A route to de novo domestication of wild allotetraploid rice

Graphical Abstract

De novo domestication of wild allotetraploid rice

Potential agronomic genes

Genome assembly

Germplasm screen

Genome editing system

Improvement of targeted traits

- Seed shattering
- Avn length
- Plant height
- Grain length
- Storm thickness
- Heading date

Highlights

- Establishing efficient transformation and genome editing system in allotetraploid rice
- The high-quality assembly of the genome of allotetraploid rice
- Identification of O. alta homologs of agronomically important genes from diploid rice
- Achieved targeted improvement of various traits in O. alta through genome editing

Authors

Hong Yu, Tao Lin, Xiangbing Meng, ..., Caixia Gao, Chengzhi Liang, Jiayang Li

Correspondence

hyu@genetics.ac.cn (H.Y.), cxgao@genetics.ac.cn (C.G.), jyli@genetics.ac.cn (J.L.), cliang@genetics.ac.cn (C.L.)

In Brief

Li and colleagues developed a breeding route to de novo domestication of wild allotetraploid rice that provides a rational strategy for creating novel crops and generated a series of allotetraploid rice lines edited in domestication-related and agronomically important genes.

Yu et al., 2021, Cell 184, 1–15
March 4, 2021 © 2021 Elsevier Inc.
https://doi.org/10.1016/j.cell.2021.01.013
A route to de novo domestication of wild allotetraploid rice

Hong Yu,1,15,* Tao Lin,1,14,15 Xiangbing Meng,1,15 Huilong Du,1,2,15 Jingkun Zhang,1,2,15 Guifu Liu,1 Mingjiang Chen,1 Yanhui Jing,1 Liquan Kou,1 Xixiu Li,1,2 Qiang Gao,1 Yan Liang,7 Xiangdong Liu,3,4 Zhilan Fan,5 Yuntao Liang,6 Zhukuan Cheng,1,2 Mingsheng Chen,1,2 Zhixi Tian,2,7 Yonghong Wang,1,2 Chengcai Chu,1,2 Jianru Zuo,1,2 Jianmin Wan,8 Qian Qian,9 Bin Han,1 Andrea Zuccolo,11,12 Rod A. Wing,11,13 Caixia Gao,2,7,* Chengzhi Liang,1,2,* and Jiayang Li1,2,3,16,*

1State Key Laboratory of Plant Genomics, and National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China
2University of Chinese Academy of Sciences, Beijing 100049, China
3Guangdong Laboratory of Lingnan Modern Agriculture, Guangzhou 510642, China
4State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, South China Agricultural University, Guangzhou 510642, China
5National Field Genebank for Wild Rice (Guangzhou), Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China
6Rice Research Institute, Guangxi Academy of Agricultural Science, Nanning 530007, China
7State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China
8National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
9State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310006, China
10National Center of Plant Gene Research Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences and CAS Center of Excellence for Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai 200233, China
11Center for Desert Agriculture, Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
12Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa 56127, Italy
13Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
14Present address: College of Horticulture, China Agricultural University, Beijing 100193, China
15These authors contributed equally
16Lead contact
*Correspondence: hyu@genetics.ac.cn (H.Y.), cxgao@genetics.ac.cn (C.G.), cliang@genetics.ac.cn (C.L.), jyli@genetics.ac.cn (J.L.)
https://doi.org/10.1016/j.cell.2021.01.013

SUMMARY

Cultivated rice varieties are all diploid, and polyploidization of rice has long been desired because of its advantages in genome buffering, vigorosity, and environmental robustness. However, a workable route remains elusive. Here, we describe a practical strategy, namely de novo domestication of wild allotetraploid rice. By screening allotetraploid wild rice inventory, we identified one genotype of Oryza alta (CCDD), polyploid rice 1 (PPR1), and established two important resources for its de novo domestication: (1) an efficient tissue culture, transformation, and genome editing system and (2) a high-quality genome assembly discriminated into two subgenomes of 12 chromosomes apiece. With these resources, we show that six agronomically important traits could be rapidly improved by editing O. alta homologs of the genes controlling these traits in diploid rice. Our results demonstrate the possibility that de novo domesticated allotetraploid rice can be developed into a new staple cereal to strengthen world food security.

INTRODUCTION

To meet the world summit on food security, 50% more food is required by 2050, putting enormous pressure on increasing the yield per unit area (Bailey-Serres et al., 2019; Godfray et al., 2010; Long et al., 2015). This challenge becomes even greater owing to rapid climate changes, which results in global warming, extreme high and low temperatures, more droughts and floods, new emergence of pests and diseases, and has already become an immense threat to agriculture and food security (Rosenzweig and Parry, 1994; Wheeler and von Braun, 2013). Environmental adaptation of major crops will involve a huge cost, and there is
a great need to develop new high yield crops that are more resilient to climate variability and disasters (Nelson et al., 2009). Polyploidy, mostly resulting from whole genome duplications or inter-specific hybridizations, is a common mode of evolution in flowering plants, and polyploid plants often have significant advantages in terms of biomass, vigor, and robust adaptation to environmental changes (Comai, 2005; Fang and Morrell, 2016; Lichman et al., 2020; Van de Peer et al., 2017). Therefore, crop polyploidization may play an important role in next-generation crop improvement aimed at facing food security challenges.

The domestication of cultivated rice from its ancestral progenitor wild diploid rice Oryza rufipogon is considered one of the most important developments in human history, and rice is now the staple food feeding more than half of the world’s population (Huang et al., 2012). During domestication, various important agronomical traits were selected, including seed shattering, erect plant architecture, panicle shape, awn length, grain size and quality, hull color, and so on (Chen et al., 2019). In the genomic era, diploid rice has become a model monocot species owing to its relatively simple genome and the development of efficient transformation systems (Izawa and Shimamoto, 1996). Agronomically important genes controlling plant height, tiller number, panicle morphology, seed weight, etc. are well-studied and have been utilized to breed new elite varieties by rational design (Qian et al., 2016; Sasaki et al., 2002; Zeng et al., 2017). Even so, new strategies for rice improvement are badly needed. Polyploidization of cultivated diploid rice species has also been proposed, but this presents difficulties mainly because of their restricted genetic backgrounds (Cai et al., 2007; Wu et al., 2014; Xu et al., 2014).

To date, 27 species of the genus Oryza have been identified and classified into 11 distinct genome types, comprising six diploids (AA, BB, CC, EE, FF, and GG) and five allotetraploids (BBCC, CCDD, HHJJ, HHKK, and KKLL) (Wing et al., 2018). Of these, the species with CCDD genome from South America have much larger biomass and stronger biotic and abiotic resistance than the cultivated diploid rice (Ammiraju et al., 2010a; Khush, 1997; Prusty et al., 2018). Allotetraploid rice with the CCDD genome originated from a single hybridization event, with the CC genome species (O. officinalis or O. rhizomatis) serving as the maternal parent and an extinct species with the DD genome type serving as the paternal donor (O. australiensis with the EE genome type is the extant closest relative of DD) (Ammiraju et al., 2010b; Bao and Ge, 2004; Lu et al., 2009; Wing et al., 2018). Research on allotetraploid rice has largely lagged behind the well-studied diploid rice species with AA genome because of the lack of a reference genome (Stein et al., 2018) and the difficulty of transformation and regeneration (Zhang et al., 2019).

Domestication of wild plants is believed to have enormous potential for agriculture. The traditional domestication of wild plants into commercially available crops has usually taken hundreds or even thousands of years, but recently developed genome editing technologies make it possible to achieve this in a few generations (DeHaan et al., 2020; Eshed and Lippman, 2019; Lemmon et al., 2018; Li et al., 2018; Van Tassel et al., 2020; Zsögön et al., 2018). Therefore, efficient genome editing systems are badly needed to accelerate the pace of domestication. Agrobacte-
material a model for further research into de novo domestication and named it polyploid rice 1 (PPR1). Importantly, PPR1 also had a low rate of heterozygosity (<0.2%) (Figure S2A), which is a positive feature for genome sequencing and assembly. Although it has a substantially larger biomass than cultivated rice, it has typical non-domesticated features, such as height >2.7 m (Figure 1B), broad and long leaves (Figure 1C), panicle >48 cm in length with sparse spikelets (Figure 1D), and small grain size (thousand-grain weight ~8.79 g) with awns >4 cm (Figure 1E). In summary, we selected PPR1 as the starting model material for de novo domestication and established an efficient tissue culture and transformation system of it, thus overcoming one of the biggest hurdles in the process of de novo domestication.

Assembling a high-quality genome of PPR1 and homeologous subgenomes

The second challenge was to assemble a high-quality genome of PPR1. We took an integrated genome sequencing and assembling approach by generating 121.1 Gb (~123.42X) of single-molecule real-time sequencing data, 287.1 Gb (~292.62X) of BioNano data, 103.2 Gb (~105.18X) of high-throughput chromatin conformation capture (Hi-C) data, and 85.2 Gb (~86.79X) Illumina short reads (Figures S2B and S2C), and achieved a high-quality assembly comprising 24 pseudo-chromosomes with contig N50 of 18.2 Mb, superscaffold N50 of 37.1 Mb (PPR1 V1.0) (Table 1), and a total genome size of 894.6 Mb (876.4 Mb anchored on chromosomes). We then constructed a gene expression atlas by generating 235 Gb of RNA sequencing (RNA-seq) data for ten representative tissues, which yielded an average 92.81% mapping rate to PPR1 V1.0 (Figures S3A and S3B). Of the 482,997 transcripts derived from transcriptome sequencing, 90.16% were detected in PPR1 V1.0 with more than 99% identity (Figure S3C). Finally, 99,312 protein-coding genes, including 81,421 high-confidence (HC) and 17,891 low-confidence genes with an average length of 3,021 bp, were predicted and functionally annotated (Figures S3D and S3E). In addition, we annotated 1,253 pseudogenes and 17,104 non-coding RNAs, including 12,332 microRNAs,
By aligning 229,886 BAC end sequences and 278 genes from CC and EE genome species (Table S3) (Ammiraju et al., 2006; Zou et al., 2013) to the PPR1 genome, we divided the 24 pseudo-chromosomes into Ct and Dt subgenomes, which contained 52,861 and 46,388 protein-coding genes, respectively (Table 1). Among them, we identified 19,958 homeologous gene pairs, and the transcriptome analysis indicated that 5,732 homeologous gene pairs on the homologous chromosomes showed expression dominance in one gene. Although genome sizes differed noticeably between CC (~651 Mb) and EE (~965 Mb) species (Zuccolo et al., 2007), the Ct and Dt subgenomes of O. alta had similar sizes of ~441.3 Mb and ~435.1 Mb (Table 1), which were also highly collinear with 100 syntenic blocks containing 17,639 and 17,527 genes, respectively, on the homologous chromosomes of the Ct and Dt subgenomes (Figure 1F; Table S4). These findings further demonstrate the high quality and coverage of the allotetraploid PPR1 genome sequence.

### Genomic features of allotetraploid O. alta

To compare gene similarity between O. alta and O. sativa, we compared the PPR1 HC genes (81,839 genes) and the Nipponbare MSU gene set (55,986 genes) and found that 43,599 (77.87%) genes of Nipponbare had homologs in O. alta, including 39,543 (70.41%) genes with homologous genes on both Ct and Dt subgenomes (Figure 2A). Meanwhile, we found that 10,862 and 9,982 genes on the Ct and Dt subgenomes of O. alta found no homologs in cultivated diploid rice, possibly because the highly diverged homologs could not be reliably identified. As O. alta showed strong biotic and abiotic resistance phenotypes (Khush, 1997; Prusty et al., 2018), we further examined the resistance gene analogs in O. alta (Table S5). Although O. sativa ssp. japonica/Geng Nipponbare, and indica/Xian R498 contain much more nucleotide-binding site (NBS) type genes compared to the Ct or Dt subgenome (Figure 2B) due to the expansion of NBS-leucine rich repeat (NBS-LRR) genes in the AA genome after the split of O. meridionalis and O. glumeapatula (Stein et al., 2018), O. alta contains more genes in total for all four types of resistance gene analogs, showing a potential for broader resistant spectrum. Moreover, the number of transcription factors and transcription regulators is similar among the Ct and Dt subgenomes and AA genome (Figures S4A–S4C; Table S5). These results indicate that the O. alta genome contains considerable genetic resources which have not been utilized in cultivated rice.

Furthermore, based on protein homologies among O. alta and nine other species, a total of 42,642 gene families were identified, with 19,925 and 18,519 in the Ct and Dt subgenomes, respectively (Figure 2C). Of these families, 14,125 were shared among Ct, Dt, and O. sativa and 525 Ct-specific and 363 Dt-specific. Phylogenetic analysis based on single-copy gene families showed that the O. alta Ct subgenome and O. sativa genome were the nearest neighbors, confirming the previous hypothesis that the AA genome shared a more recent common ancestor with the CC genome than with the DD/EE genome (Wing et al., 2018). The estimated divergence times between the AA and CC genomes, and between the CC and DD genomes, were 4.57 and 5.45 million years ago (MYA), respectively (Wing et al., 2018). Both the Ct and Dt subgenomes of allotetraploid

---

#### Table 1. The O. alta genome assembly and gene annotation

<table>
<thead>
<tr>
<th></th>
<th>Total genome</th>
<th>Ct subgenome</th>
<th>Dt subgenome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate of genome size (Mb)</td>
<td>789.9</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Genome assembly (Mb)</td>
<td>894.6</td>
<td>441.3</td>
<td>435.1</td>
</tr>
<tr>
<td>Number of superscaffolds</td>
<td>274</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>N50 of superscaffolds (Mb)</td>
<td>37.1</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>N50 of supercontigs (Mb)</td>
<td>18.2</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>N50 of contigs (Mb)</td>
<td>18.2</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>N90 of superscaffolds (Mb)</td>
<td>27.8</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>N90 of supercontigs (Mb)</td>
<td>8.5</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>N90 of contigs (Mb)</td>
<td>7.2</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Longest sequence length (Mb)</td>
<td>50.4</td>
<td>50.4</td>
<td>49.0</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>43.9</td>
<td>43.7</td>
<td>44.0</td>
</tr>
<tr>
<td>Number of genes</td>
<td>99,312</td>
<td>52,861</td>
<td>46,388</td>
</tr>
<tr>
<td>Number of high confident genes</td>
<td>81,421</td>
<td>43,481</td>
<td>37,908</td>
</tr>
<tr>
<td>Percentage of gene length in genome/subgenome (%)</td>
<td>33.5</td>
<td>36.4</td>
<td>32.0</td>
</tr>
<tr>
<td>Mean gene length (bp)</td>
<td>3,020.5</td>
<td>3,040.5</td>
<td>2,999.6</td>
</tr>
<tr>
<td>Gene density (per Mb)</td>
<td>111.0</td>
<td>119.8</td>
<td>106.6</td>
</tr>
<tr>
<td>Mean coding sequence length (bp)</td>
<td>910.2</td>
<td>908.8</td>
<td>912.0</td>
</tr>
<tr>
<td>Mean exon length (bp)</td>
<td>358.5</td>
<td>359.2</td>
<td>357.7</td>
</tr>
<tr>
<td>Mean intron length (bp)</td>
<td>621.4</td>
<td>629.1</td>
<td>612.7</td>
</tr>
<tr>
<td>Subgenome specific genes</td>
<td>–</td>
<td>6,283</td>
<td>4,931</td>
</tr>
<tr>
<td>Masked repeat sequence length (Mb)</td>
<td>551.5</td>
<td>261.3</td>
<td>272.3</td>
</tr>
<tr>
<td>Repeats percentage of genome size (%)</td>
<td>61.7</td>
<td>59.2</td>
<td>62.6</td>
</tr>
</tbody>
</table>

See also Figures S2 and S3 and Table S2.
**O. alta** were considerably larger than those of Nipponbare and R498. The analysis showed that 61.7% (551.5 Mb) of the **O. alta** genome consisted of repetitive sequences (Table 1), a much higher proportion than the 40.43% in Nipponbare and 42.05% in R498 (Du et al., 2017; International Rice Genome Sequencing Project, 2005). The major difference was in long terminal repeats (LTRs), which are the predominant repeats accounting for 24.1% (215.93 Mb) of the **O. alta** genome, markedly higher than the 17.55% in Nipponbare and 20.55% in R498 (Figure 2D). **Gypsy**-type LTRs were the most abundant transposable element subfamilies, covering 19.9% (177.60 Mb) of the **O. alta** genome (Figure 2D; Table S2). However, **Gypsy** and **Copia**-type LTRs were estimated to have a similar insertion burst time less than one MYA (Figure 2E). Repeats and **Gypsy**-type LTRs covered 62.6% and 21.8% of the D1 subgenome, respectively, higher than 59.2% and 18.7% of the C1 subgenome (Figure 2D). **PIF-Harbinger** transposons were enriched within the 2-kb upstream and downstream of coding regions (Figure S2D), indicating their possible roles in the evolution of cis-elements. These results suggest that the amplification of **Gypsy**-type LTRs is the major driving force for the expansion of the **O. alta** genome.

**Genetic diversity of allotetraploid rice with CCDD genome**

We resequenced a total of 44 wild rice accessions, including 28 CCDD, six EE and ten CC accessions, generating a total of 922.45 Gb of sequence data (Table S1) for population genomic analysis. Sequence reads for each accession were aligned to the PPR1 V1.0 with an average mapping rate of 98.47%, 88.20% and 57.17% for CCDD, CC and EE species, respectively (Table S1). We identified 30.77 million high-quality variants consisting of 21,281,771 SNPs and 9,490,896 indels with an average of 23.8 SNPs and 10.6 indels per kilobase. A total of 7,373,962 SNPs and 3,075,405 indels were located in genic regions, including 1,183,604 nonsynonymous and 288,439 frameshift variants. 70,779 variants had potentially large effects, including 54,340 SNPs and 16,439 indels that cause the introduction of start codons, premature stop codons, and longer-than-usual transcripts. These data indicate that substantial variations exist within the population of CCDD accessions, as well as between C1 and C, and between D1 and E genomes.
Figure 3. Genetic variation map of 44 CC, EE, and CCDD rice accessions and chromosomal rearrangements in the *O. alta* genome

(A) Principal component analysis (PCA) of 44 rice accessions. Different colors indicate accessions belonging to different groups: blue (group 1), diploid EE species; green (group 2), tetraploid CCDD species; red (group 3), diploid CC species.

(B) Representative mature plants of three CCDD *Oryza* species. Scale bars, 30 cm.

(C) Phylogenetic relationship of 44 rice accessions based on genetic distance.

(D) Structure plot for rice accessions with different numbers of clusters (K = 2, 3, and 4).

(E) Homologous exchanges and collinear blocks among non-homologous chromosomes between *Ct* and *Dt* subgenomes. Pink lines connect collinear blocks within *Ct* or *Dt* subgenomes, and orange lines connect collinear blocks between *Ct* and *Dt* subgenomes.

(F) Genome rearrangements between *O. alta* chromosome *C2* and *D2*. Grey and red boxes represent the regions origin from ancestral CC and DD genome, respectively. Back lines represent the sequence depth of the resequencing data from ten CC and six EE diploid rice accessions. Color lines between two chromosomes represent collinear blocks between *O. alta* chromosomes *C2* and *D2*.

(G) Chromatin interactions at 150 kb resolution by Hi-C data. Black pentagrams show the borders of the large segmental translocation between *O. alta* chromosomes *C2* and *D2*.

(H) Collinearity between genome sequences by BioNano maps. Upper horizontal green bars represent the genome sequences and the below blue bars represent the BioNano maps. The lines between green bars and blue bars represent the collinear restriction sites. Black pentagrams in (G) and (H) show borders of the large segmental translocation between *O. alta* chromosomes *C2* and *D2*.

See also Figure S5 and Tables S1 and S4.

Principal component analysis (PCA) showed that the 44 rice accessions were clearly clustered into three groups (Figure 3A), namely group 1 from the EE accessions, group 2 from the CCDD accessions, and group 3 from the CC accessions (Figure 3B). Neighbor-joining tree and model-based clustering analysis further divided group 2 into two clusters (Figures 3C and 3D), with all eight *O. alta*, two *O. grandiglumis*, and four *O. latifolia* accessions in cluster 1 and the other 13 *O. latifolia* accessions residing in cluster 2. Nucleotide diversity (π) in CCDD accessions was estimated to be 0.0049, which is lower than other polyploidy crops such as potato (0.0111) (Hardigan et al., 2017), but higher than in diploid rice (0.0024 in *O. sativa* and 0.0030 in *O. rufipogon*) (Huang et al., 2010), indicating that this population of allotetraploid rice can provide diverse genetic resources for future functional genomic study and breeding.

**Chromosomal rearrangements in allotetraploid *O. alta***

By aligning the resequencing data of the diploid CC and EE *Oryza* species to the PPR1 genome, we were able to identify which parts of the current *O. alta* genome are derived from the ancestral CC or DD genomes. This revealed several translocations of large segments between the two subgenomes after tetraploidization (Figure 3E; Table S4). Chromosomes *C2* and *D2* were highly syntenic, but the aligning pattern of the reads from CC and EE genomes to PPR1 clearly showed the occurrence of a large homoeologous exchange between them (Figure 3F). The continuous interaction signals on the Hi-C heatmap and the consistency with BioNano maps in these regions indicated that this was not due to assembly errors (Figures 3G and 3H).

In addition to the homoeologous exchange, we found three segments on *Ct*7 matching to EE genome reads were collinear with
C4, suggesting that these segments were translocated from the originally ancestral chromosome D4 (Figures 3E and S5A). Again, the correctness of the assembly was confirmed by Hi-C heatmap and BioNano data (Figures S5B and S5C). Similarly, we found several other segments, including one on D4, one on D3, and one on C1, were translocated from the ancestral chromosomes C7, C6, and D3, respectively. Additionally, we also identified 26 syntenic blocks located on non-homeologous chromosomes, apparently caused by ten intra-subgenome translocations, four intra-chromosomal duplications, and 12 inter-chromosomal duplications, respectively (Figure 3E; Table S4). C11, C12, and D11, D12 likely experienced multiple duplications and translocations (Figure 3E), but the other six chromosomes have experienced fewer chromosomal rearrangements. According to the matching reads of CC and EE species on each gene, we further identified all the genes originating from ancestral CC genome (ancestral CC genes) and ancestral DD genes in the PPR1 genome, and we found that 5,672 genes (13.04%) were now in the C1 subgenome were of DD origin, while 5,132 genes (13.54%) on the D1 subgenome were CC ancestral. Meanwhile, we found that a large number of genes had been lost since tetraploidization; 5,516 ancestral CC genes had no ancestral DD homologs in the current O. alta genome, and 3,954 ancestral DD genes had no ancestral CC homeologs. All these data suggest the highly dynamic evolution of the O. alta genome.

A route for rapid domestication of PPR1 by genome editing

During the domestication of Asian cultivated rice, traits beneficial for the farming rather than for natural growth, such as erect growth habit, hull color, shattering, awn, pericarp color, and grain size, were selected and improved, and their genetic mechanisms have been extensively studied (Chen et al., 2019). Wild rice O. alta shares some traits similar to those of the wild ancestors of the modern cultivars. To rapidly domesticate these traits, we first identified the O. alta homologs of the domestication-related genes in diploid rice (Figure 4A), and found that seven of ten important domestication-related genes (Chen et al., 2019), including homologs of qSH1 for shattering, An-1 and An-2 for awn length, BH4 for hull color, Rc for pericarp color, OsLG1 for panicle shape, and GWS for grain width, were present in the O. alta genome with >84% identity covering at least 87.5% of O. sativa proteins. PROG1, a key gene for the erect growth of cultivated rice appeared to have only low level homologs in the subgenomes of O. alta, possibly because O. alta already has an erect growth habit. These domestication-related genes have been exposed to a variety of evolutionary processes. Sh4 experienced inter-subgenome translocation from the D1 to the C1 subgenome, leading to the shortening of its coding regions. Both of the homologs of GAD1/RAE2 and OsLG1 are present in the C1 subgenome, owing to the translocation of a segment from the D1 subgenome to the C1 subgenome. As a proof-of-concept to show that modification of these homologs could truly achieve rapid domestication of wild O. alta, we used a CRISPR/Cas9-mutagenesis approach to edit the qSH1 homologs for shattering and An-1 homologs for awn length (Konishi et al., 2006; Luo et al., 2013). We first designed a single guide RNA (sgRNA) targeting the first exon of both OaqSH1-CC (OalC01g168290) and OaqSH1-DD (OalD01g114050) and identified one mutant, qsh1^CR-1, which harbored biallelic frameshift mutations for OaqSH1-CC and heterozygous frameshift mutations for OaqSH1-DD (Figure 4B; Table S6). We found that wild-type PPR1 had a complete layer of abscission cells between the rice grain and the pedicel, forming a longitudinal continuous line similar to wild diploid rice, whereas in qsh1^CR-1, the line of abscission cells was absent, indicating that the editing of the qSH1 homologs in O. alta can prevent seed shattering (Figure 4C). We further designed one sgRNA that targets the first exons of both OaAn-1-CC (OalC04g136090) and OaAn-1-DD (OalD04g130280) and another sgRNA that targets the first exon of OaAn-1-DD, and obtained two mutants of an-1^CR-1 and an-1^CR-2 harboring different homologous frameshift mutations in both OaAn1-CC and OaAn1-DD (Figure 4D; Table S6). We found that the average awn length of an-1^CR-1 (1.61 cm) and an-1^CR-2 (2.63 cm) were both significantly shorter than that of PPR1 (4.11 cm) (Figures 4E and 4F). Taken together, these results provide a solid basis for the rapid domestication of O. alta.

Variable sequences of O. alta homologous to agronomically important O. sativa genes

In addition to the domestication-related genes, we also identified O. alta homologs of 113 agronomically important genes in modern diploid rice cultivars, which determine important traits such as grain yield, grain quality, fertility, heading date, biotic and abiotic resistance, and nutrient-use efficiency (Figure 5A; Table S7). We found that the genes affecting different features had different similarity levels. The genes affecting sterility in O. sativa were poorly conserved in O. alta, because the best-matched homologs of qHMS7, S5, SaF, Scv-j, and S1TPR had similarities less than 40%, and there were no detectable homologs of SaM, WA352, and pms3 (Figure 5A; Table S7). The genes associated with biotic stress were also relatively poorly conserved, possibly because diseases and pests are different in Asia and America. In contrast, all genes associated with abiotic stress were highly conserved except for COLD1 (Ma et al., 2015b) and CAL1 (Luo et al., 2018), which are responsible for cold resistance and cadmium accumulation, respectively. Fortunately, we found that most of the genes related to yield, grain quality, and efficient use of nutrients were fairly conserved between the cultivated diploid and the tetraploid wild rice. Based on these findings, we further tested whether these homologs of modern breeding genes could be altered to improve wild allotetraploid rice.

The “Green Revolution” gene sd1 is one of the most important genes in modern rice breeding, because its mutation resulted in a shortened culm with improved lodging resistance and a greater harvest index (Sasaki et al., 2002). We therefore designed two sgRNAs targeting the first exons of both OaSD1-CC (OalC01g172060) and OaSD1-DD (OalD01g109880), and obtained an sd1^CR-1 mutant that harbored biallelic frameshift mutations in both OaSD1-CC and OaSD1-DD (Figure 5B; Table S6). We found that its height was dramatically reduced compared with the wild-type PPR1 (Figure 5C). In contrast to O. sativa, which usually has 4–5 elongated internodes, O. alta has 10–12 elongated internodes. We carried out a detailed comparison of the lengths of the first eight internodes from the base
of the stem and found that the length of the first and fifth internodes was relatively unchanged in sd1CR-1, whereas all six other internodes were significantly shortened (Figures 5D and 5E).

Similar results were obtained in the other five independently edited lines (Figures S6A, S6C, and S6D; Table S6). Furthermore, we also designed an sgRNA targeting the homologs of GS3, a key gene controlling seed morphology in diploid rice (Mao et al., 2010) and obtained six independent mutants, in which the grain lengths were significantly increased compared with PPR1 (Figures 5F, 5G, S6B, S6E, and S6F; Table S6). Taken together, these results demonstrated that the knowledge gained in diploid rice can provide valuable information and directly facilitate the improvement of wild allotetraploid rice.

**Improving PPR1 with base substitution editor and multiplex editing technologies**

Ideal Plant Architecture 1 (IPA1) is considered as one of the new green revolution genes, because a point mutation in the gene...
perturbs OsmiR156 target sites to regulate its expression level, leading to an increased stem diameter, improved lodging resistance, large panicles, and dramatically enhanced grain yield (Jiao et al., 2010; Miura et al., 2010). We found that the sequence of OamiR156-DD was identical to OsmiR156, which matches its target site on OaIPA1-DD (OalD08g132520), whereas OamiR156-CC had a single base change and matches OaIPA1-CC (OalC08g106170) (Figures 6A and 6B). We therefore tested whether a base editor optimized for diploid rice could generate gain-of-function base substitution mutations in O. alta (Zong et al., 2018) and designed an sgRNA targeting the miRNA target sites in both OaIPA1-CC and OaIPA1-DD. We obtained a mutant, ipa1CR-1, which contains one point mutation in OaIPA1-DD without affecting OaIPA1-CC (Figure 6B; Table S6). In-detailed examination of this mutant revealed that the stem diameter of ipa1CR-1 was significantly greater than that of PPR1 (Figures 6C–6E). We further examined the transcriptional changes of OaIPA1-CC and OaIPA1-DD and found that the expression levels of OaIPA1-DD were 10-fold greater in the shoot bases of ipa1CR-1, whereas OaIPA1-CC expression was unchanged (Figure 6F). Therefore, the use of base substitution editors can expand the strategies for gene improvement in O. alta.

Rice heading dates, unlike other traits that usually have changes in one direction, become more variable during breeding for expanded habitats, different cropping systems, and different day lengths. In diploid rice, varieties with weak alleles of Ghd7 can be grown for grain production in temperate zones (Weng et al., 2014; Xue et al., 2008), whereas the varieties with low DTH7 activities are less sensitive to increased...
day length to flower earlier in high geographical regions (Gao et al., 2014; Yan et al., 2013). PPR1 originates from tropical regions of South America, displays strong long-day sensitivity, and fails to flower in Beijing (40˚C14N). We used a multiplex CRISPR/Cas9 editing approach with three sgRNAs in one vector to edit OaGhd7-CC (OalC07g145340), OaGhd7-DD (OalD07g113810), OaDTH7-CC (OalD04g100060, a CC ancestral gene locating on chromosome D4), and OaDTH7-DD (OalD07g145130), and generated eight T0 independently transformed PPR1 lines. We found that the eight lines all contained different mutant alleles (Figure 6G; Table S6), and all four genes had been edited in six of the lines, indicating that this technique is highly effective in creating variants in PPR1. In consistence with their genotypes, heading dates varied extensively in these lines.

![Figure 6](image_url)

**Figure 6. Improvement of PPR1 with base editor and multiplex editing systems**

(A) Sequence alignment of OsmiRNA156, OamiRNA156-CC and OamiRNA156-DD.

(B) Base editor-induced point mutation in OaIPA1-DD in T0 plant.

(C and D) Comparison of third internodes from the stem bases (C) and their cross sections (D) between PPR1 and ipa1CR-1. Scale bars, 5 cm (C); 0.5 cm (D).

(E) Statistical analysis of stem thickness of the third internodes of PPR1 and ipa1CR-1. Data are means ± SD (n = 10 tillers). Asterisks represent significant difference determined by Student’s t test. **p < 0.01.

(F) Relative expression levels of OaIPA1-CC and OaIPA1-DD to OaUBI in the shoot bases of PPR1 and ipa1CR-1. Data are means ± SD (n = 4 technical replicates). Asterisks represent significant difference between ipa1CR-1 and PPR1 determined by Student’s t test. **p < 0.01; ns, not significant.

(G) Genotypes of OaGhd7-CC, OaGhd7-DD, OaDTH7-CC, and OaDTH7-DD and plant morphologies of eight edited T0 plants at 110 days after transplantation in Beijing. Red triangles indicate booting young panicles. Scale bars, 30 cm.

(H) Structure of the multiplex genome editing system with eight sgRNAs targeting 16 genes in PPR1. NLS, nuclear localization sequence; Pubi, promoter of maize ubiquitin; Tnos, terminator of nopaline synthase gene. See also Figure S7 and Table S6.
edited lines. Line 1 with all four genes mutated had a dramatically shortened heading date of 82 days after transplantation in Beijing, whereas wild-type PPR1 failed to flower after 150 days. Line 2 and line 3 headed 10 days later after line 1, and lines 4 to 8 gradually headed after 103 to 130 days. The seeds of lines 1 to 7 could be filled, while line 8 seeds failed, suggesting that lines 1 to 3 could even be planted in far northern regions. More importantly, we also developed a multiplex editing system that can target eight or sixteen genes each with a single vector with fairly high efficiency (Figures 6H and S7; Table S6). Taken together, these results showed that it is rational in theory and technology to achieve rapid de novo domestication and improvement of wild allotetraploid rice into a staple food crop in near future.

**DISCUSSION**

Polyploidy, one of the most important evolutionary events in plants, can increase genetic diversity, introduce new genetic combinations, foster adaptation to new environments, and create vigorousness effects. A great number of economically important crops are allopolyploids; these include food crops like wheat and oat, industrial crops like tobacco, cotton, and sugarcane, and fruit crops like strawberry, but most of them are natural allopolyploid species. Although the advantages of allopolyploidy are widely recognized, artificially induced allopolyploid crops are very few and usually result from a long-term breeding program to screen out frequently encountered undesirable features (e.g., triticale, an allopolyploid crop derived from a cross between wheat and rye). Instead of inducing allopolyploidy in the modern crops, we aimed at de novo domestication of a wild allopolyploid plant using genome editing. Starting from the selected genotype PPR1 of the wild allotetraploid rice O. alta, we established an efficient tissue culture and genome editing system, and generated a high-quality genome assembly. By taking advantage of the well-studied domestication-related genes in cultivated diploid rice, we identified their homologs in the O. alta genome, providing essential information for editing its potentially important agronomic genes. We then applied the CRISPR/Cas9, base editing, and multiplex editing technologies to improve PPR1 by editing a series of domestication-related and/or agronomically important genes in diploid rice, in which we obtained edited lines with desired characteristics including seed shattering, awn length, plant height, grain size, stem thickness, and heading date. The present work provides an effective solution to the technical difficulties in achieving rapid de novo domestication of wild allotetraploid rice. We present the use of wild germplasm, complete reference genomes, improved transformation protocol, and diverse genome editing tools to enable the creation of new crop species.

Hereafter, enormous efforts and investments are still needed to achieve the ultimate success in creating a new crop cultivar deployed in farmers’ field. First, although the barrier of transformation and genome editing has been overcome in wild allotetraploid rice, many alleles require weak effects rather than complete knockouts to generate the most desirable traits. Toward this, high efficient genome editing tools in generating various kinds of modification are still urgently needed. Second, the genes implicated usually have pleotropic effects, and the outcome of different combination of edited alleles is difficult to predict. Therefore, there need more genetic diversities to enable a breeding program that optimizes the engineered alleles. Third, wild rice has many advantages over cultivated rice for traits such as biotic and abiotic stress resistance, which are key factors toward combating the impending climate change challenges (Chen et al., 2019; Khush, 1997; Prusty et al., 2018; Song et al., 1995). Thus, there is huge interest in both characterizing these traits in wild rice (Huang et al., 2008; Stein et al., 2018; Zhou et al., 2011) and exploring the capability of hybridization with wild rice guided by molecular markers (Atwell et al., 2014; Menguer et al., 2017). Unfortunately, gene cloning in wild rice is much more difficult than in cultivated rice due to not only the lack of genomic information and transformation systems, but also difficulties in constructing new genetic populations and reflecting the complexity of these desirable traits. Therefore, classical breeding will be needed to maintain these traits without a full understanding of their mechanisms. Fourth, because the wild allotetraploid rice population has a higher genetic diversity, they offer natural genetic resources to be exploited in future functional genomic studies and serve as hybridization in new breeding programs.

Over the past decades, significant progress has been made in rice functional genomics studies with the cloning of many key genes in cultivated rice (Chen et al., 2019; Guo et al., 2019; Jiao et al., 2010; Kim and Buell, 2015; Li et al., 2003b; Miura et al., 2010; Wang et al., 2018a, 2018b). This has greatly benefited rice breeding by allowing the important alleles from different cultivated varieties to be combined by hybridization. One successful example of rice is the development of new elite rice varieties of both high yield and superior quality (Tian et al., 2009; Zeng et al., 2017). At present, three diploid (AA, BB, and FF) rice genomes have been assembled (Chen et al., 2013; Du et al., 2017; Goff et al., 2002; International Rice Genome Sequencing Project, 2005; Stein et al., 2018; Yu et al., 2002; Zhang et al., 2014; Zhao et al., 2018), but only AA species have high-quality pseudo-chromosome level genomes (Wing et al., 2018, Zhou et al., 2020). Generating high-quality genome assemblies for polyploid genomes represents a challenge, especially in discriminating between homeologous subgenomes (Kyriakidou et al., 2018). Many commercially important allopolyploid and autopolyploid crops have been sequenced and assembled, such as wheat (Brenchley et al., 2012; International Wheat Genome Sequencing Consortium, 2018), oilseed rape (Chalhoub et al., 2014), cotton (Li et al., 2015; Zhang et al., 2015; Huang et al., 2020), sweet potato (Yang et al., 2017), sugarcane (Zhang et al., 2018), and teff (VanBuren et al., 2020) with contig N50 ranging from ~480 bp to ~5 Mb. From our experiences in this study, excellent assemblers, sufficient extra-long-read sequencing data, a combination of different sequencing strategies, and selecting the good material with a low heterozygosity rate are critical for assembled polyploid genome. The high-quality O. alta genome will provide a powerful tool for molecular genetic studies of tetraploid rice, the evolution of Oryza species, and the molecular mechanisms underlying the polyploidization of higher plants.

In conclusion, the strategy of the de novo domestication of O. alta, a wild allotetraploid rice, demonstrates a clear path for
creating novel crops in the future by combining the advantages of (1) polyploidy, (2) functional genomics knowledge of cultivated crops, (3) desirable attributes of wild species, and (4) rapid genetic modification(s) through genome editing.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
- **METHOD DETAILS**
  - Assembly of the PPR1 genome
  - Genome annotation
  - Gene function annotation
  - Annotation of repetitive sequences
  - Annotation of non-coding RNAs
  - Annotation of resistance genes, transcription factors and transcription regulators
  - LTR-RT identification and insertion time estimation
  - Transcriptome data analysis
  - Gene family analysis
  - Population genetics analysis
  - Collinear analysis
  - Homologous genes of diploid rice in *O. alta*
  - Transgenic system of *O. alta*
  - CRISPR/Cas9 based genome editing of *O. alta*
  - RNA extraction and quantitative real-time PCR analysis
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

**Supplemental information**

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.01.013.

**ACKNOWLEDGMENTS**

We thank Prof. Yaoguang Liu (South China Agricultural University) for providing pYLCRISPR/Cas9-Pubi-H vector. The research was supported by the National Natural Science Foundation of China (31788103), the Chinese Academy of Sciences (XDA24030504 and XDA24040201), the National Key Research and Development Program of China (2016YFD0101800), and the National Transgenic Science and Technology Program of China (2019ZX08010-003 and 2019ZX08010-001).

**AUTHOR CONTRIBUTIONS**

H.Y., T.L., X.M., H.D., J. Zhang, L.K., X. Li, Q.G., M.C., G.L., Y.J., and Yan Liang performed the experiments. H.Y., T.L., X.M., H.D., J. Zhang, Z.T., C.C., J. Zuo, Y.W., C.G., Q.Q., B.H., C.L., and J.L. analyzed the data. Z.F., Y.H., C.G., X.M., C. Liu, and L. Wu collected and provided germplasm. H.Y., C.G., C.L., and J.L. designed the experiments. J.L. conceived the project. H.Y., T.L., and J.L. wrote the manuscript with input from C.G., C.L., J. Zhang, J. Zuo, X.M., A.Z., and R.W. All authors have read, edited, and approved the content of the manuscript.

**DECLARATION OF INTERESTS**

C.G. is a member of the Cell advisory board.

Received: June 7, 2020
Revised: December 2, 2020
Accepted: January 11, 2021
Published: February 3, 2021

**REFERENCES**


### STAR METHODS

#### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacterial and virus strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agrobacterium tumefaciens strain EHA 105</td>
<td>Meng et al., 2017</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Chemicals, peptides, and recombinant proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BsaI</td>
<td>NEB</td>
<td>Cat#R053SL</td>
</tr>
<tr>
<td>T4 DNA Ligase</td>
<td>NEB</td>
<td>Cat#M0202L</td>
</tr>
<tr>
<td>Phloroglucinol</td>
<td>SIGMA</td>
<td>Cat#108-73-6</td>
</tr>
<tr>
<td><strong>Critical commercial assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wizard SV Gel and PCR Clean-Up System</td>
<td>Promega</td>
<td>Cat#A9282</td>
</tr>
<tr>
<td>TRIzol Kit</td>
<td>Invitrogen</td>
<td>Cat#15596018</td>
</tr>
<tr>
<td>Maxima H Minus cDNA Synthesis</td>
<td>Thermo Scientific</td>
<td>Cat#M1682</td>
</tr>
<tr>
<td>Master Mix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SsoFast EvaGreen Supermix</td>
<td>Bio-Rad</td>
<td>Cat#1725201</td>
</tr>
<tr>
<td>VK005-01</td>
<td>Viewosolid Biotech</td>
<td>Cat#VK005-01</td>
</tr>
<tr>
<td><strong>Deposited data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allotetraploid rice: Pacbio, Bionano, Hi-C and DNA sequencing data of PPR1</td>
<td>This study</td>
<td>BioProject: PRJCA002366</td>
</tr>
<tr>
<td>Allotetraploid rice: de novo assembly genome data of PPR1</td>
<td>This study</td>
<td>BioProject: PRJCA002366</td>
</tr>
<tr>
<td>Allotetraploid rice: Transcriptome sequence data</td>
<td>This study</td>
<td>BioProject: PRJCA002366</td>
</tr>
<tr>
<td>Allotetraploid and diploid rice: DNA re-sequencing data of CCDD, CC and EE species</td>
<td>This study</td>
<td>BioProject: PRJCA002366</td>
</tr>
<tr>
<td>The original unprocessed data</td>
<td>This study</td>
<td>Mendeley Dataset: <a href="https://doi.org/10.17632/mk3ykkms44.1">https://doi.org/10.17632/mk3ykkms44.1</a></td>
</tr>
<tr>
<td><strong>Oligonucleotides</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward primer sequence for quantitative real-time PCR analysis for OaUBI (OalC01 g115320 and OalC03 g154610): AACCAGCTGAGGCCCAAGA</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Reverse primer sequences for quantitative real-time PCR analysis for OaUBI (OalC01 g115320 and OalC03 g154610): ACGATGTGATTAAACAGCTCCATGA</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Forward primer sequences for quantitative real-time PCR analysis for OaIPA1-CC (OalC08 g106170): GTTGAGCGCTTTACGCTCTGCTA</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Reverse primer sequences for quantitative real-time PCR analysis for OaIPA1-CC (OalC08 g106170): CATAAGCACCCTTCATCAAA</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Forward primer sequences for quantitative real-time PCR analysis for OaIPA1-DD (OalD08 g132520): TTCAGTTTCGTTGGGAGAG</td>
<td>This study</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(Continued on next page)
Reverse primer sequences for quantitative real-time PCR analysis for *OaIPA1-DD* (*OaD08 g132520*):

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverse primer sequences</td>
<td>This study</td>
<td>N/A</td>
</tr>
</tbody>
</table>

**Software and algorithms**

<table>
<thead>
<tr>
<th>Software and algorithms</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomescope (version 1.0)</td>
<td>Vurture et al., 2017</td>
<td><a href="http://www.genome.umd.edu/jellyfish.html#Release">http://www.genome.umd.edu/jellyfish.html#Release</a></td>
</tr>
<tr>
<td>CANU (version 1.3)</td>
<td>Koren et al., 2017</td>
<td><a href="https://github.com/marbl/canu">https://github.com/marbl/canu</a></td>
</tr>
<tr>
<td>Aigner and Assembler (version 2.0.6)</td>
<td>Shelton et al., 2015</td>
<td><a href="https://bionanogenomics.com/">https://bionanogenomics.com/</a></td>
</tr>
<tr>
<td>HERA (version 1.0)</td>
<td>Du and Liang, 2019</td>
<td><a href="https://github.com/liangclab/HERA/">https://github.com/liangclab/HERA/</a></td>
</tr>
<tr>
<td>Juicer (version 1.5)</td>
<td>Durand et al., 2016</td>
<td><a href="https://github.com/aidenlab/juicer/">https://github.com/aidenlab/juicer/</a></td>
</tr>
<tr>
<td>BUSCO (version 3.0.2)</td>
<td>Sepey et al., 2019</td>
<td><a href="https://busco.ezlab.org/">https://busco.ezlab.org/</a></td>
</tr>
<tr>
<td>AUGUSTUS (version 3.3)</td>
<td>Stanke et al., 2004</td>
<td><a href="https://sourceforge.net/projects/augustus/">https://sourceforge.net/projects/augustus/</a></td>
</tr>
<tr>
<td>GENSCAN (version 1.0)</td>
<td>Burge and Karlin, 1997</td>
<td><a href="http://hollywood.mit.edu/gendist.html">http://hollywood.mit.edu/gendist.html</a></td>
</tr>
<tr>
<td>Glimmer-HMM (version 3.0.4)</td>
<td>Majoros et al., 2004</td>
<td><a href="http://www.cbcb.umd.edu/software/glimmerm/index.shtml">http://www.cbcb.umd.edu/software/glimmerm/index.shtml</a></td>
</tr>
<tr>
<td>SNAP (version 2.3)</td>
<td>Korf, 2004</td>
<td><a href="https://github.com/KorfLab/SNAP">https://github.com/KorfLab/SNAP</a></td>
</tr>
<tr>
<td>EVidenceModeler (version 1.1.1)</td>
<td>Haas et al., 2008</td>
<td><a href="https://github.com/EVidenceModeler/EvidenceModeler/releases">https://github.com/EVidenceModeler/EvidenceModeler/releases</a></td>
</tr>
<tr>
<td>InterProScan (V5.19-58.0)</td>
<td>Jones et al., 2014</td>
<td><a href="https://www.ebi.ac.uk/interpro/">https://www.ebi.ac.uk/interpro/</a></td>
</tr>
<tr>
<td>RepeatScout (version 1.0.5)</td>
<td>Price et al., 2005</td>
<td><a href="http://bix.ucsd.edu/repeatscout/">http://bix.ucsd.edu/repeatscout/</a></td>
</tr>
<tr>
<td>RECON (version 1.0.8)</td>
<td>Bao and Eddy, 2002</td>
<td><a href="http://www.repeatmasker.org/RepeatModeler/">http://www.repeatmasker.org/RepeatModeler/</a></td>
</tr>
<tr>
<td>RepeatMasker (version 4.0.7)</td>
<td>Tarailo-Graovac and Chen, 2009</td>
<td><a href="http://repeatmasker.org/">http://repeatmasker.org/</a></td>
</tr>
<tr>
<td>RNAmeR (version 1.2)</td>
<td>Lagesen et al., 2007</td>
<td><a href="http://www.cbs.dtu.dk/services/RNAmeR/">http://www.cbs.dtu.dk/services/RNAmeR/</a></td>
</tr>
<tr>
<td>INFERNAL (version 1.1.3)</td>
<td>Nawrocki and Eddy, 2013</td>
<td><a href="http://eddylab.org/infearnal/">http://eddylab.org/infearnal/</a></td>
</tr>
<tr>
<td>LTR-FINDER (version 1.0.7)</td>
<td>Xu and Wang, 2007</td>
<td><a href="https://github.com/xhub/LTR_Finder">https://github.com/xhub/LTR_Finder</a></td>
</tr>
<tr>
<td>LTR-Harvest (version 1.5.7)</td>
<td>Ellingham et al., 2008</td>
<td><a href="https://github.com/Adamtaranto/LTR-Harvest-Refomat">https://github.com/Adamtaranto/LTR-Harvest-Refomat</a></td>
</tr>
<tr>
<td>LTR-retriever (version 2.8.5)</td>
<td>Ou and Jiang, 2018</td>
<td><a href="https://github.com/oushujun/LTR_retriever">https://github.com/oushujun/LTR_retriever</a></td>
</tr>
<tr>
<td>TopHat2 (version 2.1.0)</td>
<td>Kim et al., 2013</td>
<td><a href="http://ccb.jhu.edu/software/tophat/index.shtml">http://ccb.jhu.edu/software/tophat/index.shtml</a></td>
</tr>
<tr>
<td>Cufflinks (version 0.9.3)</td>
<td>Trapnell et al., 2010</td>
<td><a href="https://github.com/cole-trapnell-lab/cufflinks">https://github.com/cole-trapnell-lab/cufflinks</a></td>
</tr>
<tr>
<td>R (version 3.5.0)</td>
<td>R Core Team</td>
<td><a href="https://www.r-project.org/">https://www.r-project.org/</a></td>
</tr>
<tr>
<td>OrthoMCL (version 2.0.9)</td>
<td>Li et al., 2003a</td>
<td><a href="https://orthomcl.org/orthomcl/">https://orthomcl.org/orthomcl/</a></td>
</tr>
<tr>
<td>CLUSTALW (version 2.0.10)</td>
<td>Thompson et al., 1994</td>
<td><a href="https://www.ebi.ac.uk/Tools/msa/clustalw2/">https://www.ebi.ac.uk/Tools/msa/clustalw2/</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
Lead contact
Further information and requests for resources and reagents should be directed to and will be made available upon reasonable request by the Lead Contact, Jiayang Li (jyli@genetics.ac.cn).

Materials availability
The study did not generate new unique reagents.

Data and code availability
The sequence data reported in this study have been deposited in the Genome Sequence Archive in BIG Data Center (http://bigd.big.ac.cn/gsa) under accession number PRJCA002366.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

A total of 44 wild rice accessions were collected, including 28 CCDD species (eight O. alta, two O. grandiglumis, and 18 O. latifolia), ten CC species, six EE species (Table S1). Plants and the tissue culture plantlets were grown and cultivated in the experimental station and the greenhouses of Institute of Genetics and Developmental Biology, Chinese Academy of Sciences (IGDB-CAS), Beijing, China (40°14′22″N and 116°23′0″E) and/or Hainan (18°31′N and 109°62′E) in the growth season from 2016 to 2020.

METHOD DETAILS

Assembly of the PPR1 genome
Genomic DNA was extracted from young leaves using the cetyltrimethylammonium bromide (CTAB) method (Allen et al., 2006). At least 5 μg of genomic DNA was used for each accession to construct paired-end sequencing libraries with insert sizes of approximately 300 bp. For SMRT PacBio sequencing, a 20-kb insert SMRTbell library was generated using a 15-kb lower-end size selection protocol on a BluePippin (Sage Science). The PPR1 genome was sequenced on the PacBio RS II platform (Pacific Biosciences). PacBio raw reads were processed to filter low-quality reads (RQ > 0.75). The number of 19-mers for corrected PE150 reads was counted using the Genomescope program (Vurture et al., 2017). The PPR1 genome size was calculated by dividing the total number of k-mers times of the highest k-mer frequency in the whole genome. The following steps were used to assemble, improve and correct the PPR1 genome assembly. First, ~121.1 Gb of clean PacBio reads were pre-processed and assembled into contigs using the CANU software (Koren et al., 2017). Then 287.1 Gb of Bionano data were used for assisting the scaffold construction using Aigner and Assembler (Shelton et al., 2015). To correct and fill gaps for misassembly, the assembler HERA (Du and Liang, 2019) was used to

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCMCTree (version 4.9)</td>
<td>Yang, 2007</td>
<td><a href="http://web.mit.edu/6.891/www/lab/paml.html">http://web.mit.edu/6.891/www/lab/paml.html</a></td>
</tr>
<tr>
<td>CAFÉ (version 4.2.1)</td>
<td>De Bie et al., 2006</td>
<td><a href="https://github.com/hahnlab/CAFE">https://github.com/hahnlab/CAFE</a></td>
</tr>
<tr>
<td>BWA (version 0.7.17)</td>
<td>Li and Durbin, 2009</td>
<td><a href="https://bio-bwa.sourceforge.net">https://bio-bwa.sourceforge.net</a></td>
</tr>
<tr>
<td>SAMtools (version 1.5.0)</td>
<td>Li et al., 2009</td>
<td><a href="http://samtools.sourceforge.net">http://samtools.sourceforge.net</a></td>
</tr>
<tr>
<td>MCScan (version 1.1)</td>
<td>Tang et al., 2008</td>
<td><a href="https://github.com/tanghaibao/mcsan">https://github.com/tanghaibao/mcsan</a></td>
</tr>
<tr>
<td>Circos (version 0.69-9)</td>
<td>Krzywinski et al., 2009</td>
<td><a href="http://www.circos.ca/software/">http://www.circos.ca/software/</a></td>
</tr>
<tr>
<td>snpEFF (version 4.3)</td>
<td>Cingolani et al., 2012</td>
<td><a href="http://snef.sourceforge.net">http://snef.sourceforge.net</a></td>
</tr>
<tr>
<td>STRUCTURE (version 2.3.4)</td>
<td>Falush et al., 2003</td>
<td><a href="https://web.stanford.edu/group/pritchardlab/structure.html">https://web.stanford.edu/group/pritchardlab/structure.html</a></td>
</tr>
<tr>
<td>EIG (version 6.0.1)</td>
<td>Patterson et al., 2006</td>
<td><a href="https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT">https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT</a></td>
</tr>
<tr>
<td>RGAugury</td>
<td>Li et al., 2016</td>
<td><a href="https://bitbucket.org/yaanlpc/rgaugury">https://bitbucket.org/yaanlpc/rgaugury</a></td>
</tr>
<tr>
<td>iTAK (version 18.12)</td>
<td>Zheng et al., 2016</td>
<td><a href="http://itak.feilab.net/cgi-bin/itak/index.cgi">http://itak.feilab.net/cgi-bin/itak/index.cgi</a></td>
</tr>
<tr>
<td>CRISPR-GE</td>
<td>Xie et al., 2017</td>
<td><a href="http://skl.scau.edu.cn/">http://skl.scau.edu.cn/</a></td>
</tr>
</tbody>
</table>

RESOURCES AVAILABILITY

MCMCTree (version 4.9) Yang, 2007 http://web.mit.edu/6.891/www/lab/paml.html
CAFÉ (version 4.2.1) De Bie et al., 2006 https://github.com/hahnlab/CAFE
BWA (version 0.7.17) Li and Durbin, 2009 https://bio-bwa.sourceforge.net
SAMtools (version 1.5.0) Li et al., 2009 http://samtools.sourceforge.net
MCScan (version 1.1) Tang et al., 2008 https://github.com/tanghaibao/mcsan
Circos (version 0.69-9) Krzywinski et al., 2009 http://www.circos.ca/software/
snpEFF (version 4.3) Cingolani et al., 2012 http://snef.sourceforge.net
STRUCTURE (version 2.3.4) Falush et al., 2003 https://web.stanford.edu/group/pritchardlab/structure.html
EIG (version 6.0.1) Patterson et al., 2006 https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT
RGAugury Li et al., 2016 https://bitbucket.org/yaanlpc/rgaugury
iTAK (version 18.12) Zheng et al., 2016 http://itak.feilab.net/cgi-bin/itak/index.cgi
CRISPR-GE Xie et al., 2017 http://skl.scau.edu.cn/
update the scaffolds to supercontigs. The resulting supercontigs and scaffolds were integrated using 103.2 Gb of Hi-C data using Juicer (Durand et al., 2016). In the final round, the 85.2 Gb of Illumina paired-end reads were aligned to correct errors using Pilon (Walker et al., 2014).

Two types of data were retrieved for the identification of homeologous subgenomes in the PPR1 genome: 1) containing BAC sequences (Ammiraju et al., 2010b), and 2) 278 genes generated from CC and EE species (Zou et al., 2008). First, we aligned these sequences to the PPR1 genome using NCBI BLAST with an E-value cutoff of 1e^-5. Next, the best hits were filtered based on the sequence length of 100 nucleotides and greater than 80% sequence identity. Then, the information of these aligned BAC sequences and genes were used to label the numbers of 24 chromosomes of PPR1 as C1-12 and D1-12. To obtain homeologous gene pairs between both subgenomes in the PPR1 genome, we first aligned proteins obtained from the C1 subgenome against the D1 gene set using NCBI BLAST version 2.2.24 (Camacho et al., 2009) with an E-value cutoff of 1e^-5. Finally, the best hits were filtered based on 50% amino acid length and a score of more than 50.

**Genome annotation**

**Gene prediction**

Three types of gene prediction methods were combined to annotate the PPR1 genome, including *ab initio*, assembled transcripts and protein homolog prediction. Four programs, including AUGUSTUS (Stanke et al., 2004), GENSCAN (Burge and Karlin, 1997), Glimmer-HMM (Majoros et al., 2004), and SNAP (Korf, 2004) were performed for *ab initio* prediction. RNA-seq reads from multiple tissues were mapped to the PPR1 genome using TopHat2 (Kim et al., 2013), and assembled into transcripts using Cufflinks (Trapnell et al., 2010) for gene structures and new sequences. Protein sequences from the *O. sativa* ssp. *indica* and *japonica* and uniprot database were aligned to the PPR1 genome using NCBI BLAST (default parameters except for an e-value < 10^-5) for homology-based prediction. Accurate spliced sites were identified using GeneWise (version 2.1.20 stable) (Birney and Durbin, 2000). All predicted gene structure evidence was integrated with EvidenceModeler (EVM) (Haas et al., 2008) to produce high-confidence gene models. Gene models were filtered with the parameters: codon length > 150 bp and transcripts supported by homology-base methods. Finally, the resulting gene sets were checked for completeness in comparison with the BUSCO gene set (Sepey et al., 2019). The predicted gene models with FPKM > 0.01 in any of RNA-seq samples were labeled as High-Confidence (HC) genes, and other gene models as Low-Confidence (LC) genes.

**Gene function annotation**

Gene functions were assigned according to the best match for each predicted gene in alignment to multiple databases including the NCBI non-redundant (nr) protein database, SWISS-PROT, and Clusters of Orthologous Groups (COG) using the NCBI BLAST program (E-value < 10^-5). Gene ontology (GO) analysis was performed with InterProScan V5.19-58.0 using the default setting (Jones et al., 2014).

**Annotation of repetitive sequences**

The assembled genome was searched for repeats with a combination of *de novo* and homolog based strategies. The initial repeat library was built using two *de novo* programs including RepeatScout (Price et al., 2005) and RECON (Bao and Eddy, 2002). The initial repeat library and the Repbase database for the homolog repeat annotation were classified into classes, subclasses, superfamilies and families by RepeatMasker (Tarailo-Graovac and Chen, 2009).

**Annotation of non-coding RNAs**

Ribosomal RNAs (rRNAs) were identified using RNAamrmer with default parameters (Lagesen et al., 2007). Reliable tRNA structures were detected using tRNAscan-SEM (version 1.23) (Lowe and Eddy, 1997). Non-coding RNAs containing miRNA, snRNA, and snoRNA features were annotated using INFERNAL with default parameters (Nawrocki and Eddy, 2013). Noncoding RNAs were classified into different families through searching against the Rfam database.

**Annotation of resistance genes, transcription factors and transcription regulators**

The resistance genes were annotated using RGAugury (Li et al., 2016) with default parameters. The transcription factors and transcription regulators were annotated using iTAK (version 18.12) (Zheng et al., 2016) with default parameters.

**LTR-RT identification and insertion time estimation**

Long terminal repeat retrotransposons (LTR-RTs) were scanned in the PPR1 genome using LTR-FINDER (Xu and Wang, 2007) and LTR-Harvest (Ellinghaus et al., 2008). The identification of complete LTR-RTs was obtained by adjusting the ends of LTR-pair terminal sequences using LTR-retriever (Ou and Jiang, 2018). LTR index of PPR1 is 22.11. LTR insertion times were estimated according to the formula \( T = \frac{d}{2\mu} \) (d, the nucleotide distance for each pair of LTRs; \( \mu \), the nucleotide substitution rate estimated as previously described). The orthologous LTR-RTs in *O. alta* and *Brachypodium distachyon* were identified to measure an appropriate LTR nucleotide substitution rate based on the divergence time (~52.0 MYA) of both species (Bennetzen et al., 2012). The mutation rate was estimated as \( 0.65 \times 10^{-8} \) substitutions per site per year.
Transcriptome data analysis
For RNA-seq analysis, 10 representative tissues were collected in the greenhouse, including young tissues of seedling (leaf, stem and root), mature plant tissues (leaf, stem, root, seed, panicle, sheath), and callus. Three biological repeats were made for each tissue except seed and panicle. Total RNA was extracted with the Quick RNA isolation Kit according to the manufacturer's instructions (Huayueyang, China). RNA libraries (300 bp) were constructed and sequenced using the BGI platform. Raw RNA-seq reads (PE100) were processed to trim terminal low quality bases and adaptor sequences using a custom pipeline. High-quality RNA-seq reads were aligned to the assembled genome using TopHat2 (Kim et al., 2013) with default parameters. Transcript expression levels, in fragments per kilobase per million reads mapped (FPKM) for each sample, were determined using Cufflinks (Trapnell et al., 2010). Differentially expressed homologous pairs on the corresponding chromosomes of the two subgenomes were determined by Student’s t test with Bonferroni correction using 28 samples of RNA-seq.

Gene family analysis
Ten species, including O. alta, O. sativa, B. distachyon, Aegilops tauschii, Triticum urartu, T. aestivum, T. dicoccoides, Setaria italica, Sorghum bicolor, and Zea mays were selected to construct gene families on the basis of protein homologs (sequence length > 30, e-value < 10^{-5}, and alignment length coverage > 50%) using the OrthoMCL (Li et al., 2003a). Single-copy families were used to build the phylogenetic tree. For each single-copy gene family, we aligned and clustered the protein sequences from the 10 species with CLUSTALW (Thompson et al., 1994) and transformed the alignments to the phylip format. We determined the best fitting model for the construction of phylogenetic trees using Prottest (version 3.4) (Darriba et al., 2011). Phylogenetic trees were built with PhyML (Guindon et al., 2010) with 1,000 bootstrap replicates. The divergence time of each node was estimated based on the JC69 model using the MCMCTree program in the PAML package (Yang, 2007) with default parameters except for RooAge (< 2.0). Two reported divergence times were used as a calibration, including ~26 MYA for Sorghum and Maize, and ~52 MYA for O. sativa and B. distachyon (Falush et al., 2003). Gain and loss of gene families for the ten species were predicted using DOLLOP (Farris, 1977) from the PHYLIP package (Felsenstein, 1989). Expansion and contraction of gene families were determined using CAFE (De Bie et al., 2006) with the lambda value of 0.184609. A phylogenetic tree from the above study was used as a probabilistic model to infer family expansions and contractions.

Population genetics analysis
Genomic DNA from 44 wild Oryza rice accessions was extracted from fresh young leaves using the CTAB method (Allen et al., 2006). All accessions were re-sequenced on the Illumina NovoSeq 5000 or BGISEG-500 platform, generating 150-bp or 100-bp paired-end reads. All sequence reads from the 44 accessions were aligned to the PPR1 reference genome using BWA (Li and Durbin, 2009) and SAMtools (Li et al., 2009) with the default parameters. Variations (SNP and INDEL) were detected with GATK (McKenna et al., 2010) with base quality ≥ 30. The identified variations were further categorized as variations in intergenic regions, UTRs, coding sequences and introns according to the PPR1 genome annotation using snpEff (Cingolani et al., 2012). To build a neighbor-joining tree, we screened a subset of 524,933 SNPs (MAF > 5% and missing data < 40%) at fourfold-degenerate sites in the 44 rice accessions from the entire SNP dataset. A phylogenetic tree was constructed using PHYLIP with 100 bootstrap replicates. Using the same dataset, we also investigated the population structure using STRUCTURE on the basis of allele frequencies (Falush et al., 2003). In addition, principal-component analysis (PCA) was generated using all SNPs across the genome (MAF > 10%, missing < 5%) (Patterson et al., 2006). Nucleotide diversity for each group was measured based on the degree of variability (Tajima, 1983). The level of genetic diversity was calculated using a 100-kb window with a step size of 10 kb in each group.

Collinear analysis
For each pairwise alignment, the coding sequences of predicted gene models were compared to each other using all-versus-all BLASTP (e-value < 10^{-5}) within the PPR1 genome. Syntenic blocks (with at least five genes per block) were identified by MCScan (Tang et al., 2008) with the default parameters. The collinear blocks represented pairs of genomic sequences containing at least five collinear genes. Whole-genome synteny relationships between the species were visualized using Circos (Krzywinski et al., 2009).

Homologous genes of diploid rice in O. alta
To identify homologous genes related to domestication in O. sativa in the O. alta genome, the genes in diploid rice were first aligned against the PPR1 gene set using BLASTP with an E-value cutoff of 1e^{-5}. Next, the best hits were filtered based on 50% amino acid length and a score of more than 50. The top two best-matched results were assigned as the homologs for the following studies in the domestication-related and agronomically important genes.

Transgenic system of O. alta
The genetic transformation system of O. alta was established as previously described (Zhang et al., 2016) with modifications, which are briefly described as following.

Callus induction and proliferation. (1) The mature dehusked seeds without pathogen spots were surface-sterilized with ethanol and bleach solution. (2) The sterilized seeds were transferred to the M1 calli induction medium (Zhang et al., 2016) with the modification of replacing the sucrose with maltose and incubated in dark at 28°C. (3) After 15 days, calli were formed from the scutella of mature
seeds. Embryogenic calli with a granular structure, smooth surface and creamy/yellow color were selected and subcultured in fresh M1 medium in the dark at 28°C for 15 days after removing the seedling and seed structures. The selection of embryogenic calli from non-embryogenic calli is essential. (4) The subculturing process can be repeated a few times to get more calli, but should be performed for less than a three-month period.

Transformation and selection. (1) Activation of Agrobacterium tumefaciens strain EHA 105 containing constructs of interest was performed as previously described (Zhang et al., 2016). (2) The Agrobacterium was then resuspended in cooled (10°C) fresh infection medium (supplemented with freshly prepared 200 μM acetosyringone and 0.1% Pluronic F68) and adjusted to an optical density with OD₆₀₀ 0.5 and placed on ice for 10-20 minutes. (3) For infection, the Agrobacterium cell suspension was poured into the flask immersing the calli, and the flask was then gently shaken at 40 rpm for 10 minutes. The calli were placed onto a stack of sterile filter paper in the dishes and co-cultivated in the dark at 22°C for 3 days. (4) After co-cultivation, the infected calli were transferred to freshly prepared solid M2 selection medium and were cultured in the dark at 28°C for 7 days. The calli were sub-cultured by changing the fresh selection medium every two weeks for three cycles. The newly formed hygromycin B resistant and actively growing microcalli were subcultured under the same conditions.

Regeneration and hardening of transgenic plants. (1) Embryogenic resistant calli were transferred to solid M5 regeneration medium and incubated at 28°C in the dark for 1 week, and then place in a 16-hour light/8-hour dark photoperiod for 3 weeks under cold fluorescent light with an intensity 50 - 100 μE m⁻² s⁻¹. Green spots in the calli and newly formed shoots could be seen in 2 and 3 weeks, respectively. (2) After 4 weeks, regenerated shoots were transferred to M6 rooting medium and incubated at 28°C until the shoots grew to 10 cm. (3) Plants were then hardened by opening the lid of the medium at 28°C for 5 days with watering. Finally, the plants were washed with warm water and transplanted to soil in the greenhouse or field.

CRISPR/Cas9 based genome editing of O. alta
Single-guiding RNAs (sgRNAs) were designed using the CRISPR-GE tool (http://skl.scau.edu.cn/) (Xie et al., 2017). The sgRNAs used in this study are listed in Table S6. The sgRNAs were cloned by enzyme digestion and ligation methods into the binary vector VK005-01 for single gene editing and A3A-PBE for base editing, or generated into sgRNAs expression cassettes using the overlapping PCR method and then cloned by the Golden Gate cloning method into pYLCRISPR/Cas9P_uH for multiplex gene editing (Ma et al., 2015a; Zong et al., 2018). The final binary vector was transformed into PPR1 by the established Agrobacterium-mediated transformation methods. For genotyping of the T₀ transgenic lines, leaves of T₀ transgenic lines were collected from each plant and genomic DNA was isolated using the CTAB method (Allen et al., 2006). PCR amplification was performed using primers flanking the sgRNA target sites (Table S6) and Sanger sequencing was used to identify mutations in target genes of T₀ plants.

The T₀ seedlings of OaGhd7-CC, OaGhd7-DD, OaDTH7-CC, and OaDTH7-DD edited plants and the wild-type PPR1 were refined in the greenhouse of Institute of Genetics and Developmental Biology in Beijing, and then transplanted to the experimental fields of Institute of Genetics and Developmental Biology in Beijing when the plants grown up to 20 cm, and the picture of all lines was taken 110 days after transplantation. The T₀ seedlings of other edited plants and wild-type PPR1 were refined in the greenhouse of Institute of Genetics and Developmental Biology in Beijing, and then transplanted to the experimental fields of Institute of Genetics and Developmental Biology in Hainan.

The grain length was measure by the Scan Marker i560 (MICROTEK). The cross section photos of the stems were taken by stereomicroscope (OLYMPUS, SZX16) and the diameter of the third internodes of primary tillers were measured by vernier caliper. To observe abscission cells, the histological analysis was performed based on methods in Yoon et al. (2014) with some modifications. Samples were fixed in formaldehyde–acetic acid–alcohol solution. The tissue-paraffin block was cut to a thickness of 8 μm with a microtome (Leica Microsystems, RM2145). The cut samples were attached to a coated slide with sterile water on 42°C hot-plate, rehydrated with 100% xylene 10 min twice for clearing wax, and washed in 50% xylene (xylene in ethanol) and ethanol series (100, 100, 95, 85, 70, 50 and 30%). The samples were stained with chlorogaulcin (saturated chlorogaulcin in 20% HCl) until sections turned red. The sections were covered with acidic glycerin and then observed under a microscope (Leica, DMR).

RNA extraction and quantitative real-time PCR analysis
The total RNA of stem base was extracted using a TRIzol Kit by three biological repeats (Invitrogen, Cat#15596018) according to the user’s manual. About 1.5 μg of each RNA sample was used for cDNA synthesis (Maxima H Minus cDNA Synthesis Master Mix, with dsDNase, Thermo Scientific, Cat#M1682). Quantitative real-time PCR was performed using primer pairs (see KEY RESOURCES TABLE) in the SsoFast EvaGreen Supermix (Bio-Rad, Cat#1725201) with the real-time PCR detection system (Bio-Rad, CFX96). PCR reactions were performed in quadruplicate for each sample, and expression levels were normalized to OaUBI for expression detection. The experiments were repeated independently three times.
QUANTIFICATION AND STATISTICAL ANALYSIS

The statistical details of experiments can be found in the figure legends, including the statistical tests used, exact value of n, what n represents, definition of center, and dispersion measures. Comparisons for two groups were calculated using the unpaired two-tailed Student’s t tests, and statistical analysis was performed using Microsoft Excel 2016. Results are represented as mean ± s.d.. P values < 0.01 were considered significant, with levels of significance as follows: **p < 0.01; ns, not significant. We observed normal distribution and no difference in variance between groups in individual comparisons. The selection of sample size was based on extensive experience. Further methods to determine whether the data met assumptions of the statistical approach were not relevant for these analyses.
Figure S1. Transformation and regeneration efficiencies of allotetraploid rice species, related to Figure 1

(A) Callus induction and regeneration of 28 allotetraploid rice accessions. Detail information of each ID was listed in Table S1. The green shoot ratio or white shoot ratio is labeled as the number of calli producing green shoot or white shoot divided by the number of individual callus transferred to regeneration medium. (B) Hygromycin resistant calli of PPR1. (C) Regenerated plants of PPR1. (D) Detection of the vector in transformed plants. (E) Transformation and regeneration efficiencies of PPR1 in four experiments. The transformation efficiencies are calculated as the number of hygromycin resistant calli divided by the number of Agrobacterium-infected calli. The regeneration efficiencies are calculated as the number of calli producing plants divided by the number of individual callus transferred to regeneration medium.
Figure S2. Genome estimation and assembly of O. alta, related to Table 1

(A) Distribution of 31-mer in clean sequence data of the O. alta genome. The x axis and y axis indicate the 31-mer frequency and number, respectively. (B) Reads information used in the de novo sequencing of O. alta genome. (C) Heatmap of the O. alta genome using Hi-C technology. The heatmap represents the normalized contact matrix. (D) Frequencies of transposable elements (TE) in the vicinity of genes. The plots are anchored around the transcription start sites (TSS) and end sites (TES) of protein coding genes.
Figure S3. Transcriptome data and annotation of O. alta, related to Table 1
(A) Statistics of transcriptome data. (B) Principal component analysis of transcriptome data. These tissues were apparently separated and three repeats for each tissue have a relatively concentrated distribution. YL, young leaf; ML, mature leaf; YS, young stem; YR, young root; MR, mature root; MS, mature shoot; SE, seed; PA, panicle; SH, sheath; CA, callus. (C) De novo transcripts mapped to the genome assembly. (D) Summary of the genes predicted by three algorithms. (E) Statistics of transcripts annotated by different databases.
Figure S4. Transcription factors and regulators in *O. alta*, related to Figure 2

(A) Gene number of transcription factors and regulators in *O. alta* and *O. sativa* ssp. *indica* (R498) and *japonica* (Nip) genomes. (B) Gene number of ten families of transcription factors and regulators in *O. alta* and *O. sativa* ssp. *indica* (R498) and *japonica* (Nip) genomes. (C) GRAS family in *O. alta* and their expression in different tissues. Different colors on phylogenic tree indicate different GRAS families. Red and blue gene ids indicate the genes on C and D subgenomes, respectively. YL, young leaf; ML, mature leaf; YS, young stem; YR, young root; MR, mature root; MS, mature shoot; SE, seed; PA, panicle; SH, sheath; CA, callus.
Figure S5. Chromosomal rearrangements in the O. alta genome, related to Figure 3
(A) Genome rearrangements between O. alta chromosome C₄ and C₇. Grey and red boxes represent the regions origin from the ancestral CC and DD genomes, respectively. Back lines represent the sequence depth of resequencing data from ten CC and six EE diploid rice accessions. Color lines between two chromosomes represent collinear blocks between O. alta chromosomes C₄ and C₇.
(B) Chromatin interactions at 150-kb resolution by Hi-C data. The black pentagrams show the borders of the large segmental translocation between O. alta chromosomes C₄ and C₇.
(C) Collinearity between genome sequences by BioNano maps. The upper horizontal green bars represent the genome sequences and the below blue bars represent the BioNano maps. The lines between green bars and blue bars represent the collinear restriction sites. Black pentagrams in (B) and (C) show the borders of the large segmental translocation between O. alta chromosomes C₄ and C₇.
Figure S6. Genome editing of *O. alta* homologs of *SD1* and *GS3*, related to Figure 5

(A) CRISPR/Cas9-induced mutations in *sd1*<sup>CR-2</sup>, *sd1*<sup>CR-3</sup>, *sd1*<sup>CR-4</sup>, *sd1*<sup>CR-5</sup>, and *sd1*<sup>CR-6</sup> plants. (B) CRISPR/Cas9-induced mutations in *gs3*<sup>CR-2</sup>, *gs3*<sup>CR-3</sup>, *gs3*<sup>CR-4</sup>, *gs3*<sup>CR-5</sup>, and *gs3*<sup>CR-6</sup> plants. (C) Comparison of the total length of lowest eight internodes among PPR1, *sd1*<sup>CR-2</sup>, *sd1*<sup>CR-3</sup>, *sd1*<sup>CR-4</sup>, *sd1*<sup>CR-5</sup>, and *sd1*<sup>CR-6</sup> plants. Scale bars, 30 cm. (D) Statistical analysis of the data in (C). Data are means ± s.d. (n = 10 tillers). Asterisks represent significant differences for plant heights determined by Student’s t test. **p < 0.01. (E) Comparison of grain lengths among PPR1, *gs3*<sup>CR-2</sup>, *gs3*<sup>CR-3</sup>, *gs3*<sup>CR-4</sup>, *gs3*<sup>CR-5</sup>, and *gs3*<sup>CR-6</sup> plants. Scale bars, 1 cm. (F) Statistical analysis of the data in (E). Data are means ± s.d. (n = 20 grains). Asterisks represent significant differences determined by Student’s t test. **p < 0.01.
Figure S7. Multiplex genome editing in *O. alta*, related to Figure 6

(A) Structure of the multiplex genome editing system with four sgRNAs targeting eight genes in PPR1. Pubi, the promoter of maize ubiquitin; NLS, nuclear localization sequence; Tnos, the terminator of nopaline synthase gene.

(B) Genotypes of 20 mutant lines transformed with the vector in (A).

(C) Genotypes of 36 mutant lines transformed with the vector in Figure 6(H).