

RESEARCH

Open Access



# Diverse genetic mechanisms underlie worldwide convergent rice feralization

Jie Qiu<sup>1,2†</sup>, Lei Jia<sup>1†</sup>, Dongya Wu<sup>1†</sup>, Xifang Weng<sup>1</sup>, Lijuan Chen<sup>3</sup>, Jian Sun<sup>4</sup>, Meihong Chen<sup>1</sup>, Lingfeng Mao<sup>1</sup>, Bowen Jiang<sup>1</sup>, Chuyu Ye<sup>1</sup>, Guilherme Menegol Turra<sup>5</sup>, Longbiao Guo<sup>6</sup>, Guoyou Ye<sup>7</sup>, Qian-Hao Zhu<sup>8</sup>, Toshiyuki Imaizumi<sup>9</sup>, Beng-Kah Song<sup>10</sup>, Laura Scarabel<sup>11</sup>, Aldo Merotto Jr<sup>5</sup>, Kenneth M. Olsen<sup>12\*</sup> and Longjiang Fan<sup>1,13\*</sup>

## Abstract

**Background:** Worldwide feralization of crop species into agricultural weeds threatens global food security. Weedy rice is a feral form of rice that infests paddies worldwide and aggressively outcompetes cultivated varieties. Despite increasing attention in recent years, a comprehensive understanding of the origins of weedy crop relatives and how a universal feralization process acts at the genomic and molecular level to allow the rapid adaptation to weediness are still yet to be explored.

**Results:** We use whole-genome sequencing to examine the origin and adaptation of 524 global weedy rice samples representing all major regions of rice cultivation. Weed populations have evolved multiple times from cultivated rice, and a strikingly high proportion of contemporary Asian weed strains can be traced to a few Green Revolution cultivars that were widely grown in the late twentieth century. Latin American weedy rice stands out in having originated through extensive hybridization. Selection scans indicate that most genomic regions underlying weedy adaptations do not overlap with domestication targets of selection, suggesting that feralization occurs largely through changes at loci unrelated to domestication.

**Conclusions:** This is the first investigation to provide detailed genomic characterizations of weedy rice on a global scale, and the results reveal diverse genetic mechanisms underlying worldwide convergent rice feralization.

**Keywords:** Weedy rice, *Oryza sativa*, Crop feralization, Global population, De-domestication block, Parallel evolution

## Introduction

Weedy rice (*Oryza sativa* f. *spontanea*) has emerged as a serious global agricultural problem in recent decades [1]. It is distinguished from cultivated rice by a high degree of seed shattering and seed dormancy, by inferior grain quality, and by the ability to aggressively outcompete cultivated rice in paddy fields once established. The

evolutionary origin of weedy rice has been debated for decades. Recent evidence from whole-genome resequencing supports origins from cultivated rice ancestors [2–5]. However, genome resequencing investigations to date have been restricted to strains from a single country or region, providing only a limited understanding of the weed's genomic composition and the timing of weed evolution.

With the recent availability of genome sequence data from > 4500 worldwide cultivated rice varieties [6], it is now possible to infer geographical origins of global weedy rice with a high degree of resolution, and to potentially identify specific varieties that are progenitors of major weed strains. Crop-weed genome comparisons

\* Correspondence: [kolsen@wustl.edu](mailto:kolsen@wustl.edu); [fanlj@zju.edu.cn](mailto:fanlj@zju.edu.cn)

† Jie Qiu, Lei Jia and Dongya Wu contributed equally to this work.

<sup>1,2</sup>Department of Biology, Washington University in St. Louis, St. Louis, MO 63130, USA

<sup>1</sup>Institute of Crop Sciences and Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China  
Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

can also provide insights on the genetic mechanisms of weedy rice adaptation. In particular, genome scans can reveal whether adaptations in weedy rice that superficially resemble phenotypic reversions to wild rice (e.g., seed shattering and dormancy) have arisen through changes at the same loci that were targets of selection during domestication (e.g., through second-site suppression of domestication alleles), or through changes at loci that were not involved in the domestication process. Genome sequence analysis can further reveal whether adaptations for weediness in independently evolved strains have occurred through shared or independent genetic mechanisms, thereby shedding light on the extent to which the feralization process is genetically constrained [7, 8].

In this study, we analyzed whole-genome sequences of 524 weedy rice strains, representing all of the world's major rice-growing regions, together with a worldwide sample of cultivated rice genomes. These analyses reveal that weedy rice has evolved repeatedly from cultivated rice ancestors worldwide, that this process has occurred both early in the history of rice cultivation and very recently, and that despite a high degree of flexibility in the genetic mechanisms by which weediness adaptations evolve, some genomic regions have been targeted repeatedly in the parallel evolution of weed strains.

## Results

### Sampling and genomic data of weedy, cultivated, and wild rice

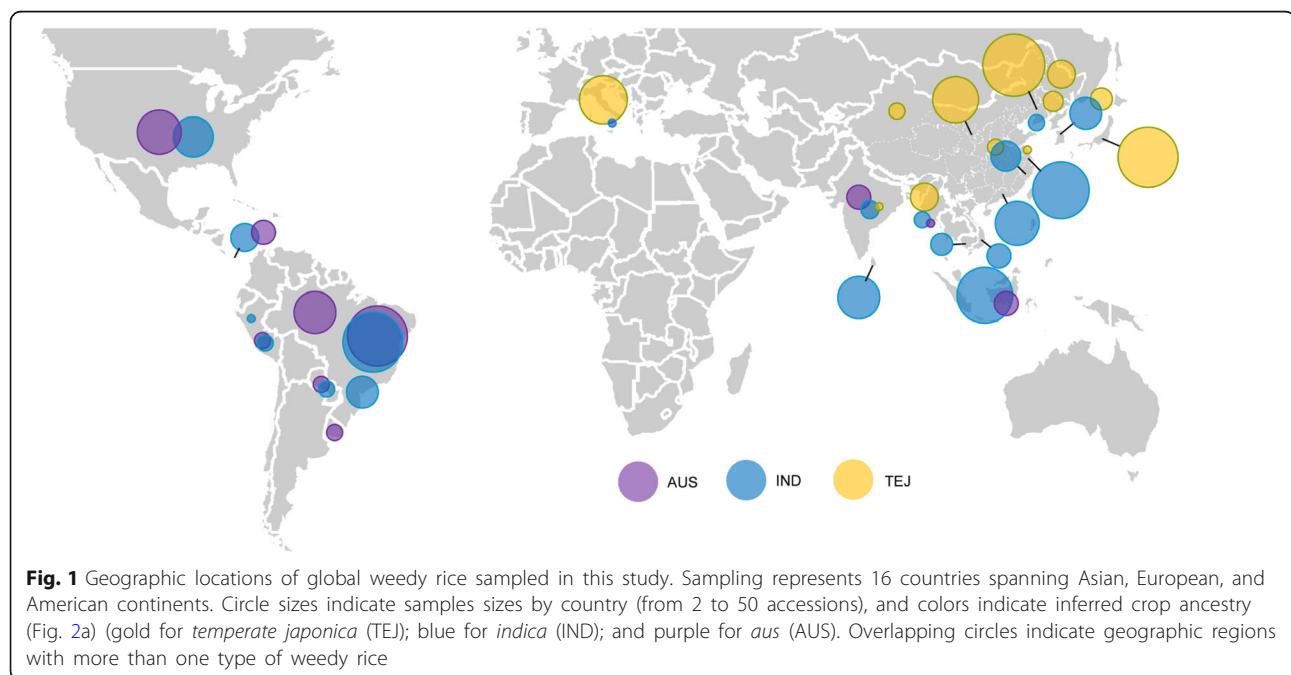
For a comprehensive investigation of the origin and evolution of global weedy rice, whole-genome sequences

were analyzed from 524 weedy rice accessions representing major rice production areas in 16 countries across Asia, Europe, North America, and Latin America (Fig. 1; Additional file 1: Table S1). Diverse types of rice are cultivated in these regions, including *temperate japonica* (predominant in northern China, Korea, Japan, and Italy), *indica* (in southern China, India, Southeast Asia and Latin America), *aus* (in upland regions of the Indian subcontinent), and *tropical japonica* (in Southeast Asia and the USA). Phenotypically, most of the weed accessions had seeds characterized by reddish-brown pericarp color and smooth spikelet bases, which are typical traits of weedy rice (Additional file 1: Table S1).

Weedy rice samples were sequenced to an average 19.9 $\times$  genome coverage. For genomic comparison, public genomic sequence data were retrieved for a worldwide sample of 426 locally cultivated rice varieties and 53 wild rice accessions [3, 9, 10]; this yielded a combined genotype dataset of 16.2 million SNPs across 1003 samples for use in the population genomic analysis described below.

### Weedy rice has evolved repeatedly from cultivated rice

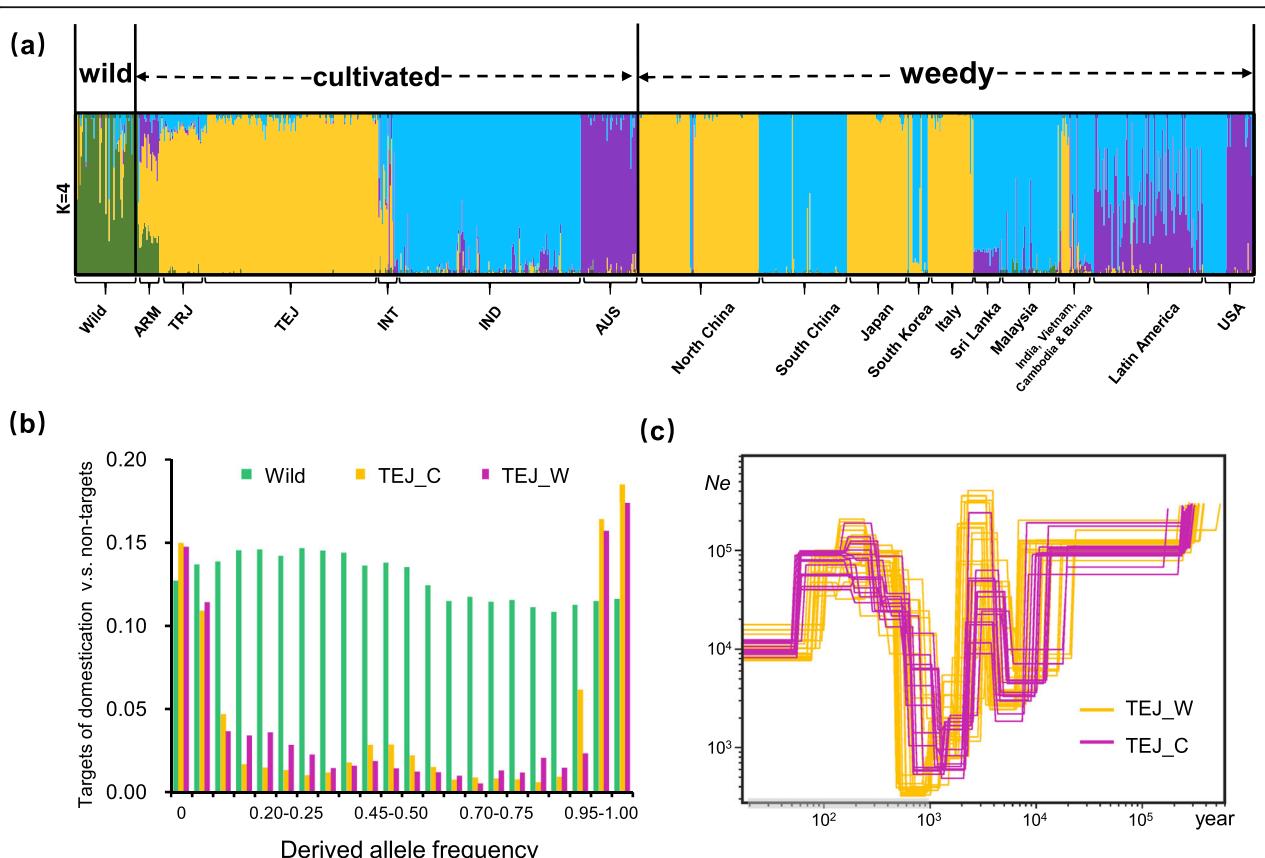
Population structure and principal component analysis (PCA) confirmed previously described subgroups within cultivated rice (*tropical japonica*, *temperate japonica*, and *aromatic* varieties within the traditional *japonica* subspecies; and *indica* and *aus* varieties within the traditional *indica* subspecies) (Fig. 2a; Additional file 2: Fig. S1). For weedy rice, though varying by region, all strains shared ancestry predominantly with cultivated rice, specifically varieties of the *indica*, *temperate japonica*, and *aus* subgroups. None of the weeds showed



closest ancestry with wild rice, although some degree of wild rice introgression was evident in Southeast Asian strains based on FastStructure and ABBA analyses (Fig. 2a; Additional file 2: Fig. S2); this is consistent with previous inferences of wild rice introgression into weed populations in this geographical region [11–13]. Assessments of genome-wide nucleotide diversity indicated that most weedy groups harbor lower genetic diversity than their respective inferred crop ancestors (Additional file 2: Fig. S3), consistent with post-domestication bottlenecks during fertilization. We also compared the ratio of the derived allele frequency spectrum of genomic regions that were targets of selection during rice domestication [9] and regions that were not, in wild, cultivated and weedy rice populations. Both *japonica* and *indica* weeds showed the domestication-associated U-shaped distribution found in cultivated rice (Fig. 2b; Additional file 2: Fig. S4) and thus bear a signature of ancestry from

domesticated ancestors. Similarly, the relative genetic diversity change of domestication and improvement genes [14] shows a similar level of reduction for weedy and cultivated rice compared to wild rice (Additional file 2: Fig. S5); this is again consistent with weedy rice descent from domesticated ancestors.

Estimates of divergence times between weeds and their respective crop ancestors revealed substantial variation among strains (Fig. 2c; Additional file 2: Fig. S6). For example, both *japonica*-derived weeds and cultivated *japonica* rice shared a very recent genetic bottleneck around 1000 years ago (assuming one generation per year) (Fig. 2c), and had similar patterns in the distribution of population effective size ( $N_e$ ) after that; this suggests that *japonica* weeds likely diverged from their cultivated counterparts < 1000 years ago. In comparison, some *indica*- and *aus*-derived weeds were inferred to have diverged from their respective crop ancestors approximately a



**Fig. 2** Genetic and geographical origin of global weedy rice. **a** Population structure of global wild, cultivated, and weedy rice. The maximum likelihood clustering with  $K=4$  is presented. The cultivated rice group includes aromatic, tropical *japonica*, temperate *japonica*, *indica*, and *aus* type (ARM, TRJ, TEJ, IND, and AUS, respectively). The genetic composition of wild, TEJ, IND, and AUS are colored predominantly with dark green, orange, blue, and purple, respectively. In the weedy rice group, accessions are grouped by country to illustrate the types of weeds in each region; **b** The ratio of allele frequency for SNPs between targets of domestication selection vs. non-targets (estimated by Huang et al. (2012)) for temperate *japonica* type, indicating a similar allele frequency spectrum for *japonica* weedy and cultivated rice. **c** Effective population size ( $N_e$ ) changes for the cultivated (C) and different weedy (W) rice populations of *japonica* type. The estimated divergence time between weedy and cultivated rice is shown with the gray-shaded bar along the x-axis and assumes one generation per year

millennium earlier (Additional file 2: Fig. S6). Taken together, these results suggest that weedy rice has evolved repeatedly and independently from cultivated ancestors at different time points during the history of rice cultivation.

#### Hybrid origin of Latin American weedy rice

Weedy rice in Latin America (e.g., Brazil, Panama, Paraguay, and Peru) is unique among worldwide samples, with over half of samples (51/95) showing admixed genetic ancestry between *indica* and *aus* (Fig. 2a; Additional file 1: Table S1, S2). Consistent with this pattern, a TreeMix analysis indicates that these putatively admixed accessions originated from hybridization between Latin America *indica*- and *aus*-type weedy rice (Additional file 2: Fig. S7). In addition, the admixed strains showed elevated nucleotide diversity compared to local *indica* or *aus* weedy rice strains (Additional file 2: Fig. S3), as well as higher observed heterozygosity than weedy rice in other world regions (Additional file 2: Fig. S8). Given that weedy rice, like cultivated rice, is predominantly self-fertilizing, these patterns suggest that many of the Latin American weeds in our sample have originated through recent hybridization of local *aus* and *indica* weeds.

Commercial varieties with ALS (acetolactate synthase) herbicide resistance (HR) have been released in many countries since 2001, and several weedy rice populations with tolerance to herbicide have been reported recently [15–17]. In our weedy rice collection, a total of 11 non-synonymous SNPs were found within ALS in 52 Latin American weedy rice accessions, mostly from Brazil (51, including 9 *indica* type, 4 *aus* type, and 38 *indica*-*aus* type) (Additional file 1: Table S3; Additional file 2: Fig. S9a). Most HR cultivars in Brazil are *indica*. Of the 11 ALS mutations observed in Latin American weeds, 3 functional mutations (Ala122Thr, Ser653Asn, and Gly654Glu) have been employed in HR cultivars [18]. These results suggest that HR weedy rice has likely acquired resistance by crop-weed hybridization and adaptive introgression—i.e., escape of resistance alleles from HR cultivars—although the possibility of parallel HR evolution in weedy rice by mutational convergence cannot be ruled out with the present data.

#### Convergent *in situ* origins of weedy rice from local cultivars

Kinship analysis was carried out to assess geographical origins of weedy rice from each sampled region or country. With the exception of the USA, where weed strains were likely introduced from Asia [2, 19], most weedy rice worldwide appears to have originated from local cultivars or varieties grown in neighboring regions (Additional file 2: Fig. S10; Additional file 1: Table S4). For example, weed accessions from southern China (Jiangsu, Guangdong, and Zhejiang) were inferred to be closest to Chinese cultivated varieties; over half of weeds from

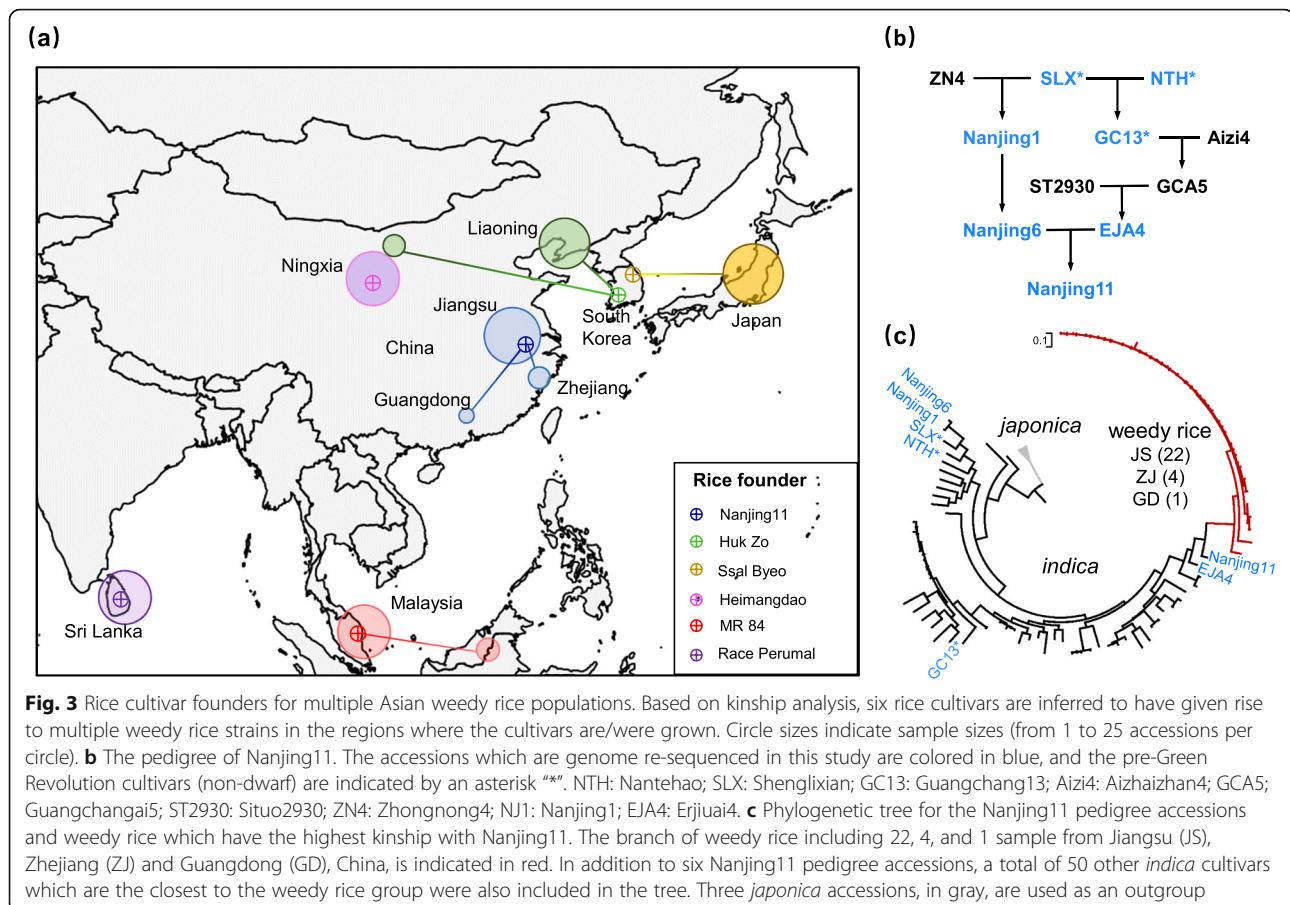
northern China (Liaoning and Jilin) have highest kinship with the cultivars from the nearby Korean peninsula; and Japanese weeds show high kinship with cultivars from South Korea and Japan. Most weeds in Southeast and South Asia also show close relationships with cultivars from local or neighboring countries. In a parallel pattern, Italian weedy rice was inferred to be most closely related to European cultivars.

Notably, the kinship analysis further revealed multiple cases where individual formerly widely-grown cultivars have apparently given rise to the major contemporary weed strains in the region where the cultivar was once grown (Additional file 1: Table S5; examples shown in Fig. 3). For example, a total of 38 weeds from Liaoning in northern China showed highest kinship with a single widely grown twentieth century cultivar “Huk Zo” while 16 Japanese weeds showed highest kinship with “Ssal Byeo.” Both Huk Zo and Ssal Byeo are Korean landraces. For Malaysia, one variety “MR 84,” which was released in 1986 and widely planted during 1980s to 1990s, has a total of 12 weedy rice accessions with closest kinship relationship. In China, one cultivar, “Nanjing11,” was found to be closest to 27 weed strains from South China (Jiangsu, Zhejiang, and Guangdong). This cultivar was bred around 40 years ago in Nanjing, Jiangsu Province; it remained one of the most popular *indica* cultivars and was broadly cultivated throughout South China until about 15 years ago, when it was replaced by newer cultivars. These patterns suggest that a large proportion of Asian weed strains are descended from commercial cultivars that were widely grown in the twentieth century, as rice agriculture shifted from smallholder farms to industrialized production.

To further document this pattern, we collected parental pedigree accessions of Nanjing11 and re-sequenced their genomes (Fig. 3b). The phylogenetic tree confirmed the group of 27 weed strains has closer kinship with Nanjing11 than its pedigree accessions (e.g., EJA4), and the topology supported that the weedy rice group is likely to be derived from Nanjing11, not its parental lines before Green Revolution (e.g., GC13, SLX, and NTH) (Fig. 3c). Extrapolating from these results, we can estimate that more than 35% (27/75) of the current weedy rice strains in southern China (Jiangsu, Zhejiang, Guangdong) are likely descended from Green Revolution cultivars. Taken together, these results indicate that widely grown twentieth century cultivars that were developed during the Green Revolution have left a legacy of weedy rice infestations throughout Asia.

#### Non-domestication genomic regions for adaptation of weedy rice

To identify genomic regions with signatures of adaptive differentiation between weed strains and their inferred cultivated ancestors, genomic scans of differentiation



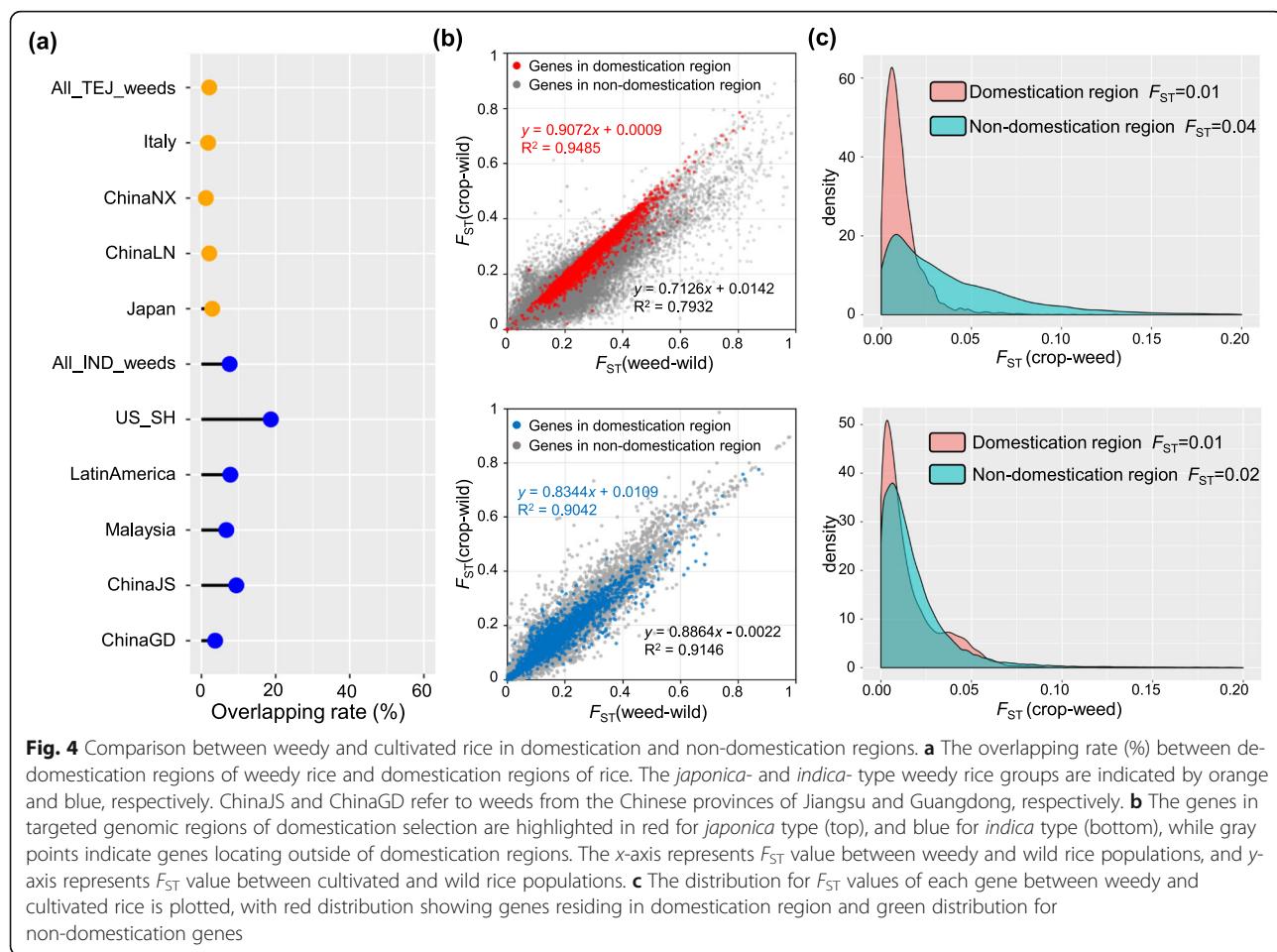
were performed ( $Z(F_{ST}) > 3$ ) (Additional file 2: Fig. S11). We further examined whether these significantly differentiated regions between weedy and cultivated rice overlap with domestication regions. Notably, a very low overlapping rate was observed for most weeds worldwide (the exceptions being regions of South and Southeast Asia where wild rice hybridization has led to adaptive introgression of wild alleles at domestication loci [11–13, 20]). An especially low rate, 2.1% (1.2–2.9%), was observed for *japonica* type weeds, while for *indica* type, the mean overlapping rate (excluding South and Southeast Asia) was 7.6% (3.7–18.7%) (Fig. 4a; Additional file 1: Table S6). In addition, we found that genes in regions not known to be associated with domestication show higher differentiation between weedy and cultivated rice, particularly for *japonica* type weeds (Fig. 4b, c); this indicates that after diverging from cultivated rice, natural selection may act more strongly on the genomic regions unrelated to domestication loci.

Consistent with this pattern of adaptive divergence, we found many mutations in weedy rice that were not observed in cultivated varieties (Additional file 2: Fig. S12). Novel variation in the *ALS* gene for herbicide resistance provides one such example. Among the 11 non-

synonymous *ALS* SNPs identified in weed strains (described above), a mutation (Ala205Val) has not been previously reported in cultivated, wild, or weedy rice (Additional file 1: Table S3). An herbicide resistance assessment revealed that the weed accessions with this *ALS* mutation showed strong herbicide (Imazamox) resistance (Additional file 2: Fig. S9b). The presence of this mutation suggests that weed populations can evolve resistance through new spontaneous mutations.

#### De-domestication blocks under parallel evolution

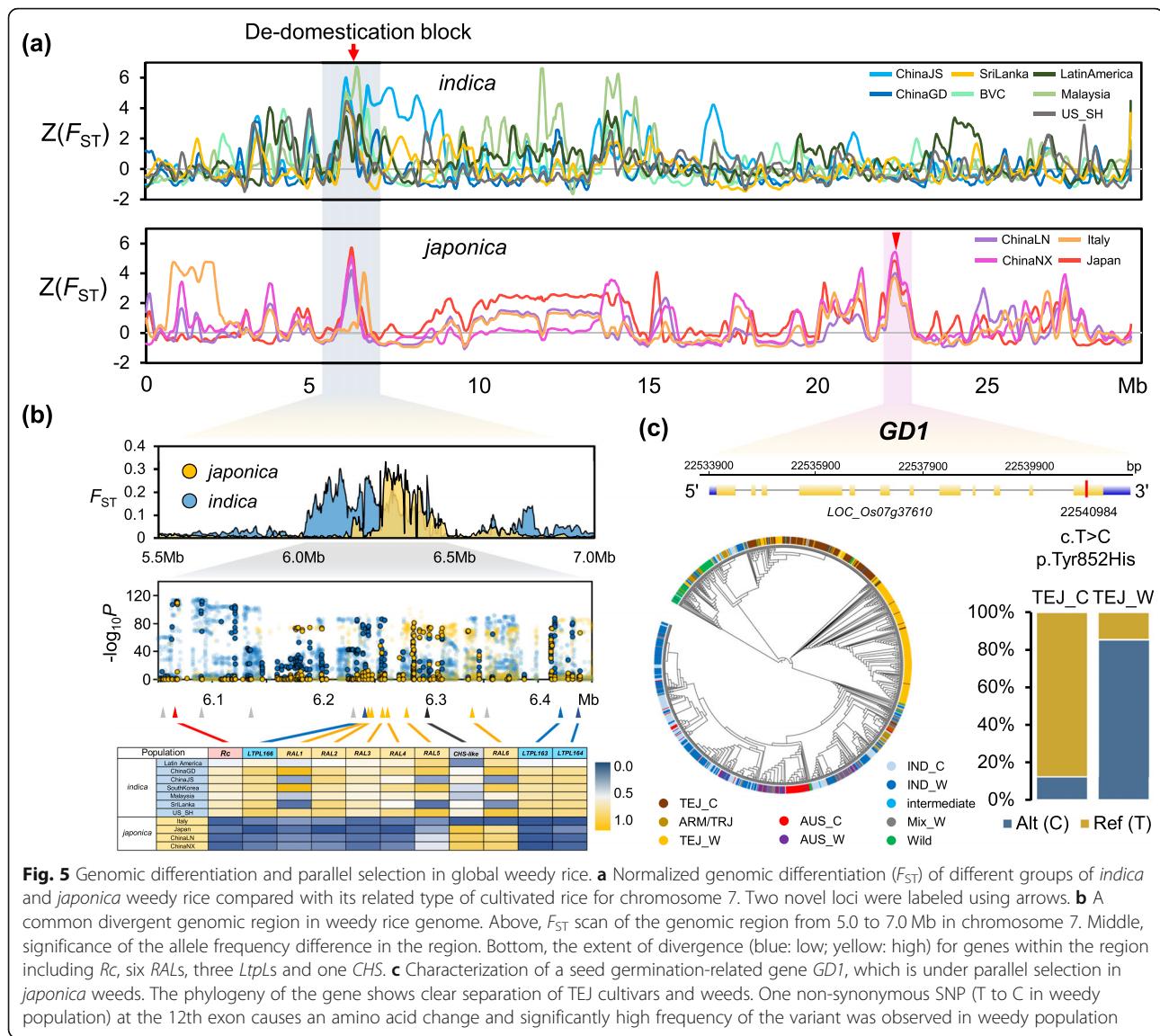
Despite the independent and repeated origins for different weedy rice populations, we can find some shared genomic regions that are highly diverged from cultivated rice, indicating that these regions may underlie shared targets of selection in weed evolution (we refer to these as de-domestication “hot blocks,” to distinguish them from more localized hotspots). One of the most significantly differentiated regions is a 0.5-Mb de-domestication hot block from 6.0 to 6.5 Mb on Chromosome 7 in both *japonica*- and *indica*-type weeds (Fig. 5a). This region harbors multiple genes (e.g., *Rc*, *RAL*, and *LtpL*) with potential functions for environmental adaptation (Fig. 5b). For example, *Rc* pleiotropically controls



both red pericarp and seed dormancy [21]. The red pigment in rice grains is caused by proanthocyanidins or condensed tannins, which could have deterrent effects on pathogens and predators [22]. Seed dormancy is a highly adaptive trait for weedy rice, as it enhances survival of seeds in the soil seed bank and allows seeds to persist in rice fields over multiple seasons [23]. Interestingly, a cluster of six *RAL* (seed allergenic protein) genes and three *LtpL* genes (encoding plant lipid transfer proteins that function in alpha amylase inhibition) are also located within this region. All nine genes harbor the protein domain PF00234 (Protease inhibitor/seed storage/LTP family). These genes are proposed to be involved in multiple roles, such as inhibiting the growth of fungal and bacterial pathogens and facilitating adaptation of plants to various environmental conditions [24], which may protect the weedy rice seeds from pathogens and predators in paddy fields for years during their dormancy. For *indica*-type weeds, each of the *RALs* showed clear differentiation from those in their cultivated ancestors ( $F_{ST} > 0.4$ ) (Additional file 2: Fig. S13). Our results thus suggest that the *RAL* genomic region has been a

repeated target of selection in the evolution of weed strains from cultivated rice.

Another de-domestication hot block occurs in the 22.5–23.1-Mb genomic region of chromosome 7, which stands out as the highest peak when comparing all combined *japonica* weedy rice with cultivars, and also with a Z-value of ~ 4 for each *japonica*-type weedy rice population compared to *japonica* cultivated rice (Fig. 5a; Additional file 2: Fig. S11). Within this region resides a gene encoding B3 domain containing transcription factor *GD1*, which participates in regulating GA and carbohydrate homeostasis, and further regulates rice seed germination and seedling development [25]. The phylogeny of *GD1* clearly shows that most *japonica* weedy rice strains are separated from the group of *japonica* cultivated rice, which is consistent with the haplotype of this gene (Fig. 5c). In addition, we found one non-synonymous SNP on the last exon, and the allele frequency is markedly different between *japonica* weedy (0.85) and cultivated (0.12) rice. The results above suggest that the seed germination-related gene is under potential parallel evolution among different *japonica* weedy rice populations and may play a crucial role for



the distinct germination behavior of weedy rice compared to cultivated rice in rice fields.

## Discussion

In this study, we have found that de-domestication is a primary mechanism for the origin of weedy rice globally, and most weeds are derived from a single genetic founder population (*indica*, *japonica*, or *aus*). In particular, some weedy rice populations can be traced to a single local contemporary cultivar that has apparently served as the founder of the local weedy rice population. Historically, a high diversity of varieties has been cultivated on a local scale in Asia [26–28]. The proliferation of modern Asian weed strains has followed widespread shifts during the Green Revolution towards industrialized rice farming and adoption of a few high-yielding elite cultivars, which our data show to be direct founders of many contemporary

Asian weed strains. Whether there are some special traits and underlying genetic basis in those candidate weed-derived founders is an interesting question for further investigation. The results will provide valuable practical implications for crop breeding and weed control.

We have also observed that the majority of the de-domestication-related loci fall outside the regions which have been targets of selection during domestication. In other words, it appears that regions unrelated to selection during domestication were differentially targeted during the de-domestication process. Interestingly, the finding is more obvious for the recently evolved *japonica* type weedy rice group than *indica* group. A possible explanation could be due to relatively higher diversity of the *indica* group, which would harbor more standing genetic variation that could facilitate adaptation during weed evolution.

Although weedy rice originated independently and repeatedly throughout the rice domestication history, we found genomic evidence for shared targets of selection in different weed populations worldwide (Fig. 5a). Just as Li et al. [2] suggested that weediness of weedy rice can emerge through selection on “genomic islands,” we believe that this “genomic island strategy” could provide weedy rice with high fitness and a rapid adaptive ability, requiring few generations and minimal genomic recombination for weed evolution.

## Conclusions

Based on the largest global weedy rice panel to date, we found weedy rice has evolved multiple times independently from domesticated ancestors, and most Old World strains have *in situ* origins within the regions where they presently occur. Despite the inbreeding nature of cultivated rice, large-scale *indica-aus* hybridization accounts for the origin many Latin American weed strains. Some of the most aggressive contemporary Asian weed strains appear to have very recent origins from now-discontinued Green Revolution elite cultivars. Genome scans suggest that feralization occurs largely through changes at loci that were not targets of selection during domestication. Moreover, independently evolved weed strains have some shared “de-domestication” genomic blocks (including regions containing candidate genes for seed dormancy and allergenic proteins), indicating that there are shared genomic targets of selection in parallel feralization events.

## Methods

### Plant sampling and DNA sequencing

A total of 524 weedy rice samples were collected from 16 countries worldwide where the rice fields are heavily infested by weedy rice. New whole-genome sequence data were generated for 327 weedy rice accessions from nine Asian countries (China, South Korea, Japan, Malaysia, Burma, Vietnam and Cambodia, Sri Lanka, and India), one European country (Italy) and five Latin American countries (Brazil, Panama, Paraguay, Peru, and Uruguay). Paired-end sequence data were generated by Illumina HiSeq4000 and Illumina X Ten. In addition, our previous genome resequencing data, including 197 accessions of weedy rice from the USA and China [2, 3], were also integrated. The average genomic coverage for the 524 weedy accessions is 19.9×. Other genomic data from 479 *Oryza* accessions were obtained from a previous study [9, 10]. These included 159 *indica* varieties; 146 *temperate japonica* varieties; 41 *tropical japonica* varieties; 17 *aromatic* varieties; 49 *aus* varieties; 14 rice varieties of intermediate classification; and 53 wild *O. rufipogon* accessions (Additional file 1: Table S1).

### Variant detection and genotyping

The raw paired-end reads were first filtered into clean data using NGSQCtoolkit v2.3.3 [29]. The cut-off value for PHRED quality score was set to 20 and the percentage of read length that met the given quality was 70. Clean reads of each accession were mapped to rice reference genome (MSUv6.1) using BOWTIE2 v2.2.1 [30] with default settings. Consecutive steps using Samtools v0.1.19 [31] and GATK v3.7 [32] were applied for variant detection. Potential PCR duplicates were removed by “Samtools rmdup.” Alignments around small indels were remapped with “ndelRealigner,” and raw variants were called based on the realigned bam file. Using the called variants as known sites, “BaseRecalibrator” and “PrintReads” in the GATK were applied for base-pair score recalibration. The resulting BAM files of each sample were used for the multi-sample variant genotyping. “UnifiedGenotyper” in GATK was applied to generate the raw variant calls with parameters “-stand\_call\_conf 30, -stand\_emit\_conf 10”. To reduce the variants false discovery rate, the SNP calls were filtered according to the following threshold: QUAL < 30, QD < 2, MQ < 30, MQ0/DP > 0.1. Potential variant annotation and effect were predicted by SnpEff v3.6 [33]. Imputation was performed by BEAGLE v4.0 [34] using the genotype likelihoods and specified the number of iterations to 10.

### Population structure and gene flow analysis

Based on the genome-wide SNPs among the 1003 wild, cultivated, and weedy rice accessions, principal component analysis was performed by SNPRelate v0.9.19 [35]. FastStructure [36] was applied to infer the ancestry of the 1003 rice accessions with *K* values ranging from 3 to 8. The phylogenetic tree was constructed using Fasttree [37] with 1000 bootstrap replicates.

### Derived allele frequency comparison for domestication and non-domestication regions

The rice domestication regions for *japonica* and *indica* rice were retrieved from Huang et al. [9] and repositioned from the genome reference of IRGSP Build 4 to MSUv6.1. The allele frequency of domestication-related and domestication-unrelated regions was calculated for wild, cultivated, and weedy rice, respectively. Genomic data of 10 *O. barthii* samples from Wang et al. [38] were used for genotyping of the SNPs identified in our 1003 rice panel. The genotype of *O. barthii*, which was determined by at least 8 individuals, was used to polarize SNPs as either ancestral or derived. Finally, the ratio of derived allele frequency between domestication-related and domestication-unrelated regions was calculated.

### Origin inference by kinship analysis

Based on the 5.23 million SNP dataset of 4591 cultivated rice which was collected and imputed by Wang et al. [6],

we implemented kinship analysis to infer the origin of each weedy rice sample. As the genome reference of Wang et al. [6] was based on MSUv7, we first transformed the 5.23 million SNP sites from MSUv7 to MSUv6.1. The 150-bp sequence up- and downstream of each SNP site was extracted and aligned to rice genome MSUv6.1 by blast, and transformed the sites by in-house PERL script. We genotyped these sites for all 524 weedy rice using GATKv3.7 [32]. Kinship was calculated with TASSEL5 using Centered IBD method [39]. Based on kinship matrix, we obtained cultivars with highest kinship for each weedy rice sample. Then, we determined the geographical origin of each weedy rice origin based on the location of cultivars.

#### Identification of genomic regions under selection

The genome was scanned in a 100-kb window size with a step size of 10 kb, and the population parameters ( $\pi$ ,  $F_{ST}$ ) were estimated for each window by VCFtools [40]. In the genomic differentiation analyses based on population differentiation index ( $F_{ST}$ ), we compared weedy rice of one group against the totality of its potential ancestors (e.g., Italian weeds vs. all temperate *japonica* cultivars). Z-transformation was applied to locate divergent regions between weedy rice and cultivated rice from the extreme tails by applying a threshold of three standard deviations.

#### Demography inference

The divergence time between cultivated rice and weedy rice of *japonica*, *indica*, and *aus* type was inferred using SMC++ v1.13 [41], which could estimate the effective population size history based on whole-genome sequence data and is powerful for recovering history for short time scales. Three lineages in each population, respectively, were randomly selected as the distinguished individuals to improve the estimation of effective population size. The divergence time range was estimated according to the distinct change of effective size between cultivated and weedy populations. The mutation rate was assumed as  $\mu = 6.5 \times 10^{-9}$  mutations  $\times$  bp $^{-1}$   $\times$  generation $^{-1}$  [42], and 1 year per generation were adopted.

#### Introgression analysis

To examine the possible introgression from wild rice into Malaysian weedy rice, ABBA-BABA statistics (also known as “Patterson’s D” or “D statistics”) were implemented. ABBA-BABA statistics tested for admixture between three closely related populations “P1,” “P2,” and “P3,” using an outgroup population “O” following the phylogeny (((P1, P2), P3), O). In our analysis, *Oryza barthii* was used as outgroup (O), while Malaysian *indica*, weedy rice, and wild rice were served as P1, P2, and P3, respectively. Genotyping data were transferred

to “geno” format using parseVCF.py within genome\_general ([https://github.com/simonmartin/genomics\\_general](https://github.com/simonmartin/genomics_general)). For genome-wide D statistics, genotyping frequency was first calculated with freq.py employed in genome\_general. Then D was calculated by custom R scripts following the equation from Green et al. [43]. The TreeMix v1.13 software [44] was also adopted to explore the potential gene flow between different groups of weedy or cultivated rice.

#### GO enrichment analysis

GO enrichment analysis was carried out using AgriGO with “*Oryza sativa* MSU6.1 nonTE” set as species background (<http://bioinfo.cau.edu.cn/agriGO/>) [45]. The  $P$  value and false discovery rate (FDR) (Yekutieli) criteria of  $< 0.0001$  and  $< 0.05$ , respectively, were used for the considered enrichment GO terms.

#### Imazamox resistance assessment

Clean seeds were sown in plastic boxes (24 cm  $\times$  30 cm  $\times$  9 cm) containing 60% silty loam soil, 15% sand, 15% perlite, and 10% peat by volume and placed in a greenhouse for germination. A population that has never been treated with herbicide was used as reference susceptible population (10–1). After 15 days, the seedlings were thinned to 20 per box and two replicates of 15–20 plants for each population were treated with Imazamox at the recommended field rate of 35 g a.i. ha $^{-1}$  (Altorex, BASF, 40 g a.i. L $^{-1}$ , solution). The herbicide was applied on seedlings at the 3–4 leaf stage, using a bench sprayer equipped with three flat-fan (extended range) hydraulic nozzles (Teejet, 11002) delivering 300 L ha $^{-1}$  at a pressure of 215 kPa. Four weeks after treatment, the number of surviving plants was assessed. Plants were considered dead if, regardless of color, they showed no active growth. Standard error (SE) was calculated for each mean.

#### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-01980-x>.

**Additional file 1: Table S1.** Information of all rice accessions used in this study. **Table S2.** Genetic subgroup assignments of global weedy rice samples used in this study. **Table S3.** Distribution of ALS non-synonymous mutations in wild, cultivar and weedy rice. **Table S3.** Kinship relationship for each weedy rice sample with cultivated rice in the panel including 4591 rice accessions. **Table S5.** The candidate ‘hub’ cultivated rice that has high number of weedy rice accessions with highest kinship. **Table S6.** The overlapping rate between de-domestication regions of weedy rice with rice domestication regions.

**Additional file 2: Figure S1.** Principal component analysis (PCA) plot of 1003 rice accessions including weedy, cultivated and wild rice by the first and second eigenvectors. **Figure S2.** Distribution of D-statistic ( $\pm$ s.e.) across 12 chromosomes for Malaysian weedy rice. **Figure S3.** Genetic diversity ( $\pi$ ) of whole-genome and domestication-related genes in wild, cultivar and weedy rice populations. **Figure S4.** The ratio of allele frequency for SNPs targets of domestication selection versus non-targets for

*indica* type. **Figure S5.** Comparisons of genetic diversity for domestication and improvement genes between weedy and cultivated rice. **Figure S6.** Effective population size change of *indica* and *aus* cultivated and weedy rice. **Figure S7.** Phylogenetic relationship of different weedy rice groups and possible introgression inferred from TreeMix analysis. **Figure S8.** Heterozygosity level evaluated by observed heterozygosity across wild, cultivated and weedy rice populations. **Figure S9.** Weedy rice with ALS-inhibiting herbicide resistance. **Figure S10.** Geographic origins of weedy rice traced by Kinship analysis. **Figure S11.** Genomic differentiation regions for different weedy populations compared to their counterparts in cultivated rice. **Figure S12.** Numbers of SNPs that are shared or unique among weedy and cultivated rice. **Figure S13.** Characterization of one of the genes (*RAL6*) encoding allergenic proteins in the de-domestication genomic block.

**Additional file 3** Review history.

## Acknowledgements

We thank Wei Tang and Yongliang Lu (China National Rice Research Institute) and Chao Bai and Guoyun Ning (Agricultural Bureau of Changxing County) for their help in field collection of weedy rice, and Xuehui Huang (Shanghai Normal University) for his insightful suggestions on how to improve this paper. The authors would like to thank members of the Olsen lab for helpful comments on earlier versions of this manuscript.

## Peer review information

Kevin Pang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the team.

## Review history

The review history is available as Additional file 3.

## Funding

This work was financially supported in part by the National Natural Science Foundation (31971865, 9143511, 31901899), Zhejiang Natural Science Foundation (LZ17C130001), Jiangsu Collaborative Innovation Center for Modern Crop Production and 111 Project (B17039), Zhejiang Key Laboratory of Crop Germplasm Resources to L.F., and the MUM-ASEAN Sustainable Development Research Grant (ASEAN-2019-01-SCI) to B.K.S..

## Availability of data and materials

Short-read sequence data of 275 samples generated by this study have been deposited into GenBank under the bioproject accession number PRJNA606132 (SRR11069989–SRR11070263) [46]. The data for 52 Japanese weedy rice generated in this study have been deposited to DDBJ under accession number PRJDB8989 (DRR209178–DRR209229) [47]. The data of 197 US and Chinese weedy rice generated from our previous research can be downloaded from NCBI with bioproject accession numbers PRJNA344937 [2] and PRJNA295802 [3].

## Authors' contributions

L.F. and K.O. managed and organized the project. J.Q., L.J., D.W., X.W., M.C., L.M., C.Y., and B.J. performed the bioinformatic analyses. A.M., L.C., D.W., L.S., B.K.S., T.I., G.Y., and T.K. collected the samples and carried out the phenotypic analyses. L.S. performed the herbicide resistance experiment. Q. Z., A.M., L.S., B.K.S., J.S., and L.G. gave insightful suggestions and comments on the manuscript. J.Q., L.F., K.O., and D.W. wrote the manuscript. The authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

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

## Author details

<sup>1</sup>Institute of Crop Sciences and Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China.

<sup>2</sup>Shanghai Key Laboratory of Plant Molecular Sciences, College of Life Sciences, Shanghai Normal University, Shanghai 200235, China. <sup>3</sup>Rice Research Institute, Yunnan Agricultural University, Kunming, China. <sup>4</sup>Rice Research Institute, Shenyang Agricultural University, Shenyang, China.

<sup>5</sup>Department of Crop Sciences, Agricultural School, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil. <sup>6</sup>State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou 310006, China.

<sup>7</sup>International Rice Research Institute (IRRI), Manila, Philippines. <sup>8</sup>CSIRO Agriculture and Food, GPO Box 1700, Canberra, ACT 2601, Australia. <sup>9</sup>National Agriculture and Food Research Organization (NARO), Tsukuba, Ibaraki 305-8666, Japan. <sup>10</sup>School of Science, Monash University Malaysia, 46100 Bandar Sunway, Selangor, Malaysia. <sup>11</sup>Istituto per la Protezione Sostenibile delle Pianta (IPSP), CNR, Viale dell'Università, 16, 35020 Legnano, PD, Italy.

<sup>12</sup>Department of Biology, Washington University in St. Louis, St. Louis, MO 63130, USA. <sup>13</sup>James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310058, China.

Received: 22 November 2019 Accepted: 2 March 2020

Published online: 26 March 2020

## References

- Ziska LH, Gealy DR, Burgos N, Caicedo AL, Gressel J, Lawton-Rauh AL, et al. Chapter three - weedy (red) rice: an emerging constraint to global rice production. In: Sparks DL, editor. Advances in Agronomy; 2015. p. 181–228.
- Li L-F, Li Y-L, Jia Y, Caicedo AL, Olsen KM. Signatures of adaptation in the weedy rice genome. *Nat Genet*. 2017;49:811–4.
- Qiu J, Zhou Y, Mao L, Ye C, Wang W, Zhang J, et al. Genomic variation associated with local adaptation of weedy rice during de-domestication. *Nat Commun*. 2017;8:15323.
- He Q, Kim K-W, Park Y-J. Population genomics identifies the origin and signatures of selection of Korean weedy rice. *Plant Biotechnol J*. 2017;15: 357–66.
- Sun J, Ma D, Tang L, Zhao M, Zhang G, Wang W, et al. Population genomic analysis and de novo assembly reveal the origin of weedy rice as an evolutionary game. *Mol Plant*. 2019;12:632–47.
- Wang DR, Agosto-Pérez FJ, Chebotarov D, Shi Y, Marchini J, Fitzgerald M, et al. An imputation platform to enhance integration of rice genetic resources. *Nat Commun*. 2018;9:3519.
- Qi X, Liu Y, Vigueira CC, Young ND, Caicedo AL, Jia Y, et al. More than one way to evolve a weed: parallel evolution of US weedy rice through independent genetic mechanisms. *Mol Ecol*. 2015;24:3329–44.
- Wedger MJ, Topp CN, Olsen KM. Convergent evolution of root system architecture in two independently evolved lineages of weedy rice. *New Phytol*. 2019;223:1031–42.
- Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature*. 2012;490:497–501.
- Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*. 2018; 557:43–9.
- Song B-K, Chuah T-S, Tam SM, Olsen KM. Malaysian weedy rice shows its true stripes: wild *Oryza* and elite rice cultivars shape agricultural weed evolution in Southeast Asia. *Mol Ecol*. 2014;23:5003–17.
- Vigueira CC, Qi X, Song B-K, Li L-F, Caicedo AL, Jia Y, et al. Call of the wild rice: *Oryza rufipogon* shapes weedy rice evolution in Southeast Asia. *Evol Appl*. 2019;12:93–104.
- Wedger MJ, Pusadee T, Wongtamee A, Olsen KM. Discordant patterns of introgression suggest historical gene flow into Thai weedy rice from domesticated and wild relatives. *J Hered*. 2019;110:601–9.
- Chen E, Huang X, Tian Z, Wing RA, Han B. The genomics of *Oryza* species provides insights into rice domestication and heterosis. *Annu Rev Plant Biol*. 2019;70:639–65.
- Roso AC, Jr AM, Delatorre CA, Menezes VG. Regional scale distribution of imidazolinone herbicide-resistant alleles in red rice (*Oryza sativa* L.) determined through SNP markers. *Field Crops Res*. 2010;119:175–82.
- Scarabel L, Cengialta C, Manuello D, Sattin M. Monitoring and management of imidazolinone-resistant red rice (*Oryza sativa* L. var. *sylvestris*) in Clearfield® Italian Paddy Rice. *Agronomy*. 2012;2:371–83.

17. Merotto A Jr, Goulart ICGR, Nunes AL, Kalsing A, Markus C, Menezes VG, et al. Evolutionary and social consequences of introgression of nontransgenic herbicide resistance from rice to weedy rice in Brazil. *Evol Appl.* 2016;9:837–46.
18. Goulart I, Borba TCO, Menezes VG, Merotto A. Distribution of weedy red rice (*Oryza sativa*) resistant to imidazolinone herbicides and its relationship to rice cultivars and wild *Oryza* species. *Weed Sci.* 2014;62:280–93.
19. Reagon M, Thurber CS, Gross BL, Olsen KM, Jia Y, Caicedo AL. Genomic patterns of nucleotide diversity in divergent populations of U.S. weedy rice. *BMC Evol Biol.* 2010;10:180.
20. Huang Z, Young ND, Reagon M, Hyma KE, Olsen KM, Jia Y, et al. All roads lead to weediness: patterns of genomic divergence reveal extensive recurrent weedy rice origins from South Asian *Oryza*. *Mol Ecol.* 2017;26:3151–67.
21. Gu X-Y, Foley ME, Horvath DP, Anderson JV, Feng J, Zhang L, et al. Association between seed dormancy and pericarp color is controlled by a pleiotropic gene that regulates abscisic acid and flavonoid synthesis in weedy red rice. *Genetics.* 2011;189:1515–24.
22. Sweeney MT, Thomson MJ, Pfeil BE, McCouch S. Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell.* 2006;18:283–94.
23. Cui Y, Song BK, Li L-F, Li Y-L, Huang Z, Caicedo AL, et al. Little white lies: pericarp color provides insights into the origins and evolution of Southeast Asian weedy rice. *G3.* 2016;6:4105–14.
24. Kader J-C. Lipid-transfer proteins in plants. *Annu Rev Plant Physiol Plant Mol Biol.* 1996;47:627–54.
25. Guo X, Hou X, Fang J, Wei P, Xu B, Chen M, et al. The rice GERMINATION DEFECTIVE 1, encoding a B3 domain transcriptional repressor, regulates seed germination and seedling development by integrating GA and carbohydrate metabolism. *Plant J.* 2013;75:403–16.
26. Peng S, Tang Q, Zou Y. Current status and challenges of rice production in China. *Plant Prod Sci.* 2009;12:3–8.
27. Morishima H, Sano Y, Oka HI. Differentiation of perennial and annual types due to habitat conditions in the wild rice *Oryza perennis*. *Plant Syst Evol.* 1984;144:119–35.
28. Neik TX, Chai J-Y, Tan S-Y, Sudo MPS, Cui Y, Jayaraj J, et al. When west meets east: the origins and spread of weedy rice between continental and island Southeast Asia. *G3.* 2019;9:2941–50.
29. Patel RK, Jain M. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One.* 2012;7:e30619.
30. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods.* 2012;9:357–9.
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25: 2078–9.
32. McKenna N, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
33. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly.* 2012;6:80–92.
34. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;81:1084–97.
35. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics.* 2012;28:3326–8.
36. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics.* 2014;197:573–89.
37. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009;26:1641–50.
38. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, et al. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat Genet.* 2014;46:982–8.
39. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics.* 2007;23:2633–5.
40. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–8.
41. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet.* 2017;49: 303–9.
42. Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, et al. Molecular evidence for a single evolutionary origin of domesticated rice. *Proc Natl Acad Sci U S A.* 2011;108:8351–6.
43. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. *Science.* 2010;328:710–22.
44. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 2012;8:e1002967.
45. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 2017; 45:W122–9.
46. Qiu J, Jia L, Wu DY, Weng XF, Chen LJ, Sun J, et al. Global weedy rice research project. *Sequence Read Archive.* (2020). <https://www.ncbi.nlm.nih.gov/sra/PRJNA606132>.
47. Qiu J, Jia L, Wu DY, Weng XF, Chen LJ, Sun J, et al. Origin of weedy rice in Japan. *Sequence Read Archive* (2020). <https://www.ncbi.nlm.nih.gov/sra/?term=PRJD88989>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](http://biomedcentral.com/submissions)

