scientific reports

Check for updates

OPEN

Evolutionary and synteny analysis of *HIS1*, *BADH2*, *GBSS1*, and *GBSS2* in rice: insights for effective introgression breeding strategies

Insu Lim[1], Yong-Jin Park[2] & Jungmin Ha[1]✉

The key genes *BADH2*, *GBSS1*, *GBSS2*, and *HIS1* regulate the fragrance, starch synthesis, and herbicide resistance in rice. Although the molecular functions of four genes have been investigated in the *Oryza sativa* species, little is known regarding their evolutionary history in the *Oryza* genus. Here, we studied the evolution of four focal genes in 10 *Oryza* species using phylogenetic and syntenic approaches. The *HIS1* family underwent several times of tandem duplication events in the *Oryza* species, resulting in copy number variation ranging from 2 to 7. At most one copy of *BADH2*, *GBSS1*, and *GBSS2* orthologs were identified in each *Oryza* species, and gene loss events of *BADH2* and *GBSS2* were identified in three *Oryza* species. Gene transfer analysis proposed that the functional roles of *GBSS1* and *GBSS2* were developed in the Asian and African regions, respectively, and most allelic variations of *BADH2* in *japonica* rice emerged after the divergence between the Asian and African rice groups. These results provide clues to determine the origin and evolution of the key genes in rice breeding as well as valuable information for molecular breeders and scientists to develop efficient strategies to simultaneously improve grain quality and yield potential in rice.

Rice is a staple food crop for half of the world population, contributing to nearly 20% of the total calorie intake of humans[1]. As the global population is predicted to reach almost 10 billion by 2050[2], the development of novel high-yield and superior-quality rice cultivars is urgently required to meet the future global food demand[3].

Rice is one of the most extensively studied crop species because of its social and economic importance as well as its environmental impact[4]. Many genetic factors related to grain production such as starch biosynthesis and abiotic stress resistance, have captured the interest of researchers[5], including the genes betaine aldehyde dehydrogenase (*BADH2,* Os08g0424500), granule-bound starch synthase 1 (*GBSS1,* Os06g0133000), granule-bound starch synthase 2 (*GBSS2,* Os07g0412100), and HPPD (4-hydroxyphenylpyruvate dioxygenase) inhibitor sensitive 1 (*HIS1,* Os02g0280700). *BADH2* encodes for the enzyme that biosynthesizes 2-acetyl-1-pyrroline, which is a potent rice flavor compound[6]. Fragrant type of rice cultivars have been identified to be determined by allelic variations of the *BADH2* gene[6]. *GBSS1* and *GBSS2* are responsible for the amylose content of rice by converting ADP-glucose into amylose instead of amylopectin in the starch biosynthesis pathway[7]. The amylose to amylopectin ratio is the most important physiological indicator of rice grain quality, particularly regarding the cooking and eating qualities[8,9]. *HIS1* imparts resistance to bTH benzobicyclon and other b-Triketone herbicides that are widely applied to weed control in rice paddy fields[10]. Because weed control of large-scale farming is highly dependent on herbicides, the wide-spectrum tolerance to multiple herbicides are essential traits for increased rice production with reduced labor.

Oryza sativa and *O. glaberrima* were independently domesticated in Asia and Africa, respectively, from different wild ancestors[1]. In general, wild species obtain various alleles for resistance or tolerance to environmental stresses as adaption to a broad biogeographical diversity[11,12]. This rich diversity of wild relatives allows for introgression breeding to be a prominent approach in rice to enhance agricultural traits such as resistance to biotic and abiotic stresses[13–15]. Several previous studies have reported that introgression of alleles from wild

[1]Department of Plant Science, Gangneung-Wonju National University, Gangneung, South Korea. [2]Department of Plant Sciences, Kongju National University, Yesan 340-702, Korea. ✉email: j.ha@gwnu.ac.kr

germplasms enhanced the productivity and grain quality in soybeans and tomatoes[16,17]. However, in rice, most introgression breeding strategies have focused on conferring pest and disease resistance to *O. sativa* species[13–15].

The genus *Oryza* comprises of two cultivated species and 22 wild species with 11 representative genome types: six diploids (AA, BB, CC, EE, FF, and GG) and five polyploids (BBCC, CCDD, HHJJ, HHKK, and KKLL)[18]. *Oryza sativa* and *Oryza glaberrima*, cultivated species, belong to the AA group with six wild species (*Oryza nivara*, *Oryza rufipogon*, *Oryza barthii*, *Oryza glumaepatula*, *Oryza longistaminata*, and *Oryza meridionalis*). Wild *Oryza* species have been considered as potential genetic resources that carry valuable alleles which are not present in the cultivated species[19]. Regarding abiotic and biotic stress resistance, many rice breeders have characterized favorable alleles in wild *Oryza* species[11,12,20], and numerous alleles conferring stress resistance have been introduced from wild *Oryza* species into elite cultivars[18,21], as demonstrated by the chromosomal introgression of iron resistance from *O. meridionalis* into *O. sativa*[22]. These results propose that the stability and productivity of elite rice cultivars could be improved by gene transfer from wild germplasms. However, the genes related to grain quality and herbicide resistance in rice, such as *BADH2*, *GBSS1*, *GBSS2*, and *HIS1*, have only been investigated within the *O. sativa* group.

This study aimed to investigate the evolution and divergence of the *BADH2*, *GBSS1*, *GBSS2*, and *HIS1* gene families within the *Oryza* genus, including both cultivated and wild species. To provide a robust foundation for the evolution analysis of gene families, a phylogenetic tree was constructed among 10 *Oryza* species and macrosynteny was analyzed in four Asian *Oryza* species, including *O. sativa* ssp. *japonica* and *indica*, *O. nivara*, and *O. rufipogon*. Through a comprehensive phylogenetic and syntenic network approach, the evolutionary events and selection pressures were explored. The findings in this study contribute to a better understanding of the evolution of these target gene families and provide valuable insights for breeders in the selection of beneficial germplasm to develop more adaptive and productive rice cultivars.

Results

Phylogenetic analysis

We constructed two phylogenetic trees using peptide and nucleotide sequences, respectively, using 50 true orthologs from 10 *Oryza* species with *A. thaliana* and *G. max* as outgroups (Fig. 1). There were no significant differences in topology between two phylogenetic trees (Fig. 1 and Fig. S2). The eight AA-genome *Oryza* species were clustered together, thereby separating from *O. punctata* (BB-genome) and *O. brachyantha* (FF-genome) (Fig. 1). Within the AA-genome clade, the African and Asian species were grouped separately and the Asian species were further separated into two groups, namely the cultivated (*O. sativa* ssp. *japonica* and *O. sativa* ssp. *indica*) and wild groups (*O. rufipogon* and *O. nivara*) (Fig. 1). Using the divergence time between *O. brachyantha* and the other *Oryza* species at 15 million years ago (mya) as the calibration point[1], the divergence times for the branching points was calculated in the phylogenetic tree (Fig. 1). The estimated divergence time between the AA- and BB-genomes was 14.8 mya, 9.2 mya between the African and Asian clades, 3.6 mya between the Asian wilds, and 0.4 mya between the Asian cultivars (Fig. 1).

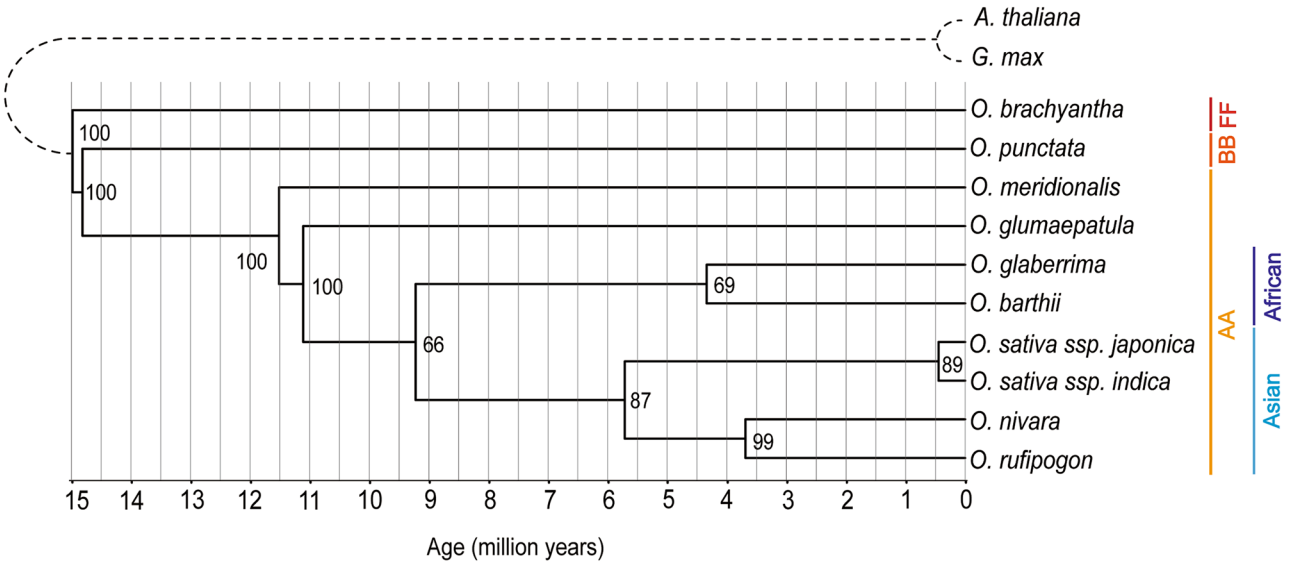

Figure 1. Phylogeny of 10 *Oryza* species and two dicot plants. The phylogenetic tree was constructed using the amino acid sequences of 50 true orthologs among 10 *Oryza* species and two dicot species as outgroups. The number of each node is the bootstrap value between nodes which is obtained from 1000 bootstrap replications. Divergence times within the genus *Oryza* were estimated by assuming 15 million years ago for the divergence point between *O. brachyantha* and the other *Oryza* species. *Mya* million years ago.

Synteny analysis

To further understand the evolutionary process during the domestication of cultivated rice, *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica*, we further investigated the macrosynteny among the four Asian species at the chromosomal-level (Fig. 2A). *O. sativa* species had a more conserved synteny with *O. rufipogon* than *O. nivara* (Fig. 2B). *O. sativa* ssp. *japonica* and *O. rufipogon* showed the most conserved synteny among the four species and its breakpoint was identified only once on chromosome 3 (Fig. 2). The conservation of synteny between *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* was slightly lower than that of synteny between *O. sativa* ssp. *indica* and *O. rufipogon* (Fig. 2B).

Identification of the *BADH2*, *GBSS1*, *GBSS2*, and *HIS1* gene families

Orthologous genes of *BADH2*, *GBSS1*, *GBSS2*, and *HIS1* were identified in the 10 *Oryza* species using clustering analysis based on the annotation information of the *O. sativa* ssp. *japonica* reference genomic sequence[23]. A total of 8, 10, 8, and 43 orthologous genes were identified as *BADH2*, *GBSS1*, *GBSS2*, and *HIS1* gene family members, respectively (Table 1). For *BADH2*, *GBSS1*, and *GBSS2*, in general, single-copy genes remained across the *Oryza* species (Table 1). *BADH2* was lost in *O. meridionalis* and *O. punctata* and *GBSS2* was lost in *O. brachyantha* and *O. meridionalis* (Table 1). While *GBSS1* was uniformly identified in every *Oryza* species without any loss events (Table 1). The *HIS1* family showed high variation in their copy number ranging from two (*O. brachatia* and *O. glumaetuala*) to seven (*O. sativa* ssp. *japonica* and *O. puctata*) (Table 1).

Gene synteny analysis of *HIS1*

The phylogenetic tree was constructed using 43 *HIS1* orthologous genes and was further clustered into five sub-classes consisting of *HIS1* and *HSL 1–4* based on the annotation data of *O. sativa* ssp. *japonica* (Fig. 3). In each clade, 9, 4, 5, 9, and 16 genes were grouped as *HIS1*, *HSL1*, *HSL2*, *HSL3*, and *HLS4*, respectively (Fig. 3). *HIS1* and *HSL* genes are tandemly located on chromosomes 2, 3, and 6 (Fig. 4A). *HIS1* and *HSL3* genes are located on chromosome 2 in all *Oryza* species except for *O. brachyantha* (Fig. 4A). *HSL1*, *HSL2*, and *HSL4* are mostly located on chromosome 6 in *Oryza* species and additional two *HSL4* are located in chromosome 3 in *O. punctata* (Fig. 4A). In *O. sativa* ssp. *japonica*, one copy of *HSL1* and *HSL4* were identified in chromosome 6, and the *HSL1* gene was only detected in *O. glaberrima*, *O. barthii*, and *O. sativa* ssp. *japonica* (Fig. 4A). The synteny analysis showed that all *HIS1* and *HSL3* orthologs had syntenic relationships in a pair-wise manner, while the synteny of *HSL1*, *HSL2*, and *HSL4* were not maintained among most of the *Oryza* species (Fig. 4B).

Gene synteny analysis of *GBSS1*, *GBSS2*, and *BADH2*

Among the four Asian species, synteny analysis was conducted for *GBSS1*, *GBSS2*, and *BADH2*, which had a low copy number variation (Table 1). Among all the *Oryza* species with orthologous genes, including orthologs that were identified based on their sequence similarity, the synteny was highly conserved between the target genes

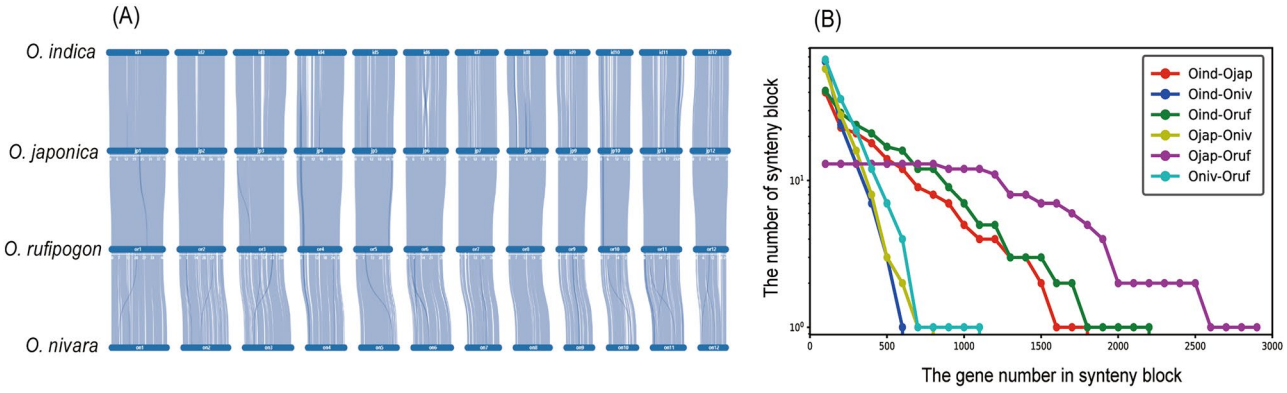

Figure 2. Synteny analysis between Asian rice groups. (**A**) Macrosynteny visualization of four Asian species ordered by *O. indica*, *O. japonica*, *O. rufipogon*, and *O. nivara*. (**B**) The distribution of the synteny block size of six species pairs show the degree of syntenic conservation. The x-axis indicates the number of genes required to call a synteny block and the y-axis indicates the number of synteny blocks between each species pair.

	O. bar	*O. bra*	*O. gla*	*O. glu*	*O. ind*	*O. jap*	*O. mer*	*O. niv*	*O. pun*	*O. ruf*
HIS1	4	2	6	2	3	7	4	5	7	3
GBSS1	1	1	1	1	1	1	1	1	1	1
GBSS2	1	0	1	1	1	1	0	1	1	1
BADH2	1	1	1	1	1	1	0	1	0	1

Table 1. Number of orthologs (*HIS1*, *BADH2*, *GBSS2*, and *GBSS2*) in the 10 *Oryza* species.

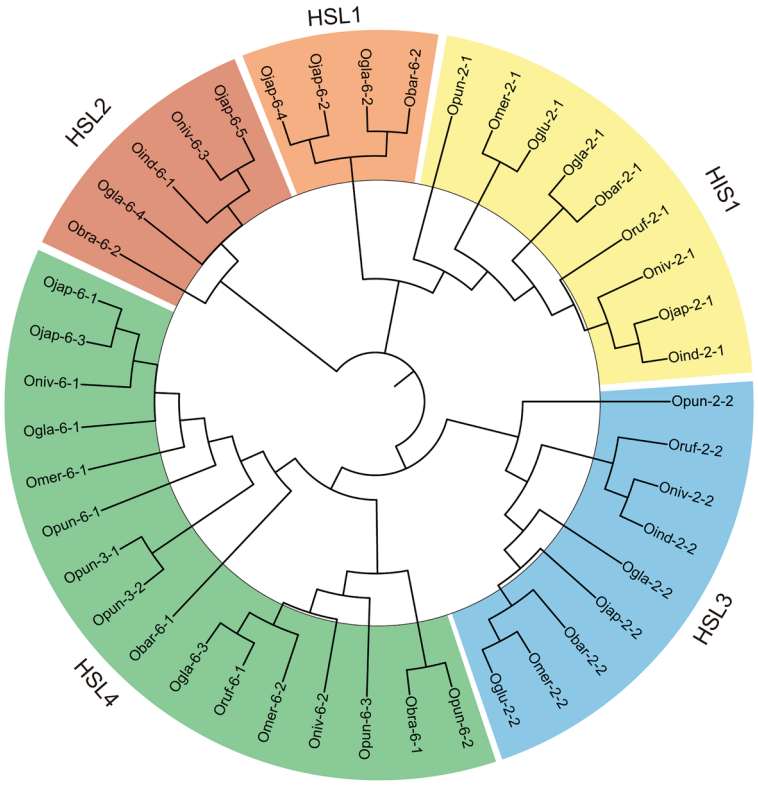

Figure 3. Phylogenetic tree of the *HIS1* and *HSL* gene families. The phylogenetic neighbor-joining tree was constructed using the amino acid sequences of *HIS1* and *HSL* orthologs in 10 *Oryza* species. In the *HSL* family, four subclasses (*HSL1*, *HSL2*, *HSL3*, and *HSL4*) were identified. Gene classes are indicated by the label colors, where 9 *HIS1* were labeled in yellow, 4 *HSL1* in orange, 5 *HSL2* in red, 9 *HSL3* in blue, and 16 *HSL4* in green.

(Fig. 5A, Table S1). The synteny blocks harboring *GBSS1*, *GBSS2*, and *BADH2*, were identified on chromosomes 6, 7, and 8, respectively (Fig. 5B). In the synteny blocks of *GBSS1*, there are 671 genes when comparing *O. sativa* ssp. *indica* with *O. sativa* ssp. *japonica*, 1766 genes when comparing *O. sativa* ssp. *japonica* with *O. rufipogon*, and 133 genes when comparing *O. rufipogon* with *O. nivara* (Fig. 5B). The *GBSS2* block contained 54, 1634, and 235 gene numbers. The block of *BADH2* had 758, 1493, and 184 gene numbers (Fig. 5B).

Selection pressure

The K_a/K_s ratio of the homologs was calculated to determine whether *BADH2*, *GBSS1*, *GBSS2*, and *HIS1* underwent negative or positive selection (Fig. 6). The mean K_a/K_s values for *HIS1*, *GBSS1*, *GBSS2*, and *BADH2* were 0.92, 0.1, 0.69, and 0.16, respectively, indicating that these genes evolved under purifying selection (Fig. 6). Moreover, the mean K_a/K_s value of the *GBSS1* and *BADH2* gene pair were lower than those of the means other two families, which suggests that the *GBSS1* and *BADH2* duplicates evolved at a slower rate (Fig. 6). Additionally, we calculated the K_a/K_s values by comparing target orthologs from two monocot plants, *Sorghum bicolor* (*S. bicolor*) and *Lessia perrie* (*L. perrie*), with those in *Oryza* species. The comparisons between *S. bicolor* and *Oryza* species showed K_a/K_s values of 0.2, 0.13, and 0.14 for *GBSS1*, *GBSS2*, and *BADH2*, respectively (Table S2). When comparing *L. perrie* with *Oryza* species, the K_a/K_s values were 0.12 and 0.11 for *GBSS1* and *BADH2*, respectively (Table S2). These results indicate that the target genes had been undergone purifying selection within cereal plants.

Discussion

The evolution and speciation of Asian rice species have remained unclear due to their high ortholog sequence similarity[1,24]. To date, there are two hypotheses concerning the origin and domestication of the *O. sativa* species[24]. The single-domestication hypothesis posits that the *O. sativa* species originated from a wild ancestor and the differentiation between *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* occurred after domestication of the cultivated species[25–28]. This single-domestication hypothesis is mainly supported by the molecular evidence of the identical sequences of the key domestication genes between the *O. sativa* subspecies, including *sh4* that reduces shattering and *prog1* which is associated with erect growth[25–27]. In contrast, the multiple-domestication hypothesis postulates that *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* were domesticated separately from different wild ancestors[29,30]. This multiple-domestication hypothesis has gained support through the phylogenetic analyses which shows that the *O. sativa* subspecies are separated into distinct clades and are closer to the different wild accessions than each other[29,30]. The phylogenetic tree that was constructed in this study using the true

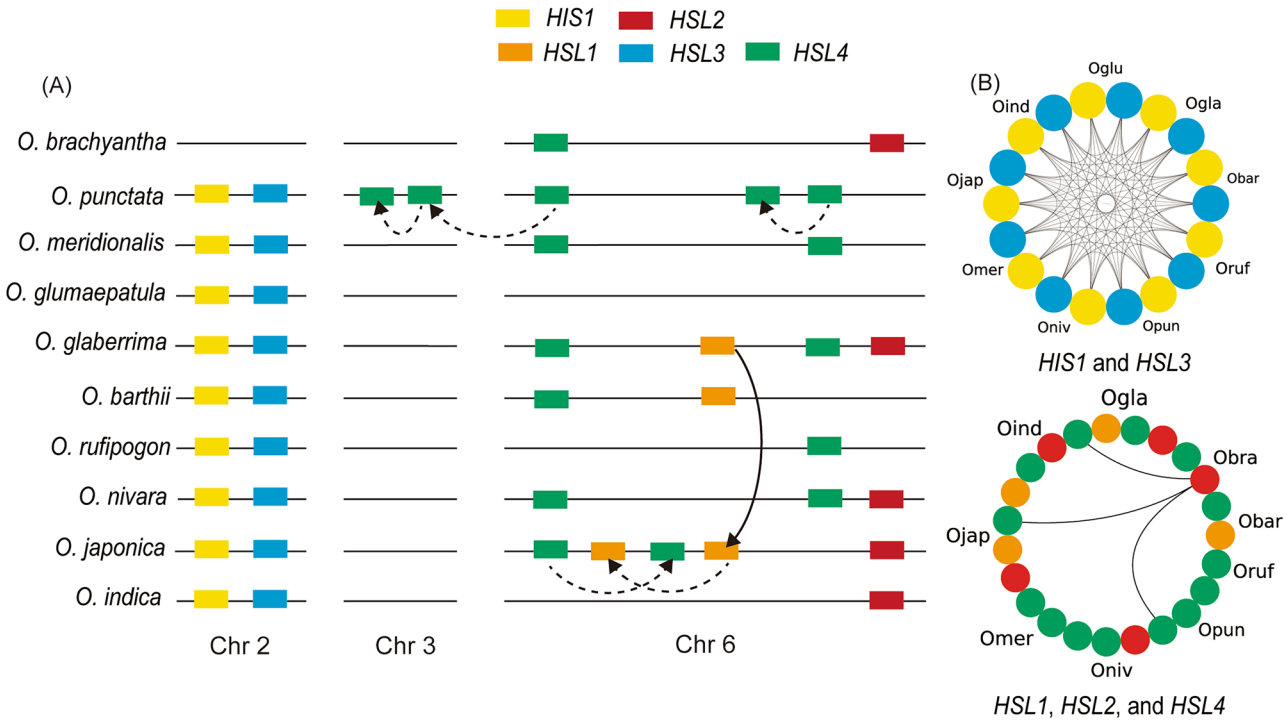

Figure 4. Chromosomal distribution and syntenic network of *HIS1* and *HSL* gene families. (**A**) The location of *HIS1* and *HSL* gene families at rice chromosomes 2, 3, and 6. Rectangles represented genes and their colors indicate the gene classes. The dotted and solid lines indicate the gene duplication and transfer events, respectively. (**B**) Syntenic network of *HIS1* and *HSL3* (upper) and *HSL1*, *HSL2*, and *HSL4* (lower). The lines represent the syntenic connection between the two genes and the circle color indicates the subclass of the gene family.

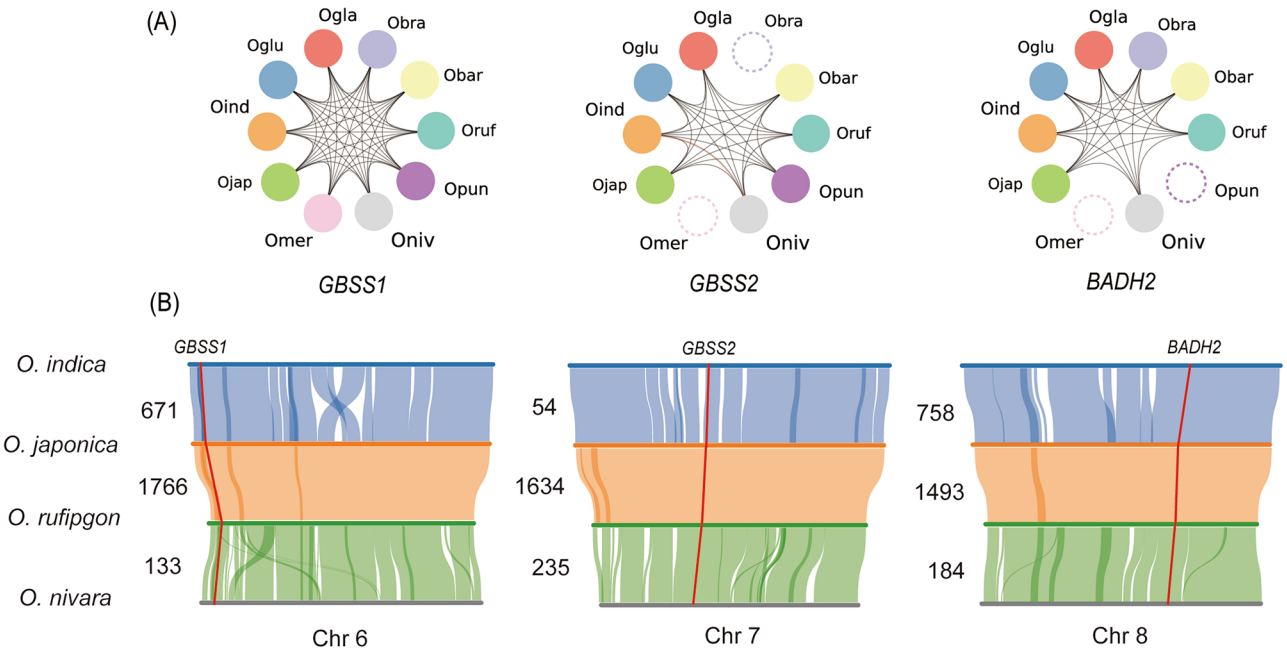

Figure 5. Syntenic conservation of *GBSS1*, *GBSS2*, and *BADH2*. (**A**) Syntenic network of the three genes across the *Oryza* species. The circles represent the orthologous genes detected by sequence similarity, dashed circles indicate the absence of the gene, and the lines represent a syntenic connection between the two genes. The line colors indicate whether the two genes are connected, and gray or red indicates the presence or absence thereof. (**B**) The feature of synteny blocks harboring the three genes among the Asian species including *O. sativa* ssp. *indica*, *O. sativa* ssp. *japonica*, *O. rufipogon*, and *O. nivara*. The red lines show the location and syntenic linkage of the orthologs. The numbers beside each synteny represent the number of genes in the synteny block of the focal genes.

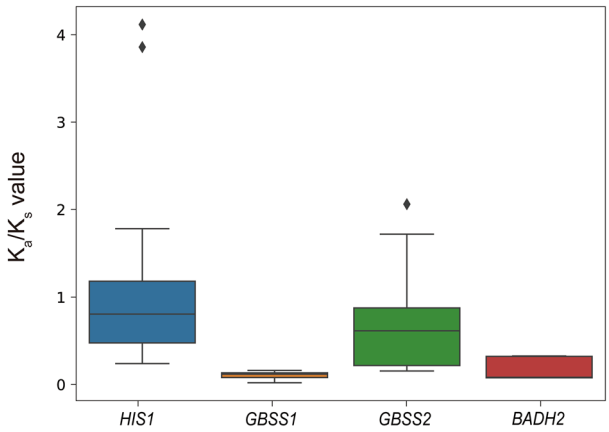

Figure 6. Comparison of the pair-wise K_a/K_s values of *HIS1*, *GBSS1*, *GBSS2*, and *BADH2* genes. The K_a/K_s values of the ortholog pairs were calculated in each of the focal genes to determine the selection pressure. The ortholog genes were obtained based on the synteny block among the 10 *Oryza* species.

orthologs based on syntenic relationship was consistent with the single-domestication hypothesis as the *O. sativa* subspecies were grouped into the same clade (Fig. 1). The synteny analysis revealed that the genomic structure of the *O. sativa* species was more conserved with *O. rufipogon* than those of *O. nivara* (Fig. 2B). This result is in agreement with a previous study which reported that the *O. sativa* species may originate from *O. rufipogon* and that *O. nivara* is one of the ecological varieties of *O. rufipogon*[25–27].

Genomic similarity is a key factor that determines genetic compatibility which enables the transfer of desired traits between two species through interspecific crossing[31]. In rice, several reproductive isolations had been observed in interspecific hybrids, which resulted in inviability, weakness, and sterility[32]. Several genetic models have been suggested to explain the mechanisms of reproductive isolations in plants and structural variations were identified as one of leading factors of reproductive isolation[32]. In rice and Arabidopsis, pollen incompatibility and inviability had been reported in their hybrids, respectively, due to a change of the gene locus caused by reciprocal gene loss of duplicated genes[20,33]. Other studies have reported that chromosomal rearrangements enhance the reproductive isolation by suppressing recombination[31,34,35], which results in unbalanced gametes that may be inviable[31,34].

Suppressed recombination also increased the extent of linkage disequilibrium (LD) block, thereby restricting gene flow in potentially larger genomic regions[34,36,37]. When introgression of a favorable gene from an external germplasm into a cultivar occurs, other genes that confer undesirable traits can also be transferred if the genes are located in the same LD block[38,39]. This phenomenon is also known as the linkage drag problem and it is a major concern in introgression breeding which prevents breeders from introducing desirable traits into elite cultivars[40–42]. Linkage drag that leads to the negative relationships among the yield potential, grain quality, and environmental resistance have been reported in *O. sativa* species[40,41]. The synteny analysis in this study identified that *O. rufipogon* would be a more genetically compatible germplasm for *O. sativa* breeding with reduced reproductive isolation and linkage drag problems (Fig. 2).

Gene evolution analysis has been widely used to investigate gene expansion, domestication process, genetic background, etc. Copy number variation (CNV) is a structural variation that alters the dosage of genes, which could result in phenotypic changes[43,44]. In plants, most resistance traits are polygenic and highly affected by CNV[43,44]. A previous study on durum wheat reported that frost resistance levels were determined by the CNV of the *CBF-A14* gene family[45]. In rice, the CNV of 28 functional genes was identified to be involved in insect resistance and response to salt stress[43]. In *Brassica napus*, 563 resistance genes experienced 1137 CNV events including 704 deletions and 433 duplications[46]. Based on the phylogenetic clustering, a total of 43 *HIS1* genes were further clustered into five subclasses including *HIS1* (9), *HSL1* (4), *HSL2* (5), *HSL3* (9), and *HLS4* (16) (Fig. 3). We identified that the *HIS1* and *HSL* families experienced multiple duplication events (Fig. 4). In *O. punctata*, a tandemly duplicated *HSL4* gene was identified on chromosome 6 and an additional pair of tandemly duplicated *HSL4* genes were detected on chromosome 3 (Fig. 4). Because these *HSL4* genes of *O. punctata* were not found in other *Oryza* species, they might be duplicated after the *O. punctata* speciation event (Fig. 4). The *HSL1* genes were only identified in *O. glaberrima*, *O. barthii*, and *O. sativa* ssp. *japonica* (Fig. 4). Considering that *O. glaberrima* was domesticated from *O. barthii*, the *HSL1* gene probably originated from *O. barthii*, and then moved to *O. satvia* ssp. *japonica* via *O. glaberrima* or directly from *O. barthii* (Fig. 4). In *O. sativa* ssp. *japonica*, duplication events of *HSL1* and *HSL4* were identified on chromosome 6 (Fig. 4). Because four of the five duplicated *HSL1* and *HSL4* genes are located close to each other (less than 50 kb interval), we propose that the *HIS1* and *HSL* families were mainly expanded through tandem duplication events in the *Oryza* species (Fig. 4). These results are in agreement with a previous study where tandem duplication events were frequently identified in CNVs[47]. Our gene evolution analysis can facilitate the improvement in herbicide resistance of rice cultivars through gene transferring from wild germplasm that has a high copy number of *HIS1* and *HSL* genes such as *O. punctata* (Fig. 4).

While *HIS1* has diverse CNVs across 10 *Oryza* species, *GBSS1*, *GBSS2*, and *BADH2* genes have at most one copy in the 10 *Oryza* species and the syntenic relationship of their orthologs was deeply conserved in pair-wise

comparison among *Oryza* species (Table 1 and Fig. 5A). These results suggest that these three genes descended from a common *Oryza* ancestor to present-day cultivars (Fig. 5). The *BADH2* and *GBSS2* loss events were identified in *O. brachyantha*, *O. punctata*, and *O. meridionalis*, which suggests that the functional role of the *BADH2* and *GBSS2* genes were developed after speciation from *O. meridionalis* (Table 1). Meanwhile, no loss event of *GBSS1* was identified in the *Oryza* species (Table 1). Plants have some critical genes that play essential roles in their survival such as photosynthesis, cell division, and reproduction[48–50]. In general, the genetic diversity of essential genes is highly conserved among related species, because the malfunction of these genes directly affects their fitness in the population[48]. In rice, a previous study reported that *GBSS1* is expressed in the endosperm and pollen grains, while the expression of *GBSS2* is limited to the vegetative tissues and pericarp[7]. The starch content of the endosperm serves as the primary source for seed germination and seedling growth[7]. Therefore, the prevalence of *GBSS1* may be the product of selection pressure for its critical role in seed germination vigor, as wild relatives and landraces with lower germination rates were extinguished in nature or removed from the breeding pool during rice evolution and domestication. This result is consistent with our selection pressure analysis which showed that *GBSS1* had the lowest K_a/K_s ratio indicating that the sequence diversity of *GBSS1* is highly conserved during evolution (Fig. 6).

Gene exchange is a key evolutionary mechanism that enhances the adaptability of the population against environment stresses[51]. Natural introgression between the Asian cultivated and wild species is common because they are often sympatric[24,51,52]. The African cultivated species have a relatively limited gene pool in the wild species compared to Asian rice and several historic introgression events from the Asian species are reported[24]. *GBSS1*, *GBSS2*, and *BADH2* genes are located in highly conserved synteny blocks as single-copy genes over the 10 *Orzya* species, which indicates they are true orthologs in the genus *Orzya*. Using the true orthologs of *GBSS1*, *GBSS2*, and *BADH2*, reconciled gene trees were constructed and estimated divergence time was calculated between orthologous gene pairs, to investigate gene transfer events of the target genes across the *Oryza* species (Fig. S1 and Table 2). Our results proposed that three transfer events had occurred in *GBSS1* from the Asian groups into the African group (*O. glaberrima* and *O. barthii*) and other wild species (*O. glumaepatula* and *O. meridionalis*) (Fig. 7). In contrast, the *GBSS2* was transferred from *O. barthii* into the Asian species including *O. sativa* ssp. *japonica* and *O. nivara* (Fig. 7). This gene flow in the opposite direction between Asian and African groups indicates that *GBSS1* and *GBSS2* gained their subfunctions independently in the Asian and African rice populations, respectively, and they were spread to other regions during the domestication process. For *BADH2*,

Genes	Orthologous pair	K_a	K_s	Estimated divergence time (mya)
GBSS1	*O. nivara–O. meridionalis*	0.0015	0.0234	1.80
	O. rufipogon–O. meridionalis	0.0022	0.0234	1.80
	O. sativa–O. barthi	0.0007	0.0091	0.70
	O. sativa–O. glaberrima	0.0007	0.0091	0.70
	O. sativa–O. glumaepatula	0.0000	0.0045	0.35
GBSS2	*O. barthi–O. glumaepatula*	0.0014	0.0087	0.67
	O. barthi–O. sativa ssp. *japonica*	0.0058	0.0068	0.52
	O. barthi–O. nivara	0.0027	0.0065	0.50
BADH2	*O. barthi–O. rufipogon*	0.0000	0.0055	0.42

Table 2. The estimated divergence time for pairs of transferred orthologs.

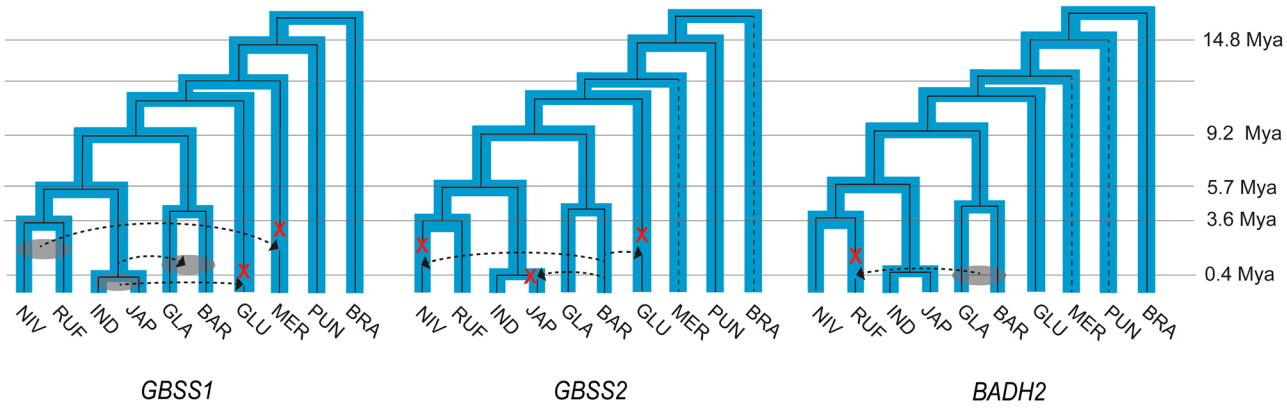

Figure 7. Schematic illustration of phylogenetic inference to detect gene transfer events. The species tree is depicted by the thick blue area and the inner black lines represent the phylogeny of each gene. The dotted line indicates the absence of the gene in the node. The red x symbol indicates the loss of the gene in the node. The translucent oval between two species represent that either a species or another species contributes to transfer event. The arrows indicate the gene transfer between two species or node.

a gene from either *O. barthii* or *O. glaberrima* was moved into *O. rufipogon*, which suggests that most *BADH2* alleles in *japonica* rice had been developed after divergence between the Asian and African groups (Fig. 7)[6,53].

Overall, this study enhances our knowledge of the gene family evolution in rice and offers practical implications for rice breeding efforts, which ultimately supports the development of improved rice varieties with enhanced adaptability and productivity.

Material and methods

Data source

The peptide sequences, nucleotide sequences, and gene location information (GFF format) of three domesticated species (*O. satvia* ssp. *japonica*, *O. sativa* ssp. *indica*, and *O. glaberrima*), seven wild species (*O. rufipogon*, *O. nivara*, *O. barthii*, *O. brachyantha*, *O. glumaepatula*, *O. meridionalis*, and *O. punctata*), and two dicot plants (*A. thaliana* and *G. max*) were downloaded from EnsemblPlants (https://plants.ensembl.org, accessed on 17 April 2023) and Rice Genome Hub (https://rice-genome-hub.southgreen.fr, accessed on 17 April 2023) (Table S3). The protein and nucleotide sequences were filtered out and the longest isoform per gene was retained for downstream analysis.

Species phylogenetic tree

Synteny analysis was conducted to obtain true orthologs across the 12 species (see section "Synteny detection" for detail). In our study, single-copy orthologous genes located in synteny blocks among 10 *Oryza* species were defined as true orthologous gene. A total of 50 syntenic orthologs, which are shared between *O. sativa* ssp. *japonica* and the 11 other species, were selected as true orthologs for phylogenetic analysis. Peptide and nucleotide sequences of the 50 true orthologs were aligned using ClustalOmega v1.2.4[54] and then concatenated. Phylogenetic trees were built using RAxML v8.2.12 with 1000 bootstrap replicates[55]. Two maximum-likely hood models, JJT and GTRGAMMA, were used for phylogenetic tree of peptide and nucleotide sequences, respectively[55]. The phylogenetic tree was visualized using Interactive Tree Of Life[56]. The divergence time between the nodes was estimated by the MCMCtree package embedded in PAML v4.10.6[57].

Synteny detection

The macrosynteny blocks among the 12 species were identified using BLASTP version 2.9.0+ and software MCScanX[58,59]. The protein sequences of the homolog pairs were obtained by all-against-all BLASTP search with the standard parameter setting for MCScanX analysis (evalue: 1e−10, num alignments: 5, and outfmt: 6). The BLASTP outputs with the gene location data were imported into the MCScanX to identify the synteny blocks with the default parameters (match score: 50, gap penalty: − 1, match size: 5, evalue: 1e−5, and max gaps: 25)[58]. The synteny was visualized using SynVisio[56].

Identification of *BAHD2*, *GBSS1*, *GBSS2*, and *HIS1* orthologs

The protein sequences of the 12 species were clustered into orthologous groups using OrthoFinder v2.5.4 with sequence similarity searches conducted using DIAMOND[60]. Based on the gene annotation of *O. sativa* spp. *japonica*, we obtained the names of the four target genes: *BADH2* (Os08g0424500), *GBSS1* (Os06g0133000), *GBSS2* (Os07g0412100), and *HIS1* (Os02g0280700). The four orthologous group containing the focal genes of *O. satvia* ssp. *japonica* were selected as orthologs corresponding to each gene. To classify the *HIS1* orthologs into subclasses, their protein sequences were aligned and used to construct a neighbor-joining phylogenetic tree using MEGAX software with 1000 bootstrap replications[61].

Gene evolutionary analysis

The "add_ka_and_ks_to_collinearity.pl" script in MCScanX was used to determine the non-synonymous (K_a) and synonymous (K_s) substitution values of the homologs pairs[58]. The selection pressures were determined based on the *K*a/*K*s values of syntenic orthologs pairs in each focal orthologous group. The gene synteny of the focal orthologous groups was identified based on the macrosynteny analysis using an in-house developed python script. The functional annotation of genes in synteny block was conducted using grameen database (https://www.gramene.org/, accessed on 7 February 2024). To construct gene tree, protein sequences of true orthologs of *GBSS1*, *GBSS2*, and *BADH2* were aligned using ClustalOmega v1.2.4 with the default parameters[54]. Gene trees were constructed using RAxML v8.2.12 with 1000 bootstrap replicates[55] and the trees were reconciled into the species tree to identify the transfer event using Notung v2.6.1.5 (Fig. S1)[62]. Divergence time (T) between orthologous pair was calculated based on a synonymous substitutions per year (λ) as $T = Ks/2\lambda$ ($\lambda = 6.5 \times 10^{-9}$ for rice)[63].

Data availability

The datasets supporting the conclusions of this article are included within the article and its Supplementary Information.

Received: 16 January 2024; Accepted: 26 February 2024
Published online: 04 March 2024

References

1. 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).
2. UN. *World Population Prospects 2022: Ten Key Messages*. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2022_wpp_key-messages.pdf (2022).

3. Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci.* **108**, 20260–20264 (2011).
4. Muthayya, S., Sugimoto, J. D., Montgomery, S. & Maberly, G. F. An overview of global rice production, supply, trade, and consumption. *Ann. N. Y. Acad. Sci.* **1324**, 7–14 (2014).
5. Zeng, D. *et al.* Rational design of high-yield and superior-quality rice. *Nat. Plants* **3**, 1–5 (2017).
6. Kovach, M. J., Calingacion, M. N., Fitzgerald, M. A. & McCouch, S. R. The origin and evolution of fragrance in rice (*Oryza* sativa L.). *Proc. Natl. Acad. Sci.* **106**, 14444–14449 (2009).
7. Seung, D. Amylose in starch: Towards an understanding of biosynthesis, structure and function. *New Phytol.* **228**, 1490–1504 (2020).
8. Juliano, B. O. Varietal impact on rice quality. *Cereal Foods World* **43**, 207–222 (1998).
9. Liu, Q. Q. *et al.* Field performance of transgenic indica hybrid rice with improved cooking and eating quality by down-regulation of Wx gene expression. *Mol. Breed.* **16**, 199–208 (2005).
10. Maeda, H. *et al.* A rice gene that confers broad-spectrum resistance to β-triketone herbicides. *Science* **365**, 393–396 (2019).
11. Atwell, B. J., Wang, H. & Scafaro, A. P. Could abiotic stress tolerance in wild relatives of rice be used to improve *Oryza sativa*?. *Plant Sci.* **215–216**, 48–58 (2014).
12. Giuliani, R. *et al.* Coordination of leaf photosynthesis, transpiration, and structural traits in rice and wild relatives (Genus *Oryza*). *Plant Physiol.* **162**, 1632–1651 (2013).
13. Wairich, A. *et al.* Introgression from *Oryza meridionalis* into domesticated rice *Oryza sativa* results in shoot-based iron tolerance. *bioRxiv.* https://doi.org/10.1101/2020.06.05.135947 (2020).
14. Zhao, K. *et al.* Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS ONE* **5**, e10780 (2010).
15. Zhang, Y. *et al.* A genetic resource for rice improvement: Introgression library of agronomic traits for all AA genome *Oryza* species. *Front. Plant Sci.* **13**, 856514 (2022).
16. Li, D., Pfeiffer, T. W. & Cornelius, P. L. Soybean QTL for yield and yield components associated with *Glycine soja* alleles. *Crop Sci.* **48**, 571–581 (2008).
17. Lippman, Z. B., Semel, Y. & Zamir, D. An integrated view of quantitative trait variation using tomato interspecific introgression lines. *Curr. Opin. Genet. Dev.* **17**, 545–552 (2007).
18. Khush, G. S. Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35**, 25–34 (1997).
19. Harlan, J. R. & de Wet, J. M. J. Toward a rational classification of cultivated plants. *Taxon* **20**, 509–517 (1971).
20. Mizuta, Y., Harushima, Y. & Kurata, N. Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. *Proc. Natl. Acad. Sci.* **107**, 20417–20422 (2010).
21. *Genetics and Genomics of Rice*. (Springer, 2013).
22. Wairich, A. *et al.* Chromosomal introgressions from *Oryza meridionalis* into domesticated rice *Oryza sativa* result in iron tolerance. *J. Exp. Bot.* **72**, 2242–2259 (2021).
23. Kawahara, Y. *et al.* Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).
24. Vaughan, D. A., Lu, B.-R. & Tomooka, N. The evolving story of rice evolution. *Plant Sci.* **174**, 394–408 (2008).
25. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
26. Lin, Z. *et al.* Origin of seed shattering in rice (*Oryza sativa* L.). *Planta* **226**, 11–20 (2007).
27. Tan, L. *et al.* Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**, 1360–1364 (2008).
28. Molina, J. *et al.* Molecular evidence for a single evolutionary origin of domesticated rice. *Proc. Natl. Acad. Sci.* **108**, 8351–8356 (2011).
29. Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S. & McCouch, S. Genetic structure and diversity in *Oryza sativa* L. *Genetics* **169**, 1631–1638 (2005).
30. Zhu, T. *et al.* Phylogenetic relationships and genome divergence among the AA-genome species of the genus *Oryza* as revealed by 53 nuclear genes and 16 intergenic regions. *Mol. Phylogenet. Evol.* **70**, 348–361 (2014).
31. Zhang, L., Reifová, R., Halenková, Z. & Gompert, Z. How important are structural variants for speciation?. *Genes* **12**, 1084 (2021).
32. Nadir, S. *et al.* An overview on reproductive isolation in *Oryza sativa* complex. *AoB Plants* **10**, 060 (2018).
33. Bikard, D. *et al.* Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* **323**, 623–626 (2009).
34. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
35. Searle, J. B. Speciation, chromosomes, and genomes. *Genome Res.* **8**, 1–3 (1998).
36. Korunes, K. L. *How Linkage Disequilibrium and Recombination Shape Genetic Variation Within and Between Species* (Springer, 2019).
37. Palaisa, K., Morgante, M., Tingey, S. & Rafalski, A. Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci.* **101**, 9885–9890 (2004).
38. Chitwood-Brown, J., Vallad, G. E., Lee, T. G. & Hutton, S. F. Characterization and elimination of linkage-drag associated with Fusarium wilt race 3 resistance genes. *Theor. Appl. Genet.* **134**, 2129–2140 (2021).
39. Huang, K. *et al.* The genomics of linkage drag in inbred lines of sunflower. *Proc. Natl. Acad. Sci.* **120**, e2205783119 (2023).
40. Vikram, P. *et al.* Linkages and interactions analysis of major effect drought grain yield QTLs in rice. *PLoS ONE* **11**, e0151532 (2016).
41. Xiao, N. *et al.* Genomic insight into balancing high yield, good quality, and blast resistance of japonica rice. *Genome Biol.* **22**, 283 (2021).
42. Olsen, K. M. *et al.* Selection under domestication: Evidence for a sweep in the rice waxy genomic region. *Genetics* **173**, 975–983 (2006).
43. Bai, Z. *et al.* The impact and origin of copy number variations in the *Oryza* species. *BMC Genom.* **17**, 261 (2016).
44. Dolatabadian, A., Patel, D. A., Edwards, D. & Batley, J. Copy number variation and disease resistance in plants. *Theor. Appl. Genet.* **130**, 2479–2490 (2017).
45. Sieber, A.-N., Longin, C. F. H., Leiser, W. L. & Würschum, T. Copy number variation of CBF-A14 at the Fr-A2 locus determines frost tolerance in winter durum wheat. *Theor. Appl. Genet.* **129**, 1087–1097 (2016).
46. Dolatabadian, A. *et al.* Copy number variation among resistance genes analogues in *Brassica napus*. *Genes* **13**, 2037 (2022).
47. Pös, O. *et al.* DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* **44**, 548–559 (2021).
48. Lloyd, J. P., Seddon, A. E., Moghe, G. D., Simenc, M. C. & Shiu, S.-H. Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes [OPEN]. *Plant Cell* **27**, 2133–2147 (2015).
49. Minkenberg, B., Xie, K. & Yang, Y. Discovery of rice essential genes by characterizing a CRISPR-edited mutation of closely related rice MAP kinase genes. *Plant J.* **89**, 636–648 (2017).
50. Meinke, D., Muralla, R., Sweeney, C. & Dickerman, A. Identifying essential genes in *Arabidopsis thaliana*. *Trends Plant Sci.* **13**, 483–491 (2008).
51. Chen, L. J., Lee, D. S., Song, Z. P., Suh, H. S. & Lu, B. Gene flow from cultivated rice (*Oryza sativa*) to its weedy and wild relatives. *Ann. Bot.* **93**, 67–73 (2004).

52. Kuroda, Y., Sato, Y.-I., Bounphanousay, C., Kono, Y. & Tanaka, K. Gene flow from cultivated rice (*Oryza sativa* L.) to wild Oryza species (*O. rufipogon* Griff and *O. nivara* Sharma and Shastry) on the Vientiane plain of Laos. *Euphytica* **142**, 75–83 (2005).
53. Shao, G. *et al.* Haplotype variation at Badh2, the gene determining fragrance in rice. *Genomics* **101**, 157–162 (2013).
54. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
55. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
56. Letunic, I. & Bork, P. Interactive tree of life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2006).
57. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
58. Wang, Y. *et al.* MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
59. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
60. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
61. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
62. Chen, K., Durand, D. & Farach-Colton, M. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7**, 429–447 (2000).
63. Gaut, B. S., Morton, B. R., McCaig, B. C. & Clegg, M. T. 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.* **93**, 10274–10279 (1996).

Acknowledgements

This work was supported by two National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT; Grant No. 2021R1C1C1004233 and No. 2022R1A4A1030348).

Author contributions

I.L., Y.P., and J.H. designed the analysis. I.L. collected the material and conducted analysis. I.L. illustrated figures and wrote draft manuscript. Y.P. and J.H. supervised and revised the manuscript. All authors contributed to the article and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-55581-w.

Correspondence and requests for materials should be addressed to J.H.

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

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