











Chromosome-level genome assembly of *Zizania latifolia* provides insights into its seed shattering and phytocassane biosynthesis

Ning Yan ^{1,8}, Ting Yang^{1,8}, Xiu-Ting Yu^{1,2,8}, Lian-Guang Shang ^{3,8}, De-Ping Guo⁴, Yu Zhang ¹, Lin Meng¹, Qian-Qian Qi ^{1,2}, Ya-Li Li^{1,2}, Yong-Mei Du¹, Xin-Min Liu¹, Xiao-Long Yuan¹, Peng Qin⁵, Jie Qiu ⁶, Qian Qian ⁷ & Zhong-Feng Zhang ¹

Chinese wild rice (*Zizania latifolia*; family: Gramineae) is a valuable medicinal homologous grain in East and Southeast Asia. Here, using Nanopore sequencing and Hi-C scaffolding, we generated a 547.38 Mb chromosome-level genome assembly comprising 332 contigs and 164 scaffolds (contig N50 = 4.48 Mb; scaffold N50 = 32.79 Mb). The genome harbors 38,852 genes, with 52.89% of the genome comprising repetitive sequences. Phylogenetic analyses revealed close relation of *Z. latifolia* to *Leersia perrieri* and *Oryza* species, with a divergence time of 19.7–31.0 million years. Collinearity and transcriptome analyses revealed candidate genes related to seed shattering, providing basic information on abscission layer formation and degradation in *Z. latifolia*. Moreover, two genomic blocks in the *Z. latifolia* genome showed good synteny with the rice phytocassane biosynthetic gene cluster. The updated genome will support future studies on the genetic improvement of Chinese wild rice and comparative analyses between *Z. latifolia* and other plants.

¹Tobacco Research Institute of Chinese Academy of Agricultural Sciences, Qingdao 266101, China. ²Graduate School of Chinese Academy of Agricultural Sciences, Beijing 100081, China. ³Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China. ⁴Department of Horticulture, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China. ⁵State Key Laboratory of Crop Gene Exploration and Utilization in Southwest China, Rice Research Institute, Sichuan Agricultural University, Chengdu, Sichuan 611130, China. ⁶Shanghai Key Laboratory of Plant Molecular Sciences, College of Life Sciences, Shanghai Normal University, Shanghai 200234, China. ⁷State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310006, China. ⁸These authors contributed equally: Ning Yan, Ting Yang, Xiu-Ting Yu, Lian-Guang Shang. ✉email: yanning@caas.cn; qianqian188@hotmail.com; zhangzhongfeng@caas.cn

Chinese wild rice (*Zizania latifolia*) is a diploid ($2n = 2 \times = 34$), perennial, and aquatic grass belonging to the tribe *Oryzae* Dum, family Gramineae^{1–3}. *Z. latifolia* originated in China and is distributed in China, Korea, Japan, and Southeast Asian countries^{1,2}. Chinese wild rice is one of the earliest important cereal crops in China and has been consumed as a cereal for more than 3000 years⁴. Since the Tang Dynasty, Chinese wild rice has been used as a traditional medicine food homologous grain^{1,4}. A medical book written in the Ming dynasty, Compendium of Materia Medica, recorded the use of Chinese wild rice for adjuvant treatment of diabetes and gastrointestinal diseases¹. The health-promoting effects of Chinese wild rice include atherosclerosis prevention, alleviation of lipotoxicity, and insulin resistance^{3,5–7}. In China, *Z. latifolia* infected by the endophytic *Ustilago esculenta* has been domesticated as the second-largest aquatic vegetable, known as ‘*Jiaobai*’^{8–13}. Therefore, *Z. latifolia* is an important economic crop with high nutritional and medicinal value, and worthy of further investigation.

Z. latifolia is usually grown in Asia, but *Zizania palustris*, *Zizania aquatica*, and *Zizania texana* are commonly found in North America^{2,14,15}. The perennial *Z. latifolia* and annual *Z. palustris* are used for the commercial production of Chinese and northern wild rice, respectively². Owing to long-term adaptation to environmental changes and resistance to abiotic and biotic stresses, genetic variation and useful gene resources of wild relatives of rice (e.g., *Z. latifolia*) have formed and accumulated during evolution¹⁶. Notably, *Z. latifolia* has several excellent traits not found in rice, including high protein content, high biomass productivity, deep-water tolerance, and blast resistance^{1,3,17}. The protein, dietary fiber, and total phenolic concentrations in Chinese wild rice are ~2-, ~5-, and ~6-fold of those in rice, respectively^{1,3}. Thus, *Z. latifolia* represents a potential gene donor for overcoming the bottleneck of narrow genetic resources in rice breeding^{1,3,17–19}.

Seed shattering is an important trait for wild rice to adapt to the natural environment and maintain population reproduction^{20,21}. The loss of seed shattering is a key event in rice domestication²². Because of the successful selection of varieties with low seed shattering²³, the commercial production of northern wild rice has been realized in the United States. As a type of health food with a unique flavor, high nutritional value, and high price, northern wild rice has entered people’s diets, and is also exported to China and Europe.

Biosynthetic gene clusters are critical genetic factors for the rapid environmental adaptability of plants^{24,25}. Two biosynthetic gene clusters are well characterized in the rice genome: the phytocassane biosynthetic gene cluster on chromosome 2²⁶ and the momilactone biosynthetic gene cluster on chromosome 4²⁷. The potential functions of both gene clusters involve defense against pathogens and weeds in rice^{26,28,29}. Additionally, the momilactone biosynthetic

gene cluster is involved in the biosynthesis of allelochemicals³⁰. However, it is not known whether the synthetic genes of phytocassane and momilactone are clustered in the genome of *Z. latifolia*.

Within the genus *Zizania*, the *Z. latifolia* genome was first sequenced in 2015 using next-generation sequencing technology³¹. The *Z. latifolia* genome completed by Guo et al.³¹ has been used for transcriptome analyses of the possible molecular mechanism of swollen culm formation in *Z. latifolia* induced by *U. esculenta*^{8,9}. Because of technical limitations and a lack of a linkage map, the previous genome was only assembled at the scaffold level and remained relatively fragmented, with a contig N50 of 14 kb³¹. In this study, we generated the Chinese wild rice genome using a combination of Nanopore and Illumina sequencing data sets. Genome sequences of ~547.38 Mb were assembled with a contig N50 of 4.78 Mb and placed into 17 pseudochromosomes assisted by Hi-C. The updated and improved genome facilitated the annotation of protein-coding genes and noncoding RNAs. In this study, collinearity and transcriptome analyses revealed candidate genes involved in abscission layer formation (ALF) and degradation (ALD). Additionally, the phytocassane biosynthetic gene cluster was identified in the *Z. latifolia* genome, with complementary subclusters separating in two chromosomes. The updated *Z. latifolia* genome sequence serves as an important resource for comparative genomic studies between the Gramineae family and other plant species and might facilitate the rapid domestication of Chinese wild rice.

Results

Genome assembly, anchoring, and quality evaluation. In this study, we sampled Chinese wild rice plants grown in a paddy field (Fig. 1a). The inflorescence of Chinese wild rice is a panicle with multiple branches (Fig. 1b); we found both male and female flowers on the same branch, with the female flower above the male flower. The seeds of Chinese wild rice were blackish brown, cylindrical, and tapered at both ends. One *Z. latifolia* plant (accession Hua’an) was selected for whole-genome sequencing, and two paired-end Illumina libraries were constructed and sequenced. After cleaning, 68.50 Gb of high-quality sequencing data were obtained. Genome characterization based on *K*-mer depth distribution revealed that the Chinese wild rice Hua’an genome size was ~606.13 MB, with 49.00% repeats, 0.18% heterozygosity, and 42.88% GC content. Subsequently, we constructed and sequenced a Nanopore library, and 61.56 Gb of high-quality sequencing data were obtained, representing ~112.46× of the Chinese wild rice genome. The detailed summary statistics of the Oxford Nanopore Technology and Illumina sequencing are provided in Supplementary Table 1. After correction with Illumina sequencing and Hi-C scaffolding, we generated an assembly of 547.38 Mb comprising 332 contigs and 164 scaffolds, with a contig N50 of 4.48 Mb and a scaffold N50 of 32.79 Mb (Table 1). Based on the Hi-C interaction maps, 300 sequences covering ~545.36 Mb were

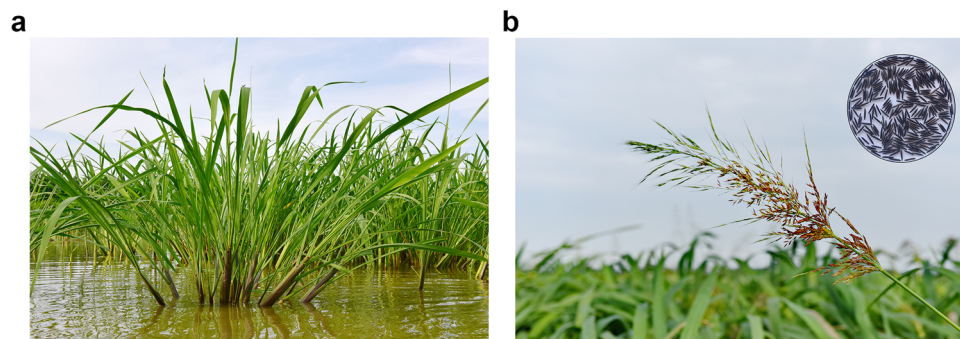


Fig. 1 Photographs of Chinese wild rice plants and inflorescence with seeds. **a** Chinese wild rice plants growing in a paddy field. **b** Inflorescence and seed morphology of Chinese wild rice. The pictures were taken by Ning Yan from Tobacco Research Institute of Chinese Academy of Agricultural Sciences.

clustered into 17 groups that corresponded to the 17 chromosomes of Chinese wild rice (Fig. 2), with the shortest being 17.01 Mb and the longest being 49.61 Mb (Table 1, Supplementary Table 2).

The Illumina reads were mapped to the assembled genome using the Burrows–Wheeler Alignment software³² to assess the quality of the genome assembly. Evaluation using CEGMA v2.5³³ with a database of 458 clusters of essential genes (CEGs) and 248 highly conserved CEGs indicated that 98.25% (450) and 94.76% (235) of the CEGs and highly conserved CEGs were present in the Chinese wild rice genome assembly, respectively. Further evaluation using Benchmarking Universal Single-Copy Orthologues (BUSCO) indicated that 97.71% (1,577) of the core genes were complete in the Chinese wild rice genome assembly, including single copies (77.39%, 1,249) and duplicated copies (20.32%, 328) (Supplementary Table 3). Additionally, 0.56% (9) of the core genes were fragmented, and only 1.73% (28) were missing. The BUSCO-based method for evaluating genome assembly integrity suggested that our Chinese wild rice Huai'an genome assembly shows better assembly integrity (Supplementary Table 4) than the previously assembled Chinese wild rice HSD2 genome³¹. This result was further supported by the long terminal repeat (LTR) assembly index (LAI)-based method³⁴ for evaluating genome assembly (Supplementary Table 5), which indicated that our Chinese wild rice Huai'an genome assembly showed an improved quality (LAI = 13.57) relative to the previously assembled Chinese wild rice HSD2 genome (LAI = 6.88)³¹ (Supplementary Table 5).

Table 1 Sequencing and assembly statistics of the Chinese wild rice genome.

Sequencing and assembly	Number	Size	N50 length
Nanopore reads	3,970,614	61.56 Gb	20.43 kb
Final assembly (contigs)	332	547.38 Mb	4.48 Mb
Final assembly (scaffolds)	164	547.40 Mb	32.79 Mb
Chromosome-anchored contigs	300	545.36 Mb	-

Genome annotation. Genome annotation resulted in the identification of 289.56 Mb (52.89%) of repetitive sequences in the assembled genome, which is substantially greater than that in the previous assembled version (227.50 Mb [37.70%] of repetitive sequences)³¹. The predominant repetitive sequences were LTR retrotransposons, which constituted 37.58% of the Chinese wild rice genome assembly (Supplementary Table 6). Among the transposable element (TE) superfamily studied, Copia (22.78%) and Gypsy (12.50%) generally occupied a relatively high proportion of the Chinese wild rice genome, whereas the Polinton superfamily, which is a unique TE type of *Z. latifolia*, only accounted for a small proportion (Supplementary Table 6). Moreover, the repetitive rate in the genome of Chinese wild rice Huai'an (52.89%) is higher than that in the *Oryza sativa japonica* (40.43%) and *Oryza sativa indica* (42.05%) groups³⁵. These results might explain why Chinese wild rice has a genome larger than *O. sativa*. Notably, the repetitive rate in the genome of northern wild rice (76.40%)³⁶ is higher than that in Chinese wild rice Huai'an (52.89%), which might explain why northern wild rice has a larger genome than Chinese wild rice. We then used three strategies (ab initio prediction, a homology-based strategy, and transcriptomic support) to predict the protein-coding genes (Supplementary Fig. 1). Finally, 38,852 protein-coding genes (136.02 Mb) were obtained (Table 2) and annotated (Supplementary Data 1). Further comparisons between northern and Chinese wild rice revealed two genomes with differential protein-coding genes, 46,491 in northern wild rice and 38,852 in Chinese wild rice. Of these predicted genes, 36,473 (93.88%) were annotated using eight functional databases. Additionally, we identified 149 microRNAs (miRNAs), 397 rRNAs, 723 tRNAs, and 1368 pseudogenes (Table 2).

Evolution of the Chinese wild rice genome. We performed comparative genomic analysis of the Chinese wild rice genome with the genome sequences of representative plant species, including eight gramineous plants (*Brachypodium distachyon*, *Hordeum vulgare*, *Leersia perrieri*, *Oryza brachyantha*, *O. sativa*, *Sorghum bicolor*, *Setaria italica*, and *Zea mays*) and one dicotyledon (*Arabidopsis thaliana*) clustered into 38,169 gene families.

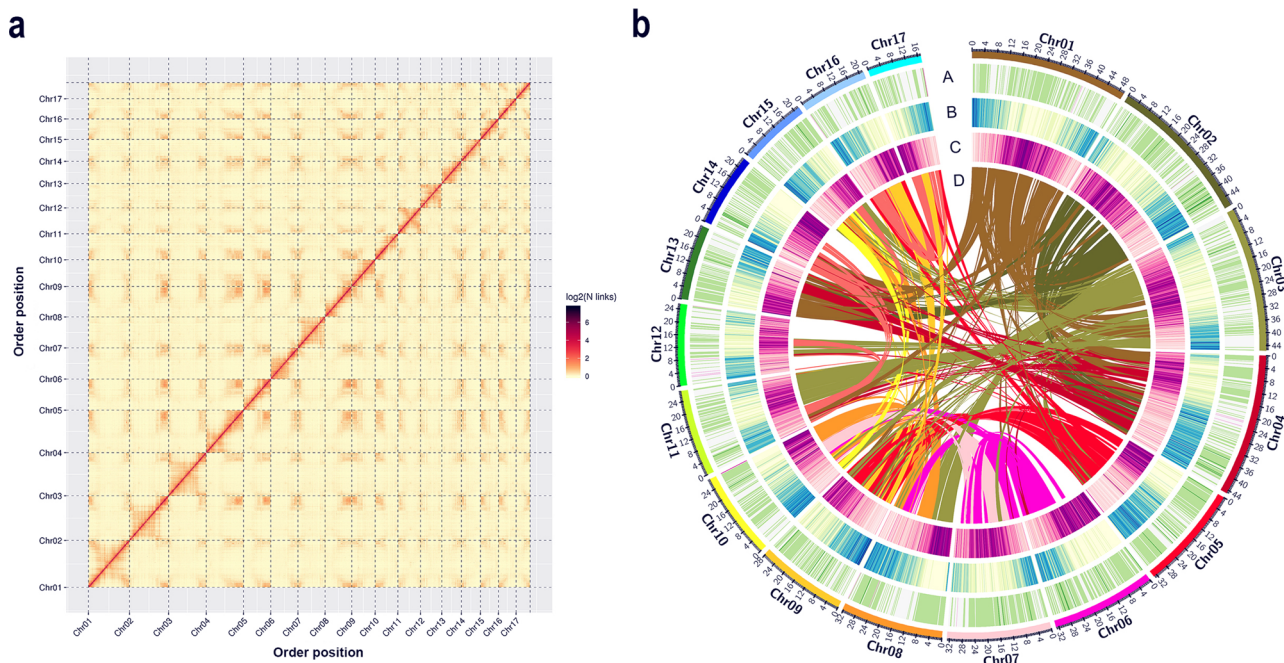


Fig. 2 Chinese wild rice genome information. **a** Hi-C contact data mapped on the updated Chinese wild rice genome showing genome-wide all-by-all interactions. **b** Overview of the Chinese wild rice genome. Ring A: distribution of GC content (green); ring B: gene density (blue); ring C: density of repeat sequences (purple); and ring D: syntenic blocks within the genome.

Table 2 Genome annotation statistics of the Chinese wild rice genome.

Genome annotation	Number	Size	Percentage (%)
Pseudogenes	1,368	4.55 Mb	0.83
miRNAs	149	-	-
rRNAs	397	-	-
tRNAs	723	-	-
Total protein-coding genes	38,852	136.02 Mb	24.85

In total, 33,924 gene families were identified in the Chinese wild rice genome, 310 of which were specific to Chinese wild rice (Supplementary Fig. 2, Supplementary Table 7). Gene family analysis revealed that the single-copy genes in Chinese wild rice accounted for 25.82% of the predicted genes, which was substantially lower than that in other gramineous species (Fig. 3a). In contrast, a higher proportion of gene families with two copies was observed in the *Z. latifolia* genome, which could be explained by a recent whole-genome duplication (WGD) event. The clustering of gene families in Chinese wild rice and the other four gramineous species (*B. distachyon*, *L. perrieri*, *O. brachyantha*, and *O. sativa*) indicated that 13,171 gene families are shared among the five grass species and that they could be the core gene families (Fig. 3b). In total, 709 gene families were specific to *Z. latifolia*, which is similar to that of *O. brachyantha*.

Based on 1,371 single-copy genes in Chinese wild rice and nine other plant species, we constructed a phylogenetic tree (Supplementary Fig. 3), which showed that *Z. latifolia* is relatively closely related to *L. perrieri*, *O. sativa*, and *O. brachyantha*. *Z. latifolia* diverged from *L. perrieri* around 19.7 Mya to 31.0 Mya, which was before the divergence of *O. sativa* and *O. brachyantha* (14.0–16.0 Mya) (Supplementary Fig. 3). In Chinese wild rice, 119 gene families showed expansion, whereas 132 gene families exhibited contraction (Fig. 3c). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis indicated that genes related to oxidative phosphorylation, photosynthesis, zeatin biosynthesis, and cyanoamino acid metabolism were enriched in the expanded Chinese wild rice gene families (Fig. 3d). The strong positive selection of genes is of great significance to the generation of new functions. The selection of positive genes in *Z. latifolia* is shown in Supplementary Data 2. Moreover, the KEGG pathway analysis indicated that the selection of positive genes was mainly related to carbon metabolism, biosynthesis of amino acids, and peroxisome (Supplementary Fig. 4).

WGD events are associated with an ancient polyploidization event predating the divergence of cereals³⁷ and are of great significance for understanding gene neofunctionalization and genome evolution³⁸. This study showed that recent WGD events occurred in the *Z. latifolia* genome after splitting from *O. sativa* (Fig. 3e), which is consistent with the previous findings³¹. *Zizania*–*Oryza* speciation events have both led to an increase in the genome size of northern wild rice (1.29 Gb)³⁶ in comparison with that of rice (390.30 Mb) and Chinese wild rice (547.38 Mb) (Table 1). Moreover, a peak centering on Ks of ~0.25 was observed between the *O. sativa* and *Z. latifolia* pairs.

Collinearity between genomes and seed-shattering-related genes of *O. sativa* and *Z. latifolia*. Significant genome collinearity has been observed between northern wild rice and rice^{36,39,40}. Similarly, in this study, significant genome collinearity was observed between Chinese wild rice and rice (Supplementary Fig. 5a). To the best of our knowledge, 10 seed-shattering-related genes have been identified in rice (i.e., *qSH1*, *OsGRF4/PT2*, *OsSh1*, *OsNPC1*, *sh4/SH4*, *SHAT*, *OsLG1*, *SH5*, *sh-h/OsCPL1*, and

SSH1) (Supplementary Table 8). Moreover, 17 seed-shattering-related genes have been identified in northern wild rice through collinearity between seed-shattering-related genes of rice and northern wild rice³⁶. Based on genome collinearity and gene homolog analyses between the genomes of Chinese wild rice and rice, we identified 29 candidate genes potentially related to seed shattering in Chinese wild rice (Supplementary Fig. 5b, Supplementary Data 3). In Chinese wild rice, collinearity was observed for two (*ZlqSH1a* and *ZlqSH1b*), seven (*ZIGRF4/PT2a*, *ZIGRF4/PT2b*, *ZIGRF4/PT2c*, *ZIGRF4/PT2d*, *ZIGRF4/PT2e*, *ZIGRF4/PT2f*, and *ZIGRF4/PT2g*), two (*Zlsh1a* and *Zlsh1b*), two (*ZINPC1a* and *ZINPC1b*), two (*ZISHATa* and *ZISHATb*), four (*ZILG1a*, *ZILG1b*, *ZILG1c*, and *ZILG1d*), two (*Zlsh4/SHA1a* and *Zlsh4/SHA1b*), two (*ZlSH5a* and *ZlSH5b*), two (*Zlsh-h/ZICPL1a* and *Zlsh-h/ZICPL1b*), and four (*ZISSH1a*, *ZISSH1b*, *ZISSH1c*, and *ZISSH1d*) genes with *qSH1*, *OsGRF4/PT2*, *OsSh1*, *OsNPC1*, *SHAT*, *OsLG1*, *sh4/SH4*, *SH5*, *sh-h/OsCPL1*, and *SSH1* in rice, respectively (Supplementary Fig. 5b, Supplementary Data 3). According to their position in the evolutionary tree, the seed-shattering candidate genes of *Z. latifolia* were divided into four categories (Supplementary Fig. 6). Moreover, the motifs of these proteins were similar in the same group and different in different groups, which confirmed the reliability of grouping (Supplementary Fig. 7). The protein sequence alignment of seed-shattering genes in *Z. latifolia* and *Oryza* species is shown in Supplementary Figs. 8–17. Notably, aligning a sequence of seed-shattering candidates across *Z. latifolia* and *Oryza* species would help to identify the potential functional polymorphism and specific candidate sites for editing in *Z. latifolia*.

Histologic, transcriptome, and phytohormone analyses of ALF and ALD tissues in Chinese wild rice. During seed abscission, one or several layers of parenchyma cells differentiate from the abscission zone to form the abscission layer. Cells in the adjacent cell layer have thick, lignified cell walls, which help provide the mechanical force needed for abscission⁴¹. The results of the histological analysis showed that the abscission layer of Chinese wild rice comprises 6–8 circles of cells radially distributed in the periphery, where the single cells were oval, demonstrating a compact and regular arrangement that could be stained red by the cell-permeant dye, Acridine Orange (Fig. 4a, b). Moreover, we observed that the ALF and ALD of Chinese wild rice were complete, and that ALD led to seed shattering (Fig. 4a, b). To elucidate the potential molecular mechanism of seed shattering in Chinese wild rice, we analyzed abscission tissues from ALF and ALD by transcriptome sequencing. Compared with ALF, ALD involved 2,827 upregulated and 3,938 downregulated genes, including nine genes related to seed shattering (Fig. 4c, Supplementary Data 4). Among them, *ZIGRF4/PT2a* (*Zla08G018880*) and *ZIGRF4/PT2g* (*Zla05G008960*) were upregulated in ALD, whereas *ZlqSH1b* (*Zla02G027130*), *ZISHATa* (*Zla07G002730*), *ZISHATb* (*Zla06G014800*), *ZILG1a* (*Zla07G002370*), *ZILG1b* (*Zla06G015090*), *ZISH5a* (*Zla01G006060*), and *ZISH5b* (*Zla13G006450*) were downregulated (Fig. 4d, Supplementary Data 4). To confirm the reliability of the transcriptome results, the expression levels of eight of these key genes were further examined using real-time PCR (qRT-PCR) evaluation. *ZlqSH1b*, *ZISHATa*, *ZISHATb*, *ZILG1a*, *ZILG1b*, *ZISH5a*, and *ZISH5b* expression was significantly downregulated in ALD ($P < 0.05$), which is consistent with the result of transcriptome sequencing (Supplementary Fig. 18).

KEGG pathway analysis indicated that genes related to plant hormone signal transduction, ribosomes, amino acid biosynthesis, starch, and sucrose metabolism, and phenylpropanoid biosynthesis were enriched among the differentially expressed genes between ALF and ALD tissues in Chinese wild rice (Fig. 4e).

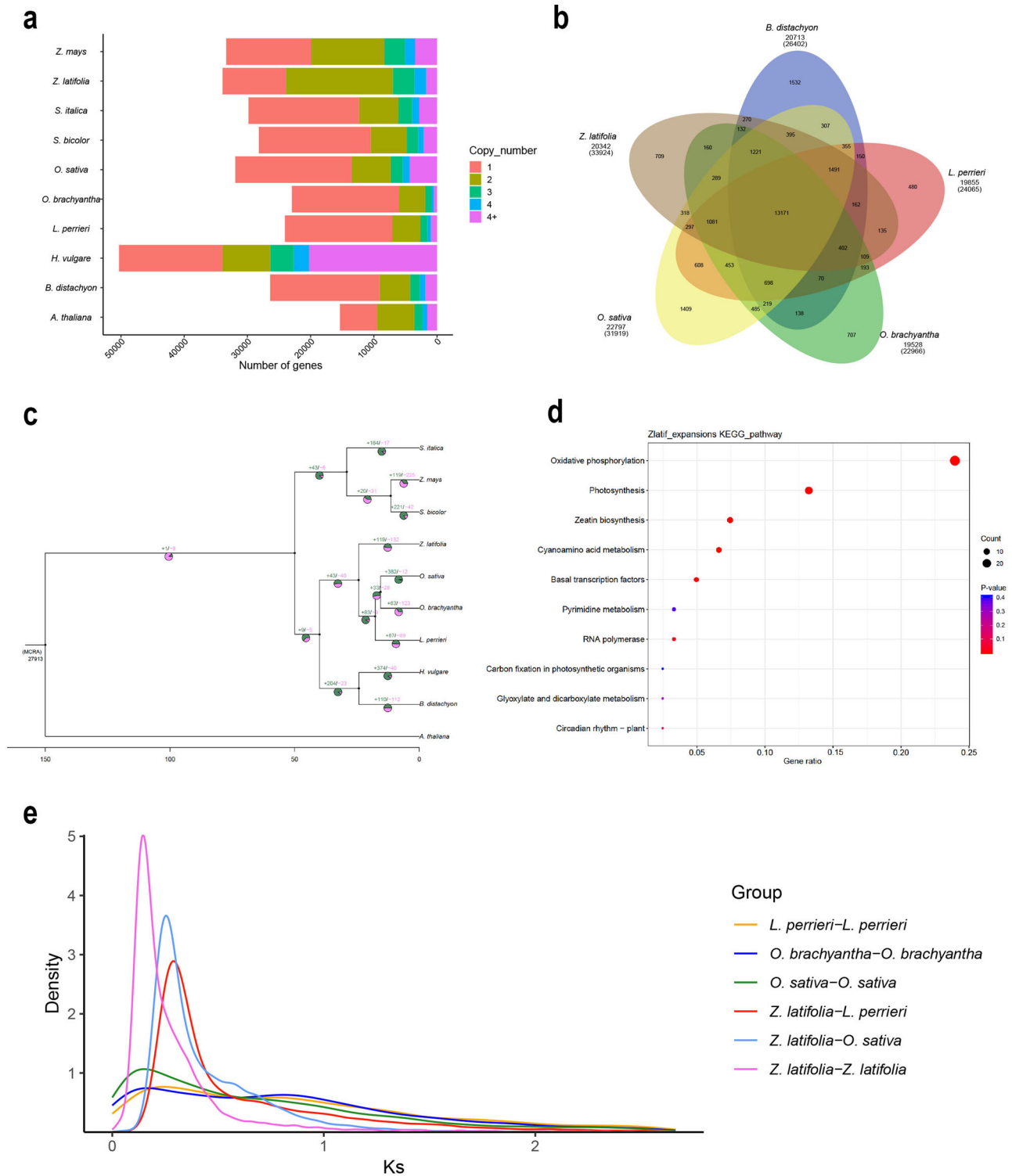


Fig. 3 Comparative genomic analyses of Chinese wild rice genome. **a** Distribution of gene copy number in Chinese wild rice and nine other species. **b** Venn diagram of shared orthologous gene families in Chinese wild rice and other four related gramineous species (*Brachypodium distachyon*, *Leersia perrieri*, *Oryza brachyantha*, and *O. sativa*). **c** Phylogenetic tree of Chinese wild rice and nine other species. “+” represents the number of gene families expanded on the node and “-” represents the number of gene families contracted on the node. The pie chart shows the proportion of the corresponding branch contraction and expansion gene families. **d**, KEGG enrichment analyses for the expanded genes in the Chinese wild rice genome. **e** Ks distribution in Chinese wild rice and other representative species.

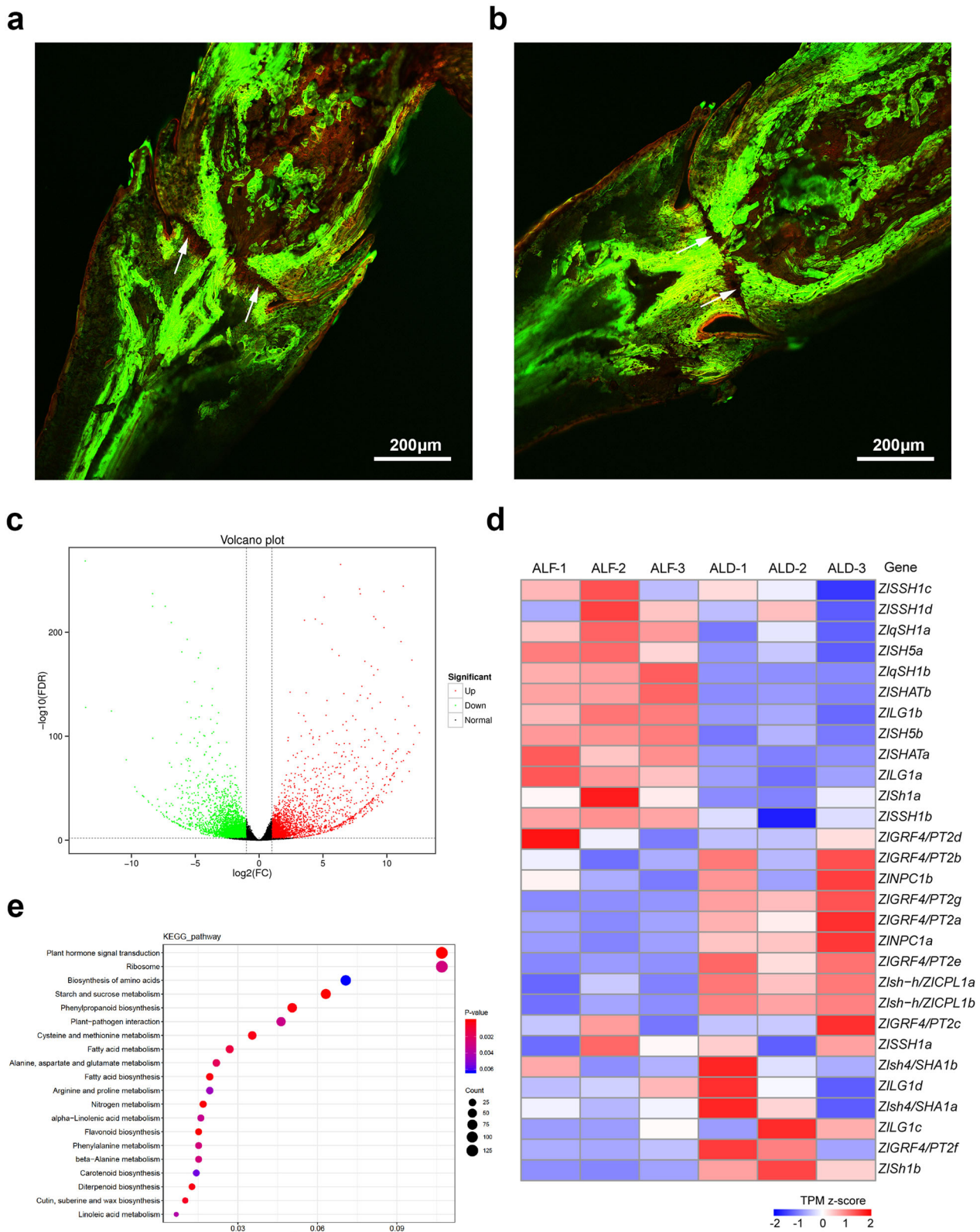


Fig. 4 Histologic and transcriptome analyses of abscission layer formation (ALF) and degradation (ALD) in Chinese wild rice. **a** ALF and **b**, ALD as revealed by staining with the cell-permeant dye Acridine Orange (green fluorescence: dye bound to dsDNA; red fluorescence: dye bound to ssDNA or RNA). The white arrows indicate the abscission layer. Scale bar = 200 μ m. **c** Volcano plot of differentially expressed genes between ALF and ALD. In this figure, green, red, and black dots represent genes with a low expression, high expression, and non-differentially expressed genes, respectively. **d** Expression levels of genes related to seed shattering between ALF and ALD. **e** KEGG enrichment bubble plot of differentially expressed genes between ALF and ALD.

To elucidate the role of phytohormones in seed shattering in Chinese wild rice, we compared the concentrations of phytohormones between ALF and ALD tissues (Supplementary Table 9). Among them, the concentrations of abscisic acid, ABA-glucosyl ester, 1-aminocyclopropanecarboxylic acid, *cis*-zeatin, *trans*-zeatin riboside, N⁶-isopentenyladenine, indole-3-acetic acid, 1-*O*-indole-3-ylacetylglucose, indole-3-carboxylic acid, methyl indole-3-acetate, salicylic acid, and salicylic acid 2-*O*- β -glucoside were significantly higher in ALD, whereas those of gibberellin A9, gibberellin A19, jasmonic acid, and methyl jasmonate were significantly lower, than those in ALF ($P < 0.05$) (Supplementary Table 9). Therefore, the concentrations of these phytohormones changed significantly during the process of ALF and ALD in Chinese wild rice.

Genomic synteny of the phytocassane biosynthetic gene cluster between *O. sativa* and *Z. latifolia*. In our assembled *Z. latifolia* genome, we found genes homologous with *MAS*, *CYP99A*, *CPS*, and *KSL* of the rice momilactone biosynthetic gene cluster. However, the genes most homologous to *MAS*, *CYP99A*, *CPS*, and *KSL* were not located nearby but were scattered among chromosomes 6, 9, 8, and 7, respectively. For the phytocassane biosynthetic gene cluster, we observed two genomic blocks in the *Z. latifolia* genome (Chr. 8: 22.53–22.62 Mb and Chr10: 53.27–53.61 Mb) that showed good synteny with the rice phytocassane biosynthetic gene cluster (Fig. 5a, b). Chromosomes 8 and 10 of *Z. latifolia* were highly colinear with rice chromosome 2, likely owing to a WGD event within the *Zizania* lineage after its divergence from *Oryza*. Upon examining the orthologous genes, we found that the candidate clusters on chromosomes 8 and 10 of *Z. latifolia* were not as complete as those in rice but generally complementary to each other (Fig. 5b, Supplementary Table 10). For each sub-cluster of genes on chromosomes 8 and 10, good collinearity was observed with those in rice, despite some rearrangement of the gene order (e.g., *CPS*) (Fig. 5b). Moreover, genes in different sub-clusters showed a highly positive co-expression pattern (Fig. 5c). This suggests that although they are separated into two chromosomes, they are still co-regulated and together play a role in the biosynthesis of phytoalexins.

Discussion

Zizania latifolia is an important aquatic vegetable in East and Southeast Asia, with a high nutritional and medicinal value. Here, we generated a quality-improved genome for Chinese wild rice Huai'an and anchored 99.63% of the sequences to 17 pseudo-chromosomes. Furthermore, our assembly of Huai'an (contig N50 = 4.78 Mb) exhibits a $367.69 \times$ longer contig N50 than HSD2 (contig N50 = 13 kb)³¹. Recently, Haas et al.³⁶ generated a high-quality genome for northern wild rice and anchored 98.53% of the sequences to 15 chromosomes. The chromosome-level genome will support future studies of molecular genetic breeding and genome evolution in wild rice and the genus *Zizania*.

As an adaptation to the natural environment and offspring propagation, losing seed shattering in rice has been a prime target during plant selection and domestication^{21,42}. Recently, Yu et al.²² successfully domesticated wild allotetraploid rice (*Oryza alta*) de novo by optimizing the genetic transformation system, assembling the wild allotetraploid rice genome de novo, and editing several genes that control key domestication-related and agronomical traits, such as seed shattering. To the best of our knowledge, *qSH1*, *OsSh1*, and *sh4/SHA1* are the major genes related to rice seed shattering²¹. In Chinese wild rice, three pairs of genes (*ZlqSH1a* and *ZlqSH1b*, *Zlsh1a* and *Zlsh1b*, and *Zlsh4/SHA1a* and *Zlsh4/SHA1b*) showed collinearity with *qSH1*, *OsSh1*, and *sh4/SHA1* in rice but showed a differentially expressed

pattern between ALD and ALF. Therefore, the genes related to seed shattering in Chinese wild rice provide a target for reducing its seed shattering and de novo domestication. Notably, genome editing using clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) facilitates targeted genetic manipulation of wild crops and can accelerate crop domestication⁴³. Furthermore, the gene editing of major seed-shattering genes by the CRISPR/Cas9 system can aid in development of Chinese wild rice materials with reduced seed-shattering.

An evolutionary history of the two biosynthetic gene clusters in rice was proposed by Miyamoto et al.⁴⁴, who compared genes in the momilactone and phytocassane biosynthetic gene clusters for different *Oryza* species. We found the momilactone biosynthetic gene cluster in the *Z. latifolia* genome, which agrees with the hypothesis of Miyamoto et al.⁴⁴, and this gene cluster evolved within the *Oryza* clade and before the divergence of the rice AA and BB genomes. As for the origin of the phytocassane biosynthetic gene cluster, it is posited that the cluster was present in the common ancestor of the *Oryza* and *Leersia* lineages, with some gene order rearrangements in *L. perrieri*, and gene deletions in some *Oryza* species (Fig. 5d). Based on our findings, the existence of the cluster in the *Z. latifolia* genome suggests that the core phytocassane biosynthetic gene cluster was available in the common ancestor of *Oryza* and *Zizania* species. The two sub-clusters in different chromosomes are likely the result of a recent WGD event within *Zizania* after its divergence from *Oryza*. Additionally, the complementary pattern of the two sub-clusters might be owing to a fractionation process after WGD. Overall, our results provide insights into the genomic evolution process of the phytocassane biosynthetic gene cluster.

In summary, our well-assembled *Z. latifolia* genome will support future basic research on and agronomic improvement of *Z. latifolia* as well as comparative genomic studies between the Gramineae family and other plant species.

Methods

DNA extraction and sequencing. The sampling site is located in Baimahu Village, Jintu County, Huai'an City, Jiangsu Province (33°11'9" N; 119°9'37" E)². Owing to the relatively closed geographical environment, Chinese wild rice in this region is highly homozygous. Leaf samples from Chinese wild rice Huai'an were collected for sequencing. After DNA sequencing, reads with adapters, low-quality reads, and reads of <2000 nt were filtered. To determine whether the sequencing data were contaminated, we randomly selected 2,000 single end reads and compared them with those in the Nucleotide Sequence Database by BLAST; there were no contaminated sequences. For Illumina sequencing, a paired-end library with an insert size of 350 bp was sequenced using the Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA) with a 150 nt layout, according to the manufacturer instructions.

Genome assembly. Nanopore third-generation sequencing data were corrected using Canu⁴⁵, the SMARTdenovo software (<https://hpc.ilri.cgiar.org/smartdenovo-software>) was used to assemble the corrected data. The Racon⁴⁶ (<https://bioinformatics.cornell.edu/tools/wga/descriptions/Racon.html>) and Pilon⁴⁷ software were used to perform three rounds of correction of the third-generation sequencing data and second-generation data, respectively.

The Burrows–Wheeler Alignment software³² was used to compare the short sequences obtained from Illumina sequencing with the reference genome in this study, and the integrity of the assembled genome was evaluated through statistical comparisons. The CEGMA v2.5³³ database containing 458 conserved eukaryotic core genes was used to evaluate the integrity of the final genome assembly. The embryophyta database in OrthoDB v10s (containing 1,614 conserved core genes) and BUSCO v4.0⁴⁸ were used to evaluate the integrity of the genome assembly. Additionally, the LAI value was used to judge the assembly quality based on repetitive genomic regions³⁴.

Hi-C analysis and pseudo-chromosome construction. Fresh young leaves collected from Chinese wild rice Huai'an plants were fixed with 1% formaldehyde. Hi-C fragment libraries were constructed using 300- to 700-bp inserts⁴⁹. The low-quality reads and adapter sequences of raw reads were removed to obtain clean data. Notably, only uniquely aligned paired reads with a mapping quality of >20

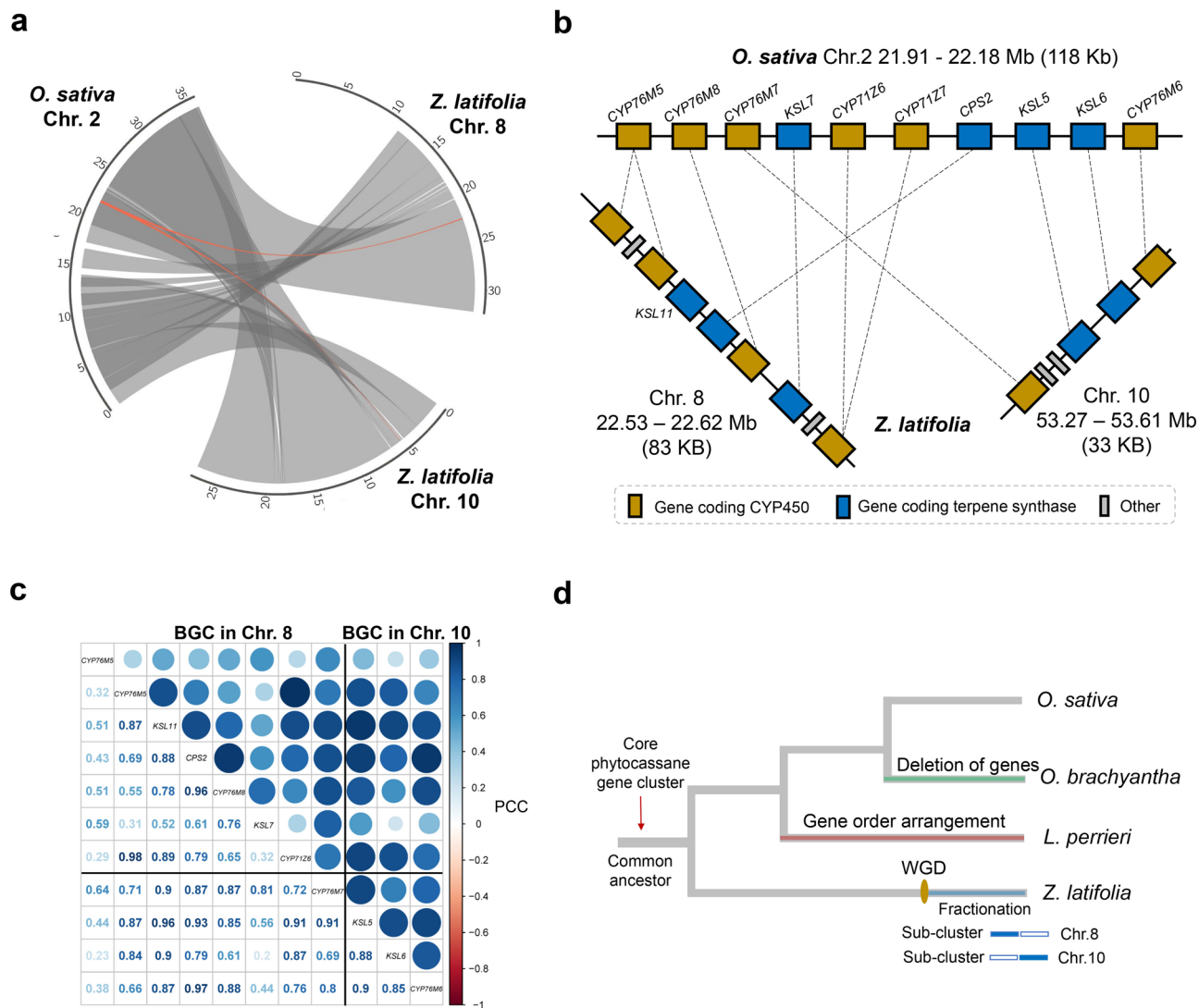


Fig. 5 Characterization of the phytocassane biosynthetic gene cluster in *Zizania latifolia* genome. **a**, Genomic synteny between chromosome 2 of *O. sativa* and chromosomes 8 and 10 of *Z. latifolia*. Syntenic genomic blocks are illustrated by the grey lines. The homologous genomic regions of phytocassane biosynthetic gene clusters between *O. sativa* and *Z. latifolia* are highlighted in red. **b** Gene-level synteny between phytocassane biosynthetic gene cluster of *O. sativa* and *Z. latifolia*; CYP450 genes are colored dark yellow; genes coding terpene synthases are colored dark blue. The genes unrelated to the cluster are in grey. **c** Gene co-expression pattern for the genes in the two sub-gene clusters in chromosomes 8 and 10. **d**, Proposed evolutionary history of the phytocassane biosynthetic gene cluster in the *Z. latifolia*, *L. perrieri*, and *Oryza* species.

were used for further analysis. Before chromosome assembly, we performed a preassembly for error correction of scaffolds, which required the splitting of scaffolds into 50-kb segments. Hi-C data were then mapped to these segments using the Burrows–Wheeler Alignment software. The uniquely mapped data were retained to perform assembly using LACHESIS⁵⁰.

Repetitive sequence and gene annotation. Using the LTR_FINDER⁵¹ and RepeatScout⁵² software, we constructed a genome repetitive-sequence database based on ab initio prediction and structure prediction. The database was classified with PASTEClassifier⁵³ and then combined with the Repbase database⁵⁴ as the final repeat-sequence database. We then used RepeatMasker⁵⁵ for repetitive-sequence prediction of the genome. Default parameters were used for LTR_FINDER, RepeatScout, and PASTEClassifier, and the '-nolow -no_is -norma -engine wublast' parameter was used for RepeatMasker. Additionally, we used the EDTA (v1.9.7) software to generate TE annotations⁵⁶.

The gene-structure prediction was performed for the *Z. latifolia* genome using ab initio prediction, prediction based on homologous species, and prediction based on Unigene analysis. The prediction results were integrated using EVM v1.1.1⁵⁷. First, we used Genscan⁵⁸, Augustus v2.4⁵⁹, GlimmerHMM v3.0.4⁶⁰, GeneID v1.4⁶¹, and SNAP⁶² for ab initio prediction. Second, we used GeMoMa v1.3.1^{63,64} for prediction based on homologous species. Stringtie v1.2.3⁶⁵ and Hisat v2.0.4⁶⁶, and GeneMarkS-T v5.1⁶⁷ and TransDecoder v2.0, were used for assembly and gene prediction, respectively. We sequenced the mixed RNA library generated from the

root, stem, leaf, leaf sheath, male and female florets, seed, and a whole un-emerged panicle for transcriptome-based predictions. Additionally, the RNA-seq reads were assembled into transcripts using Trinity v2.1.1⁶⁸, and PASA v2.0.2⁶⁹ was used to predict the Unigene based on RNA-seq reads.

Noncoding RNAs include miRNA, rRNA, tRNA, and other RNAs with known functions. Blastn was used for genome-wide alignment based on the Rfam database⁷⁰ to identify miRNAs and rRNAs, and tRNAscan-SE⁷¹ was used to identify tRNAs. The predicted protein sequences were used to search for homologous gene sequences through GenBlastA⁷² alignment, and GeneWise⁷³ was then used to search for premature stop codons and frameshift mutations that resulted in pseudogenes. For GenBlastA, an e-value of 1×10^{-5} was used; all other parameters were set to default. Additionally, default parameters were used for GeneWise. BLAST v2.2.31⁷⁴ alignment (e-value: 1×10^{-5}) was performed between the predicted gene sequence and the Non-Redundant Protein Sequence Database⁷⁵, EuKaryotic Orthologous Groups⁷⁶, Gene Ontology⁷⁷, KEGG⁷⁸, and TrEMBL⁷⁹ functional databases.

Gene families and phylogenetic analysis. Orthofinder v2.4⁸⁰ was used to classify the protein sequences of nine gramineous plants (*B. distachyon*, *H. vulgare*, *L. perrieri*, *O. brachyantha*, *O. sativa*, *S. bicolor*, *S. italica*, *Z. latifolia*, and *Z. mays*) and one dicotyledon (*A. thaliana*) into families. The PANTHER V15 database⁸¹ was used for annotation of the gene families obtained. IQ-TREE v1.6.11⁸² was used to construct a phylogenetic tree from 1,371 single-copy protein sequences.

Specifically, MAFFT v7.205 (<https://mafft.cbrc.jp/alignment/software/>) was used to align each single-copy gene family sequence, and the PAL2NAL v14 program⁸³ was then used to convert the protein alignment to codon alignment. We then used Gblocks v0.91b (parameter: -b5=h)⁸⁴ to remove regions with large differences or poor sequence alignment. Finally, the aligned gene family sequences of each species were connected end-to-end to obtain a super-gene alignment. The model testing tool ModelFinder⁸⁵, included with IQ-TREE (<http://www.iqtree.org/>), was used for model selection, with the best model identified as GTR + F + I + G4. Using this model, we applied the maximum-likelihood method to construct a phylogenetic tree, with the number of bootstraps set to 1,000. The outgroup of the obtained phylogenetic tree was set as *A. thaliana*, which gave a rooted tree, and the MCMCTREE package included in the PAML v4.9i software⁸⁶ was then used to calculate divergence times. The final phylogenetic tree with divergence times was displayed graphically using MCMCTreeR v1.1⁸⁷. CAFE v4.2⁸⁸ was used with the phylogenetic tree with divergence times and genes (after clustering into families) to estimate the number of gene family members of an ancestor from each branch through birth-death models, predicting the contraction and expansion of a gene family from each species relative to that of its ancestor. Significant expansion or contraction was defined as family-wide *p* values and viterbi *p* values (both < 0.05).

We used the CodeML module in PAML for positive-selection analysis. First, we obtained single-copy gene families common among *B. distachyon*, *H. vulgare*, *L. perrieri*, *O. brachyantha*, *O. sativa*, and *Z. latifolia*, followed by MAFFT (parameters: —localpair —maxiterate 1000) alignment of the protein sequences of each gene family and conversion to the codon alignment sequence using PAL2NAL. Finally, CodeML was used to perform likelihood ratio tests of model A and the null model using the ‘chi2’ program in PAML based on the branch-site model. An empirical Bayes method was used to obtain the posterior probability of being considered a positively selected site (>0.95 is usually considered a significantly positively selected site).

Collinearity and WGD analyses. We used Diamond v0.9.29.130⁸⁹ to compare the protein sequences of *O. sativa* and *Z. latifolia* (C-score > 0.5; *e* value < 1×10^{-5}). Subsequently, we identified the collinear blocks between the genomes of *O. sativa* and *Z. latifolia* using MCScanX⁹⁰. Finally, based on the distribution of the Ks paralogous genes, we calculated the WGD events using the WGD software⁹¹.

Identification of seed-shattering genes in Chinese wild rice. Genes related to seed shattering in *O. sativa* were obtained by querying the gene name on the website of the China Rice Data Centre (<https://www.ricedata.cn/>). Seed-shattering genes in *Z. latifolia* were obtained by comparing similar genes in *O. sativa* with the genome sequences of *Z. latifolia* in this study. The *e*-value of the sequence-alignment results was set to < 1×10^{-10} . MCScanX was used for collinearity analysis of candidate genes.

Histologic analysis of the anatomic structure of the abscission layer. The ALF and ALD tissues (1–2 mm above and below, respectively, the junction between the flower and the pedicel) were collected, and a freehand longitudinal section was prepared using a thin blade. The collected sample was stained using a 0.1% aqueous solution of Acridine Orange for 10–15 min, rinsed three times with deionized water, placed on a glass slide, and observed under a confocal laser microscope (Leica SP8; Leica Biosystems, Nussloch, Germany) at 488 and 543 nm.

Transcriptome analysis. Transcriptome sequencing analysis of the ALF and ALD tissues was performed according to the method of Yan et al.⁹². Gene-expression levels were quantified by estimating fragments per kilobase of transcript per million fragments mapped. The genes with an adjusted *P* < 0.01 according to DESeq2 were identified as differentially expressed. We used KOBAS⁹³ to test the enrichment of differentially expressed genes in the KEGG pathways.

Data validation by real-time PCR. The qRT-PCR analysis was performed on eight selected seed-shattering genes between ALF and ALD. The method of qRT-PCR was performed according to Wang et al.⁹. The primers used for detecting the expression levels of the genes are listed in Supplementary Table 11.

Biosynthetic gene clusters between *O. sativa* and *Z. latifolia*. The protein sequences encoded by the genes in two known rice biosynthetic gene clusters were identified^{26,27} and used to search for orthologous genes in *Z. latifolia*. Syntenic genomic blocks between chromosome 2 of *O. sativa* and chromosomes 8 and 10 of *Z. latifolia* were identified using the MCScan program⁹⁰ and visualized by Circos⁹⁴. Based on 12 transcriptomes from tissues of ALF, ALD, leaf, and stem, the expression levels of genes related to the phytocassane biosynthetic gene cluster were extracted, and the co-expression coefficient matrix was visualized using the R package ‘corrplot’⁹⁵.

Detection of phytohormones. Fresh plant materials (ALF and ALD tissues) were harvested, weighed, immediately frozen in liquid nitrogen, and stored at –80 °C until needed. Plant materials (50 mg fresh weight) were frozen in liquid nitrogen, ground to powder, and extracted with 1 mL of methanol/water/formic acid (15:4:1,

v/v/v). Phytohormone contents were detected using MetWare (<http://www.metware.cn/>) based on the AB Sciex QTRAP 6500 LC-MS/MS platform.

Statistics and reproducibility. In Supplementary Fig. 18 and Supplementary Table 9, we used *n* = 3 biologically independent samples. Statistical significance was assessed using a two-tailed Student’s *t*-test, **P* < 0.05.

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

Data availability

Raw reads and transcriptome sequencing data have been deposited in GenBank under the accession number PRJNA719466. The whole genome sequence data have been deposited in Genome Sequence Archive under the accession number GWHBFH100000000, which is publicly accessible at <https://ngdc.cnpc.ac.cn/gwh>.

Received: 6 May 2021; Accepted: 21 December 2021;

Published online: 11 January 2022

References

1. Yan, N. et al. Morphological characteristics, nutrients, and bioactive compounds of *Zizania latifolia*, and health benefits of its seeds. *Molecules* **23**, 1561 (2018).
2. Yan, N. et al. A comparative UHPLC-QqQ-MS-based metabolomics approach for evaluating Chinese and North American wild rice. *Food Chem.* **275**, 618–627 (2019).
3. Yu, X. et al. Wild rice (*Zizania* spp.): A review of its nutritional constituents, phytochemicals, antioxidant activities, and health-promoting effects. *Food Chem.* **331**, 127293 (2020).
4. Zhai, C. K., Tang, W. L., Jang, X. L. & Lorenz, K. J. Studies of the safety of Chinese wild rice. *Food Chem. Toxicol.* **34**, 347–352 (1996).
5. Chu, M. J. et al. Partial purification, identification, and quantitation of antioxidants from wild rice (*Zizania latifolia*). *Molecules* **23**, 2782 (2018).
6. Chu, M. J. et al. Extraction of proanthocyanidins from Chinese wild rice (*Zizania latifolia*) and analyses of structural composition and potential bioactivities of different fractions. *Molecules* **24**, 1681 (2019).
7. Yu, X. et al. Comparison of the contents of phenolic compounds including flavonoids and antioxidant activity of rice (*Oryza sativa*) and Chinese wild rice (*Zizania latifolia*). *Food Chem.* **344**, 128600 (2021).
8. Li, J. et al. Transcriptome analysis reveals the symbiotic mechanism of *Ustilago esculenta*-induced gall formation of *Zizania latifolia*. *Mol. Plant Microbe* **34**, 168–185 (2021).
9. Wang, Z. D. et al. RNA-seq analysis provides insight into reprogramming of culm development in *Zizania latifolia* induced by *Ustilago esculenta*. *Plant Mol. Biol.* **95**, 533–547 (2017).
10. Wang, Z. H. et al. Gene expression in the smut fungus *Ustilago esculenta* governs swollen gall metamorphosis in *Zizania latifolia*. *Microb. Pathogenesis* **143**, 104107 (2020).
11. Ye, C. Y. & Fan, L. Orphan crops and their wild relatives in the genomic era. *Mol. Plant* **14**, 27–39 (2021).
12. Wang, M. et al. Purification, characterization and immunomodulatory activity of water extractable polysaccharides from the swollen culms of *Zizania latifolia*. *Int. J. Biol. Macromol.* **107**, 882–890 (2018).
13. Yang, Z., Davy, A. J., Liu, X., Yuan, S. & Wang, H. Responses of an emergent macrophyte, *Zizania latifolia*, to water-level changes in lakes with contrasting hydrological management. *Ecol. Eng.* **151**, 105814 (2020).
14. Xu, X. W. et al. Phylogeny and biogeography of the eastern Asian–North American disjunct wild-rice genus (*Zizania* L., Poaceae). *Mol. Phylogenet. Evol.* **55**, 1008–1017 (2010).
15. Xu, X. W. et al. Comparative phylogeography of the wild-rice genus *Zizania* (Poaceae) in eastern Asia and North America. *Am. J. Bot.* **102**, 239–247 (2015).
16. Mao, L. et al. RiceRelativesGD: a genomic database of rice relatives for rice research. *Database* **2019**, baz110 (2019).
17. Dong, Z. Y. et al. Extent and pattern of DNA methylation alteration in rice lines derived from introgressive hybridization of rice and *Zizania latifolia* Griseb. *Theor. Appl. Genet.* **113**, 196–205 (2006).
18. Shan, X. et al. Mobilization of the active MITE transposons *mPing* and *Pong* in rice by introgression from wild rice (*Zizania latifolia* Griseb.). *Mol. Biol. Evol.* **22**, 976–990 (2005).
19. Wang, N. et al. Transpositional reactivation of the Dart transposon family in rice lines derived from introgressive hybridization with *Zizania latifolia*. *BMC Plant Biol.* **10**, 190 (2010).

20. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
21. Chen, Q., Li, W., Tan, L. & Tian, F. Harnessing knowledge from maize and rice domestication for new crop breeding. *Mol. Plant* **14**, 9–26 (2021).
22. Yu, H. et al. A route to de novo domestication of wild allotetraploid rice. *Cell* **184**, 1156–1170 (2021).
23. Kennard, W., Phillips, R. & Porter, R. Genetic dissection of seed shattering, agronomic, and color traits in American wildrice (*Zizania palustris* var. interior L.) with a comparative map. *Theor. Appl. Genet.* **105**, 1075–1086 (2002).
24. Guo, L. et al. Genomic clues for crop–weed interactions and evolution. *Trends Plant Sci.* **23**, 1102–1115 (2018).
25. Kitaoka, N. et al. Interdependent evolution of biosynthetic gene clusters for momilactone production in rice. *Plant Cell* **33**, 290–305 (2021).
26. Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B. & Peters, R. J. CYP76M7 is an ent-cassadiene C11 α -hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* **21**, 3315–3325 (2009).
27. Shimura, K. et al. Identification of a biosynthetic gene cluster in rice for momilactones. *J. Biol. Chem.* **282**, 34013–34018 (2007).
28. Hasegawa, M. et al. Phytoalexin accumulation in the interaction between rice and the blast fungus. *Mol. Plant Microbe.* **23**, 1000–1011 (2010).
29. Mennan, H. et al. Quantification of momilactone B in rice hulls and the phytotoxic potential of rice extracts on the seed germination of *Alisma plantago-aquatica*. *Weed Biol. Manag.* **12**, 29–39 (2012).
30. Kato-noguchi, H. & Peters, R. J. The role of momilactones in rice allelopathy. *J. Chem. Ecol.* **39**, 175–185 (2013).
31. Guo, L. et al. A host plant genome (*Zizania latifolia*) after a century-long endophyte infection. *Plant J.* **83**, 600–609 (2015).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
34. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126–e126 (2018).
35. Du, H. et al. Sequencing and de novo assembly of a near complete *indica* rice genome. *Nat. Commun.* **8**, 15324 (2017).
36. Haas, M. W. et al. Whole-genome assembly and annotation of northern wild rice, *Zizania palustris* L., supports a whole-genome duplication in the *Zizania* genus. *Plant J.* **107**, 1802–1818 (2021).
37. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *P. Natl Acad. Sci. Usa.* **101**, 9903–9908 (2004).
38. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009).
39. Kennard, W. C., Phillips, R. L., Porter, R. A. & Grombacher, A. W. A comparative map of wild rice (*Zizania palustris* L. $2n=2x=30$). *Theor. Appl. Genet.* **101**, 677–684 (2000).
40. Hass, B. L., Pires, J. C., Porter, R., Phillips, R. L. & Jackson, S. A. Comparative genetics at the gene and chromosome levels between rice (*Oryza sativa*) and wildrice (*Zizania palustris*). *Theor. Appl. Genet.* **107**, 773–782 (2003).
41. Estornell, L. H., Agustí, J., Merelo, P., Talón, M. & Tadeo, F. R. Elucidating mechanisms underlying organ abscission. *Plant Sci.* **199**, 48–60 (2013).
42. Fernie, A. R. & Yan, J. De novo domestication: an alternative route toward new crops for the future. *Mol. Plant* **12**, 615–631 (2019).
43. Zhang, Y., Pribil, M., Palmgren, M. & Gao, C. A CRISPR way for accelerating improvement of food crops. *Nat. Food* **1**, 200–205 (2020).
44. Miyamoto, K. et al. Evolutionary trajectory of phytoalexin biosynthetic gene clusters in rice. *Plant J.* **87**, 293–304 (2016).
45. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
46. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
47. Bruce, J. et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
48. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
49. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
50. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
51. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
52. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
53. Hoede, C. et al. PASTEC: an automatic transposable element classification tool. *PLoS ONE* **9**, e91929 (2014).
54. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
55. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **25**, 4.10.1–4.10.14 (2009).
56. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
57. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
58. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
59. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
60. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
61. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinforma.* **18**, 4.3.1–4.3.28 (2007).
62. Korf, I. Gene finding in novel genomes. *BMC Bioinforma.* **5**, 59 (2004).
63. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89–e89 (2016).
64. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinforma.* **19**, 189 (2018).
65. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
66. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
67. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78–e78 (2015).
68. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
69. Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M. & Buell, C. R. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* **7**, 327 (2006).
70. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
71. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
72. She, R., Chu, J. S. C., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
73. Birney, E. et al. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
74. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
75. Marchler-Bauer, A. et al. CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229 (2010).
76. Koonin, E. V. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7 (2004).
77. Dimmer, E. C. et al. The UniProt-GO annotation database in 2011. *Nucleic Acids Res.* **40**, D565–D570 (2012).
78. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
79. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
80. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
81. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
82. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
83. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
84. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
85. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
86. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).

87. Puttick, M. N. MCMCTreeR: functions to prepare MCMCTree analyses and visualize posterior ages on trees. *Bioinformatics* **35**, 5321–5322 (2019).
88. Han, M. V., Thomas, G. W., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997 (2013).
89. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
90. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
91. Zwaenepoel, A. & Van de Peer, Y. wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155 (2019).
92. Yan, N. et al. RNA sequencing provides insights into the regulation of solanesol biosynthesis in *Nicotiana tabacum* induced by moderately high temperature. *Biomolecules* **8**, 165 (2018).
93. Mao, X., Cai, T., Olyarchuk, J. G. & Wei, L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* **21**, 3787–3793 (2005).
94. Krzywinski, M. et al. Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
95. Wei, T. et al. Package ‘corrplot’. *Statistician* **56**, 316–324 (2017).

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Non-Profit Scientific Institution (1610232018003, 1610232020008, and 1610232021006), the Agricultural Science and Technology Innovation Program (ASTIP-TRIC05), and the National Natural Science Foundation of China (U20A2043 and 31801336).

Author contributions

N.Y., Q.Q., and Z.-F.Z. conceived and designed the study. N.Y. provided the plants. N.Y., T.Y., X.-T.Y., and L.-G.S. collected the materials, assembled the genome, and performed gene annotation, gene-family, and evolutionary analyses. N.Y. and T.Y. carried out histological observations and transcriptome analysis of the abscission layer. J.Q. carried out the biosynthetic gene cluster analysis for *Z. latifolia*. D.-P.G., Y.Z., L.M., Q.-Q.Q., Y.-L.L., Y.-M.D., X.-M.L., X.-L.Y., and P.Q. helped with data analyses. N.Y., T.Y., and X.-T.Y. wrote the manuscript with help from J.Q., Q.Q., and Z.-F.Z. All authors read and approved the final version of the manuscript.

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/s42003-021-02993-3>.

Correspondence and requests for materials should be addressed to Ning Yan, Qian Qian or Zhong-Feng Zhang.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Caitlin Karniski.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022