nature genetics

Signatures of adaptation in the weedy rice genome

Lin-Feng Li[1,2], Ya-Ling Li[3], Yulin Jia[4], Ana L Caicedo[5] & Kenneth M Olsen[2]

Crop domestication provided the calories that fueled the rise of civilization[1–3]. For many crop species, domestication was accompanied by the evolution of weedy crop relatives, which aggressively outcompete crops and reduce harvests[4–6]. Understanding the genetic mechanisms that underlie the evolution of weedy crop relatives is critical for agricultural weed management and food security. Here we use whole-genome sequences to examine the origin and adaptation of the two major strains of weedy rice found in the United States. We find that de-domestication from cultivated ancestors has had a major role in their evolution, with relatively few genetic changes required for the emergence of weediness traits. Weed strains likely evolved both early and late in the history of rice cultivation and represent an under-recognized component of the domestication process. Genomic regions identified here that show evidence of selection can be considered candidates for future genetic and functional analyses for rice improvement.

Weedy rice (*Oryza sativa* L.) is a conspecific relative of cultivated rice that occurs in rice fields worldwide. Weedy rice infestations can reduce crop yields by >80% if left unchecked, with estimated annual economic losses exceeding $45 million in the United States alone[7–9]. Widespread adoption of direct-seeded rice farming in place of hand-transplanted fields has promoted global weedy rice proliferation over the last several decades[10–12]. Key weed features include highly shattering seeds, persistent seed dormancy, rapid growth, and the ability to aggressively outcompete the crop for nutrients and light[8,13,14]. Crop–weed hybridization occurs at low frequencies[15–17], and recent agronomic shifts toward the use of herbicide-resistant rice varieties have created strong selection for introgression of resistance alleles into weed populations[18–20]. In regions of tropical Asia where rice is grown in proximity to its wild progenitor (*Oryza rufipogon*), gene flow from wild populations also contributes to the genetic composition of weed populations[21,22].

Genetic surveys around the world indicate that multiple genetically distinct weedy rice strains exist and that these have evolved independently from domesticated and wild relatives[21,23–25]. In the United States, there are two major weed morphotypes, straw hull (SH) and black hull awned (BHA). These strains are genetically distinct and are proposed to have evolved through de-domestication (endoferality) from cultivated *indica* and *aus* rice varieties, respectively, which are the two major genetic subgroups within *O. sativa* subsp. *indica*[24]. As commercially grown rice in the United States belongs to the genetically distinct *japonica* subspecies, the SH and BHA strains likely originated in Asia and were introduced as contaminants of grain stocks[24]. Because these two strains evolved independently, they provide a unique opportunity to study the genetic basis of parallel weediness evolution and the extent to which it occurs through conserved or distinct mechanisms[26–28]. Here we use whole-genome sequence analyses to examine the evolution of weedy rice within the continental United States, with a sample of Chinese weed accessions included for comparison. Analyses reveal different underlying genetic mechanisms in the emergence of weediness traits in the two US strains, with relatively few changes across the genome required for de-domestication to occur.

We generated whole-genome sequences for 18 SH weeds and 20 BHA weeds and analyzed them with 145 previously published *Oryza* genome sequences (Online Methods and **Supplementary Table 1**). Published sequences included 89 cultivated rice accessions representing the five major genetic subgroups (44 *indica*, 16 *aus*, 10 *tropical japonica*, 14 *temperate japonica*, and 5 *aromatic*); 53 accessions of the wild progenitor (43 *O. rufipogon* and 10 accessions of the annual form, *Oryza nivara*); and 3 accessions of weedy rice from central China[29]. A total of 29,408,917 raw SNPs were identified (78.8 SNPs/kb on average, 87% of which were present in wild accessions; **Supplementary Table 2**); subsets of these data were further filtered and used in the analyses described below. Of the total SNPs, 4,912,847 (16.7%) occurred in coding regions and 2,854,339 (9.7%) were nonsynonymous (**Supplementary Table 2**).

To assess the evolutionary relationships of the US weed strains to the other *Oryza* samples, we performed phylogenetic analyses based on 1,381,040 homozygous SNPs in MEGA7 (ref. 30; Online Methods). In the resulting tree, SH accessions formed a monophyletic group with high bootstrap support (96%) and clustered with Southeast Asian *indica* accessions; BHA accessions were grouped with *aus* and wild rice accessions originating from the Indian subcontinent (85% bootstrap support) (**Fig. 1** and **Supplementary Fig. 1**). The three Chinese weed accessions were phylogenetically distinct from the US weeds, clustering instead with Chinese *indica* varieties.

[1]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China. [2]Department of Biology, Washington University in St. Louis, St. Louis, Missouri, USA. [3]Key Laboratory of Molecular Epigenetics of the Ministry of Education (MOE), Northeast Normal University, Changchun, China. [4]US Department of Agriculture–Agricultural Research Service, Dale Bumpers National Rice Research Center, Stuttgart, Arkansas, USA. [5]Department of Biology, University of Massachusetts, Amherst, Massachusetts, USA. Correspondence should be addressed to K.M.O. (kolsen@wustl.edu) or A.L.C. (caicedo@bio.umass.edu).

Received 8 October 2016; accepted 2 March 2017; published online 3 April 2017; doi:10.1038/ng.3825

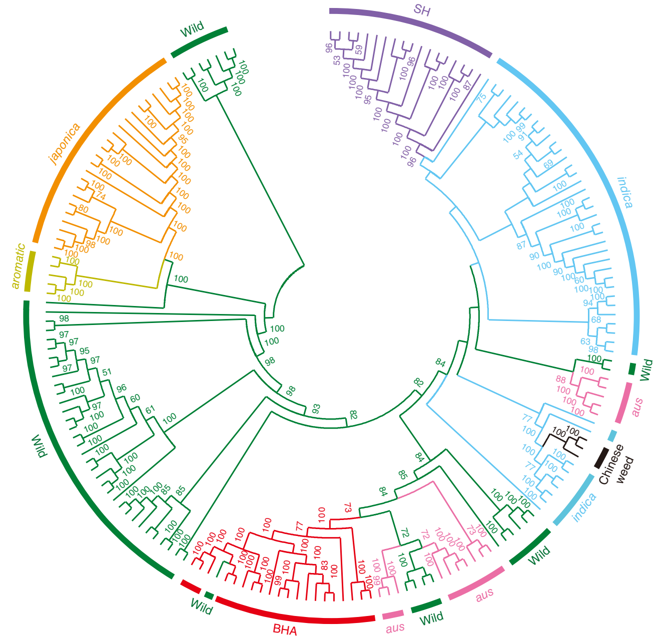

Figure 1 Neighbor-joining tree of the 183 wild, cultivated, and weedy rice accessions. Relationships of cultivated and wild rice correspond to previously observed relationships[40]. Wild rice accessions (dark green) are divided into different groups. The *japonica* (orange) and *aromatic* (light green) rice varieties form a clade. The BHA (red), SH (purple), and Chinese (black) weedy rice strains cluster with *indica* (light blue) and *aus* (pink). Bootstrap values (>50) are shown on each branch.

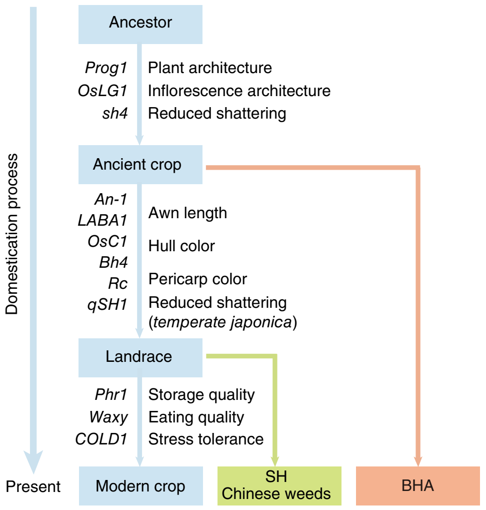

Figure 2 Origin models for SH, BHA, and Chinese weedy rice. The agronomic traits and related genes were selected according to ref. 33.

These inferred relationships were also supported by a maximum-likelihood estimation of individual ancestries performed using ADMIXTURE[31] (**Supplementary Fig. 2** and **Supplementary Table 3**). Genome-wide nucleotide diversity was lower in the US weeds than in their respective inferred crop relatives (**Supplementary Table 4**); this likely reflects demographic bottlenecks with their introductions into North America[24].

While the results above suggest that the weedy rice strains are descended primarily from domesticated ancestors, they do not provide information on how early or late in the history of rice cultivation they emerged. Weeds that diverged from early forms of domesticated rice would be expected to share a greater proportion of genetic variation with the wild progenitor than weeds descended from later, improved varieties. To assess the degree of relatedness of the weed strains to cultivated versus wild rice, we identified the set of SNPs that were present in one or more accessions of cultivated rice but not detected in wild rice (crop-specific private SNPs) and those SNPs having the opposite distribution (wild-specific private SNPs). Among the weeds, the Chinese accessions had the highest proportion of crop-specific private SNPs, a pattern that was evident across all 12 chromosomes (**Supplementary Figs. 3** and **4**, and **Supplementary Table 5**). In contrast, the US weed strains had a higher proportion of wild-specific SNPs than crop-specific SNPs, with BHA strains showing the greatest similarity to wild rice. These patterns suggest that the Chinese weeds are most closely related to modern domesticated rice and that, between the two major US weed strains, the BHA strains are the most dissimilar.

To further explore the timing of weed origins, we used BEAST[32] to estimate the relative divergence times between each weed type and its closest crop relative (**Supplementary Table 6**). Consistent with private SNP distributions, the estimated BHA–*aus* divergence time was earlier than the inferred SH–*indica* divergence (mean relative divergence values, 8.99 × 10^{-3} and 5.66 × 10^{-3}, respectively), while the Chinese–*indica* divergence was slightly later (5.02 × 10^{-3}). Moreover, while the BHA–*aus* divergence was the earliest among the three weed–crop comparisons, it was still later than the estimated *aus*–*indica* varietal divergence within cultivated *O. sativa* ssp. *indica* (mean value, 13.8 × 10^{-3}). This suggests that all three weed strains evolved after rice domestication and after the emergence of some varietal differentiation within the crop.

Crop domestication spans multiple stages, with early human selection favoring 'domestication traits' that distinguish the crop from its wild ancestor (for example, loss of seed shattering and loss of dormancy) and later improvement/diversification stages involving selection for traits that are found in particular varieties (for example, specific environmental adaptations, and flavor and pigment variation)[1,33–35]. This information can be used to assess how early or late in the trajectory of crop domestication a weedy relative diverged from the crop lineage: the more domestication-related genes that show evidence of descent from improved varieties, the later the weed strain is likely to have evolved. Using this reasoning, we selected 12 domestication and improvement genes representing the different stages of rice domestication, all with known mutations or haplotypes for domestication-related traits (**Fig. 2**, **Table 1**, and **Supplementary Table 7**). Both of the US weed strains, as well as the Chinese weed accessions, were fixed for crop-like alleles at three domestication genes that were likely targets of early selection during rice domestication: *PROG1* (erect plant architecture), *sh4* (reduced seed shattering), and *OsLG1* (closed-panicle architecture). The occurrence of domestication alleles at these loci supports de-domestication (endoferality) origins for all three weedy rice groups.

For most of the widely selected improvement genes, the SH and Chinese strains were characterized by the domestication allele while BHA weeds were fixed or nearly fixed for the wild allele (**Table 1** and **Supplementary Table 7**). These distributions are consistent with the inferences above from genome-wide SNPs that BHA weeds evolved earlier in the history of rice domestication than the other strains. Interestingly, allele distributions at the pericarp color gene *Rc* differed from the other widely selected improvement genes; all weeds were characterized by the functional wild allele. This may reflect selection to maintain seed dormancy in weed populations, as *Rc* pleiotropically

Table 1 Causative mutations of domestication genes in the three types of weedy rice

		Causative mutation				
Gene name	Gene function	Wild type	Domestication or improvement allele	BHA	SH	Chinese weeds
Domestication traits						
sh4	Seed shattering	G	T	T	T	T
OsLG1	Closed panicle	G/C/C	A/T/T	A/T/T	A/T/T	A/T/T
Prog1	Plant architecture	Polymorphic	Monomorphic	Monomorphic	Monomorphic	Monomorphic
Widely selected Improvement traits						
OsC1	Apiculus color	WT	10-bp deletion	Mostly WT	10-bp deletion	WT
An-1	Awn length	WT	1-bp deletion	WT	1-bp deletion	Unknown loss of function
LABA1	Awn barb	WT	1-bp deletion	Mostly WT	1-bp deletion	1-bp deletion
Bh4	Hull color	WT	22-bp deletion	WT	22-bp deletion	22-bp deletion
qSH1	Seed shattering (*temperate japonica*)	G	T	G	G	T
Rc	Pericarp color	WT	14-bp deletion	WT	WT	WT
Varietal-specific improvement traits						
Phr1	Phenol reaction	WT	18-bp deletion	WT	WT	WT
Waxy	Amylose content	G	T	G	G	G
COLD1	Chilling tolerance	T/G	A	T	G	G

Mutations are shown for each domestication gene. Detailed information is shown in **Supplementary Table 7**. WT, wild type.

controls both pericarp pigmentation and dormancy[36,37]. For *qSH1*, which controls reduced shattering specifically within *temperate japonica* rice, the presence of the reduced-shattering allele in the Chinese weeds supports a previous conclusion that these strains likely evolved through *indica–japonica* hybridization[29]. For the remaining three varietal-specific improvement genes, none of the weeds carried domestication alleles, indicating that the weeds most likely did not evolve from crop varieties that underwent selection at these loci.

To examine how adaptation has shaped the weedy rice genome, we performed genome-wide selection scans comparing nucleotide variation in 100-kb windows of each United States strain to its inferred crop ancestor. Regions of differentially low nucleotide diversity potentially indicate positive selection associated with weediness adaptation. Notably, the two weeds differ almost entirely in genomic regions of low nucleotide diversity (**Fig. 3**, **Supplementary Fig. 5**, and **Supplementary Tables 8** and **9**). Out of a total of 3,729 windows across the genome, differentially low diversity was detected in 121 (3.3%) and 118 (3.2%) of the SH-*indica* and BHA-*aus* comparisons, respectively, with only 12 windows (0.3%) detected in both weed-crop comparisons. Thus, weediness adaptation appears to be occurring largely through different genetic mechanisms in the two strains.

To identify potential candidate genes in genomic regions with evidence of weed-specific adaptation, we performed selection scans at a finer-scale 10-kb resolution and focused on the subset of windows showing both differentially low nucleotide diversity in a weed strain and differentially high genetic differentiation (F_{ST}) between the weed and its crop ancestor. This yielded 100 and 186 10-kb windows for the SH and BHA strains, respectively, containing a total of 178 and 307 annotated genes (**Supplementary Tables 10** and **11**). The biological processes associated with these genes are diverse, and we found no statistically significant enrichment for specific processes in comparison to genome-wide proportions (**Supplementary Figs. 6** and **7**). However, a non-significant trend toward enrichment of genes involved in tissue development and stress response (**Supplementary Tables 10** and **11**) is consistent with observations that weedy rice grows faster and shows higher stress tolerance than cultivated rice[14,23,38].

Interestingly, most of the candidate genes for weed adaptation were clustered as 'genomic islands' rather than randomly distributed across the SH and BHA genomes (**Supplementary Tables 10** and **11**). This suggests that selection during weed evolution may have acted on clusters of loci in relatively few genomic regions. To assess whether these genomic islands of adaptation colocalize with known loci that control weediness traits, we compared their locations with previously mapped quantitative trait loci (QTLs) for traits that distinguish the weeds from cultivated rice[26,28]. Several of the genomic regions overlapped with our previously mapped weediness QTLs (**Fig. 3**); this included, for example, both shattering QTLs previously identified in SH (*qSH2S* and *qSH11S*) and three of six shattering QTLs for BHA (*qSH1Bb*, *qSH3Bb*, and *qSH5Bb*). These results are further supported by substantial QTL overlap with genomic regions identified as showing selective sweeps in a SweeD[39] analysis (**Fig. 4**). Together, these findings suggest that novel mutations associated with weedy rice adaptive traits have been selected on during the de-domestication process.

In summary, our genome sequence analyses indicate that de-domestication from cultivated ancestors accounts for the origins of the two

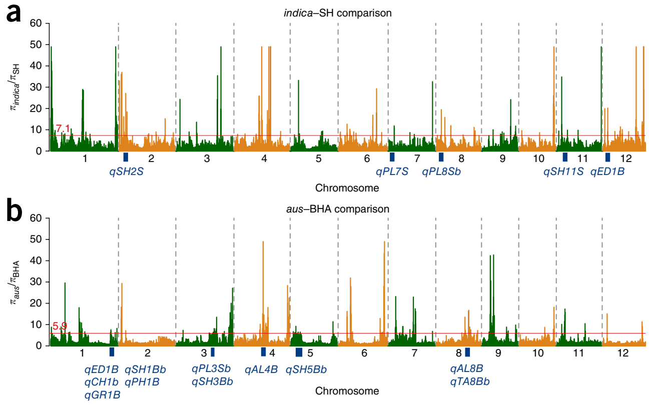

Figure 3 Candidate genomic regions under selection in weedy rice. (**a**,**b**) Decreased nucleotide diversity (π) in SH (**a**) and BHA (**b**) weedy rice genomes in comparison to their crop ancestors. The *y* axis represents the ratio of nucleotide diversity between cultivated (*indica* and *aus*) and weedy (SH and BHA) rice. Each bar is a 100-kb window, and the solid red line indicates the cutoff for the top 5% of windows. The number on the red line is the exact threshold. The blue bars on the bottom correspond to our previously identified QTLs that overlap candidate genomic regions. The length of each bar corresponds to its confidence interval on each chromosome.

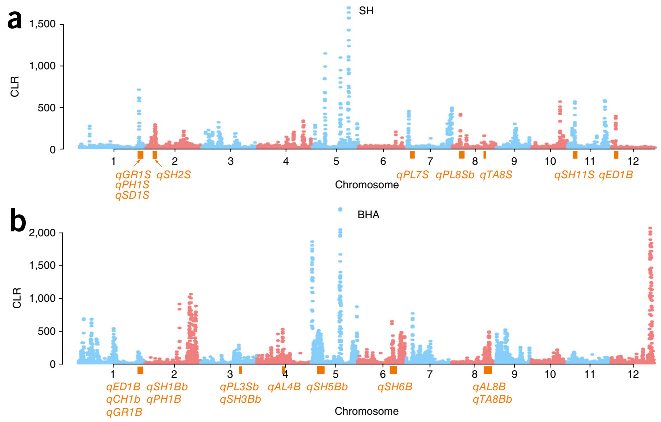

Figure 4 Selective sweeps in the weedy rice genome. (**a**,**b**) Genome-wide scan of selective sweeps in SH (**a**) and BHA (**b**) weedy rice. Each chromosome was divided into 2,000 windows, and each dot corresponds to one window. The *y* axis shows the composite likelihood ratio (CLR) value for each window in tests of deviation from neutrality in SweeD analysis. The orange bars along the *x* axis correspond to our previously identified QTLs that overlap regions with high CLR values. The length of each bar represents its confidence interval on each chromosome.

major US weedy rice strains, as well as the Chinese weeds included here. Genome-wide SNPs and domestication genes together suggest that weed evolution has occurred both early and late in the domestication process and that weediness can emerge through selection on relatively few genomic regions. The apparent ease with which this aggressive agricultural weed can repeatedly evolve should sound a note of caution as global rice agriculture continues shifts toward the mechanized production practices that promote its persistence and proliferation[10–12].

URLs. FastQC, https:/www.bioinformatics.babraham.ac.uk/projects/fastqc/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Dyer and staff of the University greenhouse for assistance in growing plants and L. Small for technical laboratory assistance. We are grateful to members of the Olsen laboratory for helpful comments on this manuscript. This work is supported by the National Science Foundation Plant Genome Research Program (IOS-1032023). The USDA is an equal opportunity provider and employer.

AUTHOR CONTRIBUTIONS

K.M.O., Y.J. and A.L.C. designed the experiments. L.-F.L. and Y.-L.L. analyzed the data. L.-F.L., K.M.O. and A.L.C. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
2. Ross-Ibarra, J., Morrell, P.L. & Gaut, B.S. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci. USA* **104**, 8641–8648 (2007).
3. Larson, G. *et al.* Current perspectives and the future of domestication studies. *Proc. Natl. Acad. Sci. USA* **111**, 6139–6146 (2014).
4. Vigueira, C., Olsen, K. & Caicedo, A. The red queen in the corn: agricultural weeds as models of rapid adaptive evolution. *Heredity* **110**, 303–311 (2013).
5. De Wet, J.M. & Harlan, J.R. Weeds and domesticates: evolution in the man-made habitat. *Econ. Bot.* **29**, 99–108 (1975).
6. Ellstrand, N.C. *et al.* Crops gone wild: evolution of weeds and invasives from domesticated ancestors. *Evol. Appl.* **3**, 494–504 (2010).
7. Oerke, E.C. Crop losses to pests. *J. Agric. Sci.* **144**, 31–43 (2006).
8. Estorninos, L.E. Jr., Gealy, D.R., Gbur, E.E., Talbert, R.E. & McClelland, M.R. Rice and red rice interference. II. Rice response to population densities of three red rice (*Oryza sativa*) ecotypes. *Weed Sci.* **53**, 683–689 (2005).
9. Gealy, D.R. & Yan, W. Weed suppression potential of 'Rondo' and other *indica* rice germplasm lines. *Weed Technol.* **26**, 517–524 (2012).
10. Chauhan, B.S. Strategies to manage weedy rice in Asia. *Crop Prot.* **48**, 51–56 (2013).
11. Hill, J., Smith, R.J. & Bayer, D. Rice weed control: current technology and emerging issues in temperate rice. *Anim. Prod. Sci.* **34**, 1021–1029 (1994).
12. Ziska, L.H. *et al.* Chapter three—weedy (red) rice: an emerging constraint to global rice production. *Adv. Agron.* **129**, 181–228 (2015).
13. Basu, C., Halfhill, M.D., Mueller, T.C. & Stewart, C.N. Weed genomics: new tools to understand weed biology. *Trends Plant Sci.* **9**, 391–398 (2004).
14. Burgos, N.R., Norman, R.J., Gealy, D.R. & Black, H. Competitive N uptake between rice and weedy rice. *Field Crops Res.* **99**, 96–105 (2006).
15. Gealy, D.R. in *Crop Ferality and Volunteerism* (ed. Gressel, J.) 323–354 (CRRC Press, 2005).
16. Shivrain, V.K. *et al.* Gene flow between Clearfield™ rice and red rice. *Crop Prot.* **26**, 349–356 (2007).
17. Shivrain, V.K., Burgos, N.R., Gealy, D.R., Moldenhauer, K.A. & Baquireza, C.J. Maximum outcrossing rate and genetic compatibility between red rice (*Oryza sativa*) biotypes and Clearfield™ rice. *Weed Sci.* **56**, 807–813 (2008).
18. Burgos, N.R. *et al.* The impact of herbicide-resistant rice technology on phenotypic diversity and population structure of United States weedy rice. *Plant Physiol.* **166**, 1208–1220 (2014).
19. Lu, B.R., Yang, X. & Ellstrand, N.C. Fitness correlates of crop transgene flow into weedy populations: a case study of weedy rice in China and other examples. *Evol. Appl.* **9**, 857–870 (2016).
20. Merotto, A. *et al.* Evolutionary and social consequences of introgression of nontransgenic herbicide resistance from rice to weedy rice in Brazil. *Evol. Appl.* **9**, 837–846 (2016).
21. Song, B.K., Chuah, T.S., Tam, S.M. & Olsen, K.M. Malaysian weedy rice shows its true stripes: wild *Oryza* and elite rice cultivars shape agricultural weed evolution in Southeast Asia. *Mol. Ecol.* **23**, 5003–5017 (2014).
22. Pusadee, T., Schaal, B.A., Rerkasem, B. & Jamjod, S. Population structure of the primary gene pool of *Oryza sativa* in Thailand. *Genet. Resour. Crop Evol.* **60**, 335–353 (2013).
23. Londo, J. & Schaal, B. Origins and population genetics of weedy red rice in the USA. *Mol. Ecol.* **16**, 4523–4535 (2007).
24. Reagon, M. *et al.* Genomic patterns of nucleotide diversity in divergent populations of US weedy rice. *BMC Evol. Biol.* **10**, 180 (2010).
25. Grimm, A., Fogliatto, S., Nick, P., Ferrero, A. & Vidotto, F. Microsatellite markers reveal multiple origins for Italian weedy rice. *Ecol. Evol.* **3**, 4786–4798 (2013).
26. Thurber, C.S., Jia, M.H., Jia, Y. & Caicedo, A.L. Similar traits, different genes? Examining convergent evolution in related weedy rice populations. *Mol. Ecol.* **22**, 685–698 (2013).
27. Vigueira, C., Li, W. & Olsen, K. The role of *Bh4* in parallel evolution of hull colour in domesticated and weedy rice. *J. Evol. Biol.* **26**, 1738–1749 (2013).
28. Qi, X. *et al.* More than one way to evolve a weed: parallel evolution of US weedy rice through independent genetic mechanisms. *Mol. Ecol.* **24**, 3329–3344 (2015).
29. Qiu, J. *et al.* Genome re-sequencing suggested a weedy rice origin from domesticated *indica–japonica* hybridization: a case study from southern China. *Planta* **240**, 1353–1363 (2014).
30. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
31. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
32. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
33. Meyer, R.S. & Purugganan, M.D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
34. Olsen, K.M. & Wendel, J.F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
35. Olsen, K.M. & Wendel, J.F. Crop plants as models for understanding plant adaptation and diversification. *Front. Plant Sci.* **4**, 1–16 (2013).
36. Cui, Y. *et al.* Little white lies: pericarp color provides insights into the origins and evolution of southeast Asian weedy rice. *G3* **6**, 4105–4114 (2016).
37. Gu, X.Y. *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* **189**, 1515–1524 (2011).
38. Reagon, M., Thurber, C.S., Olsen, K.M., Jia, Y. & Caicedo, A.L. The long and the short of it: *SD1* polymorphism and the evolution of growth trait divergence in US weedy rice. *Mol. Ecol.* **20**, 3743–3756 (2011).
39. Pavlidis, P., Živković, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224–2234 (2013).

ONLINE METHODS

Sampling and sequencing. We sequenced the genomes of 38 US weedy rice accessions (18 SH and 20 BHA) at >16× genome coverage (**Supplementary Table 1**). For previously published sequences, a total of 145 *Oryza* genomes were downloaded from GenBank according to published references[29,40,41]. We selected these accessions on the basis of geographical distribution and high genome coverage; 85% of the selected accessions had >5× average genome coverage. Genomic DNA was extracted from mature leaves for each accession using DNeasy Plant Mini kits (Qiagen) following the manufacturer's instructions. DNA libraries were constructed by Novogene and sequenced with the Illumina HiSeq 2000 platform.

Sequence alignment and genotype calling. Raw reads for the 183 accessions were assessed for quality using FastQC. The rice reference genome and gene annotations of Nipponbare (temperate *japonica*; MSU6.0 assembly) were downloaded from the Rice Genome Annotation Project website (http://rice.plantbiology.msu.edu/). The clean reads (base quality > 20) were then mapped onto the Nipponbare (*temperate japonica*) reference genome (MSU 6.0 release) using BWA[42] with the parameter set as 'bwa aln -n 0.05'. Indels were realigned with Genome Analysis Toolkit (GATK) IndelRealigner v.2.6 (ref. 43). Raw variants (SNPs and indels) were called using SAMtools[44] with parameters 'mpileup --Dsugf' and 'bcftools view --Ncvg'. A consensus sequence was generated for each accession using a series of Perl scripts designed to eliminate variant (SNP and indel) calls that are artifacts of sequencing or mapping errors. We obtained a total of 29,408,917 raw SNPs from the 183 rice accessions, 25,680,309 of which were generated from the 53 wild rice accessions (**Supplementary Table 2**).

Data quality control. Because the number of SNPs obtained in our study was much higher than those from previous studies[40,45], we filtered the raw variants with the following three strategies to ensure data quality: (1) strategy I: filtering of the raw variants using three different stringency criteria. We noted that the genome coverage of the 183 rice accessions used in this study varied widely, ranging from ~4× in the accession W1757 to ~122× in the accession IRGC81940 (**Supplementary Table 1**). To eliminate any systematic bias caused by variable genome coverage, we filtered the SNPs and indels obtained according to three different criteria corresponding to increasing stringency levels: (i) mapping quality (MQ) ≥ 10, read depth (DP) ≥ 1; (ii) MQ ≥ 20, DP ≥ 3; and (iii) mapping quality ≥ 10, read depth ≥ 1, excluding all variants with rare (<2%) heterozygosity. For the first criterion, mapping quality was used to filter the reported variants and read depth was not considered. In this case, we obtained a total of 32,043,559 raw variants (SNPs and indels) from the 183 rice accessions. Of these variants, 28,043,559 and 12,370,713 were generated from the 53 wild and 89 cultivated rice accessions, respectively (**Supplementary Table 12**). For the second criterion, only the variants with mapping quality ≥20 and read depth ≥3 were retained. This filtering criterion generated 23,474,413 and 11,075,988 raw variants from the wild and cultivated accessions, respectively (**Supplementary Table 12**). For the third criterion, the raw variants (MQ ≥ 10, DP ≥ 1) with low-frequency (< 2%) heterozygosity were excluded. This criterion yielded 22,217,427 and 11,479,934 variants from the wild and cultivated accessions, respectively (**Supplementary Table 12**). To test whether different filtering criteria had significant effects on the subsequent data analyses, we calculated the nucleotide diversity (π) and genetic differentiation (F_{ST}) between wild, cultivated, and weedy rice for non-overlapping 100-kb windows using VCFtools[46]. Distribution patterns and bar plots of F_{ST} and π ratios were generated using R scripts (**Supplementary Figs. 8–13**). Our results showed that the three data sets produced very similar patterns of nucleotide diversity and genetic differentiation. These observations suggested that, although stricter filtering criteria yielded fewer variants, all the criteria generated very similar patterns.

(2) Strategy II: application of filtering methods used in previous studies. As the two higher-stringency criteria did not change the results in preliminary analyses, results are reported for the data set generated with the first criterion (MQ ≥ 10, DP ≥ 1). However, we noted that our pipeline generated far more raw variants than those of previous studies, especially for the wild rice accessions. For example, Xu *et al.*[45] sequenced 10 wild and 40 cultivated rice accessions with >10× genome coverage, which yielded ~15 million candidate SNPs. Of those raw SNPs, the 10 wild rice accessions yielded ~5.2 million high-quality SNPs. The lower SNP numbers obtained by Xu *et al.* reflect the fact that they removed variants if any data were missing in any of their 50 rice accessions. Using this approach, we also filtered out raw variants with missing data in any of the 183 accessions. Only 4,606,171 and 1,895,403 variants were retained in the wild and cultivated accessions, respectively (**Supplementary Table 12**). To then test whether the missing data had detectable effects on subsequent population genetic inference, we performed estimation of individual ancestries based on both the full and filtered (no missing data) data sets separately. Our results showed that the two data sets generated very similar genetic assignments for the 183 rice accessions (**Supplementary Fig. 2**). This indicates that inclusion of sites with missing data had no obvious effects on the subsequent data analyses. Given this result and because exclusion of variants with missing data would lead to differential loss of variants from wild accessions, our analyses were performed using the data set that included variants with missing data.

As part of their analysis, Huang *et al.*[40] also sequenced 446 wild and 1,083 cultivated rice accessions with low genome coverage, which yielded a total of 7,970,359 non-singleton SNPs across all rice accessions. We therefore removed low-frequency (<1%) variants (including SNPs and indels) from our raw data set. This yielded 11,945,007 and 8,457,444 variants in the wild and cultivated rice accessions, respectively (**Supplementary Table 12**). Given that most of the rice accessions we included in our study had high genome coverage, we believe that it is reasonable that these data yielded more variants than in previous publications[40,45]. We further noted that more than half of the variants identified in wild rice occurred at low frequency (<1%). To examine whether low-frequency alleles can lead to data bias for subsequent inference, we compared the nucleotide variation pattern ($\pi_{wild}/\pi_{cultivar}$) generated from our raw variants with published data[40] (**Supplementary Fig. 14** and **Supplementary Table 13**). We found that the genomic regions that showed decreases in nucleotide diversity in cultivated rice were similar to previous observations[40]. Shifts of some genomic regions are mainly due to differences in the rice reference genomes used in analyses. These comparisons together suggested that, although relatively more raw variants were identified in our study, these low-frequency variants showed no effects on subsequent data analyses.

(3) Strategy III: comparison of the Illumina data set with Sanger sequences. To further check the quality of the reported variants, we converted our VCF (MQ ≥ 10, DP ≥ 1) data set for the *Rc* gene (~7 kb in length) to FASTA alignment and then compared the Illumina data set with our previous Sanger sequences for the same samples of US weeds (**Supplementary Data**). We found that the identified SNPs for BHA and SH were identical in our Illumina and Sanger sequencing data sets. In addition, some low-frequency SNPs for the wild rice reported in the Illumina data set were also found in the Sanger sequences. These results together suggested that the SNPs for the *Rc* gene reported in this study are real.

On the basis of the results of implementing these three strategies, we selected the MQ ≥ 10, DP ≥ 1 data set and then customized downstream filtering methods for the different analyses to minimize analysis-specific risks of systematic bias. Phylogenetic analyses are especially susceptible to biases introduced by low sequence coverage. In the case where stricter filtering criteria (for example, DP ≥ 3) were employed, the mapped variants with low read depth (DP < 3) were prone to be erroneously assigned the sequence of the reference *japonica* genome. At the same time, exclusion of low-depth variants altogether can lead to a loss of phylogenetic resolution. Thus, for the phylogenetic analysis, only the 1,381,040 SNPs that were homozygous and without missing data across the 183 rice accessions were used to construct the neighbor-joining tree. The same data set was also applied to estimate the divergence times between the wild, cultivated, and weedy rice groups. 6,360,057 and 2,053,581 private SNPs (MQ ≥ 10, DP ≥ 1) were identified in wild and cultivated rice, respectively. Of these private SNPs, only 377,082, 454,078, and 121,701 high-frequency (>1%) SNPs were used to infer the nucleotide variation patterns of SH, BHA, and Chinese weeds, respectively (**Supplementary Table 5**). For the selective sweep analysis performed using SweeD[39] only the variants with MQ ≥30, base quality ≥30, and DP ≥3 were used to detect selection signals across the SH and BHA genomes.

The above comparisons demonstrated that the different filtering criteria showed no substantial effects on estimation of nucleotide diversity and

doi:10.1038/ng.3825

genetic differentiation. Equally notably, the SNPs for the *Rc* gene generated in the Illumina data set were identical to their Sanger sequences. Furthermore, our results showed that the number and frequency of variants detected in weedy and cultivated rice were reasonable in comparison to previous studies. We therefore employed the variants with MQ ≥10 and DP ≥1 to calculate nucleotide diversity and differentiation between weedy rice (SH and BHA) and their closest crop ancestors (**Fig. 3**, **Supplementary Fig. 5**, and **Supplementary Tables 8**–**11**).

Population and phylogenetic analyses. Genome-wide nucleotide diversity (π) and SNP density were calculated using VCFtools v0.1.12 (ref. 46) for non-overlapping 100-kb windows across the genome. Watterson's[47] estimator of nucleotide diversity (θ_W) was calculated using Perl scripts. The program ADMIXTURE[31] was used to infer the optimum number of clusters and to assess the ancestry of the 183 rice accessions. The ancestry populations were inferred with K values from two to five, and the cross-validation error was used to select the best ancestry population (**Supplementary Table 14**). In addition, we converted all reported SNPs in the 183 accessions from variant call format (VCF) into FASTA format using Perl scripts. SNPs that were homozygous and without missing data across the 183 rice samples were used to construct a neighbor-joining tree with MEGA7 (ref. 30). For domestication gene analyses, full-length contigs of the selected loci were generated that included all SNPs and indels. Alignments of these genes were generated using Clustal X v2.0 (ref. 48).

Identification of genomic regions under selection. A series of Perl scripts was used to translate the nucleotide matrix into an amino acid matrix. Then, synonymous and nonsynonymous sites were identified on the basis of distributions of inferred amino acid replacements. Similarly, the locations of the variants in UTRs, exons, introns, and intergenic regions were determined on the basis of annotated gene information (MSU6.0). Pairwise genetic differentiation values for total sites were generated for each window across the genome among weedy and cultivated rice using VCFtools[46]. The reduction of nucleotide diversity between weedy rice and their crop ancestors ($\pi_{\mathrm{crop}}/\pi_{\mathrm{weed}}$ and $\theta_{\mathrm{crop}}/\theta_{\mathrm{weed}}$) was estimated for non-overlapping windows across the genome using VCFtools[46] and Perl scripts.

To test whether allele frequency has measurable effects on nucleotide diversity, we calculated the ratio of π and θ_W between the two types of weedy rice and their crop ancestors for each 100-kb window. We found that almost all windows with a high value for the π ratio ($\pi_{\mathrm{crop}}/\pi_{\mathrm{weed}}$) also showed a high value for the θ_W ratio ($\theta_{W\ \mathrm{crop}}/\theta_{W\ \mathrm{weed}}$) (**Supplementary Fig. 15**), indicating that there was no obvious systematic bias within our data set. To further eliminate data bias, only windows that were among the 5% with the highest π and θ_W ratio values were defined as candidate regions (**Supplementary Tables 8** and **9**). To identify candidate genes, we calculated the ratios of π and θ_W between weedy and cultivated rice for each 10-kb non-overlapping window across the genome. Our phylogenetic and population genetic analyses showed the de-domestication origins of the three weedy rice strains. Under this hypothesis, the decreased nucleotide diversity might have been inherited from their crop ancestors rather than been acquired during weed adaptation processes. We therefore estimated the genetic differentiation (F_{ST}) between weedy and cultivated rice for each 10-kb non-overlapping window. The top 5% of windows with the highest F_{ST} values were treated as candidates. The overlaps between windows with low nucleotide diversity and high genetic differentiation were selected as candidate genomic regions (**Supplementary Tables 10** and **11**). Candidate genes within these selective genomic regions and their biological functions were retrieved according to annotations from MSU6.0. We also employed a more stringent cutoff (2.5% of top windows) to detect candidate windows across the SH and BHA genomes. This stringent criterion identified only 66 (0.18% of the BHA genome) and 46 (0.12% of the SH genome) 10-kb windows in the two weedy strains (**Supplementary Tables 10** and **11**). To further determine whether these candidate windows had undergone selective sweeps before the evolution of weedy rice, we examined whether these candidate windows (**Supplementary Tables 10** and **11**) have been identified as undergoing selective sweeps in the cultivated rice genome (**Supplementary Table 13**). Our analyses found only one candidate window in SH (chromosome 1: 25,690,000–25,700,000 bp) adjacent to a selective sweep region in the cultivated rice genome (chromosome 1: 25,700,000–25,800,00 bp), suggesting that most of the 10-kb candidate windows were associated with the adaptive evolution of weedy rice rather than inherited from crop ancestors.

Distribution patterns, bar plots of F_{ST} values, and ratios of nucleotide diversity were generated using R scripts. The functions of the identified candidate genes were determined according to MSU annotations (http://rice.plantbiology.msu.edu/). In addition, we also employed SweeD[39] to detect selective sweeps in the SH and BHA weedy rice genomes. This program employs the CLR statistic[49] and identifies signals of selective sweeps by significant deviations from the neutral site frequency spectrum (SFS).

Divergence time and private SNP calculations. The 183 rice accessions were divided into wild, cultivated, and weedy rice groups. We estimated the numbers of crop-specific and wild-specific private SNPs in each 100-kb window across the SH, BHA, and Chinese weedy rice genomes. Distributions of these private SNPs were visualized by plotting the log value of the ratio between crop- and wild-specific private SNPs using an R script. A positive value indicates that there are more crop-specific private SNPs than wild-specific private SNPs within the genomic window. To estimate the divergence time between wild, weedy, and cultivated rice, we selected a total of 54 rice accessions on the basis of geographical location and phylogenetic position. A neighbor-joining tree of the 54 accessions was constructed using MEGA7 (ref. 30) (**Supplementary Fig. 16**). The topologies of the 54 rice accessions are similar to those of the 183 rice accessions. Divergence times between crop and weedy rice were estimated on the basis of homozygous SNPs using BEAST[32].

Data availability. All Illumina data for US weedy rice that support the findings of this study have been deposited in the Sequence Read Archive (SRA) under accession SRR4334499.

40. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
41. The 3,000 Rice Genomes Project Group. The 3,000 Rice Genomes Project. *Gigascience* **3**, 7 (2014).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. McKenna, A. *et al.* The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
44. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
45. Xu, X. *et al.* Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2012).
46. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
47. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
48. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
49. Nielsen, R. *et al.* Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* **19**, 838–849 (2009).