nature plants

Article

https://doi.org/10.1038/s41477-023-01476-z

Multiple domestications of Asian rice

Received: 16 August 2022

Accepted: 4 July 2023

Published online: 7 August 2023

Check for updates

Chun-Yan Jing[1,2,5], Fu-Min Zhang[1,2,5], Xiu-Hua Wang[1,2,5], Mei-Xia Wang[1,2,5], Lian Zhou[1], Zhe Cai[1], Jing-Dan Han[1], Mu-Fan Geng[1,2], Wen-Hao Yu[1,2], Zi-Hui Jiao[1,2], Lei Huang[1], Rong Liu[1,3], Xiao-Ming Zheng[1,3], Qing-Lin Meng[1,2], Ning-Ning Ren[1,2], Hong-Xiang Zhang[1,2], Yu-Su Du[1,2], Xin Wang[1,2], Cheng-Gen Qiang[1,2], Xin-Hui Zou[1,2], Brandon S. Gaut[4] & Song Ge[1,2] ✉

The origin of domesticated Asian rice (*Oryza sativa* L.) has been controversial for more than half a century. The debates have focused on two leading hypotheses: a single domestication event in China or multiple domestication events in geographically separate areas. These two hypotheses differ in their predicted history of genes/alleles selected during domestication. Here we amassed a dataset of 1,578 resequenced genomes, including an expanded sample of wild rice from throughout its geographic range. We identified 993 selected genes that generated phylogenetic trees on which *japonica* and *indica* formed a monophyletic group, suggesting that the domestication alleles of these genes originated only once in either *japonica* or *indica*. Importantly, the domestication alleles of most selected genes (~80%) stemmed from wild rice in China, but the domestication alleles of a substantial minority of selected genes (~20%) originated from wild rice in South and Southeast Asia, demonstrating separate domestication events of Asian rice.

The elucidation of crop domestication is critical for gaining insights into the process of evolution and crop improvement and for understanding the progression of human civilization[1–6]. Asian rice has been studied extensively given its importance as one of the most important staple foods worldwide and as an excellent model for biological research[7–13]. However, the fundamental question of whether the two major cultivar groups of Asian rice (*japonica* and *indica*) were domesticated independently or only once remains unsolved, with essentially two alternative hypotheses[7,9,14,15]. The multiple-domestication hypothesis argues that *japonica* and *indica* were domesticated independently from different wild lineages, and this viewpoint has been supported by numerous genome-wide studies[16–22]. In these studies, *japonica* and *indica* represent two deeply differentiated gene pools and are more closely related to different wild lineages than to each other. In contrast, the single-domestication hypothesis posits that rice domestication was a single event associated with a single wild lineage, because studies of domestication genes and alleles (that is, the alleles that contribute to domestication phenotypes[1]) have found that *japonica* and *indica* share the same domestication alleles[9,14,15,23]. However, as outlined by two competing models in Sang and Ge[15], the shared domestication alleles imply a single domestication event only if they all originated from the same wild lineage. Alternatively, if the alleles arose from one wild lineage for some genes and from another wild lineage for some other genes, this implies multiple domestication events followed by introgression between two cultivar groups. The debate therefore hinges on the question whether all domestication alleles shared by *japonica* and *indica* arose from a single or multiple wild lineages[9,13–15]. Although there have been in-depth studies of rice domestication genes, it seems unlikely that this issue can be resolved by investigating a limited number of such genes from an incomplete representation of wild rice[13–15].

Several additional factors further complicate inferences of rice domestication history. First, there is no consensus as to whether the Asian rice progenitors (*O. rufipogon* and *O. nivara*) were two distinct species and whether these species are geographically and genetically subdivided[13,14,24–26]. Many studies have treated them as a homogenous gene pool[27–29], leading to potential inferential errors (Supplementary

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. [4]University of California, Irvine, Irvine, CA, USA. [5]These authors contributed equally: Chun-Yan Jing, Fu-Min Zhang, Xiu-Hua Wang, Mei-Xia Wang. ✉e-mail: gesong@ibcas.ac.cn

Section 1). Second, the paucity of both wild and cultivated samples from some geographic areas, coupled with the use of misidentified or admixed wild accessions, have undoubtedly affected the inference of domestication history (Supplementary Section 2). Third, continuous and extensive gene flow or introgression has been reported both among cultivar groups and between wild and domesticated rice, causing additional complications for evolutionary inference[14,15,24,30–32]. All of these factors have complicated our understanding of domestication history and may explain why contradictory conclusions have been reached, even by analysing the same dataset[18,27]. Finally, the archaeological record can be interpreted as supporting either single or multiple domestication events[9,11,33,34], both because there have been fewer archaeological explorations in South and Southeast Asia and because it is difficult to distinguish between cultivar groups on the basis of archaeobotanical remains[33,34].

Rapid advances in genomics technology, combined with more functional studies on rice domestication genes and accumulating archaeological evidence, have provided an unprecedented opportunity to settle controversies surrounding crop domestication[4,9,12,23,33,34]. Here we conduct evolutionary analyses at the genome scale based on a dataset of 459 newly resequenced and 1,119 publicly available genomes of wild and domesticated accessions. Our dataset includes a substantial expansion of the sampling of wild accessions from throughout the geographic range of *O. rufipogon* and *O. nivara* (Fig. 1a and Supplementary Table 1), all of which have been verified as wild by phenotypic examination. This expansion, coupled with the removal of mislabelled and admixed accessions from previous datasets, helps infer domestication history more accurately. Collectively, our findings provide evidence that rice domestication began independently from divergent wild lineages in different areas and was then completed with continuous selection and the exchange of beneficial alleles among different cultivar groups.

Results

Identification and sequencing of wild and cultivated accessions

We began this study by extensively phenotyping wild rice accessions in experimental fields (Supplementary Section 2). On the basis of 15 diagnostic characters (Extended Data Fig. 1), we delineated accessions as either wild or cultivated and reclassified those that were mislabelled in the initial designations. Remarkably, we found that 25.5% of *O. rufipogon* and 22.5% of *O. nivara* accessions requested from seed banks were either mislabelled or admixed with cultivars (Supplementary Table 2). These studies not only facilitated sample selection but also emphasize the importance of sample validation in research and germplasm management.

On the basis of this morphological validation, we resequenced 422 wild accessions (245 *O. rufipogon* and 177 *O. nivara*) to an average depth of 11.74× per accession (Supplementary Table 3). We also resequenced 37 rice landraces (Supplementary Table 4), including 14 representing the rayada group, which was distinct genetically and featured unique agronomic characters[10,24,35] but has been ignored in most previous studies except for Wang et al.[36]. These newly sequenced genomes complemented existing genome resources and substantially improved the resolution and accuracy of our analyses, given that the sequenced genomes of wild accessions in previous studies were either low-coverage (<2×) (refs. 18,27,30,31) or limited in number (<25 accessions) (refs. 21,22,32,37,38). We combined these data with a carefully chosen resequencing dataset of published wild and cultivated accessions (Methods), resulting in a complete dataset of 1,578 whole-genome sequences (Supplementary Table 5). The dataset included 457 wild accessions that covered the entire range of two rice progenitors (Fig. 1a and Supplementary Table 1) and 1,121 rice landraces from 25 countries (Supplementary Table 6). We adopted the workflow by DePristo et al.[39] and the Broad Institute (https://gatk.broadinstitute.org/hc/en-us) for read mapping, variant discovery, genotyping and variant quality recalibration using the rice Nipponbare genome (IRGSP Build 5) (ref. 40) as the reference (Methods). We finally identified a total of ~17.2 million single nucleotide polymorphisms (SNPs), subsets of which were used in specific analyses after additional filtering procedures where necessary.

Strong genetic structure of wild and domesticated rice

To assess population genetic structure, we analysed all 457 wild rice genomes using a neighbour-joining (NJ) tree, principal component analysis (PCA) and ADMIXTURE. These analyses identified four major lineages, including two *O. rufipogon* (Ruf1 and Ruf2) and two *O. nivara* (Niv1 and Niv2) lineages (Fig. 1b, Extended Data Fig. 2a,b and Supplementary Fig. 1). Remarkably, the two *O. nivara* lineages (Niv1 and Niv2) did not form a monophyletic group and instead were more closely related to different *O. rufipogon* lineages (Ruf1 and Ruf2). We further analysed a panel of 404 wild samples by reclassifying mislabelled accessions and excluding admixed accessions (Supplementary Table 5). ADMIXTURE analyses showed that the deepest split ($K = 2$) occurred between two groups, with one mainly distributed in China and northern South Asia and the other in Southeast and South Asia (Fig. 1d). From $K = 4$ to 12, each of the two groups was divided into two subgroups, one consisting of Ruf1 and Niv1 and the other comprising Ruf2 and Niv2 (Fig. 1d and Extended Data Fig. 2c), consistent with the PCA results (Fig. 1c and Extended Data Fig. 2b). The deep divergence between these two groups has been overlooked in most previous studies, which may bias inferences about the number and location of domestication events.

We further investigated subdivision and classification within rice germplasm. Opinions differ as to whether there are as few as three or as many as nine major groups (Supplementary Section 4). To avoid confusion, we refer to two Asian rice subspecies as *Japonica* and *Indica* (italic and the first letter capitalized), to distinguish them from the commonly used cultivar groups of *japonica* (that is, *temperate japonica* and *tropical japonica*) and *indica* (italic and all lowercase letters) (Supplementary Section 1). Our preliminary population genetic analyses (Extended Data Fig. 3) on 1,121 rice landraces identified 147 misclassified and admixed accessions (Supplementary Table 7), leaving a panel of 1,089 pure landraces (Supplementary Table 5) for subsequent analyses. The NJ tree formed two separate monophyletic groups representing the two subspecies. The subspecies *Indica* included two cultivar groups (*indica* and *aus*), whereas *Japonica* contained four cultivar groups (*aromatic*, *rayada*, *temperate japonica* and *tropical japonica*) (Fig. 1e,f). Similarly, ADMIXTURE at $K = 2$ uncovered two major groups corresponding to the two subspecies and identified distinct cultivar groups within each subspecies as K increased (Fig. 1f), consistent with PCA (Fig. 1g). Notably, three minor cultivar groups (*aus*, *aromatic* and *rayada*) were genetically distinct, suggesting that they merit further attention despite their cultivation in limited areas in South and Southeast Asia[10,24,32,35].

Population dynamics associated with domestication

We estimated nucleotide diversity at different hierarchical levels and detected comparable levels of diversity both for four wild lineages and for six cultivar groups (Fig. 2a). Overall, the diversity of domesticated rice (Watterson's estimator (θ), 0.0026; average number of pairwise nucleotide differences (π), 0.0033) was roughly 50–70% that of wild rice ($\theta = 0.0053$; $\pi = 0.0046$) (Supplementary Table 8), highlighting relatively high diversity in wild rice sensu lato. Lower genetic diversity of cultivars was also reflected in patterns of linkage disequilibrium (LD) (Extended Data Fig. 4). Strikingly, the diversity levels of three minor cultivar groups (*aus*, *rayada* and *aromatic*) were close to or even higher than those of two *japonica* groups (Fig. 2a, Supplementary Fig. 2 and Supplementary Table 8) despite their limited geographic distributions. The relatively high diversity and genetic distinctiveness of these minor groups suggest the possibility of their separate domestications, as

https://doi.org/10.1038/s41477-023-01476-z

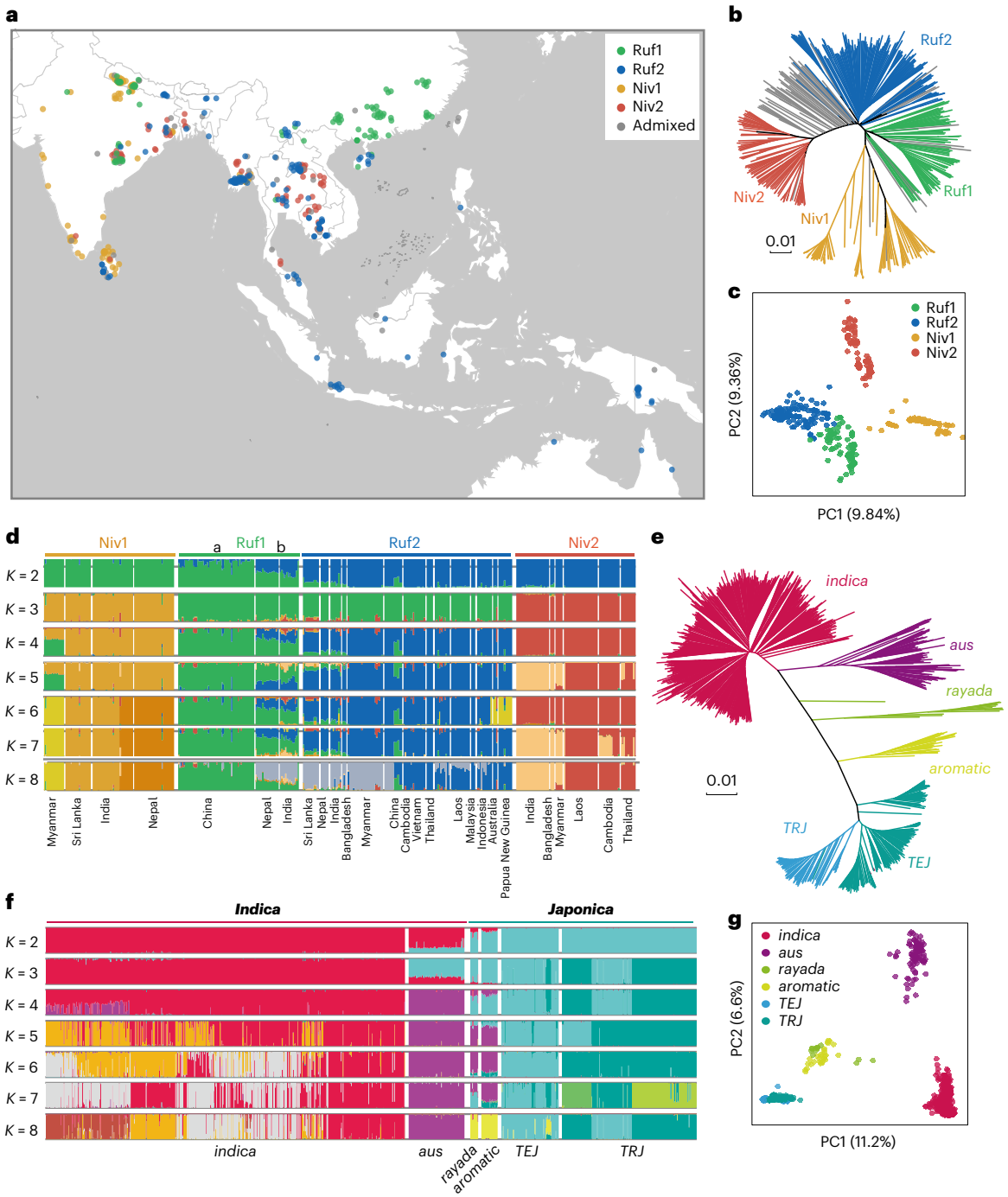


Fig. 1 | Population genetic structure and phylogenetic relationships of the wild and domesticated rice. a,**b**, Geographic distribution (**a**) and unrooted NJ tree (**b**) of the 457 wild rice accessions sampled, with admixed accessions indicated in grey. **c**,**d**, PCA (**c**) and ADMIXTURE plot (**d**) for 404 wild rice accessions excluding the admixed accessions. The columns in ADMIXTURE represent the accessions with their lineages above the plot and their origins below the plot. **e**–**g**, Unrooted NJ tree (**e**), ADMIXTURE plot (**f**) and PCA (**g**) for 1,089 rice landraces excluding admixed accessions. The columns in ADMIXTURE represent the landraces with their subspecies above the plot and cultivar groups below the plot. The scale bars in the NJ trees show substitutions per site. The dot and line colours indicate the wild lineages and cultivar groups. TEJ, *temperate japonica*; TRJ, *tropical japonica*.

claimed in previous studies[18,32,36,37], and highlight their potential value for rice breeding[35,36,41].

We then reconstructed the genealogy of four wild lineages and six cultivar groups using genome-wide SNPs. Both the NJ tree (Fig. 2b) and PCA (Fig. 2c) showed that the cultivated accessions did not form a monophyletic group but were divided into several distinct groups that were closely related to different wild lineages, consistent with previous studies based on genome-wide neutral markers[17–22,27,32,37,38]. Specifically, *Indica* (*indica* + *aus*) formed a group with one *O. nivara* lineage in South and Southeast Asia (Niv2), whereas three *Japonica* groups (*aromatic*, *temperate japonica* and *tropical japonica*) clustered with one *O. rufipogon* sub-lineage in Southern China (Ruf1a) (Fig. 2b and Supplementary Fig. 3), suggesting that *O. rufipogon* in northern South Asia (Ruf1b) was not directly involved in rice domestication. We obtained the

https://doi.org/10.1038/s41477-023-01476-z

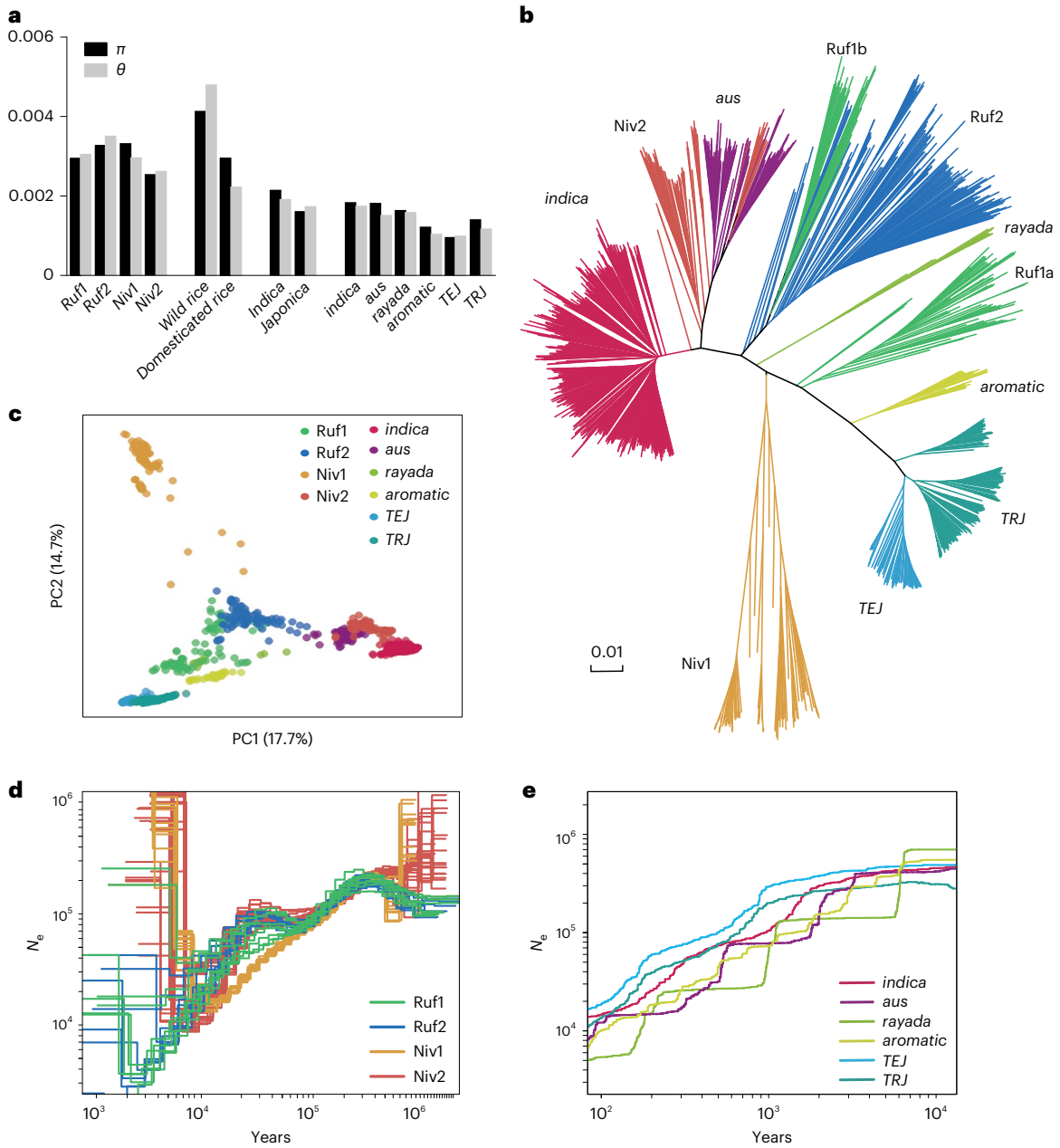


Fig. 2 | Genetic variation patterns and population dynamics of wild and cultivated rice and domestication timeframes of cultivar groups. a, Genetic diversity of wild and domesticated rice at different hierarchical levels. **b**,**c**, Unrooted NJ tree (**b**) and PCA (**c**) of 404 wild and 1,089 cultivated accessions. The line and dot colours represent different wild lineages and six cultivar groups. The scale bar in the NJ tree shows substitutions per site. **d**,**e**, N_e of wild (**d**) and domesticated (**e**) rice, inferred using PSMC (**d**) and Stairway Plot 2 (**e**). The line colours represent different wild lineages and cultivar groups.

same result when using putatively neutral loci (SNPs from intergenic regions) (Extended Data Fig. 5a,b) and after removing three minor cultivar groups (*aus*, *rayada* and *aromatic*) (Extended Data Fig. 5c). We further reconstructed single-gene trees based on 34,791 annotated genes (Methods); we found that a majority of gene trees supported a cladogram in which *Japonica* clustered with Ruf1 (*O. rufipogon*) and *Indica* with Niv2 (*O. nivara*) (Extended Data Fig. 5d,e). Together, our genome-wide analyses of SNPs and genes demonstrate distinct evolutionary histories or genomic backgrounds for *Japonica* and *Indica*.

To better understand domestication history, we estimated population size changes over time for wild and domesticated rice by applying the pairwise sequentially Markovian coalescent (PSMC)[42], Stairway Plot 2 (ref. 43) and the SMC++ method[44] (Methods). PSMC analysis indicated that the two wild species have exhibited continuous declines in effective population size (N_e) since 0.4 million years ago and gradually diverged in the time interval of 100,000 to 30,000 years ago (Fig. 2d), which is consistent with previous estimates about the origin of *O. nivara*[45]; it is possible that this population decline contributed to the transition of society from hunter-gatherers to farming, as hypothesized for African rice[46]. Domesticated rice had a similar pattern to wild rice before ~10,000 years ago (Supplementary Fig. 4), suggesting that rice domestication did not become genetically detectable until ~9,000 years before present[9,33]. We further analysed domesticated rice using Stairway Plot 2 (ref. 43), which is designed to infer demographic changes on more recent timescales. We found that domesticated rice has undergone a continuous decline in N_e since 10,000 years ago (Fig. 2e).

SMC++ analysis[44], another method used for inferring demographic history on recent timescales (Methods), provided largely congruent results (Supplementary Fig. 4). A long and protracted N_e decline in wild rice and a continuous decrease in N_e agree with archaeological evidence suggesting that Asian rice progressed from semi-domesticated to fully domesticated over several thousand years[11,33]. These observations are consistent with the protracted model of crop domestication[2,4,6] but cannot disprove a more rapid transition from wild to domesticated.

Analysis of selective sweep regions and selected genes

To test the single- versus multiple-domestication hypothesis, we chose to analyse the major cultivar groups *japonica* (*temperate japonica* and *tropical japonica*) and *indica* both because they have been focal groups of all previous studies involving rice domestication and because three minor cultivar groups (*aus*, *aromatic* and *rayada*) are cultivated in very limited areas with uncertain origins[18,24,32,37]. We proposed a strategy that included three steps: (1) identifying the genomic regions/genes under selection in both *japonica* and *indica*; (2) determining whether the selected genomic regions/genes generate a tree that forms a monophyletic group combining *japonica* and *indica*—if they do, we consider them selected regions/genes of single origin; and (3) sorting the genealogies of single-origin selected genomic regions/genes into individual categories (Methods and Extended Data Fig. 6). More specifically, if the same wild lineage was sister to all of the single-origin selected regions/genes, we considered it as strong support for a single domestication event. Alternatively, if single-origin selected regions/genes derived from more than one wild lineage, this suggested multiple domestication events. It is worth noting that our approach differs from previous studies[27,47–49], because it analyses genealogical relationships with the explicit goal of testing alternative hypotheses based on whether a single or multiple wild lineages contributed to domestication alleles (Methods).

We first screened for the genomic regions with selective sweeps in both *japonica* and *indica*. By scanning the genome in a 100 kb sliding window with a step of 10 kb, we identified 98 putative selective sweep regions (PSRs), representing regions with severe reduction of genetic diversity in cultivars and altered allele frequencies in both *japonica* and *indica*. These PSRs occurred in all 12 chromosomes and accounted for 7.7% of the Nipponbare reference genome (Fig. 3a, Extended Data Fig. 7, Supplementary Fig. 5 and Supplementary Table 9). By reconstructing NJ trees for four wild lineages and two cultivar groups using the SNPs of each PSR, we identified 71 PSRs that generated genealogies on which *japonica* and *indica* formed a monophyletic group (Table 1), suggesting that 72.4% of PSRs arose only once from either an *O. rufipogon* lineage or a *O. nivara* lineage. We defined these PSRs as single-origin selective sweep regions (SORs). The 71 SORs accounted for 6.5% of the reference genome and were located across all chromosomes (Fig. 3a, Supplementary Fig. 5 and Supplementary Table 9). As expected, many functionally characterized domestication genes were found among SORs, including *sh4*, *PROG1*, *Rc*, *An-1*, *Bh4* and *LABA1* (Fig. 3a and Supplementary Table 10). We classified the 71 SOR trees into seven categories according to the wild lineages that were sister groups to the cultivar group. We found that Ruf1 was the sister group to the cultivar group for 55 (77.4%) of the SOR trees, while Niv1 and/or Niv2 was the wild sister group for 10 (14.1%) SOR trees (Table 1 and Supplementary Table 10). These observations suggest that both *O. rufipogon* and *O. nivara* contributed domestication alleles to rice germplasm, although we do not have the power to discriminate whether these alleles arose from standing variants or de novo mutations.

Next, we analysed all the annotated genes across the genome using a strategy similar to the analysis of selective sweep regions (Methods and Extended Data Fig. 6). We characterized all 34,791 annotated genes and identified 1,882 putatively selected genes (PSGs) (Supplementary Section 8), representing 5.4% of all genes. Phylogenetic analyses of individual PSGs revealed that 993 (52.8%) generated trees on which *japonica* and *indica* clustered together (Table 1, Supplementary Tables 11 and 12 and Supplementary Fig. 6), suggesting that the domesticated alleles of these genes originated only once. We defined these genes as single-origin selected genes (SOGs). It is remarkable that over 50% of PSGs were common to *japonica* and *indica* (that is, SOGs), because the proportion of SOGs shared by divergent cultivar groups is very low in other crops with multiple domestication events, such as common bean (3–8%) (ref. 47), melon (8.4–9.7%) (ref. 48) and buckwheat (12.2–12.7%) (ref. 49). The high proportion of SOGs could reflect an extremely high level of gene flow/introgression between *japonica* and *indica*, a major factor that complicates the inference of rice domestication history[9,13,14,33].

By sorting the SOG trees, we observed that Ruf1 was the nearest to the cultivar group on 784 (79%) trees, while Niv1 and/or Niv2 were the wild lineages sister to the cultivar group on 176 (17.7%) trees (Table 1). These analyses again suggest that both *O. rufipogon* and *O. nivara* were associated with the origins of domestication genes in Asian rice, and the SOG trees are consistent with the analysis of SORs. Most of the well-known domestication genes, such as *Sh1*, *An-1*, *Bh4*, *sh4*, *PROG1* and *Rc*, are closely related *O. rufipogon* (Ruf1), while several domestication genes (such as *LABA1* and *GLO4*) appear to have originated from *O. nivara* (Niv1/Niv2) (Table 2, Fig. 3a and Supplementary Table 12). The preponderance of genes associated with *O. rufipogon* explains why a single domestication seems more likely, especially when only a limited number of domestication genes are analysed. Collectively, our genome-wide exploration of SORs and SOGs suggests that different wild lineages have contributed domestication genes, which argues against the single-domestication hypothesis.

A sweep region might include selected genes with different histories

We compared the results from region-based and gene-based scans and made two observations. First, 1,400 (74.4%) of the 1,882 PSGs were included in the 98 PSRs (Supplementary Table 13). Similarly, 791 (79.7%) of the 993 SOGs were found in the 71 SORs (Supplementary Table 12). These results show that our different approaches to identify putatively selected regions largely agree, despite the fact that it can be difficult to identify selected genes due to complications such as size, the extent of diversity reduction, soft sweeps, extensive gene flow and complicated demographic processes[5,6,12,50]. A second and unexpected finding was that the SOGs with different origins (that is, from *O. rufipogon* or *O. nivara*) collocated in 6 of the 71 SORs (Supplementary Table 10). To explore this phenomenon further, we analysed two SORs (SOR33 and SOR48) that included SOGs that originated from different wild ancestors (Fig. 3b and Supplementary Table 10). For SOR33, on which *LABA1* and *LCBK2* were located, Niv1 or Niv2 was sister to the cultivar group on the *LABA1* tree, consistent with the SOR33 tree, whereas Ruf1 was most closely related to the cultivar group on the *LCBK2* tree that differed from the SOR33 tree. For SOR48, where *PROG1* and *GLO4* coexisted in close proximity, Ruf1 was sister to the cultivar group on the *PROG1* tree, in contrast to the *GLO4* tree, on which Niv1 was nearest to the cultivar group (Fig. 3c and Extended Data Fig. 8). These observations indicate that a selected region may represent a mosaic of multiple fragments/genes that embody different evolutionary histories; we do not know yet whether this phenomenon exists in other domesticated species.

Our findings suggest that the practice of inferring domestication history by reconstructing a single phylogeny based on concatenating all PSRs may be misleading. Huang et al.[27] used this approach and concluded that Asian rice was domesticated once from Chinese *O. rufipogon*. To investigate further, we performed analyses on the 55 selective sweep regions identified in Huang et al.[27] using our strategy (Supplementary Section 7). We found that 47 selective sweep regions generated genealogies on which *japonica* and *indica* clustered together (Supplementary Table 14)—that is, 47 of 55 sweep regions were single-origin. However, *O. nivara* was sister to the cultivar group on 11 (23.4%) of the 47 trees, while *O. rufipogon* clustered with the

https://doi.org/10.1038/s41477-023-01476-z

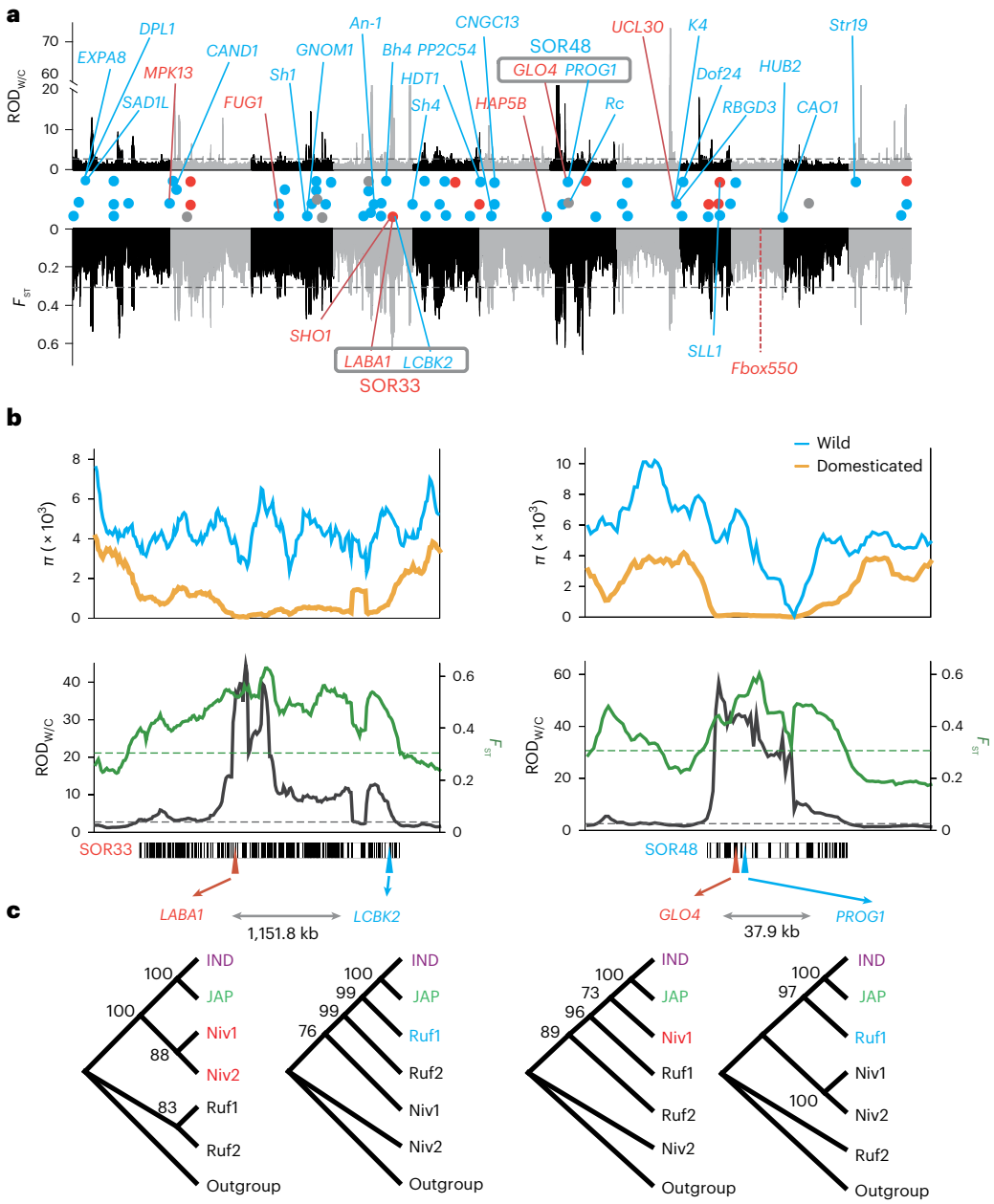


Fig. 3 | Whole-genome scans of PSRs and SORs in domesticated rice and phylogenetic analyses of domestication genes within the sweeps. a, Genome-wide distribution of the 71 SORs and the genomic locations of 30 domestication genes used for network analyses. The ratio of the nucleotide diversity of wild populations to that of domesticated populations ($ROD_{W/C}$) (top) and the genetic divergence (F_{ST}) between the wild and domesticated populations (bottom) are plotted for 100 kb windows against the position on each of the 12 chromosomes. The black dashed lines represent the thresholds of $ROD_{W/C}$ and F_{ST} values for a PSR. The solid circles between the two panels show the approximate locations of 55 SORs that originated from *O. rufipogon* (blue) and 10 SORs that originated from *O. nivara* (red). The solid grey circles represent 6 SORs with uncertain origins. The 30 domestication genes used for haplotype network analysis (Table 2) include 22 genes (blue) and 8 genes (red) that included domestication alleles from *japonica* and *indica*, respectively. Genes within and outside the 71 SORs are indicated by solid and dashed lines, respectively. Note that the genes with domestication alleles from *O. rufipogon* (or *O. nivara*) may occur in the SORs that originated from *O. nivara* (or *O. rufipogon*), and the same SORs (for example, SOR33 and SOR48) may comprise two types of genes (that is, genes with domestication alleles from *O. rufipogon* and those with domestication alleles from *O. nivara*). **b**, Diversity patterns of two SORs and their 100 kb flanking regions. SOR33 (left) is 1,950 kb and includes 252 genes, while SOR48 (right) is 600 kb and includes 62 genes. The bars along the bottom of the panel represent the chromosome structure of the two SORs, with the annotated genes indicated by black lines. **c**, NJ trees of *LABA1* (with domestication alleles from *O. nivara*) and *LCBK2* (with domestication alleles from *O. rufipogon*) within SOR33 and of *PROG1* (with domestication alleles from *O. rufipogon*) and *GLO4* (with domestication alleles from *O. nivara*) within SOR48. The NJ trees were constructed using SNPs of the fragment spanning from 3 kb upstream to 3 kb downstream of the gene. Bootstrap supports over 70% are shown near the nodes.

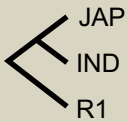
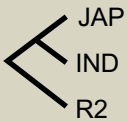
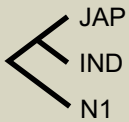
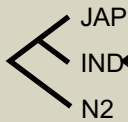
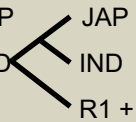
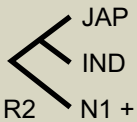
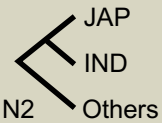
cultivar group on the remaining trees (36) (Supplementary Table 14). The sweep regions that were previously concatenated to provide proof of a single domestication event thus bear evidence of multiple domestication events.

Haplotype analysis demonstrates multiple domestications

To make full use of the available information about rice domestication genes, we compiled a subset of 192 genes that were found to be under selection in both *japonica* and *indica* and also have known functions

https://doi.org/10.1038/s41477-023-01476-z

Table 1 | Summary of NJ trees reconstructed on the basis of the SNPs of 71 SORs and 993 SOGs across the genome

Category of topology								No. of total SORs/ SOGs
No. of SORs (%)	55 (77.4)	0 (0.0)	1 (1.4)	7 (9.9)	1 (1.4)	2 (2.8)	5 (7.1)	71
No. of SOGs (%)	784 (79.0)	9 (0.9)	87 (8.8)	50 (5.0)	18 (1.8)	39 (3.9)	6 (0.6)	993

JAP, *japonica* (*temperate japonica* + *tropical japonica*); IND, *indica*; R1, Ruf1; R2, Ruf2; N1, Niv1; N2, Niv2; others, no resolution for wild lineages. A SOR/SOG was defined as a PSR/PSG that generated a phylogenetic tree on which *japonica* and *indica* formed a monophyletic group.

(Supplementary Table 15). By referring to this database, we chose a subset of 36 domestication genes that belong to SOGs and are located across all chromosomes (Supplementary Section 10). We then performed haplotype network analyses (Methods and Supplementary Fig. 7) to investigate the number and location of domestication centres. We focused on the analysis of common haplotypes—that is, haplotypes with frequencies greater than 0.5%. For a specific gene, we analysed two types of haplotype on the network: (1) the haplotype that was dominant (the most numerous) in the sample and shared by *japonica* and *indica*, which we defined as the domestication haplotype (H1); and (2) the haplotype from a wild lineage or lineages that was most closely related to H1, defined as haplotype W-H1. We were thus able to identify the potential domestication haplotype of each gene and its putative wild contributor(s).

Of all 36 domestication genes analysed, 30 genes generated resolved networks on which H1 and W-H1 could be identified unambiguously (Table 2 and Supplementary Figs. 8 and 9). Among them, 21 genes had W-H1s that occurred exclusively in Ruf1; we hypothesize that the H1s for these genes arose in *japonica* and spread to *indica*. For example, the seed hull gene *Bh4* had 13 common haplotypes and one H1 (Fig. 4a), and the network suggests that domesticated alleles associated with hull colour change had a single origin. The *Bh4* W-H1 haplotype was found in two taxa: 65.9% of the W-H1 haplotypes were found in Ruf1, and the remaining 34.1% were found in *japonica* (Fig. 4a,b and Table 2). These observations suggest that the *Bh4* H1 haplotype arose in Chinese *O. rufipogon* populations (Fig. 4b). Using this same approach, we detected eight domestication genes with H1s that arose from *O. nivara* in South and Southeast Asia, because either Niv1 or Niv2 was the main component of W-H1 (Table 2 and Supplementary Figs. 8 and 9). For example, the awn length gene *LABA1* and the stress resistance gene *Fbox550* had 15 and 11 common haplotypes, respectively (Fig. 4a), and the W-H1 haplotypes consisted mainly of Niv1 and Niv2, which were distributed in Southeast and South Asia, respectively (Fig. 4a,b and Table 2). These results show that the H1 haplotypes of these two genes arose from *O. nivara* and not *O. rufipogon*.

By plotting the geographic locations of the wild accessions with W-H1 haplotypes for all 30 domestication genes, we found that southern China and northern India were the primary sources, suggesting they are likely centres of Asian rice domestication. Southeast Asia and southern India/Sri Lanka are also common locations for putative domestication alleles, suggesting that they may be additional areas associated with rice domestication (Fig. 4c and Table 2). These results agree with archaeological evidence[9,33,34], which supports multiple events associated with Asian rice domestication.

Reciprocal gene flow between *japonica* and *indica* rice

Our results suggest that domestication genes arose from different geographic localities and are associated with different wild taxa. Yet, we also found that the selected genes are often detectable in both *japonica* and *indica*, which implies introgression between the two groups, as hypothesized previously in all major models of Asian rice domestication[13–15,27,33,37]. In particular, the scenario in which *indica* emerged from crosses between *japonica* and *O. nivara* in South and Southeast Asia has been widely presumed to prove the simple domestication hypothesis[27,29,31,33,38], but it has never been tested using empirical data (Supplementary Section 11). To further characterize introgression among cultivar groups and wild lineages, especially to test whether gene flow occurs between *japonica* and *O. nivara*, we calculated *D*-statistics and performed the three-population test (f_3) (ref. 51) (Methods). We calculated *D*-statistics for pairwise comparisons of two cultivar groups and their progenitors and detected substantial gene flow both between *japonica* and *indica* and between Ruf1 and either *indica* or *O. nivara*, while no gene flow was detected between *japonica* and *O. nivara* (tests 2, 3, 6 and 7, Extended Data Table 1a). Consistently, the f_3 test did not detect introgression from *japonica* into *O. nivara* (Extended Data Table 1b).

Analysis of the haplotype networks of the representative domestication genes was consistent with these introgression tests. The largest haplotypes in network plots of all 30 domestication genes were the haplotypes shared by *japonica* and *indica* (that is, H1s), which probably reflects a single origin followed by introgression between two cultivar groups (Table 2 and Supplementary Table 16). Across all 30 domestication genes, the H1 haplotype contains no more than 7% frequency from wild taxa (Table 2, Supplementary Table 16 and Supplementary Figs. 8 and 9), suggesting that introgression from domesticated rice to wild species was limited in scale for this gene set. Together, our analyses of gene flow did not find evidence supporting a single-domestication hypothesis that requires substantial introgression from *japonica* to *O. nivara*[27,31,33,37,52].

Discussion

We have addressed at least four shortcomings that have probably hampered the inference of domestication history in Asian rice: (1) the ambiguous population structure and genetic relationship of two wild rice species, *O. rufipogon* and *O. nivara*; (2) a lack of accuracy, due to mislabelling and admixture, of samples commonly used to infer domestication history; (3) complex patterns of gene flow and introgression among cultivar groups; and (4) the need for an effective approach to test for alternative hypotheses about domestication history. By overcoming these shortcomings, we have contradicted other recent studies[27,31,33,37,38,52] by concluding that Asian rice has been domesticated at least twice, in southern China and India (Fig. 5). This conclusion, which is based solely on genomic data, agrees with recent archaeological findings[34] indicating that *indica* domestication may have started ~8,000 years before present, much earlier than the ~4,000–5,000 years before present speculated previously[9,11,14,33]. Hence, domestication was probably underway in northern India before the arrival of *japonica*. Moreover, the origins of minor cultivar groups (*aus*, *aromatic* and *rayada*) remain unsolved or disputed[18,37,53,54]. Further in-depth exploration of their domestication histories will promote a better understanding of Asian rice domestication and facilitate their utilization in rice breeding.

https://doi.org/10.1038/s41477-023-01476-z

Table 2 | Summary of haplotype network analysis of 30 representative domestication genes

No.	Gene name (chromosome)	Phenotype involved	No. of total haplotypes	No. of types of common haplotypes[a]	Frequency (%) of H1	Wild component over 1% (%) in H1	W-H1	Wild and cultivar components (%) in W-H1
Genes with H1s that originated from *O. rufipogon*								
1	*OsEXPA8* (1)	Root architecture and plant height	1,540	14	1,097 (71.2)	No	H2	Ruf1 (100)
2	*DPL1* (1)	Hybrid incompatibility	1,602	14	1,183 (73.8)	Ruf1 (6.7%)	H2	Ruf1 (100)
3	*SAD1L* (1)	Shoot branching	1,444	16	1,103 (76.4)	No	H2, H3	Ruf1 (100)
4	*OsCAND1* (2)	Crown root emergence	1,414	15	802 (56.7)	Niv2 (1.0%)	H2, H3	Ruf1 (78.0); *JAP* (22.0)
5	*OsSh1* (3)	Seed shattering	1,290	12	1,122 (87.0)	No	H2	Ruf1 (100)
6	*OsGNOM1* (3)	Root formation	1,558	15	1,043 (66.9)	Niv2 (1.9%)	H6	Ruf1 (100)
7	*An-1* (4)	Awn development, grain size and number	1,400	17	936 (66.9)	Ruf1 (1.3%)	H2, H4	Ruf1 (29.5); *IND* (70.5)
8	*Bh4* (4)	Straw-white seed hull	1,346	13	802 (59.6)	No	H2	Ruf1 (65.9); *JAP* (34.1)
9	*OsLCBK2* (4)	Disease resistance	1,450	17	519 (35.8)	No	H2	Ruf1 (100)
10	*sh4* (4)	Seed shattering	1,432	13	1,140 (79.6)	No	H5	Ruf1 (100)
11	*OsHDT1* (5)	Heading date, seed germination and resistance to abiotic stress	1,432	12	978 (68.3)	Niv2 (2.2%)	H2	Ruf1 (100)
12	*OsPP2C54* (6)	Tolerance to salt stress	1,584	13	1,015 (64.1)	No	H2	Ruf1 (100)
13	*OsCNGC13* (6)	Seed-setting rate	1,506	13	1,015 (67.4)	Niv2 (1.2)	H4	Ruf1 (100)
14	*PROG1* (7)	Prostrate growth and plant architecture	1,344	12	1,036 (77.1)	Niv2 (2.3)	H8	Ruf1 (100)
15	*Rc* (7)	Pericarp colour and seed dormancy	1,274	12	881 (69.2)	No	H2	Ruf1 (100)
16	*OsK4* (8)	Heading date	1,450	15	944 (65.1)	Ruf1 (1%)	H2	Ruf1 (100)
17	*OsDOF24* (8)	Leaf senescence	1,516	7	1,132 (74.7)	No	H2	Ruf1 (100)
18	*OsRBGD3* (8)	Tolerance to cold stress	1,462	12	1,123 (76.8)	Ruf1 (1.3%)	H2, H3, H4	Ruf1 (100)
19	*SLL1* (9)	Leaf rolling, hull fate, grain yield and quality	1,554	14	974 (62.3)	Ruf1 (1.1%)	H3	Ruf1 (100)
20	*OsHUB2* (10)	Heading date, anther development and yield	1,432	16	1,080 (75.4)	Niv2 (2.6%)	H2, H3, H4, H5, H6	Ruf1 (94.7); Ruf2 (5.3)
21	*OsCAO1* (10)	Grain yield and quality	1,500	12	976 (65.1)	Niv2 (2.9%); Ruf1 (2.3%)	H2, H4	Ruf1 (100)
22	*Str19* (12)	Resistance to abiotic stress	1,356	17	1,084 (79.9)	Ruf1 (1.4); Niv2 (1.3%)	H2	Ruf1 (94.4); *JAP* (5.6)
Genes with H1s that originated from *O. nivara*								
1	*OsMPK13* (2)	Resistance to brown planthopper	1,486	11	1,066 (71.7)	Niv2 (2.1%)	H3	Niv1 (86.3); Ruf1 (13.7);
2	*OsFUG1* (3)	Panicle architecture and fertility	1,426	15	567 (39.8)	No	H14, H15	Niv1 (100)
3	*SHO1* (4)	Shoot apical meristem formation and leaf development	1,462	12	890 (60.9)	No	H10	Niv1 (89.5); Niv2 (10.5)
4	*LABA1* (4)	Awn length and loss of barbs	1,532	15	1,099 (71.7)	No	H9, H13	Niv2 (100)
5	*OsHAP5B* (6)	Heading date	1,482	9	1,080 (72.3)	No	H3	Niv1 (95.1); Ruf1 (4.9)
6	*GLO4* (7)	Environmental stress or stimuli and indoleacetic acid biosynthesis	1,530	10	1,179 (77.1)	Niv2 (2%)	H8	Niv1 (95.2); Niv2 (2.4); Ruf1 (2.4)
7	*UCL30* (8)	Stress tolerance and photosynthesis	1,472	13	1,190 (80.8)	Ruf1 (3%); Niv2 (1%)	H13	Niv1 (100)
8	*OsFbox550* (10)	Resistance to abiotic stress	1,530	11	816 (53.3)	Ruf1 (1.2%)	H2	Niv1 (98.4); Ruf1 (1.6)

[a]Common haplotypes are haplotypes with frequencies over 0.5%.

This study also has important implications for studies of other domesticated species. Our strategy for inferring domestication history (Methods) overcame some of the limitations of analysing selective sweep regions/selected genes across the genome (Supplementary Sections 7 and 8) and may be widely applicable, particularly for domesticated species with confounded evolutionary histories defined by widespread and continuous gene flow among cultivars and lines. We recognize that the study of individual

genomic regions has caveats, particularly that genealogical relationships in single genes may be distorted by lineage sorting and that the timing of selection events is inexact. However, lineage sorting is not the most parsimonious explanation for our observations, because phylogenies based on entire genomes are consistent with at least two origins of Asian rice and because haplotype analyses consistently show that the closest wild haplotype has the highest frequency in the expected (that is, closest) wild relative (Fig. 4 and Supplementary Figs. 8 and 9).

Finally, we found that selective sweep regions occasionally consist of mosaics of genes representing different evolutionary origins. Assuming we have correctly mapped the origins of these genes, such a pattern can occur only when there is introgression between distinct rice germplasm, followed by recombination between the selected genes, thus creating a haplotype with multiple favourable alleles. One cannot help but speculate that, in some situations, these favourable combinations may greatly speed the rate of evolution by additive or potentially synergistic interactions. Moreover, this process is reminiscent of ecological speciation theory, which predicts that, once formed, these regions define islands of divergence containing genes that contribute to local adaptation (in this case via artificial selection). Paradoxically, these islands of divergence are predicted to eventually retard gene flow between populations[55,56], which in this case is between wild and domesticated populations but could also help explain the formation of a partial reproductive barrier between *japonica* and *indica*[8,13,14]. If this model holds, it suggests a mechanism that can slow introgression between wild and domesticated populations progressively over time.

Methods

Sampling and resequencing of wild and domesticated rice

Whole-genome sequencing data for a total of 1,578 accessions containing 457 accessions of wild rice (*O. rufipogon* and *O. nivara*) and 1,121 rice landraces were included in this study. The 422 newly sequenced wild accessions included samples from some areas/countries that were not represented in previous studies and representative individuals from natural populations collected by the authors (Supplementary Tables 1, 3 and 17). All these samples were carefully chosen on the basis of our morphological verification in experimental fields (Supplementary Section 2). To capture as much variation of the wild species as possible, we downloaded 35 published genomes with sequencing depths over 4× (Supplementary Table 3).

For domesticated rice, we chose to use the 1,014 rice landraces from the 3,010 published rice genomes[16]. These landraces represent the primitive varieties that were cultivated across Asian countries and had genome sequencing depths over 4× (Supplementary Table 19). In addition, we downloaded the raw sequence data of an additional 70 rice landraces published elsewhere (Supplementary Table 4) and resequenced 37 rice landraces to maximize the representation of different groups (Supplementary Table 4). In total, these 1,121 resequenced landraces, including 387 *O. sativa* ssp. *japonica* and 734 *O. sativa* ssp. *indica*, were chosen by considering the country of origin, varietal classification and eco-cultural type (Supplementary Tables 6 and 19) and have an average sequencing depth of 16× per accession.

Sequencing, SNP calling and quality control

Genomic DNA was isolated from silica-gel-dried or fresh leaves of a single individual. The leaves were frozen in liquid nitrogen and milled with ceramic beads. The DNAsecure Plant Kit (Tiangen Biotech, product no. DP320) was used to extract genomic DNA from the frozen leaf powder following the manufacturer's instructions. The quantity and quality of the genomic DNA were checked using Nanodrop and agarose gel (0.8%), respectively. Individual libraries were constructed following the manufacturer's instructions (Illumina) for 500-bp insert size and sequenced at 100-bp or 150-bp paired-end on an Illumina HiSeq2000, a HiSeq2500 or a HiSeqXten. The raw sequencing data were cleaned by removing adapter sequences, trimming low-quality ends and filtering reads with low quality (average Phred quality score < 20) using Trimmomatic (release 0.36) (ref. 57).

For the complete dataset of 1,578 genomes, we adopted the workflow proposed by DePristo et al.[39] and developed at the Broad Institute (https://gatk.broadinstitute.org/hc/en-us) for variant discovery, which included three stages: (1) read mapping, (2) variant discovery and genotyping and (3) variant quality recalibration (Supplementary Fig. 10). In the workflow, the known SNP sites from the high-coverage (>30×) set of 130 genomes, including 59 wild accessions covering the main distribution areas of two wild species (Supplementary Table 3) and 71 rice landraces representing all six cultivar groups (Supplementary Table 19), were used for conducting recalibrations. In the read mapping stage, we first used BWA (release 0.7.10) (ref. 58), Picard-tools (release 1.119; http://broadinstitute.github.io/picard/) and SAMtools (release 1.1)[59] to build an index for the reference sequence of the *O. sativa* Nipponbare genome (IRGSP Build 5)[40] and aligned the short reads of each sample to this reference using the BWA-MEM algorithm (Supplementary Fig. 10). Applying Picard-tools, we then transformed the format of the initial alignment file of each sample from SAM to BAM, sorted the aligned reads according to their positions on chromosomes, masked duplicate reads and built an index for each BAM file. Next, using Genome Analysis Toolkit (GATK) (release 3.6), we conducted local realignment with the tools RealignerTargetCreator and IndelRealigner to refine the aligned reads around indels and obtained an initial recalibrated alignment of each sample.

To avoid potential biases in variant discovery and genotyping resulting from differences in sequencing depth and sources, we picked out the initial recalibrated alignments of 130 samples with a sequencing depth higher than 30× to construct SNP sites as 'known' variants. Using UnifiedGenotyper in GATK, we conducted multi-sample variant calling and obtained raw SNP sites of variation for the 130 samples. The raw SNP data were filtered with VariantFiltration in GATK at a strict standard (FS < 10, QD > 10, MQ > 40, AN > 156, ReadPosRankSum > −0.5, BaseQRankSum > −0.5, MQRankSum > −0.5) following Ronco et al.[60]. On the basis of the known SNPs, we finally used BaseRecalibrator and PrintReads in GATK to conduct a base quality score recalibration for the initial recalibrated alignments of each sample and got an analysis-ready alignment of each sample, which we also used to determine the coverage distribution of mapped reads in each sample using DepthOfCoverage in GATK.

In the second stage, we used HaplotypeCaller in GATK to discover variants by applying a minimum base quality score of 20 and generated an intermediate GVCF file for each sample. According to their relationships, the GVCF files of all samples were combined into different groups of GVCF files with CombineGVCFs in GATK, and the combined GVCF files were genotyped with GenotypeGVCFs in GATK and produced variants and genotype data for all samples. We used SelectVariants in GATK to pick out SNPs and indels.

In the third stage for variant quality recalibration and for acquiring reliable data on indels, we removed indel variants with low quality (FS > 30, MQ < 20, QD < 5) using VariantFiltration in GATK. To obtain reliable SNPs, we filtered the raw SNPs using a two-step process called variant quality score recalibration (https://gatk.broadinstitute.org/hc/en-us). First, the VariantRecalibrator tool in GATK was applied to known variant sites, to the raw SNPs and to the annotations from FS, QD, DP, MQ, ReadPosRankSum, MQRankSum and BaseQRankSum to build a recalibration model for scoring variant quality. Second, depending on the recalibration table produced, we used ApplyVQSR in GATK with a score cut-off of 90 to filter raw SNPs. In addition, SNPs with AN < 1,578 (that is, with less than half of samples with genotypes) were excluded.

https://doi.org/10.1038/s41477-023-01476-z

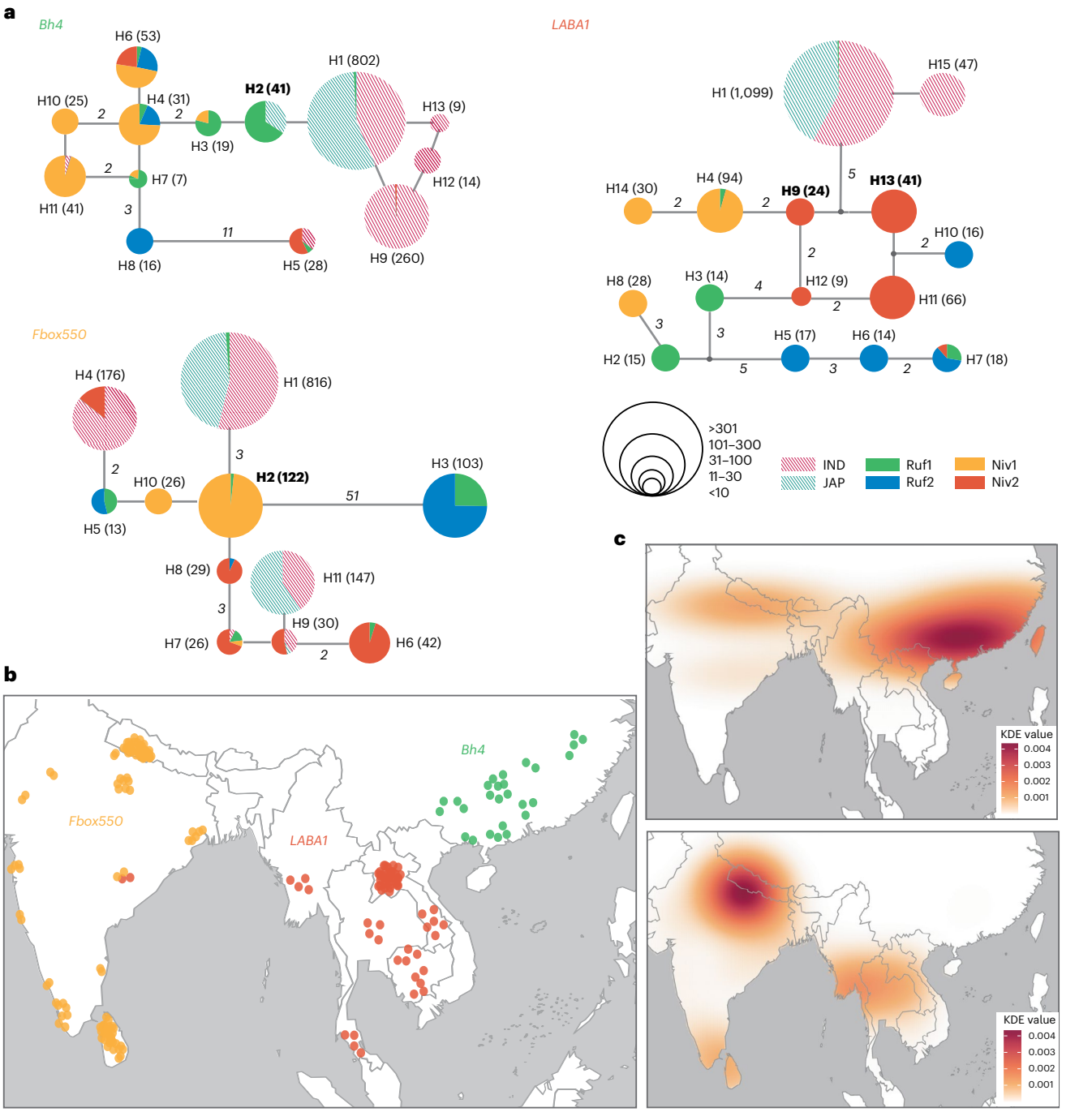


Fig. 4 | Inference of origins of domestication alleles based on haplotype network analyses. a, Haplotype network plots of three representative domestication genes, including *Bh4* (which includes domestication alleles from *japonica*) and *LABA1* and *Fbox550* (which include domestication alleles from *indica*). Each circle represents a haplotype, and its size is proportional to the number of haplotypes. The dots between haplotypes denote inferred haplotypes not recovered in the dataset. Haplotype names are shown beside the circles with the exact number of haplotypes in parentheses. H1 in the plots is the domestication haplotype that is common to *japonica* and *indica* in sequence and the largest in number. Haplotypes in bold are W-H1, a haplotype that is mainly composed of the components of wild lineages and is the nearest to H1. The number of substitutions next to a branch is shown in italic for branches with more than one substitution. Haplotypes with frequencies less than 0.5% are not included in the network. **b**, Geographic locations of the wild accessions that have the wild haplotypes most closely related the H1s of the genes *LABA1* (red), *Bh4* (green) and *Fbox550* (yellow). **c**, Heat maps showing the geographic densities of the *O. rufipogon* accessions for 22 genes that include domestication alleles from *japonica* (top) and of the *O. nivara* accessions for 8 genes that include domestication alleles from *indica* (bottom). KDE, kernel density estimation.

Inference of demographic history

We used the PSMC model[42,61], Stairway Plot 2 (ref. 43) and the SMC++ method[44] to infer the demographic history of wild and domesticated rice. PSMC is a standard method based on a high-coverage resequenced diploid individual or two haploids; Stairway Plot 2 and SMC++ are more powerful than PSMC for inference of demographic history on very recent timescales and are based on sequences of population samples[44,46,62]. Moreover, on the basis of folded SNP frequency spectra,

https://doi.org/10.1038/s41477-023-01476-z

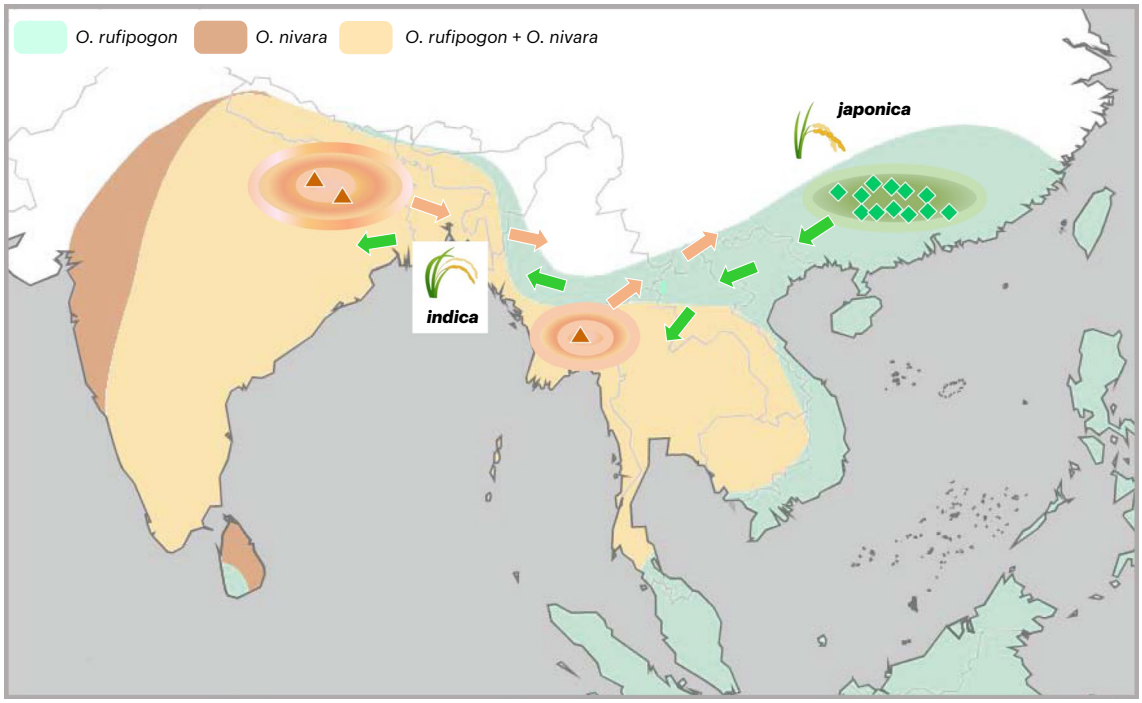


Fig. 5 | Hypothesized domestication centres of Asian rice. The geographic distributions of two ancestral species (*O. rufipogon* and *O. nivara*) are shaded in different colours. The origin and dispersal of the domesticated alleles shared by *japonica* and *indica* are depicted, with green diamonds indicating those domesticated in *japonica* (~80%) and brown triangles indicating those domesticated in *indica* (~20%). Note that the domestication event plotted in Southeast Asia might be relevant to the origin of the minor cultivar groups as suggested previously[18,37,53,54] and needs further investigation.

Stairway Plot 2 performed better than PSMC-like approaches for demographic inference on recent timescales[43].

To infer changes in N_e of wild and domesticated rice, we first performed PSMC using 74 high-coverage genomes (18–240×) (40 wild and 34 cultivated accessions), selected on the basis of phylogenetic and population genetic analyses (Supplementary Table 20). For each of the 74 genomes, we used reads with a minimum mapping quality of 20 and bases with a minimum score of 20 to call genotypes with the mpileup command in SAMtools (release 1.2) (ref. 59). The sites were retained when they had a depth of coverage of 0.5-fold to 2-fold relative to the mean genomic coverage. Because *O. nivara* and domesticated rice are both predominantly selfing, the decreased heterozygosity and longer runs of homozygosity in these species may skew demographic inference. We therefore created pseudodiploid genomes as applied in previous studies[62–64] and conducted PSMC to infer N_e changes of *O. nivara* and Asian rice with bamCaller.py and generate_multihetsep.py in MSMC-Tools (https://github.com/stschiff/msmc-tools) and MSMC2 (version 2.0.0) (ref. 61).

To infer the demographic history of domesticated rice, whose domestication started ~9,000 years ago[9,33], we performed both Stairway Plot 2 (version 2) and SMC++ (version 1.15.3) analysis using the SNP set of 1,493 samples excluding admixed accessions (Supplementary Table 18). The VCF files for each group were converted as input using the vcf2smc subcommand, and the long uncalled or runs-of-homozygosity regions were masked by filtering the sliding windows without called SNPs.

A new strategy to test for alterative hypotheses of domestication

We proposed a new strategy to test for alternative hypotheses by identifying the selective sweep regions or selected genes common to both *japonica* and *indica* (that is, single origin) (Extended Data Fig. 6 and Supplementary Section 7). We chose *japonica* (*temperate japonica* + *tropical japonica*) and *indica* in the analysis partly because they represent two distinct groups with deep divergence and have been focal groups of all previous studies involving rice domestication and partly because three minor cultivar groups (*aus*, *aromatic* and *rayada*) were cultivated in very limited areas with uncertain origins[14,15,18,24,37]. First, we conducted genome-wide scans to identify the selective sweep regions/selected genes that are common to *japonica* and *indica* by calculating $ROD_{W/C}$ and genetic differentiation (F_{ST})[65] between wild and domesticated populations. We defined such a selective sweep region/selected gene as a PSR or PSG (Extended Data Fig. 6a). Unlike many studies that detected selective sweep regions by scanning each cultivar group separately for a crop[16,47–49], we performed the scan using the entire wild rice gene pool as the wild population and the combined *japonica* and *indica* gene pool as the domesticated population. This method could avoid some of the complexity caused by widespread gene flow between rice cultivars[14,15,30,31,33], because gene flow can lead to a high proportion of false positives in genome scans[50,66]. It should be mentioned, however, that this strategy was unable to identify all the selective sweep regions/genes associated with domestication because those unique to specific cultivar groups are undetectable using this approach.

Second, unlike previous studies that generated phylogenetic trees using the concatenated selective sweep regions[18,27], we reconstructed phylogenetic trees using each PSR/PSG to identify PSR/PSGs that generated a tree on which *japonica* and *indica* formed a monophyletic group (hereafter referred to as a cultivar group). Such a PSR/PSG would originate only once because it is unlikely for such a selective sweep region/gene to be common in sequence by being selected independently from *O. rufipogon* and *O. nivara*. We defined such a PSR/PSG as a SOR/SOG (Extended Data Fig. 6b). Finally, we classified the SOR/SOG trees into categories according to the wild lineage sister to the cultivar group so that we were able to effectively test alternative hypotheses of rice domestication (Extended Data Fig. 6c).

To identify the PSRs, we first calculated $ROD_{W/C}$ and F_{ST} in 100 kb sliding windows with a step of 10 kb along chromosomes. Of 38,167 total windows, we retained 36,454 windows with more than ten SNPs

for subsequent analyses to ensure that each window included sufficient polymorphism information. To determine the thresholds of $ROD_{W/C}$ and F_{ST} for a PSR, we calculated the mean values of $ROD_{W/C}$ and F_{ST} and their standard deviations by excluding the 935 windows with $ROD_{W/C}$ values more than two standard deviations from the mean (that is, the top and bottom 5%). Then, using the mean values and standard deviations, we Z-transformed the $ROD_{W/C}$ and F_{ST} distributions and chose $Z > 2$, corresponding to 2.615 for $ROD_{W/C}$ and 0.307 for F_{ST}, as the thresholds for a window under selection. Finally, windows with $ROD_{W/C}$ and F_{ST} values above the thresholds were determined to be PSRs (Supplementary Section 7). Adjacent windows with $ROD_{W/C}$ and F_{ST} values over the thresholds were merged into a larger PSR.

Similarly, we performed genome-wide scans for PSGs on the basis of $ROD_{W/C}$ and F_{ST} values for each of the 34,791 genes that were defined according to gene structure information for the rice reference genome (Nipponbare)[67] (Supplementary Section 8). To trace the evolutionary path of the PSRs/PSGs identified, we built NJ trees for wild lineages and cultivar groups using phylip (version 3.696) (ref. 68) based on F_{ST} values calculated using the SNPs of PSRs/PSGs. We used the *O. rufipogon* populations from Papua New Guinea and Australia as outgroups because these populations are distinct from the rest of the species[25,69]. Similarly, to infer the origin of the PSGs, we reconstructed NJ trees based on F_{ST} values calculated using the SNPs extracted from three fragments spanning the PSGs—that is, the gene only, the gene and its 1 kb upstream and downstream regions (gene+1k), and the gene and its 3 kb upstream and downstream regions (gene+3k). For each gene, we performed bootstrap resampling 10,000 times and generated 10,000 NJ trees, which were used to calculate bootstrap supports using the program consense in phylip[68].

Population genetic and phylogenetic analyses

Nucleotide diversity, including π (ref. 70) and θ (ref. 71), was estimated at different hierarchical levels through scanning the whole genome with non-overlapping 10 kb windows and 100 kb sliding windows (10 kb steps), respectively. We calculated the number of private alleles, defined as the alleles with frequency >5% in a cultivar group but absent in other groups. We estimated the LD decay using PopLDdecay (version 3.29) (ref. 72) for each wild lineage and cultivar group, as well as for the combined wild and cultivated accessions. SNPs with minor allele frequency (MAF) >0.5% were used to calculate the squared correlation coefficient (r^2) with the maximum distance between pairwise SNPs set to 300 kb.

The population genetic structure of wild and domesticated rice was explored using ADMIXTURE (version 1.3.0) (ref. 73) on the basis of the LD-pruned pseudomolecule SNP data[64] with the parameter -indep-pairwise, 200, 100, 0.1 in PLINK (version 1.90 beta)[74]. The ancestral genetic component of each accession was inferred with fivefold cross-validation by increasing K (the number of clusters or groups) from 2 to 12 and plotted using barplotfunction in R (ref. 75). PCA was performed to study relatedness and clustering among populations or samples using GCTA software (version 1.24.4) (ref. 76) with the parameter -pca, 20 to output the first 20 and all eigenvalues. The first two PCs were plotted using the R package ggplot2 (ref. 77). To reconstruct NJ trees using SNP data, we estimated the pairwise genetic distance matrix using PLINK[74] with the parameters -distance, 1-ibs, flat-missing and reconstructed the trees on the basis of the distance matrix using MEGA (version 5) (ref. 78). We also performed analyses of ADMIXTURE, PCA and NJ trees based on the SNPs from intergenic regions that were commonly regarded as neutral sites.

We further explored the phylogenetic relationships of the wild lineages and cultivar groups by reconstructing single-gene trees using 34,791 annotated genes. NJ trees were built using phylip (version 3.696) (ref. 68) on the basis of F_{ST} values calculated using the SNPs in the region of the gene and its 1 kb upstream and downstream regions. Gene trees were displayed in DensiTree (version 2.2.7) (ref. 79).

Compilation of a database of domestication-relevant genes

To facilitate the analyses of domestication genes, we compiled a database of genes relevant to rice domestication on the basis of whether (1) they have been identified in rice with their function largely known and (2) they exhibited a signature of selection in both *japonica* and *indica* in this study. Specifically, by searching the literature and by surveying two well-known databases of rice functional genes, Q-TARO[80] and funRiceGenes[81], which consist of 1,949 and 3,713 rice genes, respectively, we obtained a subset of 192 genes (Supplementary Table 15) that overlapped with the 1,882 selected genes that were identified in this study and shared by *japonica* and *indica* (Supplementary Table 11). This gene set served as the source for in-depth analyses but was conservative with respect to domestication functions because we did not consider genes that were domesticated in and confined to a specific group.

Haplotype network analysis of domestication genes

We performed haplotype network analyses to identify the domestication alleles and their wild contributors. First, we obtained the haplotypes with frequencies >0.5%, defined as the common haplotype, following the workflow outlined in Supplementary Fig. 7 and retained them in the reconstruction of haplotype networks because the haplotypes with frequencies <0.5% were not informative. We then focused on the analysis of two types of haplotype in the networks: (1) the domestication haplotype (H1), a haplotype that was shared by *japonica* and *indica* and was the dominant haplotype in the samples of cultivars (the most numerous), and (2) haplotype W-H1, a haplotype that was mainly composed of the components of a wild lineage or lineages and was the most closely related to H1. The presence of H1 for a domestication gene indicated that there was a dominant allele that was shared by all cultivars and probably originated once due to domestication, while W-H1 provided evidence of the wild lineage(s) that may have contributed to the origin of H1.

Haplotype networks were constructed using SNPs within the fragment spanning each gene and its 1 kb upstream and downstream regions (Supplementary Fig. 7). The reason for including the 1 kb upstream and downstream regions was to obtain sufficient polymorphism for ensuring resolution while excluding uninformative haplotypes with low frequency because the number of haplotypes increased quickly due to fast decay of LD in wild rice. First, to obtain accurate haplotypes, we imputed and phased variants of the segment from 200 kb upstream to 200 kb downstream of the gene following the approach of Todesco et al.[82]. Specifically, we selected calling variants (SNPs) of the segment from 200 kb upstream to 200 kb downstream of the gene and filtered SNPs and samples to improve the reliability of imputation and phasing. To ensure the quality of variants, we retained only biallelic SNPs and removed SNPs with MAF <5% and missing rate >10% and obtained a new sample set by filtering out the accessions with a ratio of missing sites over 20% in the fragment spanning from 1 kb upstream to 1 kb downstream of the gene. Next, we implemented Beagle (version 5.0) (ref. 83) to impute and phase the variants of the segment from 200 kb upstream to 200 kb downstream of the gene for a new sample set. On this basis, we acquired the haplotypes for the fragment from 1 kb upstream to 1 kb downstream of all the genes under study. Finally, we retained the haplotypes by filtering out those with frequency <0.5% that were not informative. Each specific haplotype contained only identical sequences. Network construction was performed with the median-joining method using the software Network Publisher (version 2.1.2.5) (ref. 84). Most of these analyses were performed using customized Perl scripts.

To help infer potential areas of rice domestication, we drew a heat map of geographic density of all wild accessions that included the haplotypes nearest to the domestication haplotype, in terms of the number of mutations, for all analysed genes. The geographic density was estimated by kernel methods using the sm.density function of the R package sm[85] and visualized using the R packages sf[86] and ggplot2 (ref. 77).

Test for gene flow

To explore whether and to what extent gene flow has taken place between cultivar groups and between wild and domesticated rice, we performed two formal tests for admixture: the three-population test (f_3) (ref. 51) and *D*-statistics[51,87]. The f_3 test involving three populations (A, B and C) was used to test whether population C had ancestry from populations related to A and B. In the *D*-statistics involving four populations (W, X, Y and Z), two patterns were compared: the ABAB pattern, in which W and Y share one allele while X and Z share the other, and the ABBA pattern, in which W and Z (X and Y) share the same allele. By assigning the outgroup as population Z, we obtained a significantly positive *D*-statistic if there was gene flow between W and Y and a significantly negative *D*-statistic under the presence of gene flow between X and Y (refs. 87,88). We used ADMIXTOOLS (version 7.0.1) (ref. 51) to calculate the f_3 (C; A, B) and *D*-statistics. Because f_3 required high-confidence genotype calls, we used the same sample set as that for haplotype network analysis (Supplementary Table 21) and removed the SNPs with MAF <1% and missing rate >40% to ensure the quality of variants. We used *O. barthii*, which was the most closely related to the *O. sativa* complex[20], as the outgroup in the analyses. Two de novo assemblies of *O. barthii*[89,90] were downloaded, and the *O. barthii* alleles were treated as ancestral.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All newly resequenced genomes have been deposited in the National Center for Biotechnology Information Sequence Read Archive under number PRJNA705309. The genomic SNP data from all samples are available on Dryad (https://datadryad.org/stash/share/jcQfZcbai80MmLb6kO4_mrQLfu1tX-I_1_Yx7hAfJkI)[91].

Code availability

All code and scripts used in the analyses are available at https://github.com/zhangfumin/domestication.

References

1. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
2. Allaby, R. G., Fuller, D. Q. & Brown, T. A. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc. Natl Acad. Sci. USA* **105**, 13982–13986 (2008).
3. Glemin, S. & Bataillon, T. A comparative view of the evolution of grasses under domestication. *N. Phytol.* **183**, 273–290 (2009).
4. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
5. Larson, G. et al. Current perspectives and the future of domestication studies. *Proc. Natl Acad. Sci. USA* **111**, 6139–6146 (2014).
6. Gaut, B. S., Seymour, D. K., Liu, Q. & Zhou, Y. Demography and its effects on genomic variation in crop domestication. *Nat. Plants* **4**, 512–520 (2018).
7. Chang, T.-T. The origin, evolution, cultivation, dissemination, and diversification of Asian and African rices. *Euphytica* **25**, 425–441 (1976).
8. Oka, H. *Origin of Cultivated Rice* (Elsevier, 1988).
9. Gross, B. L. & Zhao, Z. Archaeological and genetic insights into the origins of domesticated rice. *Proc. Natl Acad. Sci. USA* **111**, 6190–6197 (2014).
10. Khush, G. S. Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35**, 25–34 (1997).
11. Fuller, D. Q. Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World. *Ann. Bot.* **100**, 903–924 (2007).
12. Olsen, K. M. & Wendel, J. F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
13. Sang, T. & Ge, S. Understanding rice domestication and implications for cultivar improvement. *Curr. Opin. Plant Biol.* **16**, 139–146 (2013).
14. Kovach, M. J., Sweeney, M. T. & McCouch, S. R. New insights into the history of rice domestication. *Trends Genet.* **23**, 578–587 (2007).
15. Sang, T. & Ge, S. Genetics and phylogenetics of rice domestication. *Curr. Opin. Genet. Dev.* **17**, 533–538 (2007).
16. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
17. Second, G. Origin of the genic diversity of cultivated rice (*Oryza* spp.): study of the polymorphism scored at 40 isozyme loci. *Jpn. J. Genet.* **57**, 25–57 (1982).
18. Civan, P., Craig, H., Cox, C. J. & Brown, T. A. Three geographically separate domestications of Asian rice. *Nat. Plants* **1**, 15164 (2015).
19. Vitte, C., Ishii, T., Lamy, F., Brar, D. & Panaud, O. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). *Mol. Genet. Genomics* **272**, 504–511 (2004).
20. Zhu, Q. & Ge, S. Phylogenetic relationships among A-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. *New Phytol.* **167**, 249–265 (2005).
21. He, Z. et al. Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genet.* **7**, e1002100 (2011).
22. 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).
23. Ishikawa, R., Castillo, C. C. & Fuller, D. Q. Genetic evaluation of domestication-related traits in rice: implications for the archaeobotany of rice origins. *Archaeol. Anthropol. Sci.* **12**, 197 (2020).
24. Vaughan, D. A., Lu, B.-R. & Tomooka, N. The evolving story of rice evolution. *Plant Sci.* **174**, 394–408 (2008).
25. Liu, R., Zheng, X.-M., Zhou, L., Zhou, H.-F. & Ge, S. Population genetic structure of *Oryza rufipogon* and *Oryza nivara*: implications for the origin of *O. nivara*. *Mol. Ecol.* **24**, 5211–5228 (2015).
26. Cai, Z. et al. Parallel speciation of wild rice associated with habitat shifts. *Mol. Biol. Evol.* **36**, 875–889 (2019).
27. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
28. Londo, J. P., Chiang, Y.-C., Hung, K.-H., Chiang, T.-Y. & Schaal, B. A. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl Acad. Sci. USA* **103**, 9578–9583 (2006).
29. Molina, J. et al. Molecular evidence for a single evolutionary origin of domesticated rice. *Proc. Natl Acad. Sci. USA* **108**, 8351–8356 (2011).
30. Wang, H., Vieira, F. G., Crawford, J. E., Chu, C. & Nielsen, R. Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. *Genome Res.* **27**, 1029–1038 (2017).
31. Choi, J. Y. & Purugganan, M. D. Multiple origin but single domestication led to *Oryza sativa*. *G3 (Bethesda)* **8**, 797–803 (2018).
32. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
33. Fuller, D. Q. Pathways to Asian civilizations: tracing the origins and spread of rice and rice cultures. *Rice* **4**, 78–92 (2011).

34. Bates, J., Petrie, C. A. & Singh, R. N. Approaching rice domestication in South Asia: new evidence from Indus settlements in northern India. *J. Archaeol. Sci.* **78**, 193–201 (2017).
35. Glaszmann, J. C. Isozymes and classification of Asian rice varieties. *Theor. Appl. Genet.* **74**, 21–30 (1987).
36. Wang, C. H. et al. Genetic diversity and classification of *Oryza sativa* with emphasis on Chinese rice germplasm. *Heredity* **112**, 489–496 (2014).
37. Choi, J. Y. et al. The rice paradox: multiple origins but single domestication in Asian rice. *Mol. Biol. Evol.* **34**, 969–979 (2017).
38. Yang, C.-C. et al. Independent domestication of Asian rice followed by gene flow from *japonica* to *indica*. *Mol. Biol. Evol.* **29**, 1471–1479 (2012).
39. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
40. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
41. Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S. & McCouch, S. Genetic structure and diversity in *Oryza sativa* L. *Genetics* **169**, 1631–1638 (2005).
42. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
43. Liu, X. & Fu, Y.-X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol.* **21**, 280 (2020).
44. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
45. Zheng, X.-M. & Ge, S. Ecological divergence in the presence of gene flow in two closely related *Oryza* species (*Oryza rufipogon* and *O. nivara*). *Mol. Ecol.* **19**, 2439–2454 (2010).
46. Cubry, P. et al. The rise and fall of African rice cultivation revealed by analysis of 246 new genomes. *Curr. Biol.* **28**, 2274–2282 (2018).
47. Schmutz, J. et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014).
48. Zhao, G. et al. A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. *Nat. Genet.* **51**, 1607–1615 (2019).
49. Zhang, K. et al. Resequencing of global Tartary buckwheat accessions reveals multiple domestication events and key loci associated with agronomic traits. *Genome Biol.* **22**, 23 (2021).
50. Nielsen, R. et al. Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
51. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
52. Gutaker, R. M. et al. Genomic history and ecology of the geographic spread of rice. *Nat. Plants* **6**, 492–502 (2020).
53. Civan, P. et al. Origin of the *Aromatic* group of cultivated rice (*Oryza sativa* L.) traced to the Indian subcontinent. *Genome Biol. Evol.* **11**, 832–843 (2019).
54. Choi, J. Y. et al. Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biol.* **21**, 21 (2020).
55. Nosil, P. *Ecological Speciation* (Oxford Univ. Press, 2012).
56. Via, S. Natural selection in action during speciation. *Proc. Natl Acad. Sci. USA* **106**, 9939–9946 (2009).
57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Ronco, F. et al. Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature* **589**, 76–81 (2021).
61. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
62. Schiffels, S. & Wang, K. in *Statistical Population Genomics* (ed. Dutheilm, J. Y.) 147–166 (Springer US, 2020).
63. Thomas, C. G. et al. Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. *Genome Res.* **25**, 667–678 (2015).
64. Meyer, R. S. et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* **48**, 1083–1088 (2016).
65. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
66. Weigand, H. & Leese, F. Detecting signatures of positive selection in non-model species using genomic data. *Zool. J. Linn. Soc.* **184**, 528–583 (2018).
67. Sakai, H. et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* **54**, e6 (2013).
68. Felsenstein, J. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
69. Banaticla-Hilario, M. C. N., McNally, K. L., van den Berg, R. G. & Sackville Hamilton, N. R. Crossability patterns within and among *Oryza* series *Sativae* species from Asia and Australia. *Genet. Resour. Crop Evol.* **60**, 1899–1914 (2013).
70. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
71. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
72. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
73. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
74. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
75. R Core Team. R: A Language and Environment for Statistical Computing version 3.2.1 (R Foundation for Statistical Computing, 2015).
76. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
77. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
78. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
79. Bouckaert, R. R. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**, 1372–1373 (2010).
80. Yamamoto, E., Yonemaru, J.-i, Yamamoto, T. & Yano, M. OGRO: the overview of functionally characterized genes in rice online database. *Rice* **5**, 26 (2012).
81. Yao, W., Li, G., Yu, Y. & Ouyang, Y. funRiceGenes dataset for comprehensive understanding and application of rice functional genes. *GigaScience* **7**, 1–9 (2018).
82. Todesco, M. et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).

83. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
84. Bandelt, H.-J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
85. Bowman, A. W. & Azzalini, A. sm: Smoothing methods for nonparametric regression and density estimation. R package version 2.2-5.7 http://www.stats.gla.ac.uk/~adrian/sm (2021).
86. Pebesma, E. Simple features for R: standardized support for spatial vector data. *R J.* **10**, 439–446 (2018).
87. Green, R. E. et al. A draft sequence of the neandertal genome. *Science* **328**, 710–722 (2010).
88. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
89. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
90. Ma, X. et al. Whole-genome de novo assemblies reveal extensive structural variations and dynamic organelle-to-nucleus DNA transfers in African and Asian rice. *Plant J.* **104**, 596–612 (2020).
91. Zhang, F.-M., Jing, C.-Y. & Ge, S. Genomic SNP data for domestication research of Asian cultivated rice. *Dryad* https://doi.org/10.5061/dryad.xksn02vjd (2022).

Acknowledgements

We thank T. Sang, Y.-L. Guo and J.-L. Li for discussion and suggestions; B.-R. Lu, W.-L. Chen and D. Ratnasekera for field collections; X.-H. Wei, C.-B. Chen, H.-Z. Zeng and other members of S.G.'s laboratory for phenotyping and lab assistance; and A.-L. Li for picture drawing. We also thank the International Rice Research Institute (Los Banos, Philippines) and the China National Rice Research Institute (CNRRI) (Hangzhou, China) for providing seed samples and the CNRRI, the Guangxi Academy of Agricultural Sciences (Nanning, China) and the CAS Field Station (Lingshui, China) for providing the experimental fields. This work was financially supported by funding from the National Natural Science Foundation of China (grant no. 91231201), the Strategic Priority Research Program of Chinese Academy of Sciences (grant nos. XDB31000000 and XDA08020103) and the Ministry of Science and Technology (grant no. 2021YFD1200101-02) to S.G.; the National Natural Science Foundation of China (grant nos. 91731301 and 32130008 to S.G., 31470332 to F.-M.Z. and 31800186 to Z.C.); and the China Postdoctoral Science Foundation (grant no. 2017M620950 to Z.C.).

Author contributions

S.G. conceived and designed the project. S.G. and F.-M.Z. supervised the research. C.-Y.J., X.-H.W., F.-M.Z., M.-X.W., L.Z. and L.H. obtained and analysed the genomic data. F.-M.Z., S.G., C.-Y.J., Z.C., X.-H.W., M.-X.W. and L.Z. performed the population genetic analyses. F.-M.Z., M.-X.W., W.-H.Y. and C.-Y.J. performed the haplotype analyses. S.G., L.Z., M.-X.W., M.-F.G., Q.-L.M., N.-N.R., X.-M.Z., R.L., L.H., Y.-S.D., X.W. and C.-G.Q. conducted the field collections and phenotyping. L.Z., M.-F.G., Q.-L.M., J.-D.H., Z.-H.J., H.-X.Z. and X.-H.Z. were involved in lab assistance. S.G., B.S.G. and F.-M.Z. wrote the manuscript with help from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

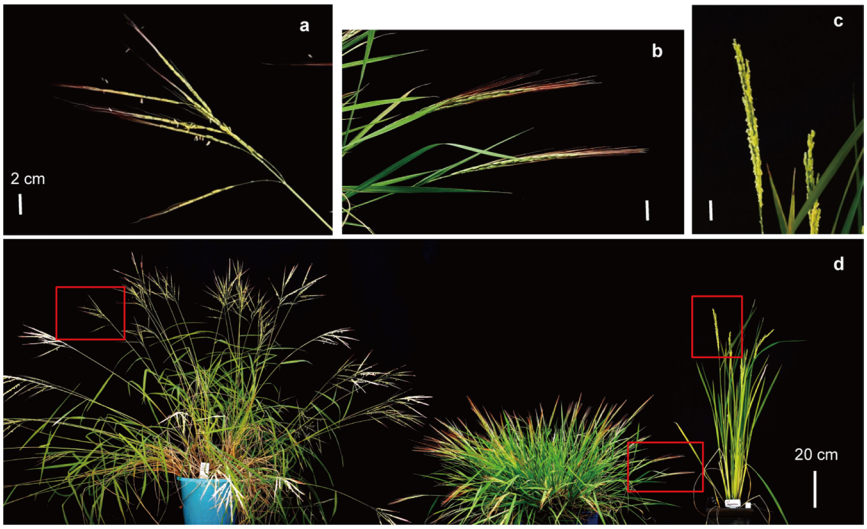
Extended data is available for this paper at https://doi.org/10.1038/s41477-023-01476-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-023-01476-z.

Correspondence and requests for materials should be addressed to Song Ge.

Peer review information *Nature Plants* thanks Feng Tian, Paul Gepts and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41477-023-01476-z



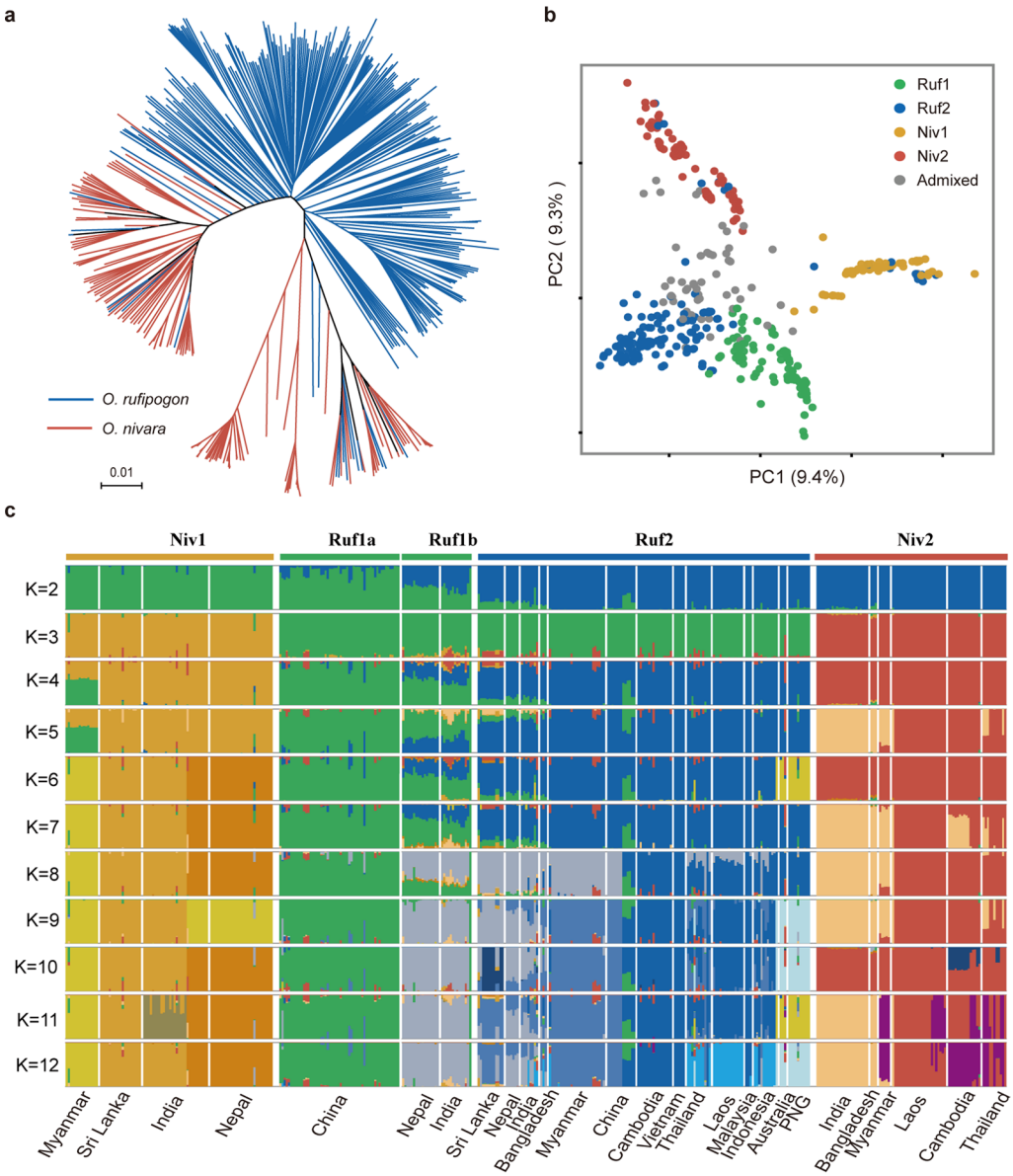
e

No.	Character	*O. rufipogon*	*O. nivara*	*O. sativa*
1	Anther length[2]	long (> 3.0 mm)	short (< 3.0 mm)	short (< 2.5mm)
2	Awn length[1]	long	long	none to short
3	Culm habit[1]	spreading (prostrated)	semierect	erect
4	Culm length[2]	long	short	varying
5	First heading (photoperiod sensitivity)[2]	late (high)	early (low)	varying (varying)
6	Life cycle	perennial	annual	annual
7	Panicle exsertion[2]	exserted	moderately inserted	varying
8	Panicle shape[1,2]	open	compact	closed
9	Outcrossing rate[2]	high	low	very low
10	Seed size[1]	small	small	large
11	Shattering[1]	Yes	Yes	No
12	Seed germination[1]	non-synchronous	non-synchronous	synchronous
13	Seed mature	inconsistent		uniform
14	Stigma exsertion	exserted	exserted	inserted
15	Grain yield	low	medium	high

Note: [1] diagnostic characters distinguishing the wild and cultivated rice; [2] diagnostic characters separating *O. rufipogon* from *O. nivara*.

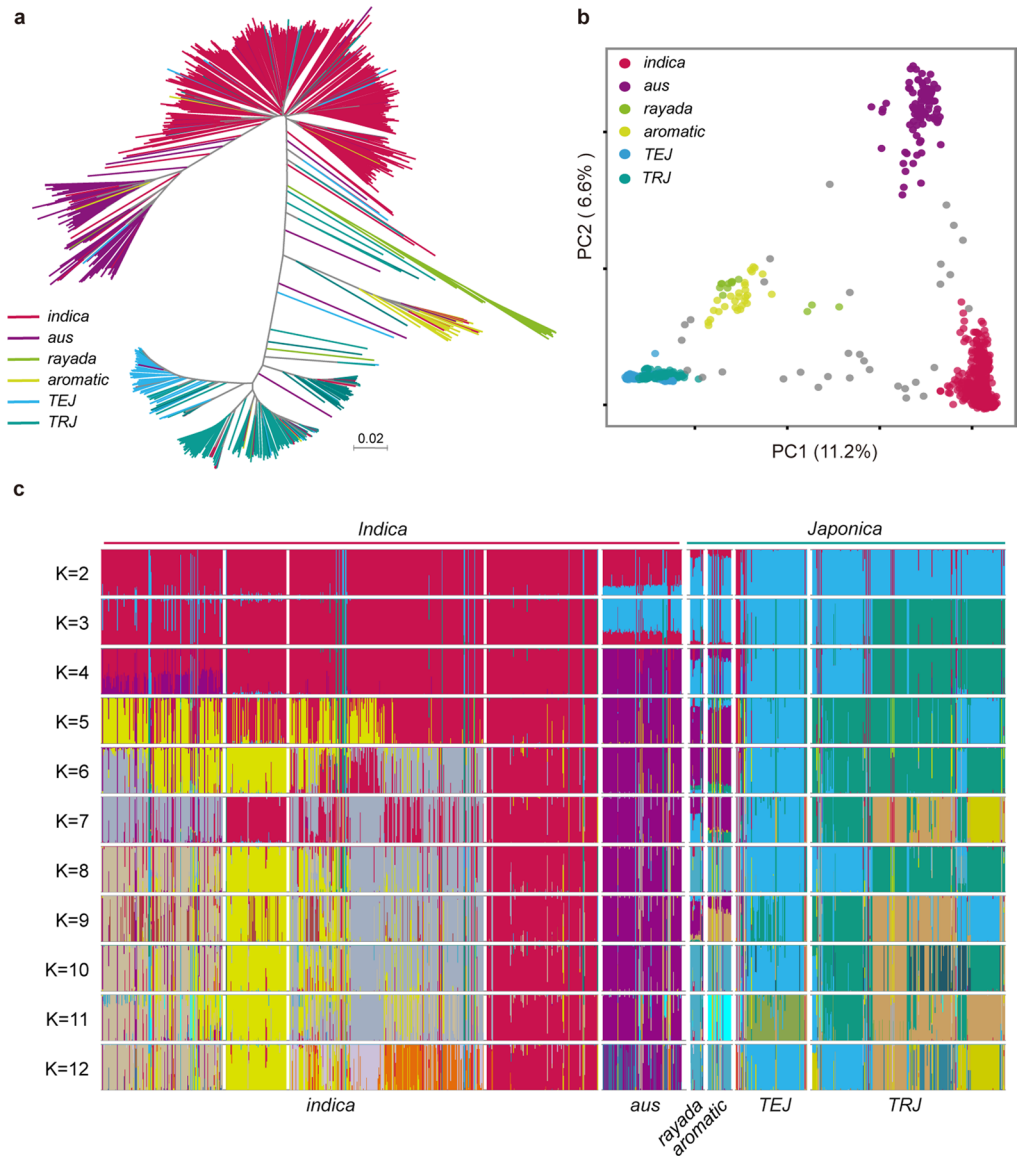
Extended Data Fig. 1 | Morphology and diagnostic features of wild rice (*O. rufipogon* and *O. nivara*) and domesticated rice (*O. sativa*). a, b, c. Panicles of *O. rufipogon*, *O. nivara* and Nipponbare (*O. sativa* ssp. *japonica*) in heading stage, respectively, which were indicated by red boxes in the bottom panel. **d**. Gross morphology of *O. rufipogon* (accession no. NEP04X4), *O. nivara* (accession no. NEP0202a) and Nipponbare. Two wild species were sampled from Nepal and were grown in the common garden in Beijing, together with Nipponbare. **e**. A list of diagnostic features to delineate wild rice and domesticated rice.

https://doi.org/10.1038/s41477-023-01476-z



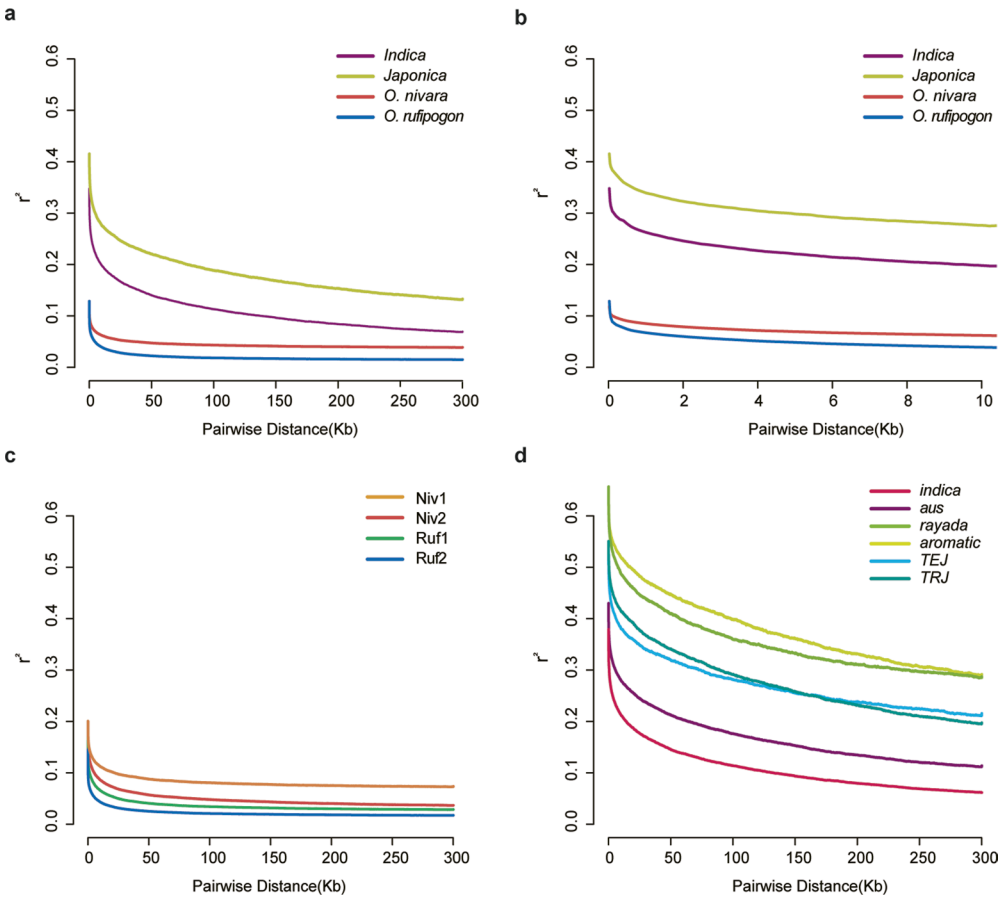
Extended Data Fig. 2 | Analyses of population genetic structure of two wild rice species (*O. rufipgon* and *O. nivara*). a. Neighbor-joining tree of 457 wild rice accessions. Lines in colors represent two species. The scale bar shows substitutions per site. **b.** Principal components of 457 wild rice accessions with mislabeled and admixed accessions indicated. Dots in colors indicate four genetic lineages and those in grey indicate admixed accessions. **c.** ADMIXTURE plots for 404 wild rice accessions excluding admixed accessions. Four different lineages are indicated above the plots. The columns represent the accessions with their origins indicated below the plots.

https://doi.org/10.1038/s41477-023-01476-z



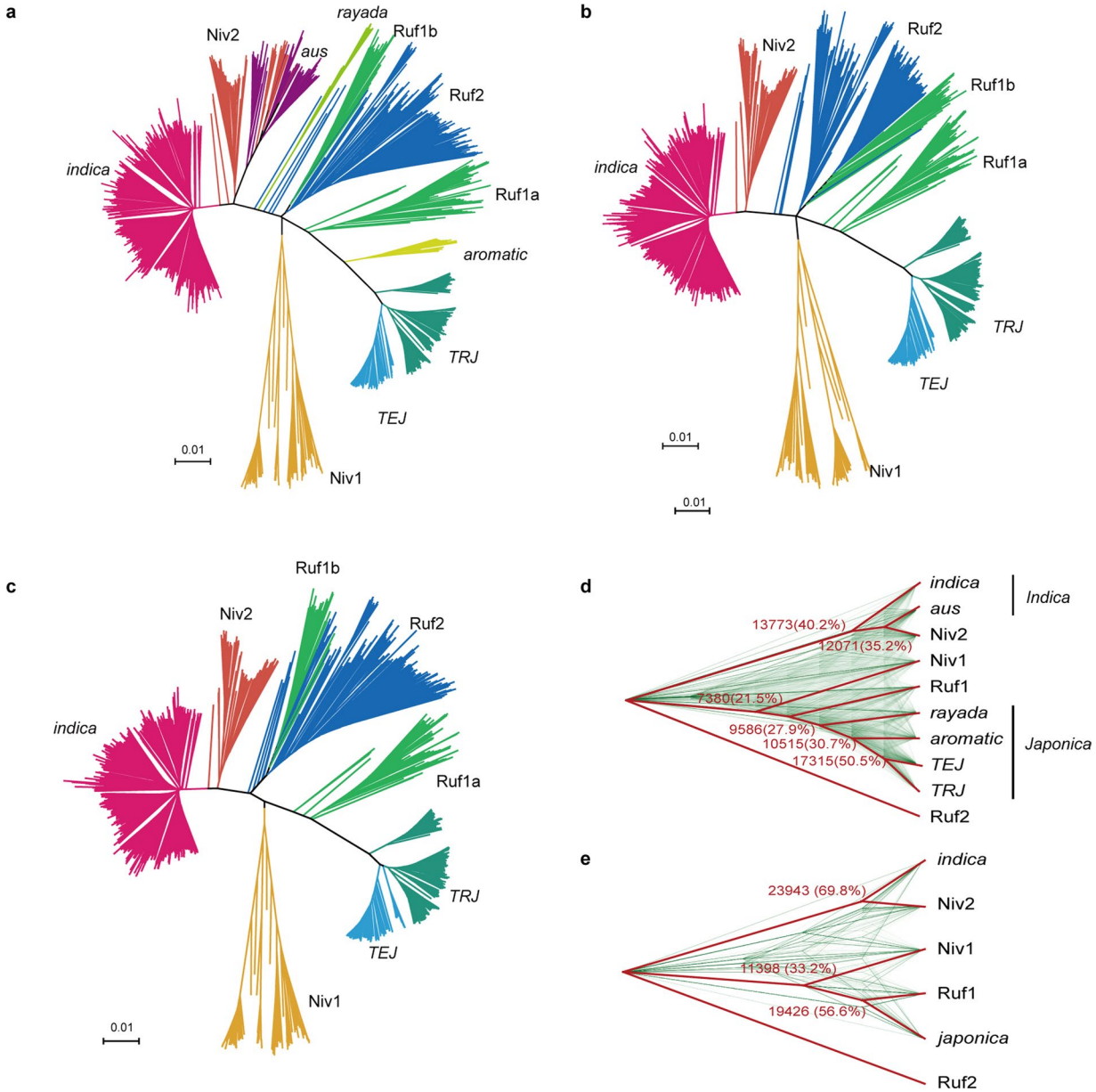
Extended Data Fig. 3 | Analyses of population genetic structure of Asian rice based on 1121 rice landraces with mislabeled and admixed accessions indicated. a. Neighbor-joining tree of 1121 rice landraces. Lines in colors represent six cultivar groups originally defined. The scale bar shows substitutions per site. **b**. Principal components of 1121 rice landraces. Dots in colors indicate six cultivar groups and those in grey indicate admixed accessions. **c**. ADMIXTURE plots for rice landraces. Two rice subspecies are indicated above the plots. Columns represent the landraces with the cultivar groups indicated below the plots.

https://doi.org/10.1038/s41477-023-01476-z



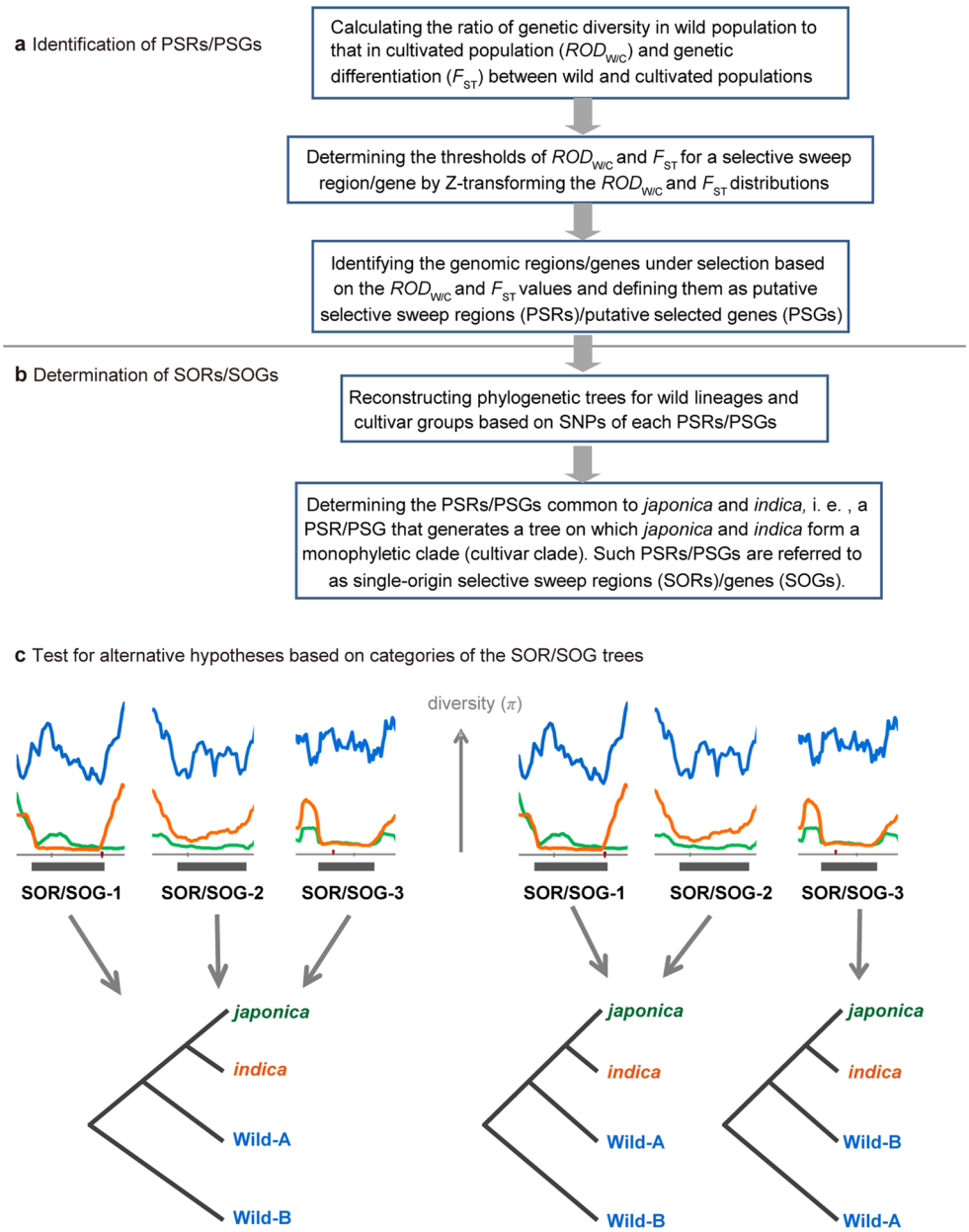
Extended Data Fig. 4 | Genome-wide linkage disequilibrium (LD) of wild and domesticated rice. a, b. Comparisons of LD patterns for two wild species (*O. rufipogon* and *O. nivara*) and two rice subspecies (*Japonica* and *Indica*). **c, d.** LD patterns of four lineages of wild species (**c**) and six cultivar groups (**d**). The x axis is the physical distance between pair of SNPs (kb) in wild and domesticated rice. The y axis is the squared allele-frequency correlations r^2.

https://doi.org/10.1038/s41477-023-01476-z



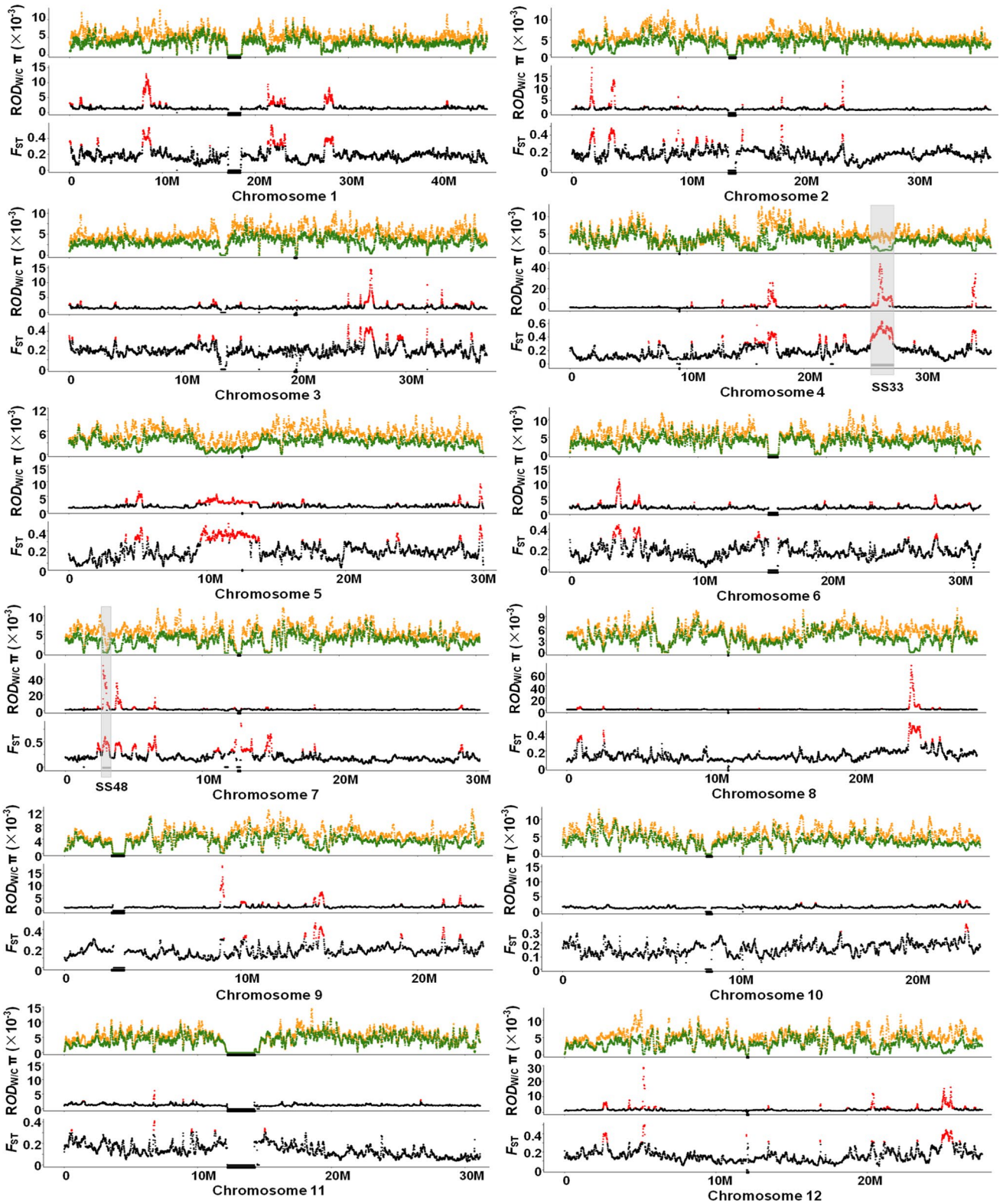
Extended Data Fig. 5 | Phylogenetic relationships of major lineages of wild rice and major cultivar groups of rice landraces. a. NJ tree of 1493 wild and domesticated rice accessions based on genetic distance calculated with neutral loci. **b**. NJ tree of 1355 wild and domesticated rice (*japonica* and *indica*) based on genetic distance calculated with neutral loci. **c**. NJ tree of 1355 wild and main cultivar groups based on all SNPs. Lines in colors represent wild lineages and cultivar groups. The scale bar shows substitutions per site. **d, e**. NJ trees constructed individually by 34,291 genes annotated based on all four wild lineages and six cultivar groups (**d**) and based on four wild lineages and three major cultivar groups (**e**). Heavy red lines indicate the cladogram supported by a majority of gene trees. Numbers near the nodes represent proportion of the genes that resolved notes on their trees. *japonica* = *temperate japonica* + *tropical japonica*.

https://doi.org/10.1038/s41477-023-01476-z



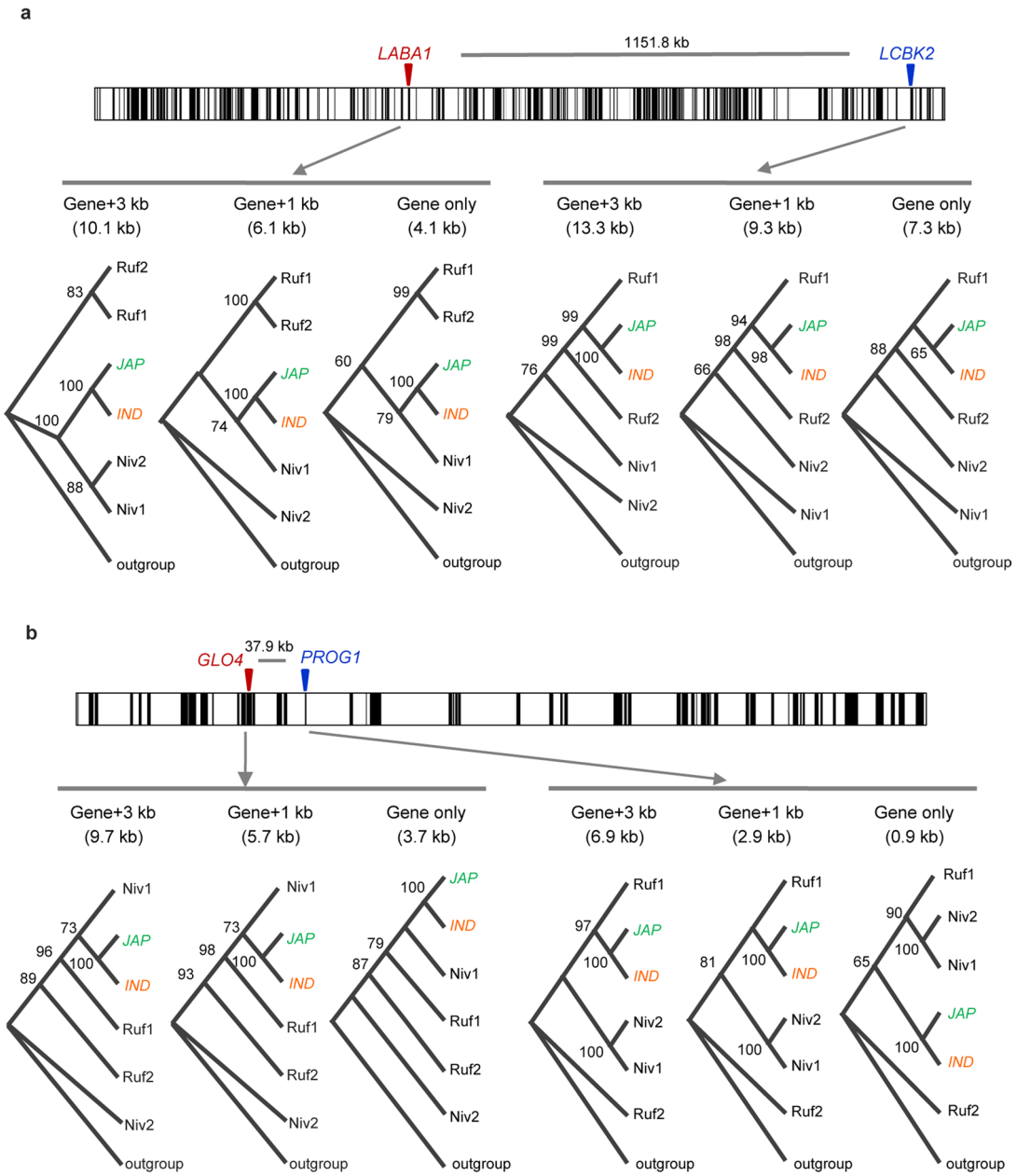
Extended Data Fig. 6 | Illustration of a new strategy to distinguish between single and multiple domestications based on phylogenetic analysis of single-origin selective sweep regions (SORs)/single-origin selected genes (SOGs). **a**. Identification of the genomic region/gene under selection common to both *japonica* and *indica*, defined as a putative selective sweep region (PSR)/putative selected gene (PSG). **b**. Determination of the PSR/PSG of single origin, that is, the PSR/PSG that generated a tree on which *japonica* and *indica* formed a group (cultivar group), defined as a single-origin selective sweep region (SOR)/single-origin selected gene (SOG). **c**. Test for hypotheses based on the phylogenetic trees of SORs/SOGs. A single domestication model is supported if the same wild lineage is sister to a cultivar group on the trees (left) and a multiple domestication model is supported if multiple wild lineages cluster with cultivar groups on different trees (right). Three lines above the panel indicate the diversity of wild (blue), *indica* (orange) and *japonica* (green) samples, respectively. Wild-A and Wild-B represent different wild lineages.

https://doi.org/10.1038/s41477-023-01476-z



Extended Data Fig. 7 | Genome-wide scan of putative selective sweep regions (PSRs) and single-origin selective sweep regions (SORs) in domesticated rice across 12 chromosomes. Nucleotide diversity (π) and its ratio ($ROD_{W/C}$) of wild to domesticated populations and divergence (F_{ST}) between wild and domesticated populations are plotted for 100kb sliding windows against the position across chromosomes. Dots in red are the outliers with $ROD_{W/C}$ and F_{ST} values over the thresholds in *japonica* and *indica*. Black rectangles on the chromosomes represent the centromeres. The shaded regions on chromosomes 4 and 7 are SOR33 and SOR48, respectively.

https://doi.org/10.1038/s41477-023-01476-z



Extended Data Fig. 8 | Schematic diagram of two single-origin selective sweep regions (SORs) (SOR33 and SOR48) and NJ trees of domestication genes within the SORs. a. SOR33 locates on chromosome 4 and includes *LABA1* (with domestication alleles from *indica*) and *LCBK2* (with domestication alleles from *japonica*) genes. **b**. SOR48 locates on chromosome 7 and includes *PROG1* (with domestication alleles from *japonica*) and *GLO4* (with domestication alleles from *indica*) genes. NJ trees were constructed using the SNPs extracted from three fragments spanning the gene, that is, the gene only, the gene and its 1 kb upstream and downstream regions (gene ± 1k), the gene and its 3 kb upstream and downstream 3 kb regions (gene ± 3k). Bootstrap supports over 60% are shown near the nodes.

https://doi.org/10.1038/s41477-023-01476-z

Extended Data Table 1 | Results of *D*-statistics (a) and three-population test (*f3*) (b) to quantify gene flow

a. *D*-statistics to assess gene flow between populations/groups

No.	Pop W	Pop X	Pop Y	Pop Z	*D*-statistics	*Z*-score	BABA	ABBA
1	**Ruf1**	*japonica*	***indica***	OG	0.0626***	3.666	126993	112032
2	**Ruf1**	*japonica*	**Niv2**	OG	0.126***	8.531	134415	104311
3	Niv2	***indica***	***japonica***	OG	-0.0848***	-3.95	101119	119791
4	Niv2	*indica*	Ruf1	OG	-0.0161	-1.495	108864	112394
5	**Ruf1**	*japonica*	***indica***	OG	0.0401***	2.43	131819	121643
6	**Ruf1**	*japonica*	**NIV**	OG	0.1015***	9.029	139637	113878
7	NIV	***indica***	***japonica***	OG	-0.1168***	-7.4	144808	183058
8	NIV	***indica***	**Ruf1**	OG	-0.0691***	-8.051	152865	175544

A significantly positive *D*-statistics indicated the presence of gene flow between populations W and Y while a significantly negative *D*-statistic indicated gene flow between populations X and Y. Two populations in bold face in each line represent the presence of gene flow between them. *Oryza barthii* was used as an outgroup (OG). NIV=Niv1+Niv2. A standard error for *D* was calculated using the weighted block jackknife. The number of standard errors is normally distributed, enabling a formal test for whether (W, X) forms a clade[51]. The corresponding two-tailed *p*-value was calculated to indicate significance level. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

b. Three-population test (*f3*) to assess gene flow between *japonica* and *O. nivara*

Source 1(A)	Source 2(B)	Target (C)	f_3	*Z*-score	No. SNPs
japonica	Niv2	*indica*	0.2584	13.394	3888754
japonica	NIV	*indica*	0.5124	15.338	4391381

Population C would be admixed with populations close to A and B if f_3 (C; A, B) was significantly negative and otherwise population C was not admixed if f_3 (C; A, B) was non-negative. NIV=Niv1+Niv2.

Population C would be admixed with populations close to A and B if f_3 (C; A, B) was significantly negative and otherwise population C was not admixed if f_3 (C; A, B) was non-negative. NIV=Niv1+Niv2. A significantly positive *D*-statistics indicated the presence of gene flow between populations W and Y while a significantly negative *D*-statistic indicated gene flow between populations X and Y. Two populations in bold face in each line represent the presence of gene flow between them. *Oryza barthii* was used as an outgroup (OG). NIV=Niv1+Niv2. A standard error for *D* was calculated using the weighted block jackknife. The number of standard errors is normally distributed, enabling a formal test for whether (W, X) forms a clade[51]. The corresponding two-tailed *p*-value was calculated to indicate significance level. [a], $p < 0.05$;

nature portfolio

Corresponding author(s): Song Ge

Last updated by author(s): 2023/6/30

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection.
Data analysis	Trimmomatic (version 0.36), BWA (version 0.7.10), Picard-tools (version 1.119), Samtools (version 1.1), GATK (version 3.6), MSMC-Tools, MSMC2 (version 2.0.0), SMC++ (version 1.15.3), Stairway Plot 2 (version 2), Samtools (version 1.2), phylip (version 3.696), PopLDdecay (version 3.29), ADMIXTURE (version 1.3.0), PLINK (version 1.90 beta), GCTA software (version 1.24.4), R package "ggplot2", MEGA (version 5), DensiTree (version 2.2.7), Network Publisher (version 2.1.2.5), R package "sm", R package "sf", ADMIXTOOLS (version 7.0.1), All code and scripts used in the analyses are available at the link: https://github.com/zhangfumin/domestication.

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All newly resequenced genomes have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under number PRJNA705309. Previously published genomic data used in this study were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under numbers ERP001143, ERP000729, ERP000106 and PRJEB19404, and NCBI Short Read Archive (https://www.ncbi.nlm.nih.gov/sra/) under number SRA023116 and SRP22609, and bioproject IDs PRJEB6180, PRJNA422249, PRJNA557122, PRJNA422249 and PRJNA557122, and PRJNA79428. The genomic SNP data of all samples are available on Dryad (https://datadryad.org/stash/share/jcQfZcbai80MmLb6kO4_mrQLfu1tX-I_1_Yx7hAfJkI).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	The dataset included 457 wild accessions that covered the entire range of two rice progenitors and 1121 cultivated rice from 25 countries. The wild accessions were chosen carefully based on 15 diagnostic characters to maximize the representativeness of wild rice gene pool. The rice accessions included only landraces by excluding advanced cultivars to avoid the interference of introgression or hybridization on inference of domestication history. The 14 newly sequenced accessions were from the rayada group that has been ignored in most previous studies. Based on population genetic analyses, we identified 147 and 53 misclassified/admixed accessions for cultivated and wild rice, respectively. Therefore, the final panel of samples used in phylogenetic and domestication analyses consisted of 1089 pure landraces and 404 true wild rice (Supplementary Table 5).
Data exclusions	We excluded 53 wild and 32 rice accessions because they were present in intermediate position between groups in the NJ tree and PCA .
Replication	The phenotyping was undertaken in three experimental fields of CNRRI in Hangzhou (N30°16'48", E120°9'0"), the Guangxi Academy of Agricultural Sciences (GAAS) in Nanning (N22°50'43", E108°14'51") and the CAS Field Station in Sanya (N18°30'45", E110°2'38"). We confirmed that all code of this study could be replicated.
Randomization	The 404 wild accessions were clustered into 4 groups and the 1089 rice landrace clustered into were two major groups corresponding to two subspecies and six cultivar groups based on the results of the Neighbor-joining (NJ) tree, principal component analyses (PCA) and ADMIXTURE.
Blinding	Blinding was not applicable to this study, as this study foucses on the origin of the two rice groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging