The Plant Journal (2015) doi: 10.1111/tpj.12912

A host plant genome (*Zizania latifolia*) after a century-long endophyte infection

Longbiao Guo[1,†], Jie Qiu[2,†], Zujing Han[3,†], Zihong Ye[4,†], Chao Chen[3,†], Chuanjun Liu[3], Xiufang Xin[5], Chu-Yu Ye[2], Ying-Ying Wang[2], Hongqing Xie[3], Yu Wang[2], Jiandong Bao[2], She Tang[2], Jie Xu[1], Yijie Gui[2], Fei Fu[2], Weidi Wang[2], Xingchen Zhang[2], Qianhua Zhu[3], Xuanmin Guang[3], Chongzhi Wang[3], Haifeng Cui[4], Daguang Cai[6], Song Ge[7], Gerald A. Tuskan[8], Xiaohan Yang[8], Qian Qian[1], Sheng Yang He[5], Jun Wang[3,*], Xue-Ping Zhou[9,*] and Longjiang Fan[2,*]

[1]*State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310006, China,*

[2]*Department of Agronomy & Zhejiang Key Laboratory of Crop Germplasm Resources, Zhejiang University, Hangzhou 310058, China,*

[3]*BGI-Shenzhen, Shenzhen 518083, China,*

[4]*College of Life Science, China Jiliang University, Hangzhou 310018, China,*

[5]*Howard Hughes Medical Institute, Department of Energy Plant Research Laboratory, and Department of Plant Biology, Michigan State University, East Lansing, MI 48864, USA,*

[6]*Department of Molecular Phytopathology, Christian-Albrechts-University of Kiel, D-24118, Kiel, Germany,*

[7]*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China,*

[8]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA, and*

[9]*State Key Laboratory of Rice Biology, Zhejiang University, Hangzhou 310058, China*

Received 24 January 2015; revised 26 May 2015; accepted 8 June 2015.
*For correspondence (e-mails: fanlj@zju.edu.cn, zzhou@zju.edu.cn or wangj@genomics.org.cn).
†These authors contributed equally to this work.

SUMMARY

Despite the importance of host–microbe interactions in natural ecosystems, agriculture and medicine, the impact of long-term (especially decades or longer) microbial colonization on the dynamics of host genomes is not well understood. The vegetable crop 'Jiaobai' with enlarged edible stems was domesticated from wild *Zizania latifolia* (Oryzeae) approximately 2000 years ago as a result of persistent infection by a fungal endophyte, *Ustilago esculenta.* Asexual propagation via infected rhizomes is the only means of Jiaobai production, and the *Z. latifolia*–endophyte complex has been maintained continuously for two centuries. Here, genomic analysis revealed that cultivated *Z. latifolia* has a significantly smaller repertoire of immune receptors compared with wild *Z. latifolia*. There are widespread gene losses/mutations and expression changes in the plant–pathogen interaction pathway in Jiaobai. These results show that continuous long-standing endophyte association can have a major effect on the evolution of the structural and transcriptomic components of the host genome.

Keywords: host–microbe interaction, genome, resistance gene analogs, *Zizania*, Jiaobai.

INTRODUCTION

The genus *Zizania* belongs to the rice tribe (Oryzeae) of the grass family Poaceae (Xu *et al.*, 2010). Phylogenetically, the genus occupies an important clade within Poaceae. Along with the genus *Leersia*, *Zizania* is the closest genus to *Oryza* (Kellogg, 2009). Two members of the genus *Zizania* have been domesticated: the annual *Zizania palustris*, which is native to North America and is called 'wild rice' (Hayes *et al.*, 1989), and the perennial *Zizania latifolia*, which is native to Asia and called 'Jiaobai' (Guo *et al.*, 2007; Xu *et al.*, 2008; Feldbrügge *et al.*, 2013). *Zizania latifolia* was domesticated as a vegetable crop approximately 2000 years ago (as noted in the Chinese first dictionary book 'Erya' in the Qin Dynasty, 207–221 BC). Because of the nutritional and economic importance of Jiaobai (Chan and Thrower, 1980), it is now widely cultivated in China (Figure S1) and other Asian countries.

The domestication of *Z. latifolia* into Jiaobai was made possible by persistent infection by a fungal endophyte, *Ustilago esculenta*, resulting in enlarged edible stems and the loss of flowering (Yu, 1962; Chan and Thrower, 1980). The endophyte-induced loss of flowering forced Jiaobai to be produced through asexual propagation for approximately 2000 years, and asexual propagation via infected rhizomes is the only means of Jiaobai production. As such, Jiaobai represents an excellent natural system in which to gain molecular insights into the impact of a defined long-term microbial infection on host genome dynamics.

Ustilago esculenta is a biotrophic basidiomycete fungus that belongs to the *Ustilago* genus of the Ustilaginaceae family. The genus includes smut fungi, such as *Ustilago maydis* and *Ustilago hordei*, which cause severe losses in *Zea mays* (maize) and *Hordeum vulgare* (barley), respectively. *Ustilago esculenta* is distinct from other members of the genus in several characteristics because of its host plants (Chung and Tzeng, 2004; Guo *et al.*, 2007; Xu *et al.*, 2008; Zhang *et al.*, 2012b): (i) to date, *Z. latifolia* is the only known host for *U. esculenta*, whereas others can attack multiple plant species; (ii) *U. esculenta* completes its entire life cycle within the host plant, whereas other species spend their life cycles in diverse ways; and (iii) *U. esculenta* is not a classical symbiotic endophyte because it always tends to escape from the host plant unless continuous artificial selection is maintained. The continuous maintenance of the *U. esculenta*–*Z. latifolia* interaction system therefore depends on long-term artificial selection pressure. In other systems, pathogens themselves have not been direct targets of artificial selection for maintenance after the arrival of fungus within the host. Moreover, until now there have been no reports of using any life stage of *U. esculenta* as an inoculum source to infest its host plant *Z. latifolia*.

In nature, many important host–microbe interactions (e.g. host–microbiome interactions, tree–symbiont interactions and chronic infections in humans) are long term, often lasting decades or longer. Despite their importance in natural ecosystems, agriculture and medicine, there are few studies on long-term (especially decades or longer) host–microbe interactions, and the impact of long-term microbial colonization on the dynamics of host genomes is not well understood. To understand the impact of the long-standing interaction of *Ustilago* with its host genome, here we sequenced and analyzed the genomes of wild and cultivated *Zizania* plants. We found features of the *Zizania* genome, such as independent genome duplication and, most importantly, a scenario of loss and mutation of plant immunity genes, in Jiaobai during the long-standing interaction of *Ustilago* with the host genome.

RESULTS

Genome sequencing, assembly and annotation

We selected a wild *Z. latifolia* plant (accession 'HSD2') lacking the *Ustilago* endophyte from the ancient 'Gu City' (Gu is a Chinese name for *Z. latifolia*) near the Taihu Lake basin of the low Yangtze region for our genome-sequencing effort. Using a whole-genome shotgun sequencing approach, we generated a total of 83.4 Gb of Illumina high-quality sequence data (130.1 Gb of raw data), representing approximately 140-fold genome coverage (Tables 1 and S1). *De novo* assembly of sequence data using ALLPATHS-LG (Gnerre *et al.*, 2011) resulted in an assembly containing 604.1 Mb with a scaffold N50 length of 604.9 Kb. Eighty per cent of the assembly falls into 761 super-scaffolds that are larger than 246.9 Kb (Table S1). The contiguous sequences of the assembly are consistent with the genome size estimated by *K*–mer analysis (594 Mb) and flow cytometry (586 Mb) (Figure S2). We further assessed the quality of the genome assembly through alignment to Sanger-derived phase 2 fosmid clone sequences. In five independent fosmid alignments (178.8 Kb in length), high

Table 1 Global metrics of the *Zizania latifolia* genomes

Genome sequencing	Insert size (bp)	Total data (Gb)	Coverage (x)
Wild species ('HSD2')	170–800	46.3	78.0
	2–20 × 10^3	37.1	62.5
	Total	83.4	140.5

Assembly	N50 (Kb)	Longest (Kb)	Size (Mb)
Scaffolds	604.9	3017.5	604.1

Annotation	Number	CDS length (bp)	Exons per gene
Genes	43 703	990.6	4.7

Genome re-sequencing	Insert size (bp)	Total data (Gb)	Coverage (x)
Jiaobai ('Zhejiao2')	170–500	24.0	43.6
Jiaobai ('Jiayou1')	170–500	21. 6	39.1

Genic variation*	Lost	Pseudogenized	PPI node number
Jiaobai ('Zhejiao2')	51	1968	17
Jiaobai ('Jiayou1')	42	1762	21
Shared	12	883	15

*Only genes with high impact variations, including lost and disabled genes (such as frame shifts and premature stops), in 'Jiaobai' are shown. The number of nodes in the KEGG plant–pathogen interaction pathway (PPI) with high-impact changes is also listed.

concordance was observed, excluding the highly repetitive regions, confirming high assembly accuracy (Figure S3; Table S1).

Based on *ab initio* and homology-based (including transcriptomic sequences generated by RNA-Seq) approaches, we predicted 43 703 protein-coding genes with a mean coding sequence length of 990 bp and an average number of 4.7 exons per gene (Table 1). Of the 43 703 predicted genes, 41 097 (94.0%) were supported by either homology in other species or *Zizania*-specific RNA-Seq data; 39 822 (81.8%) were found by homologs in the nr, InterProt, Swissprot and TrEMBL databases or functional classification using Gene Ontology (http://www.geneontology.org) and KEGG. To further validate the gene predictions, we used the predicted *Z. latifolia* gene set to search the Eukaryotic Orthologous Groups (KOG) genes by the core eukaryotic gene mapping approach (CEGMA; Parra *et al.*, 2007). The presence of 451 of 458 KOGs (98.5%) within the *Z. latifolia* gene set (Table S2) suggests that the *Z. latifolia* genome is close to complete. In addition to protein-coding genes, 285 microRNAs (miRNAs) and other non-coding RNAs in the *Zizania* genome were identified (Table S1). Based on homology and *de novo* methods, we identified a total of 227.5 Mb of repetitive elements, which represents 37.7% of the genome (Table S3). Among these sequences, long terminal repeat (LTR) retrotransposons (mainly *Gypsy* and *Copia* elements; Table S3) make up the majority of the transposable elements (29.8% of the genome), and DNA transposons comprise 7.1% of the genome.

Genome duplication and phylogenetics

Whole-genome duplication (WGD) is an important source for further gene functionalization, and has been shown to be associated with an ancient polyploidization event predating the divergence of the cereals (Paterson *et al.*, 2004). We calculated 4dTv (the transversion rate at the fourfold degenerate sites) of paralogous gene pairs between syntenic blocks of the *Zizania* genome and observed two distinct paralogous peaks, including the orthologous peak from the *Zizania*–*Oryza* speciation event (Figure 1a). The paralogous peak with 4dTv $\approx$ 0.38, shared with *Oryza sativa* (rice), corresponds to the WGD, predating the divergence of cereals. The second paralogous peak with 4dTv $\approx$ 0.07 contains 17.5% of all the *Zizania* paralogous gene pairs and supports an independent WGD event that occurred after *Zizania* separated from rice. We dated this recent WGD event to approximately 10.8–16.1 million years ago (Mya).

Using pairwise protein sequence comparisons among six members of the grass family and Arabidopsis as an out-group, we found a total of 31 144 genes representing 16 909 gene families in the *Z. latifolia* genome with 2871 genes in 594 families that appear to be *Zizania*-specific (3.5%; Table S4). Using a Markov cluster algorithm (Li *et al.*, 2003) to group putative orthologs and paralogs (OrthoMCL with BLASTP $< 1e^{-5}$), we identified orthologs that are conserved among *Brachypodium*, *Oryza*, *Sorghum* and *Zizania*, and those that are *Zizania*-specific genes (Figure 1b). As expected, *Zizania* appears to share more orthologous groups with *Oryza* (742 orthologous groups) than with *Brachypodium* or *Sorghum*; this finding is consistent with the phylogenetic tree based on their orthologous single-copy gene groups (Figure 1c). The estimated *Oryza*–*Zizania* divergence time (*c.* 26.7 Mya) is similar to previous estimations based on single genes (Ge *et al.*, 1999; Guo and Ge, 2005), and is consistent with our estimate of the *Zizania*-specific WGD event.

The *Zizania* genome shares high genomic synteny with the *Oryza* genome, with 1498 syntenic blocks shared between the two genomes (Figure 2). These blocks average 383.1 Kb in length, with a mean of 18 syntenic genes per block, or approximately 50.9% (22 288 genes) of the predicted *Zizania* genes. The reason for the synteny-poor regions (e.g. parts of chromosome 4) is mainly the result of repetitive sequences such as retrotransposons in rice (Chen *et al.*, 2013), which decrease its synteny with other members in the grass family.

Genomic changes after long-standing *Ustilago* fungal infection

To understand the molecular basis of the impact of the long-standing *Ustilago* interaction on the host genome, we sequenced the genome of a cultivated *Zizania* plant (Jiaobai cv. 'Zhejiao2') to 40-fold genome coverage (a total of 24.0 Gb of high-quality sequence data; Table S5). Compared with wild *Zizania*, 2019 (4.6%) of the 43 703 annotated genes were lost or carried loss-of-function mutations in the domesticated Jiaobai genome (Tables 1 and S6), including 53 missing genes, 1360 genes with frame shifts, 545 genes with premature stops, 125 genes with altered initiation and 112 genes with extended reading frames (details in Tables S7 and S8). It was found that more than half of the deleted/altered genes are present in a single copy in the wild genome (Table S6). We conducted a pathway enrichment analysis with the 2019 lost or mutated genes (Table S9), and found significantly enriched KEGG pathways ($Q \leq 0.05$) for genes involved in plant–pathogen interactions ($Q = 1.05 \times 10^{-2}$; Figure 3). Most strikingly, many putative plant immune receptor genes (i.e. disease resistance genes) were lost or mutated in the domesticated variety (Figure S4), including 34 resistance gene analogs (RGAs), which represent 21.8% of the 156 annotated nucleotide binding site leucine-rich repeat (NBS-LRR) genes (Tables S10 and S11; e.g. *RPM1* in Figure 3).

Additional lost/mutated genes include putative members of gene families that are associated with plant immunity in other plants: for example, orthologs of immune signal transduction kinases, such as *CDPK/CPK* (calcium-

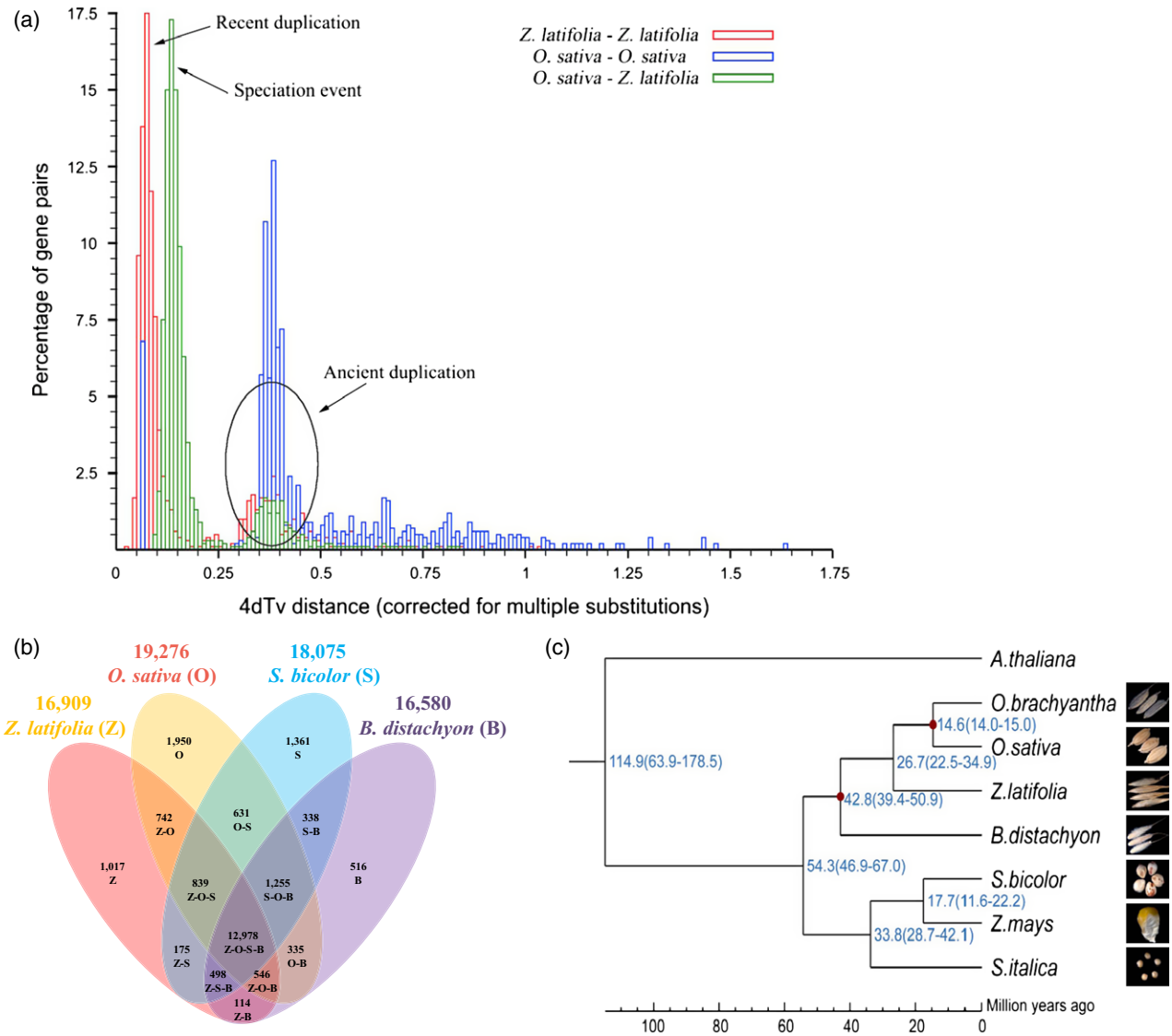

Figure 1. Evolution and phylogenetics of the *Zizania* genome.
(a) Genome duplications in cereal genomes revealed through 4dTv analyses. In addition to an ancient WGD event, predating the divergence of cereals and the *Zizania* speciation event, a recent WGD after the *Zizania*–*Oryza* divergence could be identified in the distribution of 4dTv values.
(b) Clusters of orthologous and paralogous genes in *Zizania* and four other species of the grass family. A gene family number is listed for each of the components and species.
(c) Estimation of the time of divergence (with error range shown in parentheses) of *Zizania latifolia* and four other grasses based on orthologous single-copy gene pairs with red dots indicating the calibration time.

dependent protein kinase), repressors such as the *JAZ* genes and a premature stop codon for an ortholog of *pathogenesis-related gene 1* (*PR1*) (Figure 3). Primer sequences were designed for 40 of the plant–pathogen interaction pathway genes mutated in 'Zhejiao2' (Table S12), and the mutations were confirmed by sequences of the PCR-amplified targets from genomic DNA. The selective loss/mutation of immunity genes provides evidence that the long-standing fungal infection may have driven host genome evolution, leading to a reduction of functional host immune response genes, presumably to facilitate durable fungal colonization.

To further examine this hypothesis, we sequenced the genome of a second Jiaobai cultivar ('Jiayou1') to 40-fold genome coverage (Table S5). Analysis of the sequence data found that 1804 genes were deleted or carry loss-of-function mutations (Tables 1, S6, S7 and S8). Again, the plant–pathogen interaction KEGG pathway was significantly enriched ($Q = 1.69 \times 10^{-5}$; Table S9). Furthermore, changes in 15 nodes of plant–pathogen interaction pathways in 'Zhejiao2' also occurred in the 'Jiayou1' genome (Figure 3; Table 1). Moreover, 41.7% or 754 missing or mutated genes in 'Jiayou1' were shared with 'Zhejiao2' (Table 1).

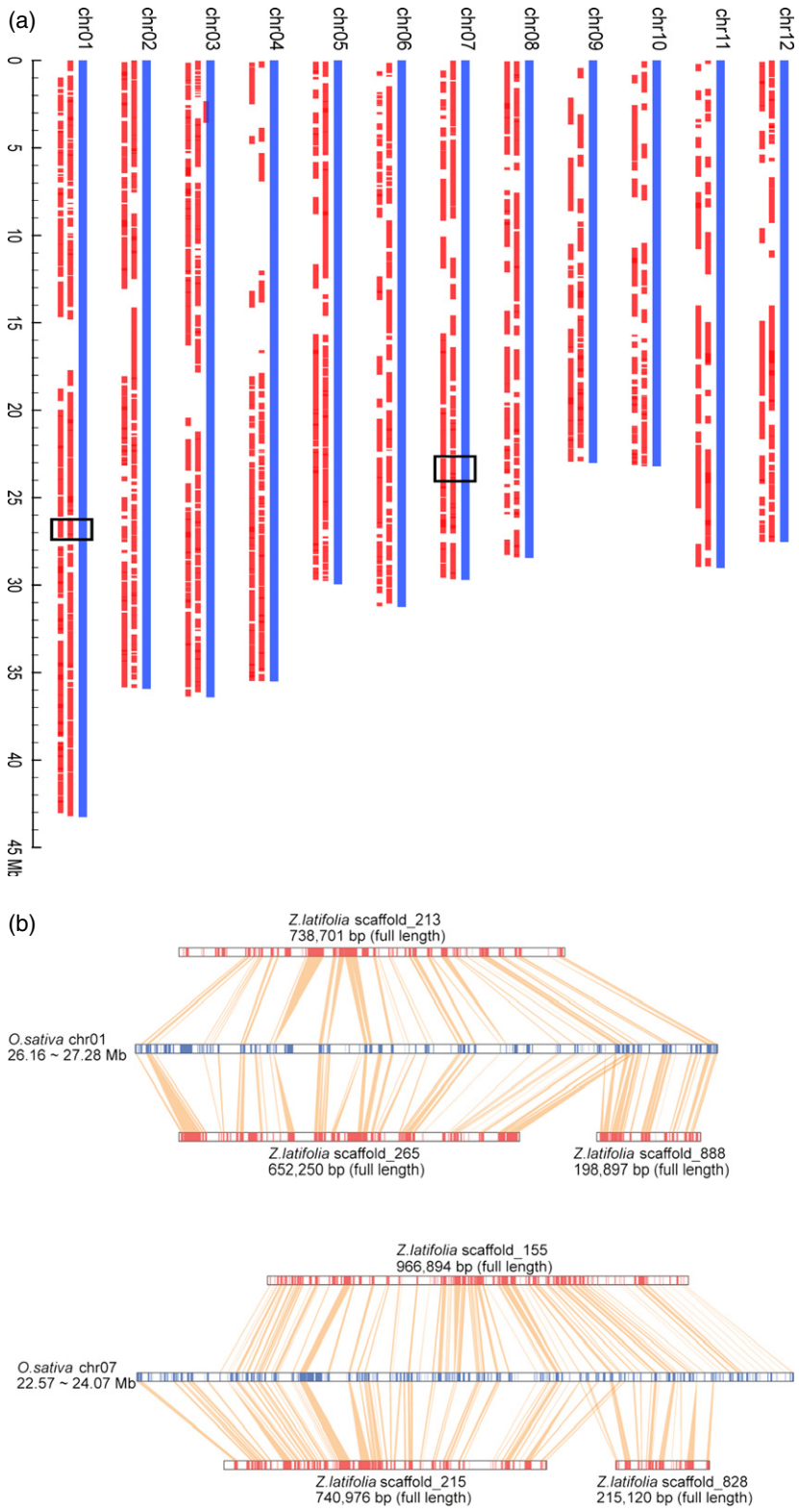

Figure 2. Genomic synteny between *Oryza* and *Zizania*.
(a) Syntenic blocks shared between the *Zizania* and rice genomes after the most recent WGD event of *Zizania*. Twelve rice chromosomes (in blue) and their *Zizania* syntenic segments (in red) are shown.
(b) Details of two examples of syntenic blocks in rice chromosomes 1 and 7 (indicated by black boxes in panel a). Genes in *Zizania* and *Oryza* are marked by red and blue boxes, respectively.

In addition to gene loss during Jiaobai domestication, gene family expansion or gene gains associated with beneficial plant–microbe interactions were further investigated. Using the wild 'HSD2' as a control and a 2–Kb sliding window size, regions with copy number variation (CNV) in the two Jiaobai cultivars were identified. CNV regions of

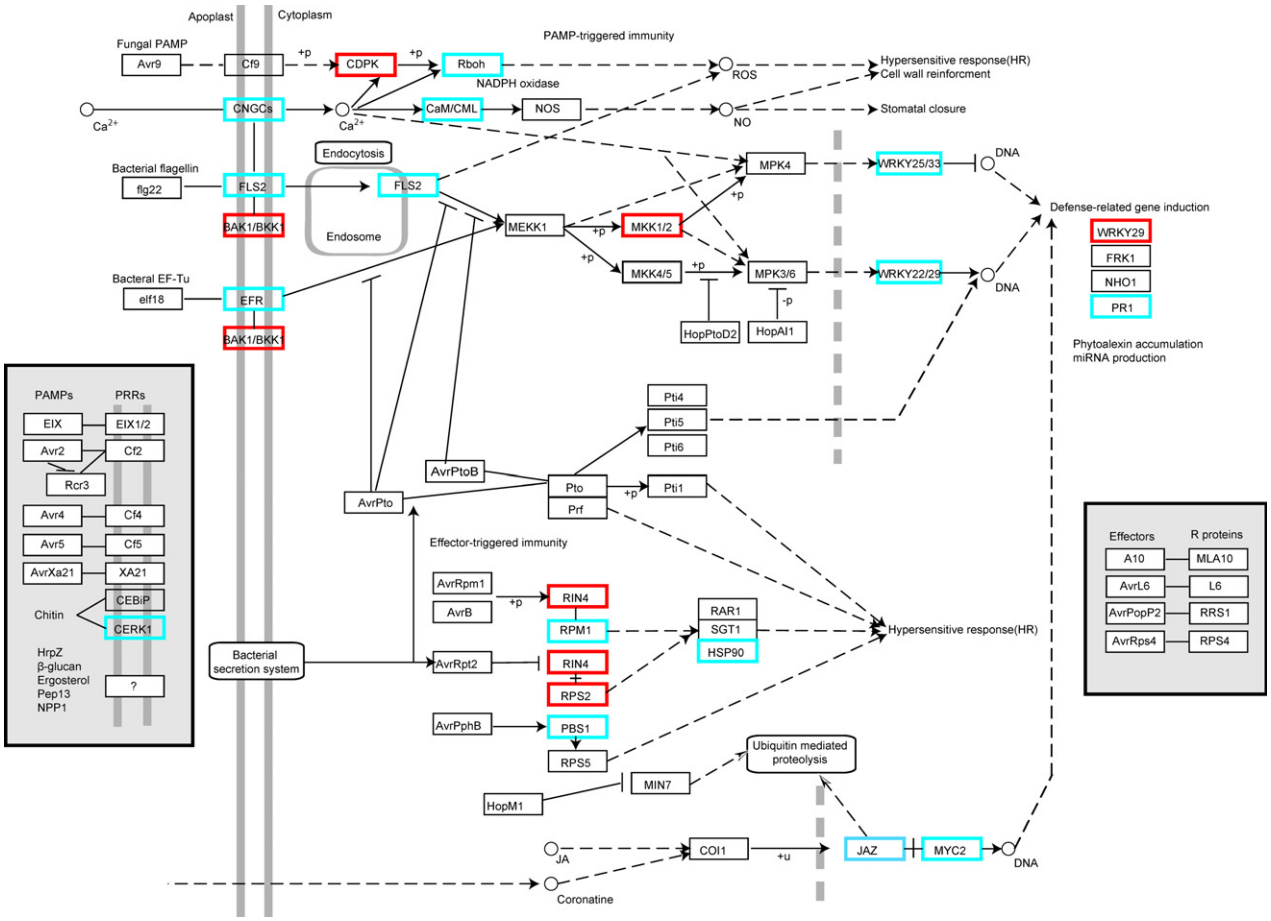

Figure 3. Gene loss and mutation in Jiaobai after long-standing *Ustilago* fungal infection.
Schematic depicting part of the plant–pathogen interaction pathway in KEGG. Blue boxes indicate lost or mutated genes belonging to the corresponding gene family in both Jiaobai cultivars ('Zhejiao2' and 'Jiayou1'), whereas red boxes indicate genes lost or mutated in one cultivar.

approximately 7 and 6 Mb, which contain 176 and 175 gained genes each, were found in 'Zhejiao2' and 'Jiayou1', respectively. Each of the two sets of genes gained has 10 orthologs of genes in the KEGG plant–pathogen interaction pathway; however, pathway-enrichment analysis indicated that the pathway was not significantly enriched in either cultivar (Table S9).

Taken together, these results suggest that a reduced plant immunity gene set is a feature of the long-term *Zizania–Ustilago* interaction during Jiaobai domestication.

Transcriptomic changes after *Ustilago* fungal affection

The transcriptomes of enlarged stems from three independent Jiaobai ('Zhejiao2') isolates were determined using RNA-Seq. As the wild-type *U. esculenta*-free of Jiaobai 'Zhejiao2' line is not available in the current community of *Zizania* biology, two normal stems from wild *Z. latifolia* ('HSD2') collected in the same field were used as controls (Table S13). A large number of differentially expressed genes (DEGs) between the wild and domesticated *Z. latifolia* were found. Gene enrichment analysis of DEGs indicated that, just as in the above results based on genomic data, the plant–pathogen interaction KEGG pathway was significantly over-represented in the three Jiaobai stems (Table S14). We further assessed the significance of gene expression differences between the cultivated and wild *Z. latifolia* groups using the NOISeq method (Tarazona *et al.*, 2011). It is interesting to note that 12/15 shared variant plant–pathogen interaction pathway nodes (Table 1) show significant changes at the transcriptional level between the cultivated and wild forms (Figure S5). Moreover, it was observed that genes putatively encoding the receptors FLS2 and CaM/CML, which are involved in the perception of pathogen effectors, were significantly downregulated in Jiaobai, compared with the wild type, suggesting that Jiaobai may fail to stimulate cell wall reinforcement (Figure S5). It is notable that there were genetic variants in the promoter or genic regions of 16 DEGs in the plant–pathogen interaction pathway, such as EFR (Zlat_10008472 and Zlat_10024724) and FLS2 (Zlat_10008470, Zlat_10012804, Zlat_10004589

and Zlat_10001197) (Table S15), indicating that the gene expression differences may be caused by those genetic alterations. Additionally, other pathways such as plant hormone signal transduction were also obviously enriched (Table S14). The NOISeq method (Tarazona *et al.*, 2011) was used to gain an overview of the significant transcriptomic changes in plant hormone signal transduction between the wild and cultivated groups (Figure S6). Genes in this pathway showed a complex pattern of regulation, with both up- and downregulation of genes in diverse biosynthesis pathways, such as zeatin, diterpenoids and carotenoids. For example, in cytokinine transduction, A–ARR, which is involved in cell division and shoot initiation, was significantly downregulated in Jiaobai compared with the wild form (Figure S6). As mentioned above, persistent infection by *U. esculenta* resulted in dramatic cell enlargement and loss of flowering for Jiaobai. Our results suggest that those pathways may also play roles in the persistence of the *Zizania*–*Ustilago* complex.

DISCUSSION

We sequenced a unique crop species in the rice tribe to gain insights into the *Zizania* genome and the genome-scale consequences of its long-standing plant–pathogen interaction with the fungal endophyte *U. esculenta*. Our within-species comparative genomic analysis suggests that the control of plant–pathogen interactions played a key role in the long-term accommodation of an agriculturally favorable fungal infection, and consequently in the domestication of Jiaobai as a vegetable crop. The long-term cultivation and artificial selection of a plant–pathogen interaction made cultivated *Z. latifolia* ('Jiaobai') have a significantly smaller repertoire of immune receptors compared with wild *Z. latifolia,* although some new members were recruited via gene duplication in Jiaobai. The 'weakened' immune system of Jiaobai is crucial for the continuous maintenance of the *U. esculenta*–*Z. latifolia* interaction. As mentioned above, *U. esculenta* is not a classical symbiotic endophyte, and its relationship with the host plant is weak. Meanwhile, under pressure from artificial selection, nucleotide sequence mutations causing, for example, frame shifts and premature stops, have decreased the repertoire of immune receptors in Jiaobai (Table 1). The genes completely lost/deleted only count for a small part of the disabled genes. Regarding the potential mechanisms responsible for these mutations, previous studies have suggested that genes involved in plant defense are hypermutagenic, and their high sequence similarity and repetitive motifs (e.g. the LRR domain) prompt non-allelic homologous recombination, which leads to frequent turnover and diversification (deletion, duplication or fusion; Karasov *et al.*, 2014). Repetitive elements are important drivers of the evolution of genes and genomes (Jiang and Ramachandran, 2013). We scanned for repetitive sequences in the genomic regions flanking the 2019 disabled genes of the plant–pathogen interaction pathway, and different distribution patterns of genomic distances to repetitive sequences were observed between the disabled gene set and the control (a randomly selected gene set). The results suggest that repeat sequences might have also played a role in the mutations.

The *Zizania* genomes reported here provide an important resource for comparative genomic studies of the rice tribe and the grass family in general. To date, the available sequenced grass genomes are all within either an extended or near-range evolutionary time frame relative to rice. For example, the recently released *Oryza brachyantha* genome (Chen *et al.*, 2013) is a sister taxon within the genus *Oryza* that separated from *O. sativa* approximately 14–15 Mya (Tang *et al.*, 2010), whereas other grasses such as *Brachypodium distachyon* and *Phyllostachys heterocycla* (bamboo) separated from *Oryza* species nearly 50 Mya (Peng *et al.*, 2013). As shown in Figure 1c, the *Zizania* species are located at an intermediate position in the evolutionary time frame (approximately 26.7 million years) between *Oryza* (Ehrhartoideae) and *Brachypodium* (Pooideae) or *Phyllostachys* (Bambusoideae) in the BEP clade.

Agronomically, relative to *Oryza*, *Z. latifolia* generally has higher resistance to abiotic (such as cold) and biotic stresses (such as rice blast) (An *et al.*, 2011; Shen *et al.*, 2011), and higher biomass production. Therefore, the *Zizania* genomes provide a genetic resource for rice breeding. Efforts to transfer the *Zizania* genetic background into *Oryza* via distant hybridization have been successfully attempted over the last 15 years (Liu *et al.*, 1999; Shen *et al.*, 2011). The *Zizania* genome will also provide a diverse gene pool for rice molecular breeding. For example, it may be possible to use the candidate gene approach and transfer *Z. latifolia* genes (such as NBS-type genes) into rice as a means of improving its abiotic and biotic stress resistance (Dangl and Jones, 2001). Those NBS genes with sequence novelty may provide the genetic basic for resistance in *Zizania*. Moreover, the *Zizania* assembly also offers a genomic/genetic resource for North American breeders who are trying to improve *Zizania palustris* (http://www.reeis.usda.gov; Hayes *et al.*, 1989; Kennard *et al.*, 2002) and for the Chinese effort to re-domesticate Asian species (Wang *et al.*, 2013).

Biologically, the Jiaobai genomic resource represented here creates a model for studying and testing the genetic basis of plant–fungus interactions. As mentioned above, Jiaobai provides a rare opportunity to address the molecular mechanisms involved in the long-term accommodation of an agriculturally favorable fungal infection. Our results suggest that in addition to the intriguing genomic changes occurring in the host genome in response to a long-standing endophytic microbe, the role of gene regulation,

including interspecific variation in promoter motifs and epigenetic regulation, merits further investigation. Finally, sequencing and analysis of the genome of the endophytic *Ustilago* strain currently underway will further facilitate the better understanding of the adaptive evolution of the *Zizania–Ustilago* complex.

EXPERIMENTAL PROCEDURES

Plant materials

The *Z. latifolia* plant 'HSD2' was collected from the ancient Gu City near the Taihu Lake basin, Zhejiang Province, and the two Jiaobai accessions ('Zhejiao2' and 'Jiayou1') are local cultivars.

Genome sequencing and assembly

Genomic DNA pair-end libraries were prepared using standard Illumina protocols, and then the libraries were sequenced using an Illumina Genome Analyzer II. 'HSD2' genome assembly was performed by ALLPATHS-LG (http://www.broadinstitute.org/software/allpaths-lg/). The two Jiaobai cultivars were deep sequenced and assembled using the same method. To exclude possible fungal sequences from the Jiaobai sample raw data, a draft genome of the endophyte *U. esculenta* generated by us (under the National Center for Biotechnology Information accession: JTLW00000000) was used to find and discard the fungal reads. The Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under accession ASSH00000000. The version described in this paper is version ASSH01000000.

Transcriptome sequencing and analysis

Enlarged stems from three independent Jiaobai 'Zhejiao2' isolates and two normal stems from wild *Z. latifolia* ('HSD2') planted in the same field were harvested directly into liquid nitrogen. RNA was extracted and used for RNA-Seq conducted by Ion Proton, which generated a total of 9.3 Gb of transcriptomic data. Mapping of the Ion Proton reads was performed by TMAP 3.4.1, followed by the quantification of gene expression using the Reads Per Kilobase per Million mapped reads (RPKM) algorithm. The programs DEGSEQ (Wang *et al.*, 2010) and NOISEQ (Tarazona *et al.*, 2011) were further applied to identify differentially expressed genes between individuals and groups. The transcriptomic data have been deposited at GenBank under accession PRJNA187578. The gene enrichment analysis was carried out using EnrichPipeline (Chen *et al.*, 2010).

Gene and repeat annotation

Tandem repeats were identified using a tandem repeat finding program (TRF; Benson, 1999). Transposable elements (TEs) were predicted using a combination of homology-based comparisons and *ab initio* approaches. The homology-based comparisons were performed by REPEATMASKER (Tarailo-Graovac and Chen, 2009) with RepBase (Jurka *et al.*, 2005) as the database, whereas LTR_FINDER (Xu and Wang, 2007) and REPEATMODELER (http://www.repeatmasker.org/RepeatModeler.html) were used as *ab initio* approaches. Finally, the predicted results were merged into a final set.

To predict genes in the 'HSD2' genome, we used both homology-based and *ab initio* methods. For the homology-based prediction, *O. sativa* (IRGS Project 2005), *Brachypodium* (Vogel *et al.*, 2010) and *Arabidopsis* (Swarbreck *et al.*, 2008) proteins were mapped onto the *Z. latifolia* 'HSD2' genome using TblastN. Homologous genomic sequences were aligned against the matching proteins using GENEWISE (Birney and Durbin, 2000) to define gene models. For *ab initio* prediction, AUGUSTUS (Stanke *et al.*, 2006) was employed using the maize matrix. Our RNA-Seq data were mapped to the 'HSD2' genome using BLAT (Kent, 2002) to obtain gene structure hits. Together, hits from these complementary analyses were merged with AUGUSTUS to produce a non-redundant reference gene set. Furthermore, the predicted genes were checked to remove possible transposon genes. First, we scanned for retrotransposon domains among the raw predicted genes based on the following profiles: reverse transcriptase (RT) by PF00078 and PF07727, integrase (INT) by PF00665, PF00552 and PF03732, aspartic proteinase (AP) by PF00026 and PF00077. Then we searched genes for homologous proteins from the Repbase with BLASTP ($E \leq 1e^{-4}$, identity $\geq$ 30%, coverage $\geq$ 30% and minimum matching length $\geq$ 30 aa). Genes with these domains but with no expression support were filtered from the original gene set. Finally, the remaining genes with neither annotation nor expression were collected, and those with a gene model score $<$ 0.7, as evaluated by AUGUSTUS, were filtered.

Comparative analyses of genomes

We used OrthoMCL to obtain paralog and ortholog gene family estimates among the 'HSD2', *Arabidopsis thaliana*, *B. distachyon*, *O. sativa*, *Sorghum bicolor* (Paterson *et al.*, 2009), *O. brachyantha* (Chen *et al.*, 2013), *Setaria italica* (Zhang *et al.*, 2012a) and *Z. mays* (Schnable *et al.*, 2009) with BLASTP ($E < 1e^{-5}$) and using the default inflation value. The *O. brachyanthaprotein* sequences were downloaded from http://archive.gramene.org/species/oryza_species/o_brachyantha.html, and other protein sequences were obtained from the Phytozome database (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/). The longest transcript of each gene in each species was used for ORTHOMCL analysis. Syntenic blocks (with more than five genes per block) were identified based on MCscan (Tang *et al.*, 2008). We calculated the 4dTv values by using a PERL script for each paralogous gene pair within the *Z. latifolia* 'HSD2' and *Oryza* genomes, and for orthologous gene pairs in their syntenic blocks. The 4dTv distribution was used for the estimation of speciation and whole-genome duplication events. The *Ks* value of each paralogous gene pair was calculated and used to estimate the time of the WGD event based on a synonymous substitution rate of 6.5×10^{-9} substitutions per synonymous site per year for cereals (Blanc and Wolfe, 2004). The divergence time was calculated using MCMCTree implemented in PAML (Yang, 2007). Soft prior constraints were set for the *O. brachyantha–O. sativa* (15.0 Mya) and *B. distachyon–O. sativa* (40.0 Mya) split times.

Genomic variation and pathway enrichment analysis

To identify variation of the Jiaobai cultivars ('Zhejiao2' and 'Jiayou1'), the Jiaobai reads were mapped onto the 'HSD2' genome with STAMPY (Lunter and Goodson, 2011). SNPs and indels were detected using both MPILEUP in SAMTOOLS (Li *et al.*, 2009) and the HAPLOTYPECALLER program in GATK (McKenna *et al.*, 2010). Concordant variants were retrieved and subjected to a further variant quality score recalibration (VQSR) and filtering process implemented by the VARIANTRECALIBRATOR and APPLYRECALIBRATION program in GATK. The results were defined as the final 'Zhejiao2' or 'Jiayou1' variants and used for further analysis. Genes with average coverage of <1x in their CDS region and extremely low expression (FPKM $<$ 1) in three different tissues of 'Zhejiao2' were defined as missing genes. To identify CNV regions, reads amounting to 30–40-fold coverage from the 'HSD2' and two Jiaobai cultivars were first

mapped onto the 'HSD2' genome assembly, and copy numbers were estimated based on alignments by Control-FREEC (Chen *et al.*, 2010) using 'HSD2' as the control and a 2–Kb window size. Genes contained completely in the gain region were designated as gain genes in Jiaobai. The gene enrichment analysis was performed using EnrichPipeline (Chen *et al.*, 2010).

ACKNOWLEDGEMENTS

The Jiaobai cultivar 'Zhejiao2' was kindly provided by Deping Guo and the authors thank Michael Timko for critical reading. This work was supported by State Key Lab of Rice Biology of China, Zhejiang Key Lab of Crop Germplasm Resources, the Gordon and Betty Moore Foundation, the National Natural Science Foundation of China (31000357 and 30921140408), and the National Science and Technology Ministry of China (2012BAD27B01).

AUTHOR CONTRIBUTIONS

L.F., L.G. and X-P.Z. conceived the study; Y.G. carried out flow cytometry work; L.F., Y.G. and Z.Y. collected the plant and fungal materials. Y.G., Y-Y.W., S.T., J.B., F.F., H.C., J.X. and Z.Y. prepared the samples. X.G. and C.S. performed the genome sequencing and assembly; C.S., L.F. and J.W. supervised the genome sequencing and assembly; J.Q., Z.H., C.C., C.L., H.X., Q.Z., C-Y. Y., W.W., X.Z., X.X., S.Y.H. and L.F. performed the analyses of genomic and transcriptomic data, and X-P.Z., S.Y.H., X.Y., G.A.T. S.G. and Q.Q. discussed the data. All authors contributed to the data interpretation. L.F., J.Q., X.Y. and S.Y.H. wrote the paper with significant contributions from L.G. and X-P.Z and input from all authors.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Domestication, production and application of *Zizania latifolia*.

Figure S2. Genome size estimation of *Zizania latifolia*.

Figure S3. Sequence alignments between five *Zizania latifolia* 'HDS2' fosmid clone sequences and the assembled genome.

Figure S4. Loss and mutation of genes associated with plant–pathogen interactions in the domesticated *Zizania latifolia* variety Jiaobai.

Figure S5. A schematic depicting transcriptomic changes of nodes in the plant–pathogen interaction pathway in KEGG.

Figure S6. A schematic depicting transcriptomic changes of nodes in the plant hormone signal transduction pathway in KEGG.

Table S1. Genome sequencing, assembly and annotation of wild *Zizania latifolia* (HSD2).

Table S2. Evaluation of the *Zizania latifolia* 'HSD2' gene set by the CEGMA pipeline.

Table S3. TEs in the *Zizania latifolia* 'HSD2' genome.

Table S4. Statistics of gene families in *Zizania latifolia* and other species.

Table S5. Statistics for the Jiaobai cv. 'Zhejiao2' and 'Jiayou1' Illumina sequence data.

Table S6. SNPs, indels and lost genes in two Jiaobai cultivars compared with the *Zizania latifolia* 'HSD2' genome.

Table S7. List of lost genes in the Jiaobai cv. 'Zhejiao2' and 'Jiayou1' genomes.

Table S8. List of disabled genes in Jiaobai cv. 'Zhejiao2' and 'Jiayou1'.

Table S9. KEGG pathway enrichment analysis of genes disabled or lost in Jiaobai cv. 'Zhejiao2' and 'Jiayou1'.

Table S10. Statistics of NBS-LRR-type immune receptor genes in *Zizania latifolia* and *Oryza sativa*.

Table S11. Mutations in NBS-LRR-type immune receptor genes in 'Zhejiao2' compared with 'HSD2'.

Table S12. Validation results of selected genetic variants and PCR primers used in this study.

Table S13. RNA-seq data from enlarged Jiaobai 'Zhejiao2' stems and from two normal wild *Zizania latifolia* stems without *Ustilago* infection ('HSD2').

Table S14. Enriched pathways ($Q < 0.05$) among differentially expressed genes (DEGs) between enlarged stems of Jiaobai and normal stems of wild *Zizania latifolia*.

Table S15. Transcriptomic changes of genes involved in the plant–pathogen interaction pathway in KEGG.

REFERENCES

An, X., Cui, H., Yu, X. and Ye, Z. (2011) Low temperature stress on membrane damage and cold response of cultivar Longjiao 2, *Zizania latifolia*. *J. China Jiliang Uni.* **2**, 171–175.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573.

Birney, E. and Durbin, R. (2000) Using genewise in the *drosophila* annotation experiment. *Genome Res.* **10**, 547–548.

Blanc, G. and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell*, **16**, 1679–1691.

Chan, Y.S. and Thrower, L. (1980) The host-parasite relationship between *Zizania caducifyora* Turcz. and *Ustilago esculenta* P. Henn. I. structure and development of the host and host-parasite combination. *New Phytol.* **85**, 201–207.

Chen, S., Yang, P., Jiang, F., Wei, Y., Ma, Z. and Kang, L. (2010) De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS ONE*, **5**, e15633.

Chen, J., Huang, Q., Gao, D. *et al.* (2013) Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nat. Commun.* **4**, 1595.

Chung, K.R. and Tzeng, D.T. (2004) Nutritional requirements of the edible gall producing fungus. *J. Biol. Sci.* **4**, 246–252.

Dangl, J.L. and Jones, J.D. (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.

Feldbrügge, M., Kellner, R. and Schipper, K. (2013) The biotechnological use and potential of plant pathogenic smut fungi. *Appl. Microbiol. Biotechnol.* **97**, 3253–3265.

Ge, S., Sang, T., Lu, B.-R. and Hong, D.Y. (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Natl Acad. Sci. USA*, **96**, 14400–14405.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P. and Sykes, S. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Guo, Y.L. and Ge, S. (2005) Molecular phylogeny of *Oryzeae* (Poaceae) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. *Am. J. Bot.* **92**, 1548–1558.

Guo, H.B., Li, S.M., Peng, J. and Ke, W.D. (2007) *Zizania latifolia* Turcz. cultivated in China. *Genet. Resour. Crop Evol.* **54**, 1211–1217.

Hayes, P.M., Stucker, R.E. and Wandrey, G.G. (1989) The domestication of American wildrice. *Econ. Bot.* **43**, 203–214.

IRGS Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Jiang, S.Y. and Ramachandran, S. (2013) Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. *PLoS ONE*, **8**, e71118.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Karasov, T.L., Horton, M.W. and Bergelson, J. (2014) Genomic variability as a driver of plant-pathogen coevolution? *Curr. Opin. Plant Biol.* **18**, 24–30.

Kellogg, E.A. (2009) The evolutionary history of *Ehrhartoideae*, *Oryzeae*, and *Oryza*. *Rice*, **2**, 1–14.

Kennard, W., Phillips, R. and Porter, R. (2002) Genetic dissection of seed shattering, agronomic, and color traits in American wildrice (*Zizania palustris* var. interior L.) with a comparative map. *Theor. Appl. Genet.* **105**, 1075–1086.

Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu, B., Liu, Z. and Li, X. (1999) Production of a highly asymmetric somatic hybrid between rice and *Zizania latifolia* (Griseb): evidence for intergenomic exchange. *Theor. Appl. Genet.* **98**, 1099–1103.

Lunter, G. and Goodson, M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. and Daly, M. (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Paterson, A.H., Bowers, J.E. and Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA*, **101**, 9903–9908.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T. and Poliakov, A. (2009) The sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.

Peng, Z., Lu, Y., Li, L., Zhao, Q., Feng, Q., Gao, Z., Lu, H., Hu, T., Yao, N. and Liu, K. (2013) The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nat. Genet.* **45**, 456–461.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L. and Graves, T.A. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Shen, W., Song, C., Chen, J., Fu, Y., Wu, J. and Jiang, S. (2011) Transgenic rice plants harboring genomic DNA from *Zizania latifolia* confer bacterial blight resistance. *Rice Sci.* **18**, 17–22.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R. and Ploetz, L. (2008) The Arabidopsis information resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

Tang, L., Zou, X.-H., Achoundong, G., Potgieter, C., Second, G., Zhang, D.-Y. and Ge, S. (2010) Phylogeny and biogeography of the rice tribe (*Oryzeae*): evidence from combined analysis of 20 chloroplast fragments. *Mol. Phylogenet. Evol.* **54**, 266–277.

Tarailo-Graovac, M. and Chen, N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, **4**, 1–14.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223.

Vogel, J.P., Garvin, D.F., Mockler, T.C. *et al.* (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.

Wang, Y., Huang, L. and Fan, L. (2013) Main agronomic traits, domestication and breeding of Gu (*Zizania latifolia*). *J. Zhejiang. Uni.* **39**, 629–635.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268.

Xu, X., Ke, W., Yu, X., Wen, J. and Ge, S. (2008) A preliminary study on population genetic structure and phylogeography of the wild and cultivated *Zizania latifolia* (Poaceae) based on Adh1a sequences. *Theor. Appl. Genet.* **116**, 835–843.

Xu, X., Walters, C., Antolin, M.F., Alexander, M.L., Lutz, S., Ge, S. and Wen, J. (2010) Phylogeny and biogeography of the eastern Asian-North American disjunct wild-rice genus (Zizania L., Poaceae). *Mol. Phylogenet. Evol.* **55**, 1008–1017.

Yang, Z.H. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Yu, Y. (1962) Study on the materials secreted by *Ustilago esculenta* P. Henn. in *Zizania latifolia*. *Acta Bot. Sin.* **4**, 339–350.

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z. and Wang, W. (2012a) Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* **30**, 549–554.

Zhang, J.Z., Chu, F.Q., Guo, D.P., Hyde, K.D. and Xie, G.L. (2012b) Cytology and ultrastructure of interactions between Ustilago esculenta and *Zizania latifolia*. *Mycol. Prog.* **11**, 499–508.