Computational and Structural Biotechnology Journal 19 (2021) 600–611

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Interpopulation differences of retroduplication variations (RDVs) in rice retrogenes and their phenotypic correlations

Haiyue Zeng [a,b,c], Xingyu Chen [c], Hongbo Li [d], Jun Zhang [e], Zhaoyuan Wei [a,b], Yi Wang [a,b,*]

[a] State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing 400715, China
[b] Biological Science Research Center, Southwest University, Chongqing 400715, China
[c] Shennong Class, Southwest University, Chongqing 400715, China
[d] College of Electronic and Information Engineering, Southwest University, Chongqing 400715
[e] College of Computer & Information Science, Southwest University, Chongqing 400715, China

ARTICLE INFO

Article history:
Received 8 September 2020
Received in revised form 29 December 2020
Accepted 31 December 2020
Available online 05 January 2021

Keywords:
Retrocopy
Retrogene
Rice
Retroposition
RDV

ABSTRACT

Retroduplication variation (RDV), a type of retrocopy polymorphism, is considered to have essential biological significance, but its effect on gene function and species phenotype is still poorly understood. To this end, we analyzed the retrocopies and RDVs in 3,010 rice genomes. We calculated the RDV frequencies in the genome of each rice population; detected the mutated, ancestral and expressed retrogenes in rice genomes; and analyzed their RDV influence on rice phenotypic traits. Collectively, 73 RDVs were identified, and 14 RDVs in ancestral retrogenes can significantly affect rice phenotypes. Our research reveals that RDV plays an important role in rice migration, domestication and evolution. We think that RDV is a good molecular breeding marker candidate. To our knowledge, this is the first study on the relationship between retrogene function, expression, RDV and species phenotype.

1. Introduction

Increasing attention has been paid to retrocopy and retroposition, which have been recently reported to be vital to the formation of new genes. Retrocopies are defined as gene copies originating by reverse transcription of mRNA and are subsequently inserted into the genome (Fig. 1A). This above process is called retroposition, which generally occurs by enzymes derived from retrotransposable elements, thereby giving rise to intronless copies of existing genes [1]. The source gene of the duplicated mRNA is generally named the parental gene. For a long time, retrocopies were considered as "junk DNA" due to a lack of regulatory elements as well as the presence of functional disruptions (including in-frame terminations, frameshift mutations and 5' truncations) [2]. However, at present, retrocopies are called "seeds of evolution" due to their significant contributions to genomic evolution [3]. Numerous studies have shown that retrocopies may gain functionality and become so-called retrogenes. Retrogenes might have intact open reading frames (ORFs) with protein-coding functions [4] or lead to new protein domain formation through fusion with other genes [5,6]. Functional retrogenes can encode proteins that may be similar or different from those encoded by their parental genes [7], and they can also be used to replace exons or produce chimeric transcripts [8]. It was found that retrogenes significantly contribute to the ongoing generation of new genes over evolutionary processes [7,9,10] and affect the expression of parental genes by retrogene-originated siRNA [11–13] and miRNA [14]. Meanwhile, parental genes are likely to become pseudogenes, and their function is subsequently completely lost and are further replaced by retrocopies, which can turn retrogenes into "orphans" (with the absence of parental genes in the genome) [15]. Therefore, retrocopies can also be utilized as useful phylogenetic markers, shedding novel light on evolutionary processes such as sex chromosome origination [16] or alterations in germline gene expression and retrotransposable element activities [7].

To study the function of retrogenes, researchers usually analyze their expression through RNA-Seq data. Baertsch et al. [17] studied human retrogenes and found that 726 retrogenes can be expressed.

Abbreviations: RDV, Retroduplication variation; Indel, Insertion-deletion; ORF, Open reading frame; FPKM, Fragments per kilobase per million mapped reads; GO, Gene ontology; MF, Molecular function; CC, Cellular component; BP, Biological process; MST, monosaccharide transporter.

* Corresponding author.
E-mail address: yiwang28@swu.edu.cn (Y. Wang).

https://doi.org/10.1016/j.csbj.2020.12.046

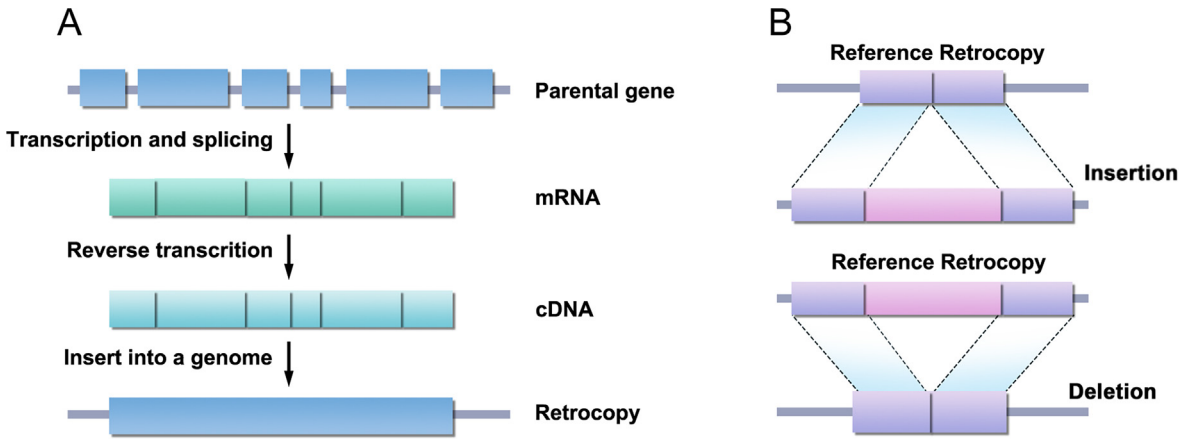

Fig. 1. (A) Formation of a retrocopy: the mRNA transcribed from the gene is reverse transcribed into cDNA and subsequently inserted into the genome. (B) Formation of an RDV: an insertion or deletion in a retrocopy.

They believe that these genes may play an important role after expression. Navarro et al. [18] found approximately 3,600 expressible retrogenes in seven primates. These expressible retrogenes are often close to a normal gene, and their expression is tissue-specific. They also believe that the expression of the retrogene and the parental gene is not collinear. Interestingly, Zhong et al. [19] found 306 retrogenes in zebrafish. Among them, the expressible retrogenes and their parental genes have strong collinearity in expression. Du et al. [20] obtained 219 expressible retrogenes in the coelacanth fish genome, of which the newly formed retrogenes had a lower expression value than the old retrogenes. In the plant genome, Sakai et al. [21] analyzed the expression of 150 retrogenes in rice in 7 tissues, and the results showed that the expression level of retrogene is often lower than that of the parental gene, but there is a certain collinearity. Abdelsamad et al. [22] believe that retrogenes in Arabidopsis are abundantly expressed in pollen and play an important role in pollen development.

Retroduplication variations (RDVs) are a type of polymorphism with the insertion or deletion of a retrocopy from individual genomes (Fig. 1B). Three studies [6,23,24] have uncovered 208 RDVs in the human genome, two of which [23,24] facilitate the reestablishment of the phylogenetic tree of human populations, confirming the significance of RDV polymorphisms as genomic markers for a population's history. Moreover, Kabza et al. [25] developed a new approach to detect novel retrocopies not annotated in the reference genome and focused on identifying RDV caused by retrocopy deletion. By studying human genomes from 17 populations, they found 193 RDVs and analyzed their distribution among populations, which has important implications for studying human evolution and migration. With the exception of these studies, there is no research on the RDVs of other species. Moreover, RDV may affect the function and expression of retrogenes, which is directly reflected in the phenotypes of the species. However, no research has been conducted on the relationship between retrogene function, expression, RDV and species phenotype.

Due to the above various unresolved questions, we aimed to identify RDVs caused by the different indels in functional retrogenes of rice. RNA-Seq data from the Rice Expression Database [26] were utilized to evaluate the expression of retrogenes and their parental genes. The 3,000 Rice Genomes Project [27–29] was adopted for assessment of the RDVs among rice populations. The mutated, ancestral and expressed retrogenes were detected in rice genomes, and their RDV influence on rice phenotypes was analyzed with statistical methods.

2. Results

2.1. Summary of data for retrocopies in rice and nonsynonymous (Ka)/ synonymous (Ks) analysis

To analyze RDVs in rice, first, we need to scan retrocopies in the genome. We obtained all 74 retrocopies of rice (*Oryza sativa japonica*) from RetrogeneDB [30]. Although previous studies have suggested that retrocopies in rice might be more numerous [21,31], retrocopies in RetrogeneDB are high quality because they were annotated based on a very rigorous analysis, followed by a manual inspection to ensure maximum data quality. Compared with the parental gene, all retrocopies met the following conditions: length of the alignment at least 150 bp, minimum of 50% coverage, minimum of 50% identity, and loss of at least two introns. Then, we obtained 37 retrocopies with a high coverage of over 90% with their parental genes (Fig. 2A), and 6 retrocopies had a high identity of over 90% (Fig. 2B). Forty-nine retrocopies were annotated as functional retrogenes whose status was classified as "KNOWN_PROTEIN_CODING" in RetrogeneDB. Twenty-five retrocopies were neither annotated as protein coding genes nor overlapped with annotated pseudogenes; they were determined to be "NOVEL" retropseudogenes (Fig. 2C). Fifty-nine retrocopies had conserved ORFs (Fig. 2D), which means that these retrocopies contained no frameshifts or stop codons.

To detect the selection pressure of protein levels for retrocopies, we calculated the Ka, Ks and substitution rate (Ka/Ks) between retrocopies and their parental genes; if Ka/Ks $\leq$ 0.5, then a retrocopy is considered to be negatively selected [32,33]. Table 1 shows the different Ka/Ks distributions of functional retrogenes and retropseudogenes. Their proportions show a clear distribution difference. For the functional retrogenes, 44.90% had Ka/Ks < 0.5, while 76.00% of retropseudogenes had Ka/Ks in a range of 0.5–1.2, indicating that most functional retrogenes were affected by negative selection and most retropseudogenes were neutrally selected. We also found that the Ka/Ks values of retropseudogenes were concentrated at 0.5–0.6, while those of functional retrogenes were concentrated at 0.1–0.3. The geometric mean of the Ka/Ks value for the rice retropseudogene was 0.5168, which was lower than the expected value (1.0) [34]. The reason why certain retropseudogenes are under significant selection pressure may be that they remove deleterious mutations by selective cutting and maintaining their functionality, or they were originally functional but later turned into pseudogenes due to harmful mutations. In

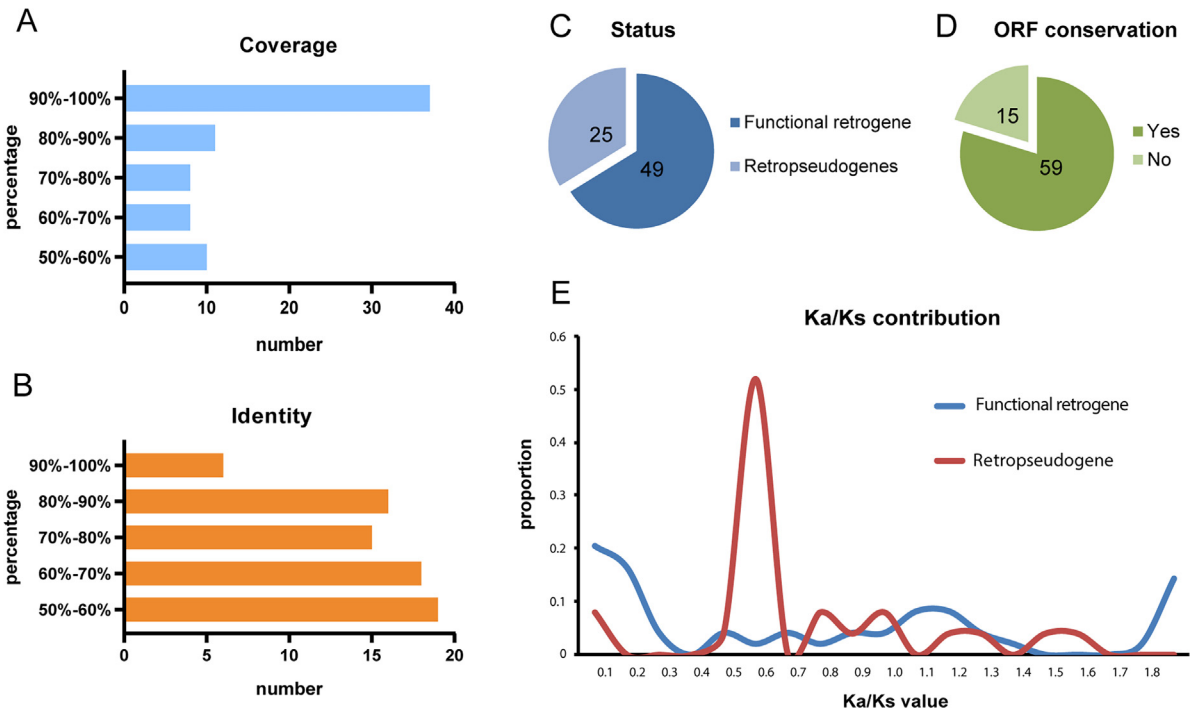

Fig. 2. Summary of 74 rice retrocopies (Table S1). (A) Parental gene protein coverage within the retrocopy-parental alignment. (B) Retrocopy-parental gene alignment percent identity. (C) Ratio of functional retrogenes and retropseudogenes. (D) Ratio of retrocopy with and without conserved ORF. (E) Ka/Ks distribution of functional retrogenes and retropseudogenes.

Table 1
. Selective pressure of rice retrocopies.

Gene	Ka/Ks < 0.5	Ka/Ks = 0.5 ~ 1.2	Ka/Ks > 1.2	Geometric mean
Functional retrogene	22/49 (44.90%)	16/49 (32.65%)	11/49 (22.45%)	0.3747
Retropseudogene	3/25 (12.00%)	19/25 (76.00%)	3/25 (12.00%)	0.5168

addition, some retrocopies had extremely large Ka/Ks values, indicating that they are strongly positively selected.

2.2. Estimates of the years of retrocopy origin and conservation analysis

To assess the formation history and identify ancestral retrocopies, we used the molecular clock to calculate the origin time. We assume the Ks of rice genes to be 6.5×10^{-9} substitutions per silent site per year [35]; then, the results show that a large number of new retrocopies in the rice genome were produced within the last 20 million years, later than the differentiation time of indica rice and japonica rice (20–40 million years ago) [36–39] (Fig. 3A). This shows that after the differentiation of indica and japonica rice, the number of retrocopies has exploded. This is consistent with previous studies in mammals [33], fish [40,41] and other plants [31], proving that retrocopy plays an important role in species formation and differentiation. Interestingly, retropseudogenes are found at both ends of the timeline, indicating that some retropseudogenes newly formed in the past 20 million years and that some represent ancient ancestral genes.

Moreover, we performed comparative analysis across 25 Tracheophyta species to identify retrocopy orthologs that originated prior to rice subspecies. OrthoVenn2 [42] was used to identify retrocopies originating in certain lineage ancestors, followed by checking orthologous genes in major plant lineages. We subsequently identified 72 retroposition events in the *Oryza sativa* ancestral genome, followed by 50 retropositions in Oryzinae, 196 in Poaceae, and 18 in commelinids. In comparison, we detected 117 retrocopies common to Pentapetalae and 65 shared by tomato and potato (Fig. 3B).

We also determined the presence of 48 retrocopies in more than one other genome, that is, 48 retrocopies of *Oryza sativa japonica* were ancestral. Of these, 4 originated in the common ancestor *Oryza sativa,* 7 originated in Oryzinae, and 16 were extremely ancient with a presence in more than one genome of analyzed Mesangiospermae (Fig. 3B). This is consistent with the results of the molecular clock analysis, proving that rice retrocopies mostly originated from very old ancestors or newly formed *Oryza sativa japonica.*

2.3. Identification of RDVs from known retrogenes in the rice genome

To analyze the retrocopy variation caused by retrogene indels in rice populations, all 74 rice retrocopies were used, and the genomic variation data were obtained from the 3,000 Rice Genomes Project [27–29]. To avoid the complexity of variation, we focused on only the indels of biallelic variation. Each RDV is an indel of at least 8 bp of retrogene sequence in certain analyzed genomes. As a result, 73 indels influencing 35 retrogenes were identified by mapping rice retrocopies to genomic variants (Table S2), and only one RDV (RDVD_osat_62) was detected in the retropseudogene (Fig. 4A). Then, the allele frequencies of detected RDVs were calculated in different populations, including geographic regions and subpopulations, with RDVs specific information, were displayed in Table S3. It was observed that the distribution differences of some

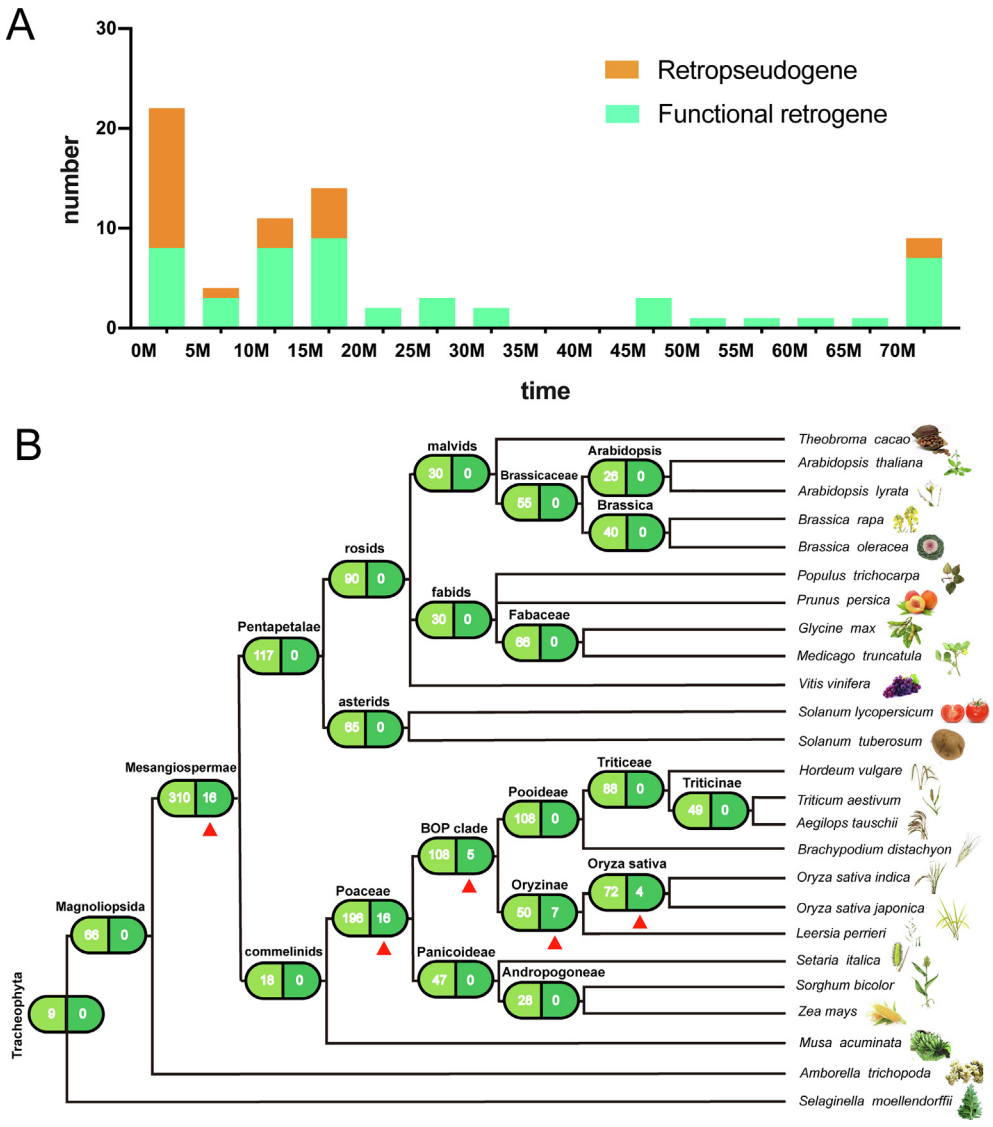

Fig. 3. (A) Origin time of retrocopy based on a molecular clock. (B) Retroposition in Tracheophyta. Left boxes indicate the number of retrocopies originating in a certain ancestral genome; right boxes reveal the number of retrocopies in the *Oryza sativa japonica* genome. All detected ancestral retrocopies are marked by red triangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

RDVs had obvious characteristics among populations. The differences in regions were manifested as the spread of frequency values from the center to the surroundings, and among the subpopulations, there were obvious differences in the magnitude of frequency values among the five populations (i.e. Xian/Ind, Geng/Jap, cA(Aus), cB(Bas) and admix).

For example, RDVD_osat_9_1 (one deletion in retro_osat_9), which was detected in the largest number of rice samples, can be observed that the Xian/Ind and cA(Aus) populations had the highest absence rate (approximately over 75%), while the Geng/Jap and cB(Bas) populations had the lowest absence rate (approximately <10%), and the admix population had a moderate absence rate (approximately 50%) (Fig. 4B). This shows that RDVD_osat_9_1 plays an important role in the differentiation of the rice populations.

Moreover, the proportion of detected RDVD_osat_9_1 in the South Asian population was the highest at 71.60%, with approximately 55%-60% of alleles in the East Asian, Southeast Asian and African populations; approximately 25%-30% of alleles in the American, West Asian and Oceanian populations; and approximately 5% in the European population (Fig. 4C). The most likely explanation for this phenomenon is the emergence of a new retroposition occurring in Asia, which spread to other continents, or it might be a deletion originating in Europe. We prefer to accept the first hypothesis, which is consistent with the origin, evolution, and migration routes of rice based on previous research that cultivated rice originated and was domesticated in Asia [43].

2.4. Gene ontology (GO) enrichment analysis of mutated, ancestral and expressed retrogenes

To study the relationship between RDV and functional retrogenes in depth, we set three criteria, mutated, ancestral and expressed, to screen for important retrogenes in the rice genome. These retrogenes are derived from ancestors and have undergone domestication and natural selection periods; they are mutated but still remain in the rice genome, so their produced RDVs may have important biological significance. We considered a retrogene to be expressed if it had a normalized expression value of over 1 FPKM (fragments per kilobase per million mapped reads) in at least

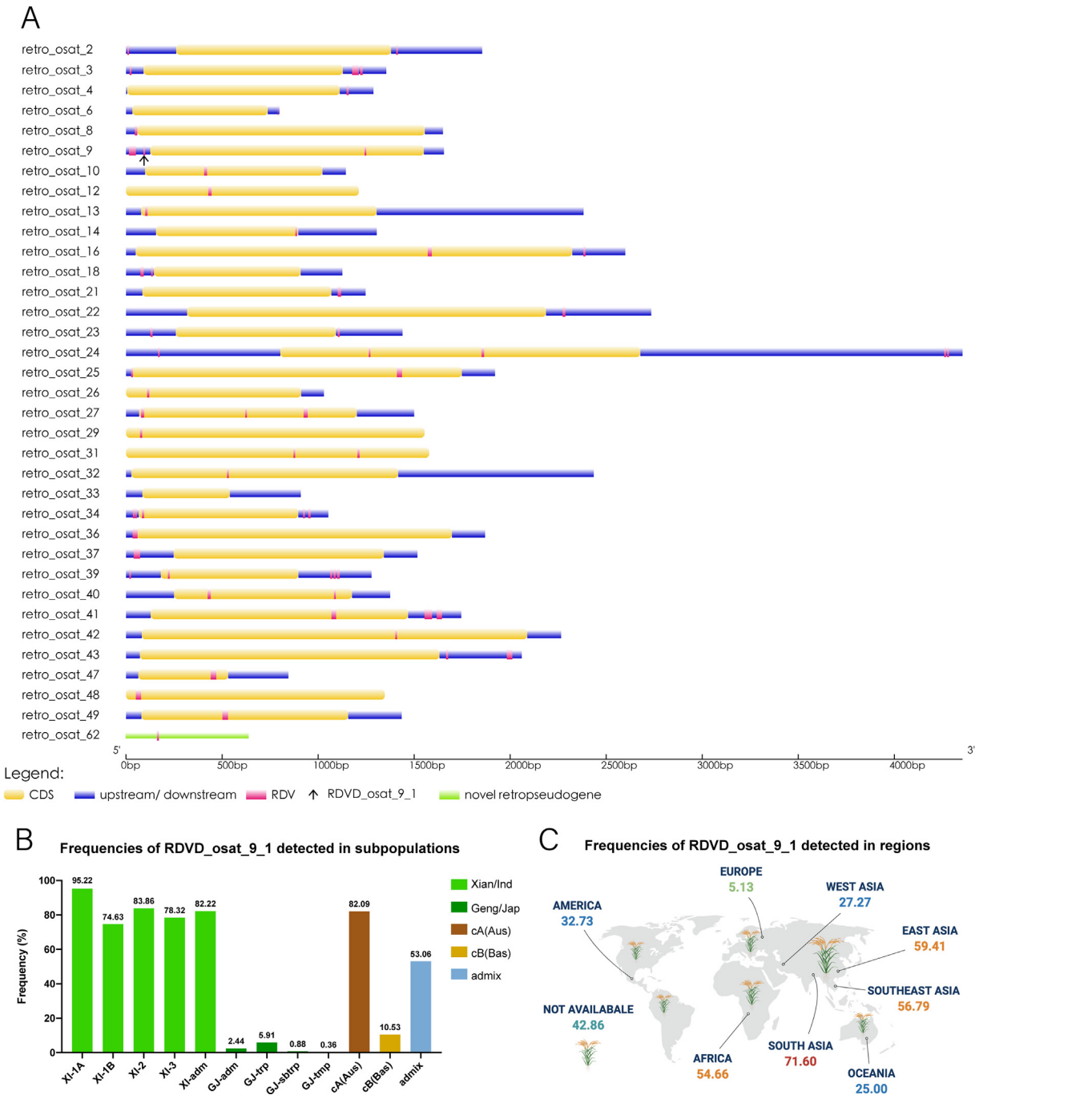

Fig. 4. (A) Seventy-three RDVs in 35 retrogenes. (B) Frequencies of RDVD_osat_9_1 in different subpopulations. (C) Frequencies of RDVD_osat_9_1 in different geographic regions.

one normal tissue. These criteria were met by 29 retrogenes, and most of the expressed retrogenes met these criteria (Fig. 5A).

To study the role of these ancestral, expressed retrogenes that had undergone insertion or deletion in the adaptive evolution of rice, the GO method was used to classify these genes functionally into three sets including biological process (BP), molecular function (MF), and cellular component (CC) based on GO analysis. These retrogenes were significantly enriched in 20 GO clusters ($p < 0.05$, FDR < 0.05, Fig. 5B). It is worth noting that the most significant GO clusters were mainly concentrated in the MF and CC categories. For the MF category, hydrogen symporter activity, sugar symporter activity and transmembrane transporter activity were the most important terms. For the CC category, cell, cell component and membrane were significantly enriched. The BP category was significantly enriched in only one term of carbohydrate transport. These results prove that RDV mainly occurs in expressed ancestral retrogenes that are primarily involved in transporter activity, cell membrane composition and transmembrane transporter activity.

2.5. Associations of the RDVs and phenotypes

The genotypes and phenotypes of rice RDVs of 29 expressed ancestral retrogenes were analyzed for statistical correlation, and RDVs with less than three samples corresponding to a single geno-

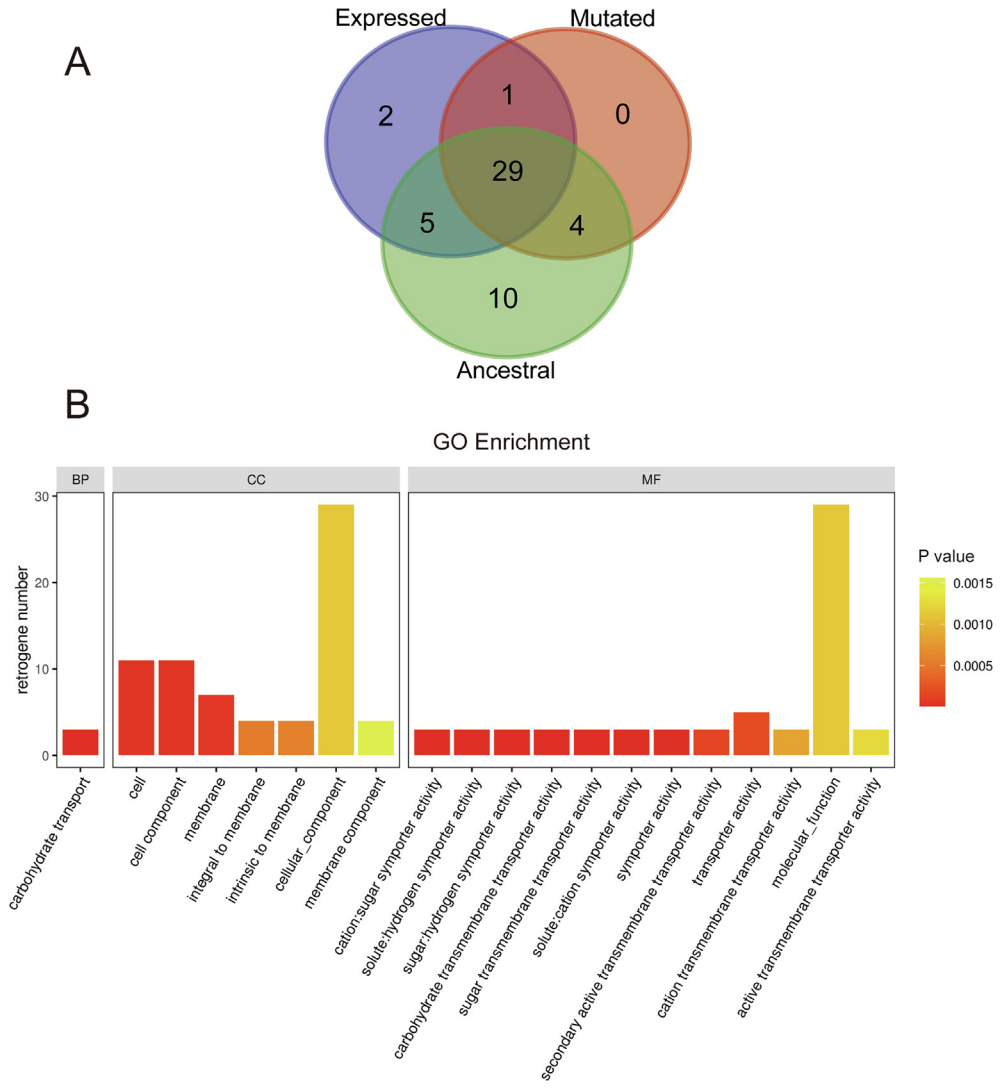

Fig. 5. (A) The Venn diagram shows the number of retrogenes that are expressed, ancestral and mutated. (B) GO enrichment analysis. BP (biological process), CC (cellular component) and MF (molecular function). Enriched GO terms were selected by Fisher's exact test and FDR < 0.05.

type were eliminated. A total of 14 RDVs were analyzed for their influence on the phenotypes. The results show that each RDV can significantly affect at least one trait of rice ($p < 0.05$), which strongly proves the effect of RDV on gene function or phenotype. All statistical results can be found in Table S4.

For example, the results of RDVs in three different retrogenes (retro_osat_9, *OsVPE4*, annotated as vacuolar-processing enzyme precursor; retro_osat_16, *OsMST5*, annotated as transporter family protein; retro_osat_27, annotated as cupin domain-containing protein) are displayed in Table 2. Among the nine traits, RDVD_osat_9_1 can significantly affect seven traits, indicating that retro_osat_9 may have abundant gene functions and may participate in numerous physiological activities and that this deletion affects the various agronomic characteristics of rice, while RDVI_osat_16 and RDVI_osat_27_1 can significantly affect four traits. The insertion variant of RDVI_osat_16 increased the culm number and ligule length and reduced the grain width and thousand grain weight compared with the DD and II samples. The insertion variation in RDVI_osat_27_1 increased the culm number, ligule length and time to flowering and reduced the grain width.

An interesting phenomenon was the three detected RDVs located on retro_osat_34 (*EXPB8*, annotated as expansin precursor); although their deletion sites were different and their genotype distributions were also different in rice samples, they had affected seven identical rice phenotypes (Table 3). The culm length, culm number, ligule length and seedling height of the mutant rice samples were higher than those of the reference genome genotype, while the grain length, grain width and thousand grain weight were lower than those of the reference genome genotype. This result shows that these three RDVs are closely linked with these seven phenotypes and have an in-depth development potential for molecular breeding.

2.6. Confirmation and evaluation of the expression of retrogenes

To validate and assess the expression level of retrogenes and their parental genes, Rice Expression Database-derived RNA-Seq data of 49 functional retrogenes were utilized, and the FPKM of all functional retrogenes and their parental genes can be found in Table S5. By analyzing the expression pattern of the above retrogenes (Fig. 6A), surprisingly, we found that although the RDVs of the above four typical retrogenes have been detected to significantly affect the grain size of rice and aleurone is usually considered to be related to grain size development, their expression level in the aleurone was very low or even not expressed. Simultaneously, it can be observed that they were highly expressed in reproductive

Table 2
Effects of three RDVs in three retrogenes (retro_osat_9, retro_osat_16, retro_osat_27) on the agronomic traits and grain size of rice from the 3,000 Rice Genomes Project based on statistical analysis. The genotype that was homozygous for deletion was DD, the genotype that was homozygous for insertion was II, and the genotype that was heterozygous was DI. These genotypes apply below.

RDVD_osat_9_1					
Phenotype	Genotypic (LSM ± SE)				
	II	DI	DD	P value	n_{II}, n_{DI}, n_{DD}
Culm length (cm)	**109.50 ± 1.11^{B}**	**110.13 ± 4.53^{AB}**	**114.83 ± 0.91^{A}**	**0.0048**	718,32,1007
Culm number (count)	**15.15 ± 0.19^{B}**	**15.56 ± 0.84^{AB}**	**17.49 ± 0.16^{A}**	**3.4615e-19**	718,32,1008
Grain length (mm)	8.61 ± 0.04	8.53 ± 0.26	8.62 ± 0.03	0.7766	827,18,1163
Grain width (mm)	**3.16 ± 0.02^{A}**	**2.87 ± 0.08^{B}**	**2.92 ± 0.01^{B}**	**1.585e-34**	827,18,1163
Ligule length (mm)	**16.91 ± 0.22^{B}**	**18.25 ± 0.95^{AB}**	**19.31 ± 0.16^{A}**	**4.3148e-12**	712,32,992
Panicle length (cm)	25.12 ± 0.17	25.28 ± 0.71	24.77 ± 0.10	0.0579	716,32,1007
Seedling height (cm)	**36.59 ± 0.42^{B}**	**38.84 ± 2.17^{AB}**	**40.06 ± 0.37^{A}**	**4.8483e-8**	710,32,994
Thousand grain weight (g)	**25.70 ± 0.17^{a}**	**24.06 ± 0.53^{b}**	**24.72 ± 0.13^{b}**	**0.0000061842**	782,48,1007
Time to flowering (day)	**96.87 ± 0.72^{B}**	**96.49 ± 3.51^{AB}**	**103.00 ± 0.71^{A}**	**0.0000044165**	845,37,1213
RDVI_osat_16					
Phenotype	Genotypic (LSM ± SE)				
	DD	DI	II	P value	n_{DD}, n_{DI}, n_{II}
Culm length (cm)	112.05 ± 0.80	120.17 ± 15.26	114.29 ± 1.45	0.322	1368,6,383
Culm number (count)	**16.21 ± 0.14^{B}**	**19.17 ± 2.51^{AB}**	**17.50 ± 0.26^{A}**	**0**	1368,6,384
Grain length (mm)	8.61 ± 0.03	8.30 ± 0.65	8.65 ± 0.05	0.6747	1535,5,468
Grain width (mm)	**3.05 ± 0.01^{A}**	**2.90 ± 0.20^{AB}**	**2.92 ± 0.02^{B}**	**4.0563e-7**	1535,5,468
Ligule length (mm)	**18.02 ± 0.15^{B}**	**19.00 ± 1.71^{AB}**	**19.35 ± 0.27^{A}**	**0.0001**	1352,6,378
Panicle length (cm)	24.89 ± 0.11	25.13 ± 0.87	25.02 ± 0.17	0.9974	1366,6,383
Seedling height (cm)	38.34 ± 0.32	39.83 ± 1.66	39.60 ± 0.60	0.1955	1353,6,377
Thousand grain weight (g)	**25.29 ± 0.12^{A}**	**24.30 ± 1.41^{AB}**	**24.56 ± 0.19^{B}**	**0.0028**	1408,12,417
Time to flowering (day)	99.71 ± 0.57	98.71 ± 9.11	102.90 ± 1.14	0.0732	1624,7,464
RDVI_osat_27_1					
Phenotype	Genotypic (LSM ± SE)				
	DD	DI	II	P value	n_{DD}, n_{DI}, n_{II}
Culm length (cm)	112.32 ± 0.83	115.06 ± 6.62	113.08 ± 1.34	0.8511	1236,16,505
Culm number (count)	**16.06 ± 0.14^{B}**	**16.81 ± 1.13^{AB}**	**17.56 ± 0.25^{A}**	**0.0000045154**	1236,16,506
Grain length (mm)	8.62 ± 0.03	8.51 ± 0.25	8.63 ± 0.04	0.7492	1351,12,645
Grain width (mm)	**3.09 ± 0.01^{a}**	**2.81 ± 0.12^{b}**	**2.87 ± 0.01^{b}**	**6.9863e-29**	1351,12,645
Ligule length (mm)	**17.95 ± 0.16^{B}**	**19.25 ± 1.50^{AB}**	**19.16 ± 0.23^{A}**	**0.0001**	1224,16,496
Panicle length (cm)	24.91 ± 0.12	25.11 ± 0.84	24.94 ± 0.15	0.952	1232,16,507
Seedling height (cm)	38.62 ± 0.34	40.19 ± 2.79	38.58 ± 0.51	0.8919	1224,16,496
Thousand grain weight (g)	25.27 ± 0.13	23.84 ± 0.93	24.81 ± 0.17	0.0864	1268,19,550
Time to flowering (day)	**97.36 ± 0.58^{B}**	**109.67 ± 5.67^{AB}**	**107.14 ± 1.00^{A}**	**9.1898e-16**	1445,18,632

organs such as anther and pistil, that retro_osat_24 and retro_osat_16 were highly expressed in anther, and that retro_osat_34 was highly expressed in panicle and pistil, indicating that the generation of RDVs might change their degree of expression in the aleurone or that these retrogenes affected the grain size traits by playing a certain role in the reproductive organs and panicles, but this explanation requires further experiments and research to verify the functions. Moreover, in anther, panicle and pistil, the expression level of retro_osat_16 in other tissues was very low, while the other three retrogenes had moderate expression in some vegetative organs, such as the root, leaf and shoot.

To compare the expression pattern between the retrogene and its parental gene, we calculated the Pearson product-moment correlation coefficient (R) between the retrogenes and parental genes using the FPKM values observed from nine tissues (Table S5). For above four retrogenes, except for retro_osat_16 (R < 0), the expression of the other three retrogenes and their parental genes in various tissues was positively correlated (R > 0) (Fig. 6B). The R value of retro_osat_34 > 0.90 showed a strong positive correlation. As a whole, the expression level of most retrogenes in a single tissue was lower than that of their parental genes, and some were no longer expressed, but the relative degree of expression in each tissue was similar. However, there were two exceptions. The parental gene of retro_osat_16 was almost not expressed in anther, but it had a high degree of expression on its own; retro_osat_27 and its parental genes were expressed in anther, but the expression level of retrogenes was obviously higher than that of the parental genes. This anomaly seems to imply a certain connection between retrogene and male reproductive organs.

For retropseudogenes, their expression information cannot be obtained from the Rice Expression Database because the database only contains gene expression data, so we downloaded the original RNA-Seq data from NCBI SRA and calculated the normalized expression value of these retropseudogenes (Table S5). It shows that 92% retropseudogenes can be expressed, and 70% of them were originated from one parental gene, Os04g0473025, which encodes plastoquinone oxidoreductase and plays energy sensing and response to abiotic stress in rice photosynthesis [44]. They are mainly expressed in anther, pistil and root at a high level and may play a role in reproductive and nutritional processes, and participate in abiotic stress response in root like their parental gene. Moreover, we compared the locus of rice retropseudogenes with rice regulatory RNAs (including siRNA, miRNA, lncRNA, etc.) data from published research [45], database [46] and China Rice Data Center (http://www.ricedata.cn/gene/), no rice retropseudogene generated by retrocopy had been identified as regulatory RNAs previously.

3. Discussion

RDVs are defined as a type of genomic sequence polymorphism related to the insertion or deletion of retrocopies in individual genomes. Because retrocopy is a single exon structure in most cases, we speculated that indels in functional retrogenes may have a greater probability of affecting gene structure and species phenotype. However, so far, there are only a limited number of studies

Table 3
Effects of three RDVs in the distinct loci of one retrogene (retro_osat_34) on the agronomic traits and grain size of rice from the 3,000 Rice Genomes Project based on statistical analysis.

RDVD_osat_34_2					
Phenotype	Genotypic (LSM ± SE)				
	II	DI	DD	P value	n_{II}, n_{DI}, n_{DD}
Culm length (cm)	**111.03 ± 0.90^{b}**	**113.89 ± 4.74^{ab}**	**115.85 ± 1.08^{a}**	**0.0227**	1186,27,544
Culm number (count)	**15.87 ± 0.15^{B}**	**17.19 ± 0.80^{AB}**	**17.84 ± 0.22^{A}**	**1.1001e-13**	1186,27,545
Grain length (mm)	**8.69 ± 0.03^{A}**	**8.55 ± 0.17^{AB}**	**8.47 ± 0.04^{B}**	**0.0000022052**	1360,22,626
Grain width (mm)	**3.07 ± 0.01^{a}**	**2.87 ± 0.06^{b}**	**2.90 ± 0.01^{b}**	**1.324e-15**	1360,22,626
Ligule length (mm)	**17.62 ± 0.16^{B}**	**20.67 ± 0.84^{A}**	**19.70 ± 0.23^{A}**	**2.8357e-12**	1173,27,536
Panicle length (cm)	25.00 ± 0.12	24.62 ± 0.58	24.78 ± 0.13	0.2194	1185,27,543
Seedling height (cm)	**37.55 ± 0.33^{B}**	**39.70 ± 2.25^{AB}**	**40.92 ± 0.53^{A}**	**0.000001111**	1173,27,536
Thousand grain weight (g)	**25.53 ± 0.13^{a}**	**23.4 ± 0.69^{b}**	**24.13 ± 0.16^{b}**	**4.0517e-10**	1314,32,491
Time to flowering (day)	99.93 ± 0.59	98.41 ± 4.40	101.57 ± 1.00	0.6254	1414,34,647
RDVD_osat_34_3					
Phenotype	Genotypic (LSM ± SE)				
	II	DI	DD	P value	n_{II}, n_{DI}, n_{DD}
Culm length (cm)	**111.14 ± 0.90^{b}**	**114.23 ± 4.35^{ab}**	**115.62 ± 1.09^{a}**	**0.0473**	1189,31,537
Culm number (count)	**15.88 ± 0.15^{B}**	**17.10 ± 0.86^{AB}**	**17.83 ± 0.23^{A}**	**8.5279e-13**	1189,31,528
Grain length (mm)	**8.69 ± 0.03^{A}**	**8.54 ± 0.20^{AB}**	**8.47 ± 0.04^{B}**	**0.0000037749**	1364,23,621
Grain width (mm)	**3.07 ± 0.01^{A}**	**2.79 ± 0.06^{B}**	**2.91 ± 0.01^{B}**	**3.3993e-16**	1364,23,621
Ligule length (mm)	**17.64 ± 0.16^{B}**	**19.68 ± 0.90^{AB}**	**19.72 ± 0.23^{A}**	**6.466e-12**	1177,31,528
Panicle length (cm)	24.97 ± 0.12	25.06 ± 0.63	24.80 ± 0.13	0.4736	1188,31,536
Seedling height (cm)	**37.47 ± 0.33^{B}**	**40.35 ± 2.28^{AB}**	**41.08 ± 0.52^{A}**	**1.0705e-7**	1177,31,528
Thousand grain weight (g)	**25.57 ± 0.13^{a}**	**23.59 ± 0.61^{b}**	**23.99 ± 0.16^{b}**	**1.8174e-12**	1321,36,480
Time to flowering (day)	99.83 ± 0.59	100.24 ± 4.22	101.73 ± 1.01	0.7257	1421,38,636
RDVD_osat_34_4					
Phenotype	Genotypic (LSM ± SE)				
	II	DI	DD	P value	n_{II}, n_{DI}, n_{DD}
Culm length (cm)	**111.18 ± 0.78^{B}**	**119.50 ± 8.98^{AB}**	**120.96 ± 1.31^{A}**	**0**	1507,6,244
Culm number (count)	**16.23 ± 0.13^{B}**	**17.67 ± 2.82^{AB}**	**18.12 ± 0.34^{A}**	**0.0000032952**	1508,6,244
Grain length (mm)	**8.66 ± 0.03^{A}**	**9.30 ± 0.35^{AB}**	**8.37 ± 0.05^{B}**	**0**	1727,4,277
Grain width (mm)	**3.04 ± 0.01^{A}**	**2.55 ± 0.17^{AB}**	**2.93 ± 0.02^{B}**	**0.0003**	1727,4,277
Ligule length (mm)	**18.04 ± 0.15^{b}**	**24.00 ± 1.48^{a}**	**19.83 ± 0.32^{a}**	**1.5847e-7**	1490,5,241
Panicle length (cm)	24.93 ± 0.10	24.33 ± 1.23	24.91 ± 0.19	0.9053	1506,6,243
Seedling height (cm)	**38.11 ± 0.30^{B}**	**37.80 ± 5.67^{AB}**	**41.77 ± 0.75^{A}**	**0**	1489,5,242
Thousand grain weight (g)	**25.29 ± 0.11^{A}**	**24.93 ± 3.49^{AB}**	**23.69 ± 0.25^{B}**	**0.0000037815**	1637,4,196
Time to flowering (day)	99.76 ± 0.53	110.71 ± 10.16	104.26 ± 1.59	0.0424	1801,7,287

concerning the significance and prevalence of the above phenomenon.

In this research, we used bioinformatics methods to comprehensively analyze rice retrogenes and RDVs. By calculating the Ka/Ks value of rice retrogenes and their parental genes, it was found that the geometric means of functional retrogenes and retropseudogenes were both <0.6, which tended to be negatively selected. The molecular clock and orthologous analysis showed that rice retrogenes were formed explosively nearly 20 million years ago, and retropseudogenes were either new genes or ancient genes. Seventy-three RDVs affecting 35 retrogenes were detected in the rice genomes in total, mainly in retrogenes that have the functions of transmembrane transport. Most RDVs were detected at a very low frequency in the populations. The distribution differences of some RDVs had obvious characteristics among geographic regions and subpopulations. Through the genotype-phenotype association analysis of 14 RDVs in ancestral retrogenes, it was found that they could all affect at least one of the nine traits of rice.

The role of retrocopy polymorphisms as markers for human population history has been clearly established, but our findings suggest that they can also provide great insight into ongoing evolutionary processes of species. RDVs seem to affect the ancestral retrogenes and the new retrogenes indiscriminately but mainly affect the expressed functional retrogenes, with only one RDV was in the retropseudogene, meaning that the functions they expressed are subject to the selection of the external environment to produce RDV. For the RDVs in the new retrogenes, their formation time was relatively short, which might be a mutation produced by the combined action of natural selection and artificial domestication. The high frequency of a small portion of rice RDVs may indicate that large-scale mutations have occurred in the rice population and have been stably inherited. On the other hand, most low-frequency RDVs may indicate that these indels were deleterious deletions and were therefore subject to negative selection pressure or newer mutations.

Retrogenes with protein coding functions participate in the physiological and biochemical activities of organisms, and the insertion or deletion of RDV can change the functions of these retrogenes, which can be visually expressed in phenotypes. For example, retro_osat_34 (*EXPB8*), which belongs to the expansin family, involves in the process of cell wall loosening [47,48] and has been proven to be widely involved in various developmental processes of plant, including cellular turgor, pollen tube entry into stigma, fruit ripening and softening, organ shedding, leaf formation and response to stress stimuli [49,50]. It was considered the mutations in expansin genes among rice cultivars might generate protein variants that differ in terms of efficiency [51]. Our study also finds that three RDV in *EXPB8* could affect seven identical rice phenotypes, including the culm length, culm number, ligule length, seedling height, grain length, grain width and thousand grain weight. This shows that RDV in a multifunctional retrogene may causes multiple phenotypes.

We performed expression analysis based on RNA-Seq data on nine tissues and found that the expression level of retrogenes in rice is generally lower than that of the parent gene but shows a certain degree of collinearity, which is consistent with the results of a previous study [21]. We detected two highly expressed retrogenes (retro_osat_16 and retro_osat_27) in the anther, and their expression level in the anther was also exceptionally higher than that of their parental gene. Among them, retro_osat_16 (*OsMST5*) belongs

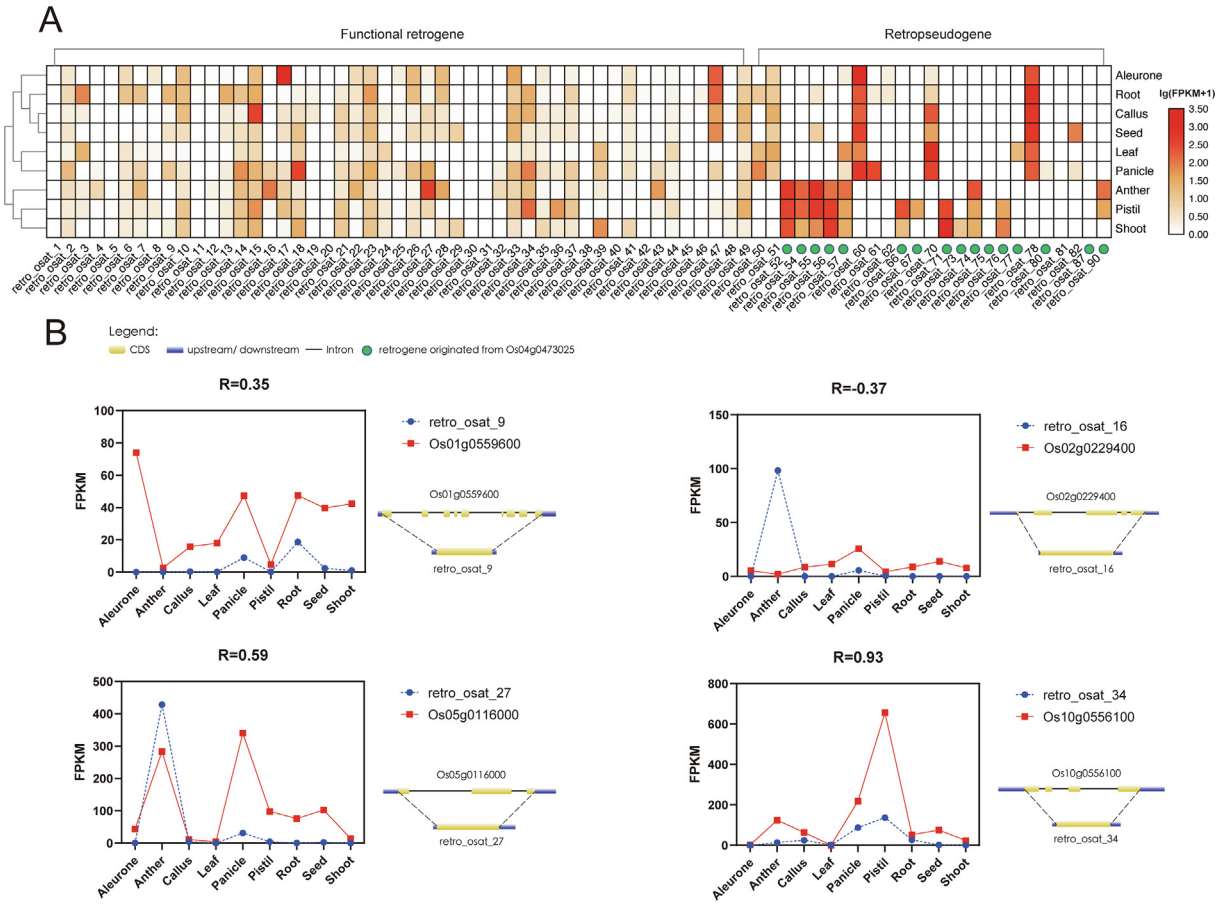

Fig. 6. Expression pattern of retrogenes with their parental genes (A) Heat map of the expression pattern of retrogenes. (B) Comparison of the expression patterns between the above four typical retrogenes and their parental genes.

to the monosaccharide transporter (MST) gene family and has been proven that it was associated with pollen development in rice [52,53]. Although the function of retro_osat_27 is unreported, its parental gene Os05g0116000 encodes globulin and has been shown to play a role in seed development [54,55], but its role in anther is unclear. And its retrogene, retro_osat_27 might has gained new function in anther for its higher expression level. This supports results from previous studies, showing that many retrogenes have broad expression patterns [56]. Numerous studies have revealed a tendency for retrogenes to be expressed in the human testis [5,16,25], which implies a certain connection between the male reproductive organs and retrogenes, and some mechanisms have been confirmed [33,57,58]. However, the set of analyzed retrogenes is relatively small and may not be fully representative.

Compared with the research results of Kabza et al. [25] on RDVs in humans, which was the only species that has been specifically studied for RDV in the past, some of our research results on RDV in rice were similar to those on human RDV. First, the frequency of most RDVs detected between populations was very low. Second, some RDVs were distributed regularly among populations, especially geographically. This shows that these RDVs spread with the migration of species. Third, the mutated, ancestral, and expressed retrogenes of both humans and rice have detected retrogenes that were highly expressed in the male reproductive organs. Human retrogenes were reflected in the testis, while rice retrogenes were reflected in anthers. However, there are still some differences between rice and human RDVs. First, the research of human RDV focuses on detecting the deletions of retrogenes relative to the reference genome, but since we cannot distinguish whether the reference genome or the genome being compared has insertion or deletion mutations [59], in our work, we scanned insertions and deletions of retrogenes, which should be more comprehensive. Second, in the study of Kabza et al. [25], human RDV is defined as a fragment deletion of at least 100 bp, but this threshold was too high for the rice genome. Previous study has shown that the length of indel varies greatly, and there are also different length distributions among different species, the reason for this phenomenon may be that the directed evolution and molecular characteristics are different between human and rice genomes [60], which implies that their retrotransposon activities are different, and it may also be related to their different selection and evolution histories [61,62]. Therefore, we chose to retain insertions or deletions of 8 bp and above as RDVs in this research. Third, our study performed an association analysis between rice RDVs and phenotypes, which has not yet been involved in human RDV research and can more intuitively reflect the impacts of RDV on gene function and species phenotype.

In recent years, molecular breeding technology has gradually become the mainstream of crop breeding worldwide. Because indel molecular markers have the characteristics of high density, high accuracy, good reproducibility, and easy operation, they have been widely used in genetic analysis and molecular-assisted breeding of rice [63–65], oilseed rate [66], wheat [67] and other crops. This is an ideal solution to improve the genetic yield potential of crops. RDV is a special kind of indel because the retrogene is almost always with a single exon structure, so RDV may have a greater probability of affecting crop traits and is more suitable for screening molecular breeding markers than normal indel markers. How-

ever, further research is needed to explore the biological function mechanism of RDV. Furthermore, the relationship between retrogenes and male organs is also worthy of in-depth study, which will help to explore the mechanism of retrogene formation.

4. Materials and methods

4.1. Ka/Ks calculation

Clustalw v2.1 [68] was used to align sequences between retrocopies and their parental genes. All sequence data were downloaded from RetrogeneDB [30]. Next, the files were converted into axt format. KaKs_Calculator v2.0 [69] was used to calculate the Ka and Ks substitution rates and Ka/Ks ratios using the MA method and 1-standard genetic code.

4.2. Retrocopy conservation analysis

We set a phylogenetic tree of all analyzed species based on NCBI Taxonomy [70,71], which is represented in Fig. 3B. To analyze conserved retrocopies, we used the retrocopy data of 25 species from RetrogeneDB [30]. A retrocopy originating in a given lineage ancestor was identified using OrthoVenn2 [42]. We considered a retrocopy to be ancestral for a given lineage if it was observed in any two species from this lineage. All the ancestral clusters (E-valu e < 1×10^{-5}) were then downloaded. Orthologs of *Oryza sativa japonica* retrocopies were identified using Diamond [72]. To be more specific, we used an *Oryza sativa japonica* retrocopy to conduct comparisons with all genes of ancestral clusters. The following parameters were used: identity > 30%; e-value < 1×10^{-10}; score > 200; overlap > 60%. The main branch, which contains *Oryza sativa japonica*, was first detected from the top node to the end node. The accessory branch was then detected.

4.3. Detecting RDVs of known retrocopies

The presently constructed set of rice retrocopies was employed to identify RDVs, and we downloaded and searched the sequence variants detected in 3,010 accessions in the 3,000 Rice Genomes Project [27–29], aiming to detect indels that overlapped with retrogene loci using BEDtools [73]. We incorporated only indels causing the insertion or deletion of at least 8 bp of retrogene sequence for further analysis. The frequency of every indel was calculated in each population. All RDVs were marked on retrogenes by GSDS 2.0 [74] (Fig. 4A).

4.4. GO enrichment analysis

PlantGSEA [75] was used for GO enrichment analysis. Using the default parameters, the clustering results with $p < 0.05$ and FDR < 0.05 were selected as the research objects. Since retropseudogenes are unannotated and only one RDV were detected in them, they were not analyzed.

4.5. Obtaining rice trait data

All phenotypic data of rice samples were downloaded from the 3,000 Rice Genome Project. Among them, the culm length, culm number, grain length, grain width, ligule length, panicle length, seedling height and thousand grain weight traits were downloaded from RFGB v2.0 [27]. Time to flowering (from sowing) trait data were downloaded from the Rice SNP-seek database [76]. The rice phenotype data in the 3,000 Rice Genomes Project have not been fully published, so the total number of statistical samples for each trait was <3,010.

4.6. Statistical analysis of phenotype-genotype correlation

We assumed that a certain RDV locus was compared with a reference genome, the genotype that was homozygous for deletion was DD, the genotype that was homozygous for insertion was II, and the genotype that was heterozygous was DI. The genotypes of each RDV were counted in each rice sample of the 3,000 Rice Genome Project (Table S3), and RDVs with a single genotype of less than three samples were eliminated because they were not sufficient for analysis of variance. Then, we performed a complete random analysis of variance (three-group comparison) between the genotypes of all detected RDV and the phenotypic data of each rice sample to analyze the impact of RDV on the rice phenotype. All analyses were carried out in MATLAB (version R2019a) and R (version 3.5.3).

4.7. Expression analysis

We used RNA-Seq data derived from the Rice Expression Database [26] to analyze the expression patterns of retrogenes. Expression data of 49 functional retrogenes derived entirely from NGS RNA-Seq data of Nipponbare (*Oryza sativa japonica*) provide information on the gene expression profiles of normal tissues. For retropseudogenes, we downloaded the original RNA-Seq data from NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) and the Nipponbare reference genome Os-Nipponbare-Reference-IRGSP-1.0 [77], then performed the same procession and calculation of the FPKM normalized value of retropseudogenes by the method of Rice Expression Database [26]. We considered a retrogene to be expressed only with a normalized expression value of at least 1 FPKM. The experimental ID and developmental stages of each rice tissue are displayed in Table S5.

Author contributions

Conceived and designed the experiments: Y.W., H.-Y.Z., W.-Z.Y. Performed the experiments: H.-Y.Z., X.-Y.C. Analyzed the data: H.-Y.Z., H.B.-L., J.Z. Wrote the paper: H.-Y.Z., Y.W., X.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (31871330), the Scientific Research Starting Foundation of Southwest University (SWU118103), Chongqing Municipal Training Program of Innovation and Entrepreneurship for Undergraduates (S201910635003) and the National Undergraduate Training Program for Innovation and Entrepreneurship (202010635111).

CRediT authorship contribution statement

Haiyue Zeng: Methodology, Visualization, Writing - original draft, Writing - review & editing. **Xingyu Chen:** Data curation, Writing - original draft, Validation. **Hongbo Li:** Validation, Data curation. **Jun Zhang:** Data curation, Investigation. **Zhaoyuan Wei:** Methodology, Investigation. **Yi Wang:** Conceptualization, Methodology, Supervision, Writing - review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Tianqing Zheng and Chunchao Wang (Chinese Academy of Agricultural Sciences, Beijing, China) for their help with the identification of rice indels. We also thank Michał Kabza and Wojciech Rosikiewicz (Adam Mickiewicz University, Poznań, Poland) for their help with the use of RetrogeneDB.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.12.046.

References

[1] Maestre J, Tchenio T, Dhellin O, Heidmann T. mRNA retroposition in human cells: processed pseudogene formation. EMBO J 1995;14(24):6333–8.
[2] Vanin EF. Processed pseudogenes: characteristics and evolution. Annu Rev Genet 1985;19(1):253–72.
[3] BRoslus J. Retroposons—seeds of evolution. Science 1991;251(4995):753.
[4] Young J, Menetrey J, Goud B. RAB6C is a retrogene that encodes a centrosomal protein involved in cell cycle progression. J Mol Biol 2010;397(1):69–88.
[5] Vinckenbosch N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci 2006;103(9):3220–5.
[6] Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, et al. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet 2013;9(1):e1003242.
[7] Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 2009;10(1):19–31.
[8] Kubiak MR, Makalowska I. Protein-coding genes' retrocopies and their functions. Viruses-Basel 2017;9(4).
[9] Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. Nat Rev Genet 2013;14(9):645–60.
[10] Long M, VanKuren NW, Chen S, Vibranovski MD. New gene evolution: little did we know. Annu Rev Genet 2013;47:307–33.
[11] Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, Rossiter JP, et al. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics 1987;1(2):113–25.
[12] Grimaldi G, Skowronski J, Singer MF. Defining the beginning and end of KpnI family segments. EMBO J 1984;3(8):1753–9.
[13] Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 2003;35(1):41–8.
[14] Damert A, Raiz J, Horn AV, Löwer J, Wang H, Xing J, et al. 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res 2009;19(11):1992–2008.
[15] Ciomborowska J, Rosikiewicz W, Szklarczyk D, Makałowski W, Makałowska I. "Orphan" retrogenes in the human genome. Mol Biol Evol 2012;30(2):384–96.
[16] Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, Kaessmann H. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 2008;6(4):e80.
[17] Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. Retrocopy contributions to the evolution of the human genome. BMC Genom 2008;9(1):466.
[18] Navarro FC, Galante PA. A genome-wide landscape of retrocopies in primate genomes. Genom Biol Evol 2015;7(8):2265–75.
[19] Zhong Z, Yang L, Zhang YE, Xue Y, He S. Correlated expression of retrocopies and parental genes in zebrafish. Mol Genet Genom 2016;291(2):723–37.
[20] Du K, He S. Evolutionary fate and implications of retrocopies in the African coelacanth genome. BMC Genom 2015;16(1):915.
[21] Sakai H, Mizuno H, Kawahara Y, Wakimoto H, Ikawa H, Kawahigashi H, et al. Retrogenes in rice (Oryza sativa L. ssp. japonica) exhibit correlated expression with their source genes. Genom Biol Evol 2011;3:1357–68.
[22] Abdelsamad A, Pecinka A. Pollen-specific activation of Arabidopsis retrogenes is associated with global transcriptional reprogramming. Plant Cell 2014;26(8):3299–313.
[23] Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genom Biol 2013;14(3):1–14.
[24] Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L, Lee C, Gerstein M, Consortium GP. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genom Res 2013;23(12):2042–52.
[25] Kabza M, Kubiak MR, Danek A, Rosikiewicz W, Deorowicz S, Polański A, et al. Inter-population differences in retrogene loss and expression in humans. PLoS Genet 2015;11(10):e1005579.
[26] Xia L, Zou D, Sang J, Xu X, Yin H, Li M, et al. Rice expression database (RED): an integrated RNA-Seq-derived gene expression database for rice. J Genet Genom 2017;44(5):235–41.
[27] Sun C, Hu Z, Zheng T, Lu K, Zhao Y, Wang W, et al. RPAN: rice pan-genome browser for~ 3000 rice genomes. Nucleic Acids Res 2017;45(2):597–605.
[28] Wang CC, Yu H, Huang J, Wang WS, Faruquee M, Zhang F, et al. Towards a deeper haplotype mining of complex traits in rice with RFGB v2. 0. Plant Biotechnol J 2020;18(1):14–6.
[29] Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 2018;557(7703):43–9.
[30] Rosikiewicz W, Kabza M, Kosiński JG, Ciomborowska-Basheer J, Kubiak MR, Makałowska I. RetrogeneDB–a database of plant and animal retrocopies. Database 2017;2017.
[31] Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, et al. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 2006;18(8):1791–802.
[32] Betrán E, Thornton K, Long M. Retroposed new genes out of the X in Drosophila. Genom Res 2002;12(12):1854–9.
[33] Emerson J, Kaessmann H, Betrán E, Long M. Extensive gene traffic on the mammalian X chromosome. Science 2004;303(5657):537–40.
[34] Torrents D, Suyama M, Zdobnov E, Bork P. A genome-wide survey of human pseudogenes. Genom Res 2003;13(12):2559–67.
[35] Gaut BS, Morton BR, McCaig BC, Clegg MT. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci 1996;93(19):10274–9.
[36] Bennetzen JL. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 2000;12(7):1021–9.
[37] Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci 2004;101(34):12404–10.
[38] Zhu Q, Ge S. Phylogenetic relationships among A-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. New Phytol 2005;167(1):249–65.
[39] Vitte C, Ishii T, Lamy F, Brar D, Panaud O. Genomic paleontology provides evidence for two distinct origins of Asian rice (Oryza sativa L.). Mol Genet Genom 2004;272(5):504–11.
[40] Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12(2):85–94.
[41] Fu B, Chen M, Zou M, Long M, He S. The rapid generation of chimerical genes expanding protein diversity in zebrafish. BMC Genom 2010;11(1):1–9.
[42] Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res 2019;47(W1):W52–8.
[43] Huang X, Kurata N, Wang Z-X, Wang A, Zhao Q, Zhao Y, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature 2012;490(7421):497–501.
[44] Locke AM, Barding Jr GA, Sathnur S, Larive CK, Bailey-Serres J. Rice SUB1A constrains remodelling of the transcriptome and metabolome during submergence to facilitate post-submergence recovery. Plant Cell Environ 2018;41(4):721–36.
[45] M ZX, J C, B PH, S L, Q G, R WJ, H QW, H W, J L, M OK et al, Genome-wide analyses reveal the role of noncoding variation in complex traits during rice domestication. Sci Adv 2019;5(12).
[46] Zhonglong G, Zheng K, Ying W, Yongxin Z, Yihan T, Chen C, et al. PmiREN: a comprehensive encyclopedia of plant miRNAs. Nucleic Acids Res 2020;48(D1).
[47] Cosgrove DJ. Enzymes and other agents that enhance cell wall extensibility. Ann Rev Plant Physiol Plant Mol Biol 1999;50(1):391–417.
[48] Huang J, Takano T, Akita S. Expression of α-expansin genes in young seedlings of rice (Oryza sativa L.). Planta 2000;211(4):467–73.
[49] Cd J. Loosening of plant cell walls by expansins. Nature 2000;407(6802).
[50] Li Y, Jones L, McQueen-Mason S. Expansins and cell growth. Curr Opin Plant Biol 2003;6(6).
[51] Magneschi L, Kudahettige RL, Alpi A, Perata P. Expansin gene expression and anoxic coleoptile elongation in rice cultivars. J Plant Physiol 2009;166(14).
[52] Deng X, An B, Zhong H, Yang J, Kong W, Li Y. A novel insight into functional divergence of the MST Gene family in rice based on comprehensive expression patterns. Genes 2019;10(3).
[53] Ngampanya B, Sobolewska A, Takeda T, Toyofuku K, Narangajavana J, Ikeda A, et al. Characterization of rice functional monosaccharide transporter, OsMST5. Biosci Biotechnol Biochem 2003;67(3):556–62.
[54] Wang J, Chen Z, Zhang Q, Meng S, Wei C. The NAC transcription factors OsNAC20 and OsNAC26 regulate starch and storage protein synthesis. Plant Physiol 2020;184(4):1775–91.
[55] Lee HJ, Jo YM, Lee JY, Lim SH, Kim YM. Lack of globulin synthesis during seed development alters accumulation of seed storage proteins in rice. Int J Mol Sci 2015;16(7):14717–36.
[56] Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 2005;3(11):e357.
[57] Kleene KC. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech Dev 2001;106(1–2):3–23.
[58] Fontanillas P, Hartl DL, Reuter M. Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin. PLoS Genet 2007;3(11):e210.
[59] Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genom Res 2006;16(9):1182–90.

[60] Britten RJ, Rowen L, Williams J, Cameron RA. Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci 2003;100 (8):4661–5.
[61] Mullaney JM, Mills RE, Pittard WS. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet 2010;Vol. 19(No.2):6.
[62] Lü Y, Cui X, Li R, Huang P, Zong J, Yao D, et al. Development of genome-wide insertion/deletion markers in rice based on graphic pipeline platform. J Integr Plant Biol 2015;57(11).
[63] Kim S-R, Ramos J, Ashikari M, Virk PS, Torres EA, Nissila E, et al. Development and validation of allele-specific SNP/indel markers for eight yield-enhancing genes using whole-genome sequencing strategy to increase yield potential of rice, Oryza sativa L. Rice 2016;9(1):12.
[64] Whankaew S, Kaewmanee S, Ruttajorn K, Phongdara A. Indel marker analysis of putative stress-related genes reveals genetic diversity and differentiation of rice landraces in peninsular Thailand. Physiol Mol Biol Plants 2020.
[65] Gull S, Haider Z, Gu H, Raza Khan RA, Miao J, Wenchen T, Uddin S, Ahmad I, Liang G. InDel marker based estimation of multi-gene allele contribution and genetic variations for grain size and weight in rice (Oryza sativa L.). Int J Mol Sci 2019;20(19):4824.
[66] Chen R, Chang L, Cai X, Wu J, Liang J, Lin R, et al. Development of InDel markers for brassica rapa based on a high-resolution melting curve. Horticult Plant J 2020.
[67] Qiu L, Wang H, Li Y, Wang W, Liu Y, Mu J, et al. Fine mapping of the wheat leaf rust resistance gene LrLC10 (Lr13) and validation of its co-segregation markers. Front Plant Sci 2020;11:470.
[68] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23(21): 2947–8.
[69] Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinform 2010;8(1):77–80.
[70] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2010;39(suppl_1):D32–D37.
[71] Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2010;39(suppl_1):D38–51.
[72] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12(1):59–60.
[73] Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protocol Bioinform 2014;47(1):11.12.11–11.12.34.
[74] Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 2015;31(8):1296–7.
[75] Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res 2013;41(W1):W98–W103.
[76] Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, et al. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res 2017;45(D1):D1075–81.
[77] Kawahara Y, Mdl Bastide, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice 2013;6(1).