequences using the AUGUSTUS algorithm (version 2.5.5)[83] packaged in the MAKER pipeline. In total, we obtained 47 ORFs out of 1 million random sequences. This translates to roughly one de novo ORF in the *O. sativa* subspecies *japonica* genome. The null hypothesis was that a high number of de novo genes detected were generated by random shuffling, which was tested under Poisson distribution (λ = 1).

Ribosomal profiling and tRNA adaptation of de novo genes. A ribosomal profiling experiment in *O. sativa* was successfully performed using a transgenic strain in which the ribosomal protein RPL18 was tagged (p35S:HF-OsRPL18), creating a translatome database (MSU-RAP) for the species[65]. This database reported transcript abundance for all genes.

The raw sequencing data for transcription in three tissues (the panicle, seedling and callus) were downloaded from the National Center for Biotechnology Information Sequence Read Archive: SRR2638777 for panicles, SRR2638779 for seedlings and SRR2638780 for calluses[65]. The three datasets were pooled together to search for TRAP-Seq signals of de novo genes. To update gene models recently found in the OGE/IOMAP 13 genomes, we assembled strand-specific transcripts out of the sequencing reads of the TRAP-Seq databases using Trinity version 2.4.0 (ref. [66]). We used the default parameters (--seqType fq --max_memory 100 G --CPU 30 --SS_lib_type RF --full_cleanup) when running the Trinity programme. We then used BLAT to align candidate de novo genes to assembled transcripts, requiring an identical sequence match in at least 20% of the gene length between transcripts and candidate genes (Supplementary File 2).

The tAI for a gene was defined as the geometric mean of adaptive values of its codons[84]. Adaptive values of codons were influenced by the tRNA gene numbers in the genome. We downloaded the *Oryza* tRNA profile as instructed (http://gtrnadb.ucsc.edu/)[85], and installed the standalone programme on a local server. The *O. sativa* genome was chosen, and tAI values were calculated for all genes, using the stAIcalc programme[67].

Rice materials and growth conditions. *O. sativa* subspecies *japonica* (Nipponbare) was used for sampling for proteomics analyses. Tissues used for analyses included flag leaves, anthers, and pistils and glumes one day before flowering. The rice plants were grown in a field in the winter season in Hainan, China. The planting density was 16.5 cm between plants in a row and 26.5 cm between rows. Field managements, including irrigation, fertilizer application and pest control, followed essentially normal agricultural practices.

Proteomics analyses. We used MRM-MS to target peptides with high sensitivity and wide dynamics, especially the peptides from low-abundance proteins in complex biological samples[61,62,86–88], as was found previously[89–92].

General procedure. To select proper peptides for MRM-MS assay, a total of 343 novel gene products underwent the computational evaluation through Skyline (version 3.7), which provided the mass values of the selected peptides and their corresponding fragment ions[61]. The Skyline analysis resulted in the peptides from 184 proteins possibly being identified by MRM, including 54 proteins with at least 3 unique peptides, 53 proteins with 2 unique peptides and 77 proteins with a single unique peptide. On the basis of expression data (mRNA or proteins) in rice and *Arabidopsis*, as well as humans, reproductive tissues such as pollen and testis were found to have plentiful gene expression products from orphan genes and other genes[93–96].

Generally, the MRM identification results require more experimental evidence to verify the annotation to mass spectra using synthetic peptides. A total of 44 peptides belonging to 35 proteins described above were synthesized and further analysed by MRM assay under the same experimental conditions. As shown in Fig. 4, two peptides ('TFFDVGSATGGGVPR' and 'FTLILLNGAPR') were identified from all of the tissues in *O. sativa* subspecies *japonica* rice. The chromatographic and mass spectrometry behaviours in retention time, as well as the transition abundance patterns in MRM, were well shared with those of the chemically synthesized peptide. Approximately 84% (37/44) of the selected peptides in rice tissues had MRM spectra that were well matched with those acquired from the correspondent peptides, which were synthesized with a retention time shift of less than ±2 min in a 3 h gradient, and a transition abundance change of less than 20% in individual transitions following the Commission Decision 2002/657/EEC (Fig. 4 and Supplementary Fig. 2)[63].

Protein extraction and digestion. The rice tissues of *O. sativa* subspecies *japonica*, such as anthers, pistils and glumes, and flag leaves, were ground in a tissue lyser (Shanghai Jingxin) for 5 min with the lysis buffer (7 M urea, 2 M thiourea, 0.2% sodium dodecyl sulfate and 20 mM Tris-HCl, pH 8.0) complemented with 10% polyvinylpolypyrrolidone, 10 mM dithiothreitol (DTT) and protease inhibitor cocktails (Roche) at 4 °C. After centrifugation, the pellets were discarded and the proteins in the supernatants were precipitated by a final concentration of 10% cold TCA/acetone with 10 mM DTT at −20 °C for 2 h followed by a thorough wash with cold acetone containing 10 mM DTT. The protein precipitates were dried in air, then suspended in the lysis buffer with sonication assistance on ice. The supernatants after debris removal were ready for measurement of protein concentrations and digestion to proteins with trypsin (Progema).

For a typical digestion, the proteins were first reduced and alkylated with 10 mM DTT at 30 °C for 2 h and 55 mM iodoacetamide at room temperature for 45 min in the dark. They were then digested with the modified approach of filter-aided sample preparation[97]. In brief, the proteins were put onto a filter unit with a 10 kD cut-off (Sartorius), followed by complete exchange of the denatured solution with 0.5 M triethylamonium bicarbonate (TEAB), and finally digestion by trypsin on the unit top at 37 °C with shaking overnight. The digested peptides were collected in the unit bottom for further analysis with a mass spectrometer.

MRM-MS assay. The 343 candidate de novo novel genes in candidate de novo gene set 2 (of which set 1 is a subset) in fasta format were analysed by Skyline software to achieve the proper peptides for MRM detection based on two criteria: (1) without any missed cleavage site or methionine residue; and (2) with a peptide length of 8–25 amino acids[61,98,99]. The selected peptides (1,183 peptides corresponding to 321 novel genes) were further filtered using 5 criteria based on theoretical mass spectrometry parameters: (1) precursor charges at 2 or 3; (2) ion charges at 1; (3) ion type at b or y; (4) product ions from $m/z >$ precursor to 3 ions; and (5) precursor m/z exclusion window at 10 m/z. The filtration resulted in the 763 peptides being detectable for MRM assay, in which 107 proteins contained at least 2 peptides detectable by MRM.

The primary experiments of MRM-MS were conducted towards a peptide mixture of *O. sativa* subspecies *japonica* rice, pooling the peptides from varied rice tissues. The peptides acquired were filtered by mProphet[100] with a cut-off *Q* value of <0.01. The peptides with better transition signals were selected as targets for further chemical synthesis, as well as validation. A peptide obtained from the rice tissues was defined as validated, meaning that it shared the same retention time (variation < 1%) and similar transition patterns (at least three fragments per peptide) as its corresponding synthetic peptide.

MRM-MS assay was performed on a QTRAP5500 mass spectrometer (AB SCIEX) coupled with a nanoAcquity UPLC system (Waters) with a self-packed column (ID 150 μm × 30 cm; 1.7 μm particles). The iRT peptides (Biognosys) were incorporated into the tissue peptide samples for retention time calibration in all of the MRM assays[101,102]. The elution system consisted of two solutions: solvent A (5% acetonitrile with 0.1% formic acid) and solvent B (95% acetonitrile with 0.1% formic acid). The peptide mixtures were eluted through a 180 min gradient programme (8% solvent B for 10 min, 8–45% solvent B for 150 min, 45–80% solvent B for 10 min and 80% solvent B for 10 min) at a flow rate of 500 nl min^{-1}.

The MRM-MS parameters were set as the following: ionspray voltage at 2,400 V, curtain gas at 35.00, ion source gas 1 at 20.00, collision gas at high, interface heater temperature at 150 °C, entrance potential at 10.00, dwell time at 8 ms, and Q1 and Q3 at unit resolution. The collision energy (CE) used for the MRM scans was calculated according to the formula $\mathrm{CE} = \mathrm{a} \times m/z + \mathrm{b}$, based on the m/z and charge statuses of parent ions, in which the parameter pairs a and b were set as 0.036 and 8.857 for a double-charged ion and 0.0544 and −2.4099 for a triple-charged ion, respectively. The raw data were processed and integrated using Skyline software with the iRT peptides across the liquid chromatography gradient for retention time calibration.

To validate the peptide signals identified by MRM assay, 44 peptides were selected from the total 109 unique peptides with high-quality MS/MS signals and chemically synthesized by GL Biochem. They were then delivered to the same MRM assay to evaluate whether the retention time and transition patterns between rice tissue and synthesized peptide remained consistent (Supplementary Fig. 4).

Molecular evolution and genomics. We used the branch model test in PAML[60] to detect the signals of natural selection for de novo genes in their sequence substitutions at d_S, d_N and $\omega = d_N/d_S$. Since the estimation of parameters in PAML may be affected by the alignment quality[103], we used codon-based alignments generated by PRANK[104] to avoid the alignment error. Among 343 candidate de novo genes, 79 genes were excluded from the branch model analysis, as they have too few sequences (<3) available for calculating the branch model. Thus, we used the remaining 264 genes to perform model selection via the AIC scores and likelihood ratio test. The statistical frame of model comparison for each gene includes the null model (the one-ω model, assuming the same ω for all branches), alternative branch models (two-ω models, assuming different ω ratios between foreground branches and background branches) and free ω model (assuming an independent ω ratio for each branch). The appropriate model was selected according to the minimum AIC, where the AIC was calculated using $-2 \times \ln[L] + 2 \times$ (number of parameters), where L is the log-likelihood difference. The statistical significances between the models were estimated by calculating twice the log-likelihood difference following a χ^2 distribution, using the differences in the numbers of free parameters between models as the numbers of degrees of freedom. To further test whether the ω ratio of a model significantly deviated from neutral evolution ($\omega = 1$), we incorporated the neutral model, which estimates model parameters by fixing $\omega = 1$ (ref. [105]). After model selection and comparison, 24 genes fit the two-ω optimal model, whereas 4 genes fit the one-ω best-fitted model. Among these 28 genes with the best-fitted models, 15 were significantly different from the neutral prediction of $\omega = 1$ ($P < 0.05$, χ^2 test).

We computed the indel and single nucleotide polymorphism (SNP) ratio in intergenic regions of *O. sativa* subspecies *japonica* using 3K rice genome variation data[47]. VCF files including indels and SNPs for the 30 *O. sativa* subspecies *japonica* strains we sampled (B001, B002, B003, B004, B008, B014, B016, B017, B018, B023, B025, B034, B036, B037, B038, B043, B045, B046, B047, B054, B055, B056, B066, B069, B070, B071, B077, B100, B101 and B102) were downloaded from the 3K rice database[47]. A total of 403,591 indels and 5,868,200 SNPs in intergenic regions were identified based on the OGE/IOMAP gene annotation. The SNP/indel ratio is 14.54 in the intergenic regions of 30 *japonica* strains (Supplementary Table 5).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available in Supplementary Files 1 and 2, Supplementary Figs. 4 and 5, and Supplementary Tables 11 and 14.

Received: 8 May 2018; Accepted: 23 January 2019;
Published online: 11 March 2019

References

1. Chen, L., DeVries, A. L. & Cheng, C. H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl Acad. Sci. USA* **94**, 3811–3816 (1997).
2. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103**, 9935–9939 (2006).
3. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
4. Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
5. Gilbert, W. Why genes in pieces? *Nature* **271**, 501 (1978).
6. Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance* (Belknap Press, 1982).
7. Patthy, L. in *Protein Evolution* 2nd edn 108–109 (Blackwell Publishing, 2008).
8. Klasberg, S., Bitard-Feildel, T., Callebaut, I. & Bornberg-Bauer, E. Origins and structural properties of novel and de novo protein domains during insect evolution. *FEBS J.* **285**, 2605–2625 (2018).
9. Bitard-Feildel, T., Heberlein, M., Bornberg-Bauer, E. & Callebaut, I. Detection of orphan domains in *Drosophila* using "hydrophobic cluster analysis". *Biochimie* **119**, 244–253 (2015).
10. Cai, J., Zhao, R., Jiang, H. & Wang, W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**, 487–496 (2008).
11. Carvunis, A. R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
12. Xiao, W. et al. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE* **4**, e4603 (2009).
13. Wu, D. D. et al. "Out of pollen" hypothesis for origin of new genes in flowering plants: study from *Arabidopsis thaliana*. *Genome Biol. Evol.* **6**, 2822–2829 (2014).
14. Cui, X. et al. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol. Plant* **8**, 935–945 (2015).
15. Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H. & Spillane, C. Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* **11**, 47 (2011).
16. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. *Genetics* **176**, 1131–1137 (2007).
17. Chen, S. T., Cheng, H. C., Barbash, D. A. & Yang, H. P. Evolution of *hydra*, a recently evolved testis-expressed gene with nine alternative first exons in *Drosophila melanogaster*. *PLoS Genet.* **3**, e107 (2007).
18. Chen, S., Zhang, Y. E. & Long, M. New genes in *Drosophila* quickly become essential. *Science* **330**, 1682–1685 (2010).
19. Reinhardt, J. A. et al. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* **9**, e1003860 (2013).
20. Zhou, Q. et al. On the origin of new genes in *Drosophila*. *Genome Res.* **18**, 1446–1455 (2008).
21. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
22. Toll-Riera, M. et al. Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).
23. Li, C. Y. et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput. Biol.* **6**, e1000734 (2010).
24. Wu, D. D., Irwin, D. M. & Zhang, Y. P. De novo origin of human protein-coding genes. *PLoS Genet.* **7**, e1002379 (2011).
25. Zhang, Y. E., Vibranovski, M. D., Landback, P., Marais, G. A. & Long, M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* **8**, e1000494 (2010).
26. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res.* **19**, 1752–1759 (2009).
27. Murphy, D. N. & McLysaght, A. De novo origin of protein-coding genes in murine rodents. *PLoS ONE* **7**, e48650 (2012).
28. Xie, C. et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8**, e1002942 (2012).
29. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Canas, J. L., Messeguer, X. & Alba, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* **2**, 890–896 (2018).
30. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
31. Schlötterer, C. Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
32. Moyers, B. A. & Zhang, J. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Mol. Biol. Evol.* **33**, 1245–1256 (2018).
33. Zhao, Y. et al. 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 Evol. Biol.* **15**, 66 (2015).
34. Cheng, C. H. & Chen, L. Evolution of an antifreeze glycoprotein. *Nature* **401**, 443–444 (1999).
35. Husnik, F. & McCutcheon, J. P. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67–79 (2018).
36. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270 (1996).
37. Gubala, A. M. et al. The *goddard* and *saturn* genes are essential for *Drosophila* male fertility and may have arisen de novo. *Mol. Biol. Evol.* **34**, 1066–1082 (2017).
38. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
39. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* **32**, 835–845 (2015).
40. Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).
41. Sakai, H. et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* **54**, e6 (2013).
42. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
43. Long, M. Y., VanKuren, N. W., Chen, S. D. & Vibranovski, M. D. New gene evolution: little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013).
44. Zhang, C. J. et al. High occurrence of functional new chimeric genes in survey of rice chromosome 3 short arm genome sequences. *Genome Biol. Evol.* **5**, 1038–1048 (2013).
45. Zhang, Y. E., Landback, P., Vibranovski, M. & Long, M. New genes expressed in human brains: implications for annotating evolving genomes. *BioEssays* **34**, 982–991 (2012).
46. Mills, R. E. et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
47. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
48. Xu, X. et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2012).
49. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
50. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
51. Wang, M. et al. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat. Genet.* **46**, 982–988 (2014).
52. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* 4th edn 172–175; 351–354 (Sinauer Associates, Sunderland, 2007).
53. Berretta, J. & Morillon, A. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* **10**, 973–982 (2009).
54. Bornberg-Bauer, E. & Alba, M. M. Dynamics and adaptive benefits of modular protein evolution. *Curr. Opin. Struct. Biol.* **23**, 459–466 (2013).
55. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 0217 (2017).
56. Heinen, T. J., Staubach, F., Häming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr. Biol.* **19**, 1527–1531 (2009).
57. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
58. Long, M., Rosenberg, C. & Gilbert, W. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl Acad. Sci. USA* **92**, 12495–12499 (1995).
59. Sharp, P. A. Speculations on RNA splicing. *Cell* **23**, 643–646 (1981).
60. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
61. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4**, 222 (2008).
62. Ebhardt, H. A., Root, A., Sander, C. & Aebersold, R. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* **15**, 3193–3208 (2015).

63. Pecorelli, I., Bibi, R., Fioroni, L. & Galarini, R. Validation of a confirmatory method for the determination of sulphonamides in muscle according to the European Union regulation 2002/657/EC. *J. Chromatogr. A* **1032**, 23–29 (2004).
64. Wen, B. et al. IPeak: an open source tool to combine results from multiple MS/MS search engines. *Proteomics* **15**, 2916–2920 (2015).
65. Zhao, D. et al. Analysis of ribosome-associated mRNAs in rice reveals the importance of transcript size and GC content in translation. *G3 (Bethesda)* **7**, 203–219 (2017).
66. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
67. Sabi, R., Volvovitch Daniel, R. & Tuller, T. stAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* **33**, 589–591 (2017).
68. Lees, J. G., Dawson, N. L., Sillitoe, I. & Orengo, C. A. Functional innovation from changes in protein domains and their combinations. *Curr. Opin. Struct. Biol.* **38**, 44–52 (2016).
69. Davidson, A. R. & Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA* **91**, 2146–2150 (1994).
70. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
71. Vaughan, D. A., Morishima, H. & Kadowaki, K. Diversity in the *Oryza* genus. *Curr. Opin. Plant Biol.* **6**, 139–146 (2003).
72. Murat, F., Van de Peer, Y. & Salse, J. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol. Evol.* **4**, 917–928 (2012).
73. Huey, R. B. et al. Plants versus animals: do they deal with stress in different ways? *Integr. Comp. Biol.* **42**, 415–423 (2002).
74. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146 (2017).
75. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–578 (2016).
76. Zhang, Y. E., Vibranovski, M. D., Krinsky, B. H. & Long, M. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Res.* **20**, 1526–1533 (2010).
77. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* **9**, e1001179 (2011).
78. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
79. Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E. J. MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS ONE* **6**, e22594 (2011).
80. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
81. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15 (2009).
82. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
83. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
84. Dos Reis, M. et al. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
85. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93–D97 (2009).
86. Aebersold, R., Burlingame, A. L. & Bradshaw, R. A. Western blots versus selected reaction monitoring assays: time to turn the tables? *Mol. Cell. Proteomics* **12**, 2381–2382 (2013).
87. Sjostrom, M. et al. A combined shotgun and targeted mass spectrometry strategy for breast cancer biomarker discovery. *J. Proteome Res.* **14**, 2807–2818 (2015).
88. Guo, J. et al. A comprehensive investigation toward the indicative proteins of bladder cancer in urine: from surveying cell secretomes to verifying urine proteins. *J. Proteome Res.* **15**, 2164–2177 (2016).
89. Xie, Y. et al. The levels of serine proteases in colon tissue interstitial fluid and serum serve as an indicator of colorectal cancer progression. *Oncotarget* **7**, 32592–32606 (2016).
90. Zhang, S. et al. Quantitative analysis of the human AKR family members in cancer cell lines using the mTRAQ/MRM approach. *J. Proteome Res.* **12**, 2022–2033 (2013).
91. Hou, G. et al. Biomarker discovery and verification of esophageal squamous cell carcinoma using integration of SWATH/MRM. *J. Proteome Res.* **14**, 3793–3803 (2015).
92. Hou, G., Wang, Y., Lou, X. & Liu, S. Combination strategy of quantitative proteomics uncovers the related proteins of colorectal cancer in the interstitial fluid of colonic tissue from the AOM-DSS mouse model. *Methods Mol. Biol.* **1788**, 185–192 (2017).
93. Fagerberg, L. et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
94. Uhlen, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
95. Lindskog, C. The potential clinical impact of the tissue-based map of the human proteome. *Expert Rev. Proteomics* **12**, 213–215 (2015).
96. Uhlen, M. et al. Transcriptomics resources of human tissues and organs. *Mol. Syst. Biol.* **12**, 862 (2016).
97. Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
98. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
99. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).
100. Reiter, L. et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **8**, 430–435 (2011).
101. Bruderer, R., Bernhardt, O. M., Gandhi, T. & Reiter, L. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* **16**, 2246–2256 (2016).
102. Navarro, P. et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).
103. Jordan, G. & Goldman, N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* **29**, 1125–1139 (2012).
104. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014).
105. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998).
106. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).

Acknowledgements

We appreciate valuable discussions with N. Jiang at MSU, the group of M. L. at Chicago, Y. Liao and M. Chen at the Institute of Genetics and Development in Beijing, and J. P. Staley at Chicago. We are thankful for the editing done by E. Mortola. This work was supported by the USA National Science Foundation (NSF) under Plant Genome Research Program numbers 0321678, 0638541 and 0822284, the Bud Antle Endowed Chair of Excellence in Agriculture and Life Sciences, and the AXA Chair for Evolutionary Genomics and Genome Biology (to R.A.W.), NSF MCB number 1026200 (to M.L. and R.A.W.), NSF MCB 1051826 and NIH R01 GM 100768 (to M.L.), the National Key R&D Program of China 2017YFC0908400 (to S.L.) and the National Program for Support of Top-notch Young Professionals of China (to Y.O.).

Author contributions

L.Z., R.A.W., S.L. and M.L. conceived and designed the project. L.Z., Y.R., R.A.W., S.L. and M.L. wrote the manuscript, with significant contributions from C.Z., A.R.G., J.C. and Y.Z. L.Z. conducted the computational genomic analysis, with significant contributions from A.R.G., K.C., J.Z. and Y.Z. C.Z., Y.Y., J.Z., K.C., M.W., D.C. and R.A.W. generated and annotated the genome sequences. Y.R., G.H., J.Z., L.Z. and S.L. designed and conducted the proteomics experiments to detect proteins translated from de novo genes. R.Z., B.W., L.Z. and Z.P. conducted the analysis of public proteomics databases. Y.R., L.Z., J.C., M.L. and S.L. performed further evolutionary and proteomics analyses. T.Y., G.L. and Y.O. grew rice strains in Sanya (China) and dissected rice tissues. J.C., L.Z., C.Z. and M.L. conducted the evolutionary substitution analyses of de novo genes.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-0822-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.A.W., S.L. or M.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Rod A. Wing4, Siqi Liu, Manyuan Long

Last updated by author(s): Jan 17, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Wget 1.19.5
Data analysis	BLAT v35xl; MACSE v1.0.0.i5; Geneious R9; Trinity v20140717; Tophat2 v2.0.12; Cufflinks v2.2.1; Augustus v2.5.5; Trinity v2.4.0; Skyline v3.7; PAML v4.8

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

1. rice genome assemblies and gene annotations
listed in https://www.nature.com/articles/s41588-018-0040-0
associated with Figure 1, 2 and 3

2. MRM-MS raw signals
PASS01123 at peptideatlas
associtaed with Figure 4 and 5
will be released after publication

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We have identified excess japonica de novo genes by pinpointing the stepwise origination process of ORF from noncoding sequence in closely related outgroup species and provided multiple lines of evidence, including proper gene structure, transcriptional and translational evidence and selective signals, to support the potential functionality of these gene candidates.
Research sample	We focused on 11 closely related rice species with the biggest divergence time of 15 myr and most of them are only a few myrs diverged. When the divergence time is small, the chance to find remnants of ortholog noncoding sequences in outgroup species is big.
Sampling strategy	We focused on male reproductive tissues for proteomics experiments since new genes are lowly expressed but enriched in male reproductive tissues. We used the MRM-MS method to enrich new gene signals since new genes are lowly expressed.
Data collection	Rice samples were generated by Ouyang group and MRM-MS experiments were finished by Liu group. All details were described in methods section.
Timing and spatial scale	Rice samples were planted 12/10/2016 and collected at 3/22/2017.
Data exclusions	No data were excluded from the analyses.
Reproducibility	We artificially synthesized 39 peptides to validate our MRM-MS results.
Randomization	We collected rice tissues for MRM-MS experiments without spliting them into groups.
Blinding	All experimental designs were transparent to every participants.

Did the study involve field work? ☒ Yes □ No

Field work, collection and transport

Field conditions	A great location for rice planting, average 25.7 centigrade and 1540 mm rainfall.
Location	Lingshui, Hainan, China (18.5060° N, 110.0375° E).
Access and import/export	The Ouyang group worked in a field working station in Lingshui run by Huazhong Agricultural University.
Disturbance	We used fields which were originally used for rice planting.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging