Genetic control of seed shattering during African rice domestication

Shuwei Lv[1,7], Wenguang Wu[1,7], Muhua Wang[2,6,7], Rachel S. Meyer[3,7], Marie-Noelle Ndjiondjop[4], Lubin Tan[1], Haiying Zhou[1], Jianwei Zhang[2], Yongcai Fu[1], Hongwei Cai[1], Chuanqing Sun[5], Rod A. Wing[2] and Zuofeng Zhu[1]*

Domestication represents a unique opportunity to study the evolutionary process. The elimination of seed dispersal traits was a key step in the evolution of cereal crops under domestication. Here, we show that *ObSH3*, a YABBY transcription factor, is required for the development of the seed abscission layer. Moreover, selecting a genomic segment deletion containing *SH3* resulted in the loss of seed dispersal in populations of African cultivated rice (*Oryza glaberrima* Steud.). Functional characterization of *SH3* and *SH4* (another gene controlling seed shattering on chromosome 4) revealed that multiple genes can lead to a spectrum of non-shattering phenotypes, affecting other traits such as ease of threshing that may be important to tune across different agroecologies and postharvest practices. The molecular evolution analyses of *SH3* and *SH4* in a panel of 93 landraces provided unprecedented geographical detail of the domestication history of African rice, tracing multiple dispersals from a core heartland and introgression from local wild rice. The cloning of *ObSH3* not only provides new insights into a critical crop domestication process but also adds to the body of knowledge on the molecular mechanism of seed dispersal.

Crop domestication is a process of reshaping wild species to be adapted for cultivation and to meet human needs[1]. Several morphological characteristics, such as plant architecture, seed size and dispersal, are common changes during domestication[2–15]. Cereal crops, which are globally critical foods, present perhaps the most classical change: the elimination of the primary seed dispersal mechanism, known as shattering[2]. It is thought that wild progenitors of modern cereal crops shed their seeds on maturation to ensure effective reproduction, and cultivars retain seed on the plant to avoid yield loss and to improve efficient harvests. Recently, the genes controlling the seed dispersal of several cereal crops, such as Asian rice, wheat, sorghum and barley, have been characterized, promoting our understanding of the genetic mechanisms of the elimination of seed shattering[13–15]. In addition, investigating the evolutionary history of the genes controlling the loss of seed shattering provides valuable insight into the geography of artificial selection shaping domestication and the natural selection patterns on these traits in the wild. For example, a detailed analysis of the molecular evolution of the *Non-brittle rachis 1* (*BTR1*) and *Non-brittle rachis 2* (*BTR2*) genes identified the origin of cultivated barley[15].

African rice (*Oryza glaberrima* Steud.) was gradually domesticated from the African wild species (*Oryza barthii*), with a peak genetic bottleneck at ~3,000 years ago[16] that coincides in timing with an abundance of archaeological findings[17]. The crop is well adapted for cultivation in West Africa, and possesses traits for increased tolerance to biotic and abiotic stresses including high temperature, drought, soil acidity and weed competitiveness. In areas with the most adverse ecological conditions, *O. glaberrima* is favoured by farmers for its adaptability and resistance to multiple constraints[17–21]. Recent genomic studies of *O. glaberrima* and *O. barthii* pointed to one domestication event for African rice[22], and suggested the geographical routes of the spread[19]. Scholars have attributed this spread to the translocation of the Mande homeland[23] and migrations to the coast[23] that have largely been due to climate change[24]. However, the genetic mechanism and evolutionary history of major domestication genes of African rice, knowledge of which can assist in tracing dispersal routes after the onset of domestication, is largely unknown. The simple domestication history of *O. glaberrima* makes it a good model to study the evolution of domestication traits. As loss of seed dispersal is one of the most important domestication traits, unravelling the genetic mechanism and evolutionary history of genes controlling this trait in *O. glaberrima* may provide insight into the routes through which the crop diversified and adapted to new agroecologies and cultural systems.

Our previous study indicated that a single nucleotide polymorphism (SNP) in the *shattering 4* gene (*SH4*, an orthologue of *grain length 4* (*GL4*)) resulted in a premature stop codon and led to loss of seed shattering during African rice domestication[25]. However, this SNP mutation does not exist in some non-shattering African rice varieties, implying that there might be another gene or another mutation of *SH4* controlling the non-shattering trait. To identify the new gene or mutation responsible for the loss of seed shattering in cultivated African rice, we developed an F_2 segregating population derived from a cross between the African wild rice accession W1411, which exhibits a shattering phenotype, and a non-shattering cultivar of African cultivated rice, IRGC104165, containing the wild allele of *SH4* (Supplementary Fig. 1).

To distinguish precisely the differences in abscission layer anatomy between W1411 and IRGC104165, we observed longitudinal sections of spikelets using confocal microscopy. We found that the W1411 samples exhibited a complete abscission layer between the seed pedicel and the spikelet, which can be seen in a longitudinal
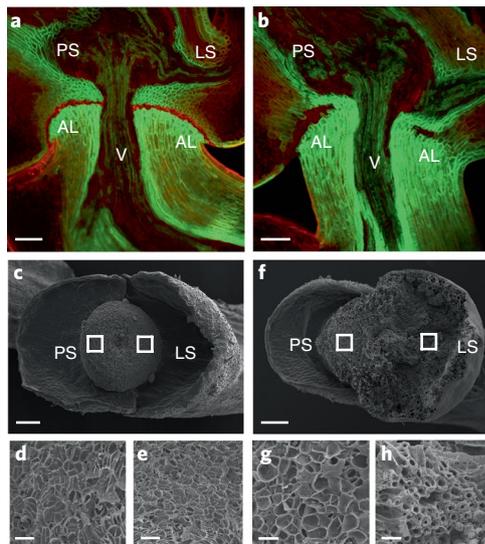
Fig. 1 | Comparison of seed shattering and floral abscission zone morphologies between W1411 (*O. barthii*) and IRGC104165 (*O. glaberrima*). a,b, Confocal microscopy images of longitudinal sections of the junction between the flower and the pedicel stained by acridine orange under fluorescence. Scale bars, 50 μm. The experiment was repeated five times independently, and similar results were obtained. **c,f,** SEM photographs of the junction after the seeds were detached. Scale bars, 100 μm. The white boxes highlight the lemma side (LS) and the palea side (PS) of the fractured surface. **d,e,g,h,** Close-up views of the areas corresponding to the white boxes in **c** (**d,e**) and **f** (**g,h**). In W1411, a smooth surface (**d,e**) is observed in both PS (**d**) and LS (**e**). In IRGC104165, a smooth palea side (PS) surface (**g**) and a rough lemma side (LS) surface (**h**) are observed. Scale bars, 50 μm. The experiment was repeated three times independently, and similar results were obtained. AL, abscission layer; v, vascular bundle.

section as continuous lines of abscission cells between the vascular bundle and the epidermis (Fig. 1a). Conversely, IRGC104165 samples had a wider and partially developed abscission layer (Fig. 1b). Further careful comparison of the abscission layers of W1411 and IRGC104165 spikelets showed that the former consist of mostly one layer of small, thin-walled cells, while the latter consist of three or four layers of cells. On the palea side, the abscission layer cells were partially developed, and a very irregularly developed abscission layer existed on the lemma side. The fracture surface of rachilla was investigated using scanning electron microscopy (SEM). We found that W1411 samples had a smooth fracture surface (Fig. 1c–e), whereas IRGC104165 samples had a smooth surface only at the peripheries in the transverse plane on the palea side (Fig. 1f–h). These results indicated that the loss of seed shattering in IRGC104165 resulted from the irregular development of the seed abscission layer.

A genetic linkage analysis of 168 F_2 individuals derived from the cross between W1411 and IRGC104165 suggested that seed shattering was controlled by a single gene lying on the long arm of chromosome 3. We designated this gene as *Oryza barthii seed shattering 3* (*ObSH3*) (Fig. 2a). Using a total of 2,650 recessive homozygote plants with the non-shattering phenotype from the F_2 population, we delimited *ObSH3* between the SNP29 and SNP31 markers (Fig. 2b). In this fine mapping region, the genomic sequence of IRGC104165 was 17-kb without a predicted open reading frame (ORF). By contrast, the genomic sequence of W1411 was 63-kb, with a 45.5-kb insertion compared with that of IRGC104165, which contained six predicted ORFs (ORF1–ORF6) (Fig. 2c; Supplementary Fig. 2). The quantitative rtPCR analyses showed that of the six ORFs, only ORF3 was expressed in the abscission layer region (Supplementary Fig. 3).
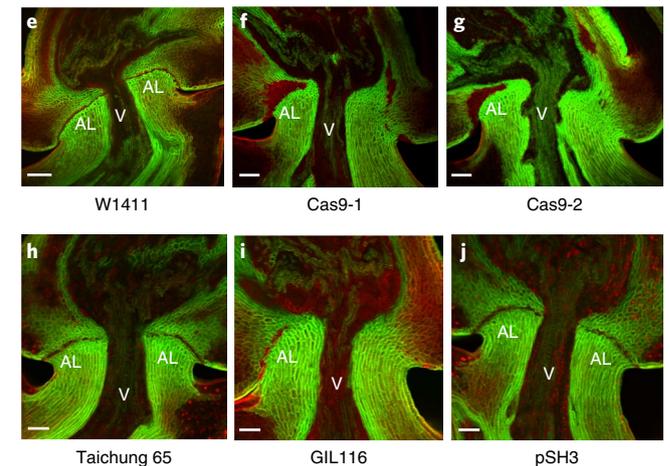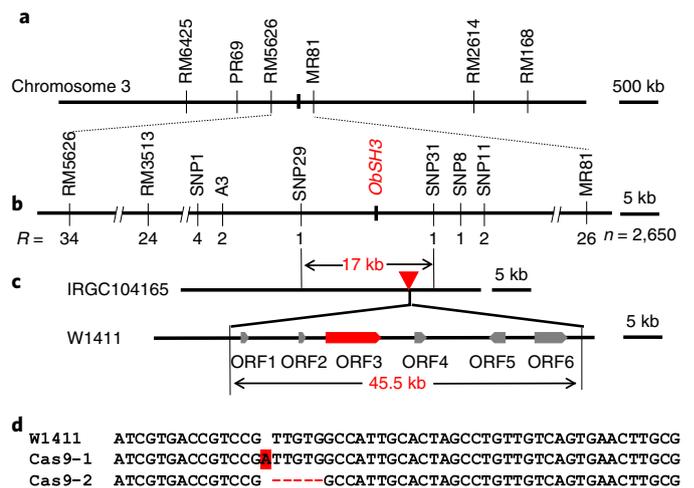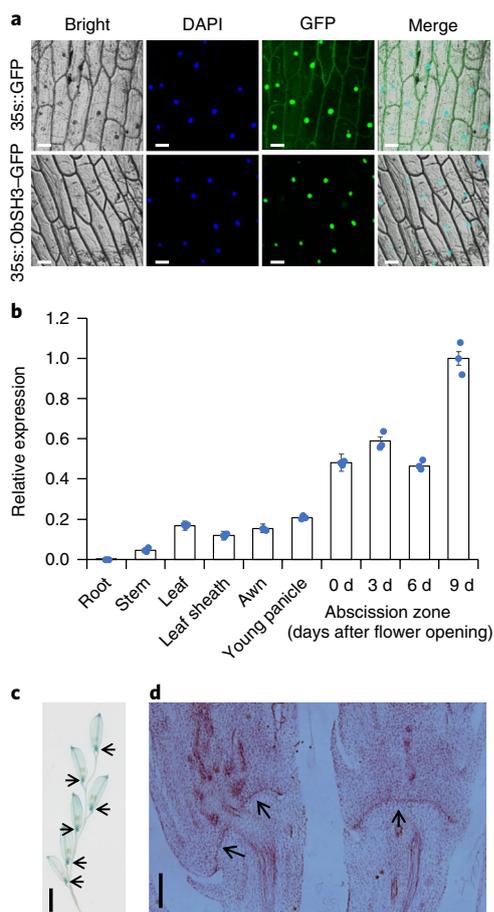

Fig. 2 | Map-based clone of *ObSH3*. a, The target gene (or genes) for seed shattering was mapped between RM5626 and MR81 on the long arm of chromosome 3 based on linkage analysis of 168 F_2 individuals. **b,** *ObSH3* was further narrowed down to a 17-kb region between the markers SNP29 and SNP31 using 2,650 recessive homozygote plants. **c,** Comparison of the predicted ORFs based on the genome sequences between mapping parents IRGC104165 and W1411. **d,** Second exon sequence of candidate genes in transgenic plants, which were edited using the CRISPR–Cas9 technique, and the control (W1411). **e–j,** Confocal microscopy images of longitude sections of W1411 (**e**), CRISPR–Cas9 knockout lines Cas9-1 and Cas9-2 (**f,g**), Taichung 65 (**h**), GIL116 (**i**), and the transgenic line of pSH3 (**j**). Scale bars, 50 μm. The experiment was repeated three times independently, and similar results were obtained.

A sequence analysis of 5′- and 3′-rapid amplification of cloned/complimentary DNA end (RACE) products indicated that the ORF3 cDNA in W1411 is 1,187-bp long, with an ORF of 561-bp, a 325-bp 5′ untranslated region (UTR) and a 301-bp 3′ UTR (Supplementary Fig. 4). ORF3 was predicted to encode a transcription factor gene belonging to the YABBY family, which plays important roles in the development of plant lateral organs such as leaves and floral organs[26–28]. Therefore, we focused on ORF3 as a candidate for *ObSH3*.

To confirm this hypothesis, we knocked out the ORF3 gene of W1411 using the CRISPR (clustered regularly interspaced short palindromic repeats)–Cas9 genome editing system. We selected a unique target site in the coding region of the ORF3 gene that did not have any similar and potentially off-target sites in the rice genome. In the T_0 generation, more than ten heterozygous transgenic plants were screened by sequencing analysis. The transgenic plants that

Fig. 3 | Expression and subcellular localization of *ObSH3*. a, The *35s::ObSH3–GFP* fusion gene and *35s::GFP* were expressed transiently in onion epidermal cells. Bright, bright-field image; DAPI, the same cells but showing the nuclei stained using DAPI (4,6-diamidino-2-phenylindole); GFP, the same cells but showing green fluorescence in the nuclei; Merge, the merged images. Scale bars, 100 μm. The experiment was repeated twice independently, and similar results were obtained. **b**, Quantitative rtPCR of *ObSH3* in W1411. Data are given as the mean ± s.d. ($n = 3$ biologically independent samples). **c**, GUS staining in young panicles of *ObSH3*_pro_*::GUS* transgenic plants. Arrows indicate abscission zones. Scale bar, 0.5 cm. The experiment was repeated three times independently, and similar results were obtained. **d**, In situ hybridization of *ObSH3* during stage sp8 (formation of ovule and pollen). Arrows indicate the abscission layer. Scale bar, 100 μm. The experiment was repeated twice independently, and similar results were obtained.

showed sequence variations in the target region were self-pollinated to generate T_1 generations and genotyped using the primers flanking the target region. Two knockout lines with homozygous mutations that included a 1-bp insertion (Cas9-1) and a 5-bp deletion (Cas9-2) in the target region were selected. These targeted mutations caused frameshifts (and therefore loss of function), and all exhibited loss of seed shattering as well as an incomplete abscission layer, similar to that of IRGC104165 plants (Fig. 2d–g; Supplementary Fig. 5). These results indicated that ORF3 is *ObSH3*, and is essential for abscission zone development below the grain.

We also generated a construct (pSH3) by placing a 12-kb genomic fragment from W1411, covering the ORF3 gene only, into the vector pCAMBIA1300. Owing to the recalcitrance of the African variety IRGC104165 to regenerate shoots from callus, we introduced the pSH3 construct into a chromosome segment substitute line GIL116. This construct includes a small region containing the *sh3* locus
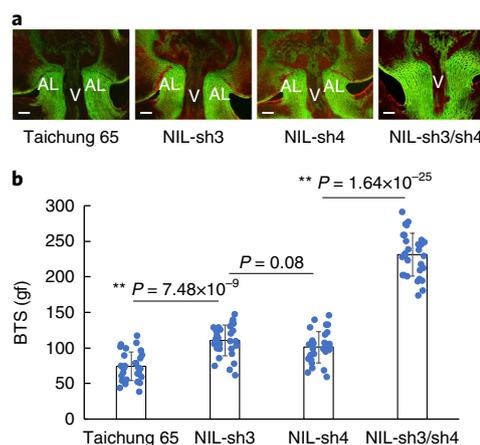

Fig. 4 | Additive effect of *SH3* and *SH4*. a, Confocal microscopy images of longitudinal sections showing abscission layer development of single (NIL-sh3, NIL-sh4) or double mutant (NIL-sh3/sh4) plants. Scale bars, 50 μm. The experiment was repeated twice independently, and similar results were obtained. **b**, Force needed to pull grains away from pedicels on the 35th day after flowering. BTS, breaking tensile strength; $n = 30$. **$P < 0.01$, two-tailed paired *t*-test.

from African rice under the Asian cultivated rice *Oryza sativa* var. Taichong 65 genetic background. GIL116 plants exhibited a partially developed abscission layer and harder seed shedding than that of Taichong 65 plants (Fig. 2h–i). All 15 of the independent transgenic lines showed abscission layers and easy shedding traits similar to that of Taichong 65, but without changes in agronomic traits such as grain length and grain weight (Fig. 2j; Supplementary Fig. 6). These results further confirmed that the ORF3 gene was *SH3*.

SH3 encodes 186 amino acids, and was predicted to contain a nuclear localization signal using the PSORT program (http://psort. hgc.jp/). To confirm this prediction, we developed a constitutively expressing construct by fusing the full-length SH3 to the carboxyl terminus of green fluorescent protein (GFP). The construct was transiently expressed in onion epidermal cells. The GFP signal was detected in the nucleus, which is in accordance with the prediction that *ObSH3* is a transcription factor (Fig. 3a).

To further examine the SH3 protein, we retrieved PSI-BLAST results using the full-length protein sequence of SH3 as a query against the non-redundant protein database (https://www.ebi. ac.uk/Tools/sss/psiblast/). A phylogenetic analysis of these putative homologues indicated that *SH3* was most closely related to genes found in other monocots, including maize (B4FY22), barley (M0YM09) and *Brachypodium* (I1GPY5). An amino acid sequence analysis also showed that the zinc finger domain and the YABBY domain of ObSH3 orthologues were highly conserved in monocots (Supplementary Fig. 7).

The expression pattern of the *ObSH3* gene was examined by real-time quantitative PCR, with the β-glucuronidase (*GUS*) reporter gene driven by the *ObSH3* promoter. We found by quantitative rtPCR that *ObSH3* was mainly expressed at the abscission layers of seeds, leaves and stems, but not in the roots (Fig. 3b). The GUS signal correlated well with the quantitative rtPCR results, which exhibited an intense signal in the pedicel abscission zone and the apiculus of the spikelet hull (Fig. 3c; Supplementary Fig. 8). RNA in situ hybridization results corroborated the expression of *ObSH3* at the abscission layer of the spikelet pedicel (Fig. 3d), which is consistent with its role of controlling abscission layer cellular development.

Our previous study showed that a SNP in *SH4* caused both loss of seed shattering and reduced grain sizes in *O. glaberrima*[25]. To compare the genetic effect of seed shattering between plants with
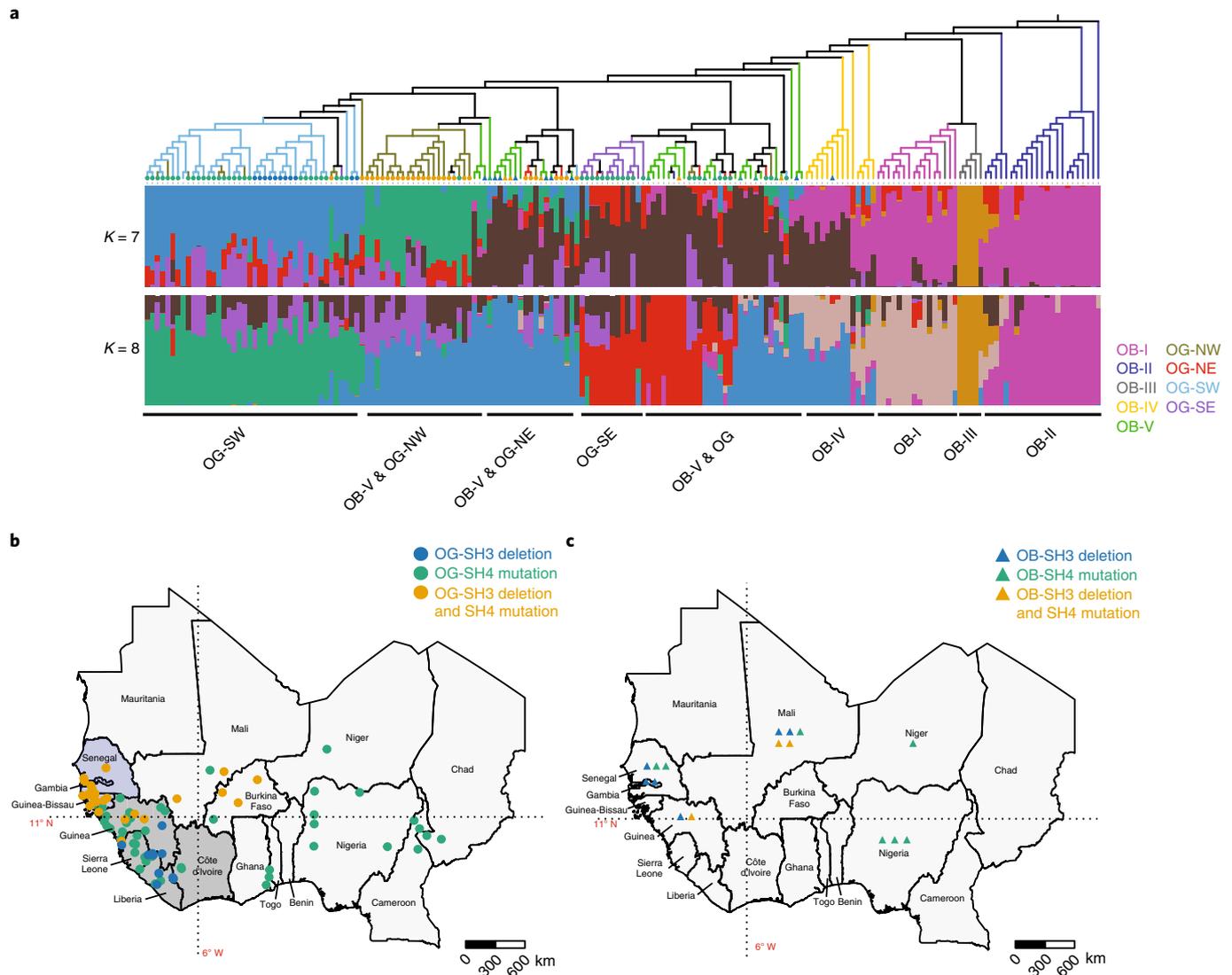
Fig. 5 | Evolution of the non-shattering trait in African rice. a, Genetic relationship of 93 *O. glaberrima* and 94 *O. barthii* accessions carrying different genotypes at *SH3* and *SH4* genes. The approximate maximum likelihood tree of *O. glaberrima* and *O. barthii* accessions was plotted on top of the results of the ADMIXTURE analysis with proposed ancestral populations *K* = 7 and *K* = 8. *O. glaberrima* and *O. barthii* accessions from different populations are indicated using different colours of phylogenetic tree branches and bars under the ADMIXTURE plots. The genotypes of *SH3* and *SH4* of *O. glaberrima* and *O. barthii* accessions are indicated using coloured solid circles and triangles at the tips of phylogenetic tree. The genotypes of *SH3* and *SH4* in *O. glaberrima* and *O. barthii* accessions with missing data at either *SH3* or *SH4* as well as *O. barthii* accessions that did not have mutations at *SH3* and *SH4* are not indicated. **b**, Biogeographical analysis of genotypic variations at *SH3* and *SH4* genes in *O. glaberrima*. The collection sites of *O. glaberrima* accessions are indicated as solid circles on the map. The 11° N latitude line divides the arid north from the tropical south, and the 6° W longitude line separates the coastal region from the inland region of West Africa. The countries in the southwestern forest region are highlighted in dark grey, and countries in the northwestern arid region are highlighted in light blue. **c**, Biogeographical analysis of mutations in the *SH3* and *SH4* genes in *O. barthii*. The countries of origin of *O. barthii* accessions carrying mutations in either *SH3* or *SH4* or both are indicated as triangles on the map. The latitude and longitude lines divide the region in the same manner as in **b**.

mutations in *sh3*, *sh4* or both (*sh3* and *sh4*), we developed three near-isogenic lines (NIL-sh3, NIL-sh4 and NIL-sh3/sh4) under the Asian cultivated rice *O. sativa* var. Taichong 65 genetic background. These near-isogenic lines contained a very small sh3 (NIL-sh3) or sh4 (NIL-sh4) region from African cultivated rice *O. glaberrima*. Longitudinal sections of spikelets at the anthesis stage were compared using confocal microscopy. We found that Taichong 65 samples showed a nearly complete abscission zone between the grain and the pedicel. Both NIL-sh3 and NIL-sh4 exhibited partially developed abscission layers. The double mutant NIL-sh3/sh4 samples had no abscission layer (Fig. 4a). Consistent with the severity

of abscission layer loss, the seeds of the NIL-sh3/sh4 double mutant was harder to shed than the seeds of NIL-sh3 and NIL-sh4 (Fig. 4b), indicating that combinations of these mutations can tune the threshability of rice.

To elucidate the evolutionary history of seed shattering in *O. glaberrima*, we first investigated the genetic relationship of 93 *O. glaberrima* and 94 *O. barthii* accessions using previously published resequencing data[16,22]. Our results were consistent with those of previous studies[16,22], in which *O. glaberrima* can be partitioned into the following four geographical quadrants with landraces sharing some genetic proximity: northwest (OG-NW) and southwest

(OG-SW) coastal populations, northeast (OG-NE) and southeast (OG-SE) inland populations. *O. barthii* can be grouped into four populations (OB-I, OB-II, OB-III and OB-V) as well as one admixture population (OB-IV) (Fig. 5a; Supplementary Figs. 9–11). Results from a previous study[16] demonstrated that after domestication, *O. glaberrima* diversified into east and west populations, and those later diversified into north and south populations. Our phylogenetic analysis of 93 *O. glaberrima* and 94 *O. barthii* accessions matched that result and supported the findings of another study[22] that demonstrated that the OB-V population shared most ancestry with *O. glaberrima*. While details of the domestication trajectory of African rice are beyond the scope of this work, we point out that the OG-NE accessions, although less represented in number, are spread across several clades in the phylogeny, and share more alleles with *O. barthii* according to the ADMIXTURE plot for $K = 7$ (where K is the number of populations assumed for analysis). This result is consistent with the hypothesis that *O. glaberrima* was domesticated in the middle Niger River delta of Mali and then spread across West Africa[23,29,30].

A spatial analysis of the distribution of *SH3* and *SH4* or combined genotypes showed that most accessions with both mutations occur in arid regions north of 11° N (Fig. 5b). However, the addition of the phylogeny and ADMIXTURE analysis (Fig. 5a) demonstrates that this double mutation form evolved twice: once in the NE and once in the NW. Principal component analysis (PCA) confirmed the isolation of NE double mutation accessions (Supplementary Fig. 12). Both the NE and NW innovations could have arisen from exploiting standing variation during artificial selection for stronger resistance to shattering. For *O. glaberrima* plants that carry either of the mutations but not both, there is widespread occurrence of the SNP mutation in *SH4*, whereas there is geographical and phylogenetic restriction of the *SH3* deletion. This result suggests that the *SH3* deletion is a derived form that is only maintained in certain agroecologies, possibly unpopular elsewhere because of the additional phenotype of smaller grain size that reduces yield[25]. That is, *O. barthii* carrying the *SH4* mutation are far more widespread (Fig. 5c). The OG-SE population also has the *SH4* mutation only, as does the adjacent OB-V accessions in the phylogeny (Fig. 5a).

Selection on pre-existing standing genetic variations has been shown to contribute to the adaptation of several organisms in nature[31], and has been well documented in other domesticated plants[32,33].

The prevalence of either or both non-shattering genotypes in the wild progenitor of African rice, *O. barthii*, is of great significance regarding how domestication is generally characterized. It may be that several of the classical traits that became fixed over time, and that are part of the "domestication syndrome"[2,34], are in fact also under natural selection in the same direction[35]. Only in recent years have large numbers of wild relatives of crops been sequenced and evaluated for domestication-associated mutations. The loss of the seed shattering trait is a textbook example of a change that is expected to have dire consequences on fitness in the wild. The ecological significance of non-shattering phenotypes in the wild merits future investigation, especially as diversity panels of wild relatives of crops are being assembled for future crop improvement[36,37]. Likewise, the agroecological and cultural significance of having these different non-shattering genotypes merits future investigation, especially as rice breeding programmes in Africa must consider the adoptability of new varieties into existing farming regimens. What advantage does complete removal of the grain abscission zone have for farmers in arid climates? Harvest lengths and storage time before threshing may be location-specific practices set to optimize thresh ability for different varieties. A better understanding of the ecological context of these genotypes, spanning pre-harvest and postharvest abiotic and biotic risks, presents opportunities to explain the process of co-evolution between crop species and people.

Methods

Plant materials and growth conditions. For the cloning of *ObSH3*, W1411 and IRGC104165 plants were used. W1411, an African wild rice accession (*O. barthii*), was collected from Sierra Leone. IRGC104165, a cultivar of African rice (*O. glaberrima*), was collected from Guinea. The F_2 segregating population was derived from the cross between W1411 and IRGC104165. The other cultivars and wild-rice accessions used in this study are listed in Supplementary Table 1. All plants were grown in field conditions in Beijing or Sanya, Hainan province, China.

Evaluation of shattering. The degree of shattering was measured after each panicle was shaken gently by hand. Plants with >80% of the grains removed were classified as shattering, whereas those with few grains removed were classified as non-shattering. The panicles of plants were harvested 35 days after heading and were kept at room temperature. Each panicle was attached vertically upside down to a digital force gauge (FGP-1; Nidec-Shimpo), and each grain was pulled down using forceps. The maximum tensile strength measured at the moment when the grain detached from the pedicel was recorded in gram-force (gf) units. For each plant, a total of 50 grains from three panicles were measured.

Primers. The primers used in this study are listed in Supplementary Table 2.

Histological analysis and SEM. Approximately ten spikelet samples were gathered at the flowering stage. A longitudinal section was made through the junction between the flower and pedicel by hand cutting, and the sections were stained with acridine orange. Sections were observed using an Olympus FV1000 laser scanning microscope. A 488 nm and a 543 nm laser lines were used. For SEM, the pedicel junctions after detachment of mature seeds were fixed in 2.5% glutaraldehyde solution, then gold plated, and observed using a Hitachi S-2460 scanning electron microscope.

Fine mapping of *ObSH3*. An F_2 population was bred from the cross between W1411 and IRGC104165. DNA was extracted from fresh leaves according to the cetyl-trimethyl-ammonium bromide method[38]. The genomic location of the *ObSH3* locus was defined by two molecular markers SNP29 and SNP31. Details of the markers used for fine mapping are given in the Supplementary Table 2.

Preparation of constructs for gene editing and *ObSH3* promoter:*GUS* fusion. Constructs for CRISPR–Cas9-based genome editing of *ObSH3* (the target sites are shown in Supplementary Fig. 4) were performed as previously described[39]. Rice transformation was performed using the *Agrobacterium*-mediated method. To construct the *ObSH3* promoter:*GUS* fusion plasmid, an ~2-kb DNA fragment comprising the promoter sequence of *ObSH3* in W1411 was amplified and cloned into the pCAMBIA1301-GUS-nos vector. The *SH3* coding sequence together with the 2,208-bp upstream and 696-bp downstream flanking region was amplified using the primers pSH3-2F and pSH3-2R from W1411 by PCR and recombined with the pCAMBIA1300 vector digested by *Kpn* I and *Sal* I to generate the pSH3 vector. Transgenic plants were generated by *Agrobacterium*-mediated transformation and then transformed into ZH17. Relevant PCR primer sequences are given in Supplementary Table 2.

5′- and 3′-RACE and quantiative real-time PCR. Total RNA was extracted using TRIzol reagent (Qiagen). We conducted 5′- and 3′-RACE using a SMARTer RACE 5′/3′ kit (TaKaRa) following the manufacturer's instructions. The product of first-strand cDNA was used as the template for the PCR. Real-time (rt)PCR was performed using an ABI Prism 7900 Sequence Detection System (Applied Biosystems). Quantitative rtPCR was carried out using SYBR Green Master Mix (Bio-Rad). Three replicates were performed. Rice *ACTIN1* was used as the internal control. Primers used for quantitative real-time PCR are listed in Supplementary Table 2.

Subcellular localization of *ObSH3*. The coding sequences of *ObSH3* were amplified from W1411 to generate the *35S::ObSH3–GFP* vector. We bombarded the resulting plasmid into onion epidermal cells using a helium biolistic device (Bio-Rad PDS-1000). The bombarded tissues were examined using a confocal laser scanning microscope (Olympus FV1000).

RNA in situ hybridization. Young panicles from W1411 were fixed in formaldehyde–acetic acid–ethanol fixation solution, subjected to a series of dehydration and infiltration, and embedded in paraffin. The tissues were sliced into 8–10-μm sections with a microtome (Leica RM2265). A 258-bp gene-specific region of *ObSH3* cDNA was amplified by PCR to generate sense and antisense RNA probes. Digoxigenin-labelled RNA probes were prepared using a DIG Northern Starter Kit (catalogue no. 2039672; Roche) according to the manufacturer's instructions. Primers used for the probes are listed in Supplementary Table 2.

Identification of *SH3* genotype in *O. glaberrima* and *O. barthii* accessions. To determine the genotype of *SH3* deletion in *O. glaberrima* and *O. barthii* accessions, the sequence of the bacteria artificial chromosome (BAC)-spanning *SH3* deletion (GenBank accession no. KF284072) was downloaded from the NCBI GenBank

database, as the genomic region containing the *SH3* gene was missing in the *O. glaberrima* CG14 genome assembly[22]. Raw sequencing reads of 93 *O. glaberrima* and 94 *O. barthii* accessions were aligned onto the BAC sequence using Burrows–Wheeler Aligner (BWA)[40] v.0.7.10. The presence or absence of the *SH3* deletion in *O.glaberrima* and *O. barthii* accessions was determined using the Eukaryotic Pan-genome Analysis Toolkit (EUPAN)[41] v.0.43 and manual curation using Integrative Genomics Viewer[42] (IGV) v.2.4.

Evolutionary analysis of *SH3* and *SH4*. Previously published resequencing reads of 93 *O. glaberrima*[16] and 94 *O. barthii*[22] accessions were downloaded from the NCBI Short Read Archive (SRA) database. Raw sequencing reads were aligned onto the *O. glaberrima* CG14 genome assembly[22] using BWA[40] v.0.7.10. PCR duplicates in the aligned reads were masked using the MarkDuplicate function of Picard Tools v.1.128 (https://broadinstitute.github.io/picard/). SNP calling and filtering were performed using Genome Analysis Toolkit (GATK)[43] v.3.4. In total, 6,640,731 SNPs were identified and used in the evolutionary analysis.

The phylogenetic tree of 93 *O. glaberrima* and 94 *O. barthii* accessions was inferred using SNPs from the 12 assembled chromosomes. To eliminate the effect of rare genetic variants, only SNPs with minor allele frequency (MAF) values greater than 0.05 were retained for the analysis. In total, 2,216,796 SNPs were used for the phylogenetic reconstruction. Due to computational limitations, the approximate maximum likelihood tree of *O. glaberrima* and *O. barthii* accessions was constructed using FastTree[44] v.2.1.8 and the GTR+CAT approximation model with 20 rate categories. The phylogenetic tree of *O. glaberrima* and *O. barthii* was plotted and annotated using the interactive tree of life (iTOL) online tool v.3[45].

The population structure of 93 *O. glaberrima* and 94 *O. barthii* accessions was inferred using ADMIXTURE[46] v.1.23. SNPs from unanchored scaffolds and contigs of *O. glaberrima* CG14 assembly were removed for the analysis. In addition, only SNPs with MAF values greater than 0.05 were retained for the analysis to eliminate the effect of rare genetic variants. SNP pruning was performed for the SNP dataset using plink[47] v.1.90 with the parameter "–indep-pairwise 400 50 0.3". In total, 193,454 SNPs were inputted into the ADMIXTURE program. The ancestry of each population was inferred from $K=2$ to $K=8$ (Supplementary Table 3). The result of the ADMIXTURE analysis was plotted using a custom R script.

PCA of 93 *O. glaberrima* and 94 *O. barthii* accessions was performed using the smartpca program implemented in the EIGENSOFT package[48] v.6.0.1. All the SNPs in the *O. glaberrima* and *O. barthii* SNP dataset (6,640,731) were inputted into the smartpca program. After filtering out SNP positions that were missing substantial numbers of genotype calls, smartpca used 2,161,274 SNPs to perform PCA analysis. PCA was performed for *O. glaberrima*, and *O. barthii* individuals were projected onto the principal component space of *O. glaberrima* to better elucidate the relationship of these two species (Supplementary Table 4). The results of the PCA were plotted such that *O. glaberrima* or *O. barthii* accessions with different genotypes of *SH3* and *SH4* genes were indicated by different colours using a custom R script.

The global positioning system coordinates of 93 *O. glaberrima* accessions were downloaded from a previous study[16]. Two *O. glaberrima* accessions (IRGC103993 and IRGC104573) with missing calls at the causative SNP of *SH4*[25] were excluded from the biogeographical analysis. The countries of origin of 94 *O. barthii* accessions were downloaded from a previous study[22]. The map of the biogeographical analysis was drawn using a custom R script.

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

Code availability. The R script that supports the findings of this study is available in the Dyrad Digital Repository (https://doi.org/10.5061/dyrad.qh5r649).

Data availability. The data that support the findings of this study are available in the Dyrad Digital Repository (https://doi.org/10.5061/dyrad.qh5r649). The gene sequence of *ObSH3* has been deposited in GenBank with the following accession code: MH159201.

References

1. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
2. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
3. Tan, L. et al. Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**, 1360–1364 (2008).
4. Jin, J. et al. Genetic control of rice plant architecture under domestication. *Nat. Genet.* **40**, 1365–1369 (2008).
5. Mao, H. et al. Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. *Proc. Natl Acad. Sci. USA* **107**, 19579–19584 (2010).
6. Li, Y. et al. Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. *Nat. Genet.* **43**, 1266–1269 (2011).
7. Che, R. et al. Control of grain size and rice yield by *GL2*-mediated brassinosteroid responses. *Nat. Plants* **2**, 15195 (2015).
8. Duan, P. et al. Regulation of *OsGRF4* by OsmiR396 controls grain size and yield in rice. *Nat. Plants* **2**, 15203 (2015).
9. Wang, Y. et al. Copy number variation at the *GL7* locus contributes to grain size diversity in rice. *Nat. Genet.* **47**, 944–948 (2015).
10. Si, L. et al. *OsSPL13* controls grain size in cultivated rice. *Nat. Genet.* **48**, 447–456 (2016).
11. Simons, K. J. et al. Molecular characterization of the major wheat domestication gene *Q. Genetics* **172**, 547–555 (2006).
12. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
13. Konishi, S. et al. An SNP caused loss of seed shattering during rice domestication. *Science* **312**, 1392–1396 (2006).
14. Lin, Z. et al. Parallel domestication of the *Shattering1* genes in cereals. *Nat. Genet.* **44**, 720–724 (2012).
15. Pourkheirandish, M. et al. Evolution of the grain dispersal system in barley. *Cell* **162**, 527–539 (2015).
16. Meyer, R. S. et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* **48**, 1083–1088 (2016).
17. Agnoun, Y. et al. The African rice *Oryza glaberrima* Steud: knowledge distribution and prospects. *Int. J. Biol.* **4**, 158–180 (2012).
18. Linares, O. F. African rice (*Oryza glaberrima*): history and future potential. *Proc. Natl Acad. Sci. USA* **99**, 16360–16365 (2002).
19. Rhodes, E. R., Jalloh, A. & Diouf, A. *Review of Research and Policy for Climate Change Adaptation in the Agriculture Sector of West Africa* (AfricaInteract, 2014).
20. Jones, M. P., Dingkuhn, M., Aluko, G. K. & Semon, M. Interspecific *Oryza sativa* L. X *O. glaberrima* Steud. progenies in upland rice improvement. *Euphytica* **94**, 237–246 (1997).
21. Li, X. M. et al. Natural alleles of a proteasome alpha2 subunit gene contribute to thermotolerance and adaptation of African rice. *Nat. Genet.* **47**, 827–833 (2015).
22. Wang, M. et al. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat. Genet.* **46**, 982–988 (2014).
23. Carney, J. A. *Black Rice: the African Origins of Rice Cultivation in the Americas* (Harvard Univ. Press, Cambridge, MA, 2001).
24. Vydrin, V. On the problem of the Proto-Mande homeland. *J. Lang. Relat.* **1**, 107–142 (2009).
25. Wu, W. et al. A single-nucleotide polymorphism causes smaller grain size and loss of seed shattering during African rice domestication. *Nat. Plants* **3**, 17064 (2017).
26. Cong, B., Barrero, L. S. & Tanksley, S. D. Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nat. Genet.* **40**, 800–804 (2008).
27. Siegfried, K. R. et al. Members of the *YABBY* gene family specify abaxial cell fate in Arabidopsis. *Development* **126**, 4117–4128 (1999).
28. Yamaguchi, T. et al. The *YABBY* gene *DROOPING LEAF* regulates carpel specification and midrib development in *Oryza sativa*. *Plant Cell* **16**, 500–509 (2004).
29. Portères, R. in *Papers in African Prehistory* (eds Fage, J. D. & Oliver, R. A.) 43–58 (Cambridge Univ. Press, Cambridge, 1970).
30. Portères, R. in *Origins of African Plant Domestication* (eds Harlan, J. R., De Wet, J. M. & Stemler, A. B.) 409–452 (De Gruyter Mouton, Berlin, 1976).
31. Barrett, R. D. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
32. Stetter, M. G., Gates, D. J., Mei, W. B. & Ross-Ibarra, J. How to make a domesticate. *Curr. Biol.* **27**, R896–R900 (2017).
33. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* **43**, 1160–1163 (2011).
34. Hammer, K. Das Domestikationssyndrom. *Kulturpflanze* **32**, 11–34 (1984).
35. Mercuri, A. M., Fornaciari, R., Gallinaro, M., Vanin, S. & di Lernia, S. Plant behaviour from human imprints and the cultivation of wild cereals in Holocene Sahara. *Nat. Plants* **4**, 71–81 (2018).
36. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
37. Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
38. Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325 (1980).
39. Ma, X. et al. A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Mol. Plant* **8**, 1274–1284 (2015).
40. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

41. Hu, Z. et al. EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics* **33**, 2408–2409 (2017).
42. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
43. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
44. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 — approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
45. Letunic, I. & Bork, P. Interactive tree of life (iTOL)v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
46. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
47. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
48. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

Author contributions

Z.Z. designed and supervised this study. S.L. conducted the map-based cloning, genetic transformation and gene expression analyses. S.L., W.W. and H.Z. conducted the histological analyses of the seed abscission layers. M.W. performed the evolutionary analysis and R.S.M assisted in analysing the results. M.-N.N., L.T., H.C., Y.F., J.Z. and C.S. conducted the collection of rice germplasm and phenotypic data. Z.Z., R.S.M. M.W. and R.A.W. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0164-3.

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

Correspondence and requests for materials should be addressed to Z.Z.

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

nature research

Corresponding author(s):   Zuofeng Zhu

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	The data of phylogenetic analysis were performed by MEGA5.0. Raw sequencing reads were aligned to the O. glaberrima CG14 genome assembly using the Burrows-Wheeler Aligner (BWA) V0.7.10. PCR duplicates in the aligned reads were masked using MarkDuplicate function of Picard Tools V1.128. The presence/absence of SH3 deletion was determined using Eukaryotic Pan-genome Analysis Toolkit (EUPAN) V0.43 and manual curation using Integrative Genomics Viewer (IGV) V2.4. Approximately-maximum-likelihood tree of O. glaberrima and O. barthii accessions was constructed with FastTree V2.1.8 using the GTR+CAT approximation model The phylogenetic tree of O. glaberrima and O. barthii was plotted and annotated using interactive tree of life (iTOL) online tool V3. The population structure of 93 O. glaberrima and 94 O. barthii accessions was inferred using ADMIXTURE V1.23. SNP pruning was performed for the SNP dataset using plink47 V1.90 with parameter "--indep-pairwise 400 50 0.3". The result of ADMIXTURE analysis was plotted using a custom R script. SNP calling and filtering were performed using Genome Analysis Toolkit (GATK) V3.4 Principal Component Analysis was performed using the smartpca program implemented in EIGENSOFT package V6.0.1. The map of biogeographic analysis was drawn using custom R script.

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

Data

Policy information about <u>availability of data</u>

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

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

> The data that support the findings of this study are available from the corresponding author upon request.

Field-specific reporting

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

☒ Life sciences   ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see <u>nature.com/authors/policies/ReportingSummary-flat.pdf</u>

Life sciences

Study design

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

Sample size	Sample sizes are provided in the manuscript. Size of mapping population was extremely large in order to pinpoint candidate gene, which doesn't need statistical method to determine the size and only recombinants between markers and the target gene were useful for fine mapping . n=3 biologically independent samples in qRT-PCR experiment . n=30 grains of three panicles were measured for force required to pull grains away from pedicels.
Data exclusions	No data were excluded from the analysis.
Replication	Every experiment was repeated for at least two times, and similar results were obtained.
Randomization	The samples were randomly selected for qRT-PCR and phenotyping.
Blinding	We did not apply any blinding in phenotyping.

Materials & experimental systems

Policy information about <u>availability of materials</u>

n/a	Involved in the study
☒ ☐	Unique materials
☒ ☐	Antibodies
☒ ☐	Eukaryotic cell lines
☒ ☐	Research animals
☒ ☐	Human research participants

Method-specific reporting

n/a	Involved in the study
☒ ☐	ChIP-seq
☒ ☐	Flow cytometry
☒ ☐	Magnetic resonance imaging