Article

A pangenome reference of wild and cultivated rice

https://doi.org/10.1038/s41586-025-08883-6

Received: 12 March 2024

Accepted: 11 March 2025

Published online: 16 April 2025

Open access

Check for updates

Dongling Guo[1,2,6], Yan Li[1,6], Hengyun Lu[1], Yan Zhao[1], Nori Kurata[3], Xinghua Wei[4], Ahong Wang[1], Yongchun Wang[1], Qilin Zhan[1], Danlin Fan[1], Congcong Zhou[1], Yiqi Lu[1], Qilin Tian[1], Qijun Weng[1], Qi Feng[1], Tao Huang[1], Lei Zhang[1], Zhoulin Gu[1], Changsheng Wang[1], Ziqun Wang[1], Zixuan Wang[1], Xuehui Huang[5], Qiang Zhao[1]✉ & Bin Han[1]✉

Oryza rufipogon, the wild progenitor of Asian cultivated rice *Oryza sativa*, is an important resource for rice breeding[1]. Here we present a wild–cultivated rice pangenome based on 145 chromosome-level assemblies, comprising 129 genetically diverse *O. rufipogon* accessions and 16 diverse varieties of *O. sativa*. This pangenome contains 3.87 Gb of sequences that are absent from the *O. sativa* ssp. *japonica* cv. Nipponbare reference genome. We captured alternate assemblies that include heterozygous information missing in the primary assemblies, and identified a total of 69,531 pan-genes, with 28,907 core genes and 13,728 wild-rice-specific genes. We observed a higher abundance and a significantly greater diversity of resistance-gene analogues in wild rice than in cultivars. Our analysis indicates that two cultivated subpopulations, intro-*indica* and *basmati*, were generated through gene flows among cultivars in South Asia. We also provide strong evidence to support the theory that the initial domestication of all Asian cultivated rice occurred only once. Furthermore, we captured 855,122 differentiated single-nucleotide polymorphisms and 13,853 differentiated presence–absence variations between *indica* and *japonica*, which could be traced to the divergence of their respective ancestors and the existence of a larger genetic bottleneck in *japonica*. This study provides reference resources for enhancing rice breeding, and enriches our understanding of the origins and domestication process of rice.

Asian cultivated rice (*Oryza sativa* L.) was domesticated from its wild progenitor *O. rufipogon*, and is one of the most important food crops in the world[2]. Increasing rice yield is crucial to address global challenges and meet the food demands driven by rapid population growth and environmental changes[3]. The geographically and genetically diverse *O. rufipogon* species is a key genetic reservoir that contains valuable stress-resistance and weed-competitiveness traits for the genetic improvement of modern rice[1].

Since the draft genomes of *O. sativa* ssp. *japonica* and *O. sativa* ssp. *indica* cultivars and the first completed rice reference *japonica* Nipponbare genome were released[4–6], a large number of comprehensive genome studies on diverse varieties of cultivated rice have been performed[7–10]. A single reference genome cannot fully represent the genetic diversity of a species and might hinder functional genomics research[11]. Pangenome studies offer comprehensive insights into genetic diversity, species evolution and cultivar improvement[12]. Rice pangenome studies have been performed using large-scale next-generation sequencing[13]. Advances in sequencing, and especially in long-read technology, have improved rice pangenome research by enabling more accurate identification of structural variations (SVs) and a comprehensive depiction of genetic diversity[14–18].

Population genetics studies have revealed that ancient *japonica* rice was first domesticated from the *O. rufipogon* group IIIa (Or-IIIa) population in China, and that *indica* rice was subsequently domesticated when ancient *japonica* spread southward and westward in Asia and crossed with the local *O. rufipogon* group I (Or-I) population[19,20]. Building on this, the model of 'multiple origins but single domestication' for different rice subspecies was proposed[21]. The construction of the first syntelogue-based rice pangenome further supported the single domestication hypothesis[22]. A multiple-origins model of rice was also proposed, which suggested separate and independent domestications for *japonica*, *indica* and *O. sativa* ssp. *aus*[23,24]. However, current rice pangenome research focuses mainly on cultivated populations, and investigations of diverse wild resources, especially *O. rufipogon*, are lacking. Therefore, constructing a high-quality *O. rufipogon* pangenome is essential for guiding future rice breeding strategies and understanding the evolutionary and domestication pathways.

In this study, we construct a comprehensive *O. rufipogon*–*O. sativa* pangenome reference of 145 chromosome-level genomes, mainly using PacBio high-fidelity (HiFi) sequencing. We accurately identify a wide range of sequence variations and fully annotate gene models and transposable elements (TEs) across the genomes of both cultivated

[1]National Center for Gene Research, State Key Laboratory of Plant Trait Design, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Plant Genetics Laboratory and Comparative Genomics Laboratory, National Institute of Genetics, Mishima, Japan. [4]State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou, China. [5]College of Life Sciences, Shanghai Normal University, Shanghai, China. [6]These authors contributed equally: Dongling Guo, Yan Li. ✉e-mail: zqiang@ncgr.ac.cn; bhan@ncgr.ac.cn

and wild rice. We examine sequence diversity by capturing alternate assemblies that include heterozygous information not present in the primary assemblies of 133 HiFi genomes. We perform a deep analysis of complex gene flows both within cultivars and between cultivars and wild rice, delineating the evolutionary and domestication pathways of various rice types. Our study provides strong evidence supporting the hypothesis that the initial domestication regions of all Asian cultivated rice had a monophyletic origin from Or-IIIa, the ancestor of *japonica*. Our findings shed light on the genomic changes that occurred in cultivated rice during its origin and domestication processes, and provide valuable resources for improving rice varieties in terms of yield, quality, environmental adaptability and resistance to disease and stress.

De novo assembly of 149 rice genomes

To capture the inherent genetic diversity across cultivated and wild rice, we selected a representative set of 149 samples from a previous study[19] (Supplementary Table 1). This set comprised 145 genetically and geographically diverse varieties, including 16 *O. sativa* cultivars and 129 *O. rufipogon* accessions from a broad geographical range spanning approximately 20 countries (Fig. 1a and Supplementary Table 2). By comparing highly conserved chloroplast genomes, we re-identified three accessions as *Oryza longistaminata* (OL2296, OL3101 and OL3102) and one as *Oryza meridionalis* (OM1952), which were considered as outgroups for subsequent evolutionary studies (Supplementary Fig. 1).

Of the selected samples, 133 accessions were deeply sequenced using PacBio HiFi technology, and the remaining 16 accessions were sequenced using Oxford Nanopore Technologies (ONT). The average sequencing depths reached approximately 103.0-fold for ONT and 24.2-fold for PacBio HiFi (Supplementary Table 3). The raw data obtained from both sequencing methods were carefully assembled using various strategies to obtain high-quality genomes (Methods), as evidenced by an average contig N50 of 14.95 Mb and an average LTR assembly index (LAI)[25] of 24.13. The robustness of these assemblies was further validated by BUSCO (benchmarking universal single-copy orthologues) assessments, yielding an average score of 98.55% (Supplementary Table 1). The average quality value (QV)[26] assessments for the assemblies were 24.55 for ONT and 57.92 for PacBio HiFi, highlighting the high accuracy of HiFi sequencing (Supplementary Table 3). It should be noted that the Nipponbare genome, assembled with HiFi reads, was stringently quality-checked against both the previously established IRGSP-1.0 reference genome and the telomere-to-telomere (T2T) assembly[27]. Our assembly had 141,775 bp mismatches and 54,873 insertions and deletions (indels) (245,980 bp) compared with IRGSP-1.0, and only 32,731 bp mismatches and 42,154 indels (96,331 bp) compared with the T2T-NIP assembly. We also compared some assembly quality indicators of 534M (an *indica* cultivar, also called WSSM, R534) with that reported in a previous study[16] (Supplementary Table 4). These findings affirmed the precision and reliability of our sequencing and assembly approaches.

Moreover, we achieved chromosome-level assembly for 30 representative species using high-throughput chromatin conformation capture (Hi-C) sequencing data. On average, 97.25% of the contigs were accurately ordered, oriented and curated into 12 chromosome-scale scaffolds (Supplementary Fig. 2a and Supplementary Table 5). These assemblies showed significant collinearity with the reference genome (Supplementary Fig. 2b). When comparing the anchored pseudo-chromosomes with Hi-C genomes, we found that only 0.07% of regions were non-syntenic, indicating high consistency between the assemblies achieved by the two methods. This comparison confirmed the effectiveness of the collinearity-based approach for attaining chromosome-level assembly across all accessions in our study (Supplementary Fig. 3). In conclusion, we achieved reference-level genome sequences, providing a valuable resource for pangenome construction and evolutionary analysis.

Using the 7-base telomeric repeat as a sequence query in our analysis revealed an average of 19 telomeres per genome. Notably, most genomes lacked the telomere on the short arm of chromosome 9 (Supplementary Fig. 4a), indicating an unresolved gap region, owing potentially to a large rDNA array of nearly identical 45S rDNA repeats[27]. In addition, we used sequence homology to CentO satellite repeats for pinpointing CentO enriched regions in each chromosome. We discovered that the homology of CentO sequences in each species was concentrated predominantly at 155 bp and 165 bp (Supplementary Fig. 4b). Moreover, within each chromosome, there was a considerable similarity in CentO sequences (Supplementary Fig. 4c), a pattern consistently observed across both cultivated and wild rice species[28].

In our quest for a more exhaustive representation of sequence diversity, alternate assemblies (a-contigs) were also captured from 133 HiFi genomes (Fig. 1b and Supplementary Fig. 5). These a-contigs contained heterozygous information that was absent from the primary assemblies (p-contigs). Notably, the size of a-contigs in cultivated rice was significantly smaller than that in wild rice (Supplementary Fig. 6a,b), reflecting a lower heterozygosity rate in the former (Supplementary Table 6). This observation is consistent with the fact that *O. sativa* is a predominantly self-pollinating species. Furthermore, the size of these a-contigs seemed to be influenced by the degree of uniform distribution of heterozygous loci, as evidenced in Supplementary Fig. 6c,d.

Gene annotation and characterization of RGAs

Using an integrated approach of homology-based, transcriptomic and ab initio predictive methods, we annotated genes across each genome in this study (Methods). Each genome contained about 40,273 genes, with an average length of 2,566 bp and a density of around 99.15 genes per Mb (Supplementary Table 1). Compared with wild rice, cultivated rice genomes showed notable reductions in both genome size and gene number, with a higher gene density implying reduced TE content (Fig. 1c–e). On average, 92% of the genes of each variety contained functional domains, and 72% of them were expressed in at least one tissue (including roots, leaves, seedlings and panicles). In HiFi genome assemblies, gene annotation was also performed on the a-contigs to identify a range of 199 to 41,230 genes, amounting to a total of 10,521 genes missing in the p-contigs (MIP genes) (Fig. 1b,f, Supplementary Fig. 5 and Supplementary Table 6). Notably, an average of 57.26% of the genes on the a-contigs were heterozygous alleles, with differentially expressed alleles accounting for 19.61%, 18.75%, 31.54% and 20.71% in roots, leaves, seedlings and panicles, respectively (Supplementary Table 7).

Owing to long-term natural selection, wild rice is highly resistant to biotic and abiotic stress, and is thus a natural repository of resistance alleles. Notable examples from *O. rufipogon* include the bacterial blight resistance gene *Xa23* (ref. 29), the brown planthopper resistance gene *bph19(t)* (ref. 30) and the white-backed planthopper resistance gene *Bph38* (ref. 31). To assess the content of resistance-gene analogues (RGAs) in wild and cultivated rice, we predicted receptor-like protein kinases (RLKs), nucleotide-binding-site-encoding proteins (NBSs), transmembrane coiled-coil (TM-CC) proteins and receptor-like proteins (RLPs) on the basis of their conserved structural characteristics (Supplementary Table 8). We observed a slightly higher abundance but a significantly greater diversity of RGAs in *O. rufipogon*, averaging 1,710, compared with approximately 1,652 RGAs in *O. sativa* (Extended Data Fig. 1a,b). Acknowledging the propensity of RGAs to form clusters in the genome[17], we classified them as singletons, pairs or clusters, depending on their chromosomal positions across different species. We noted that singletons constituted the highest proportion in each type of these RGAs (Extended Data Fig. 1c).

To further exploit the resources of RGAs, we identified a total of 1,184 (51.12%) collinear loci in wild rice that have a higher average copy number than those in cultivated rice, with 638 (27.55%) loci being specific to wild rice (Fig. 1g, Extended Data Fig. 1d and Supplementary Fig. 7). Among these loci, we found an RLK gene from *O. rufipogon* (Extended Data Fig. 1e), *LOC_Os07g35680*, which was recently reported[32] to be a

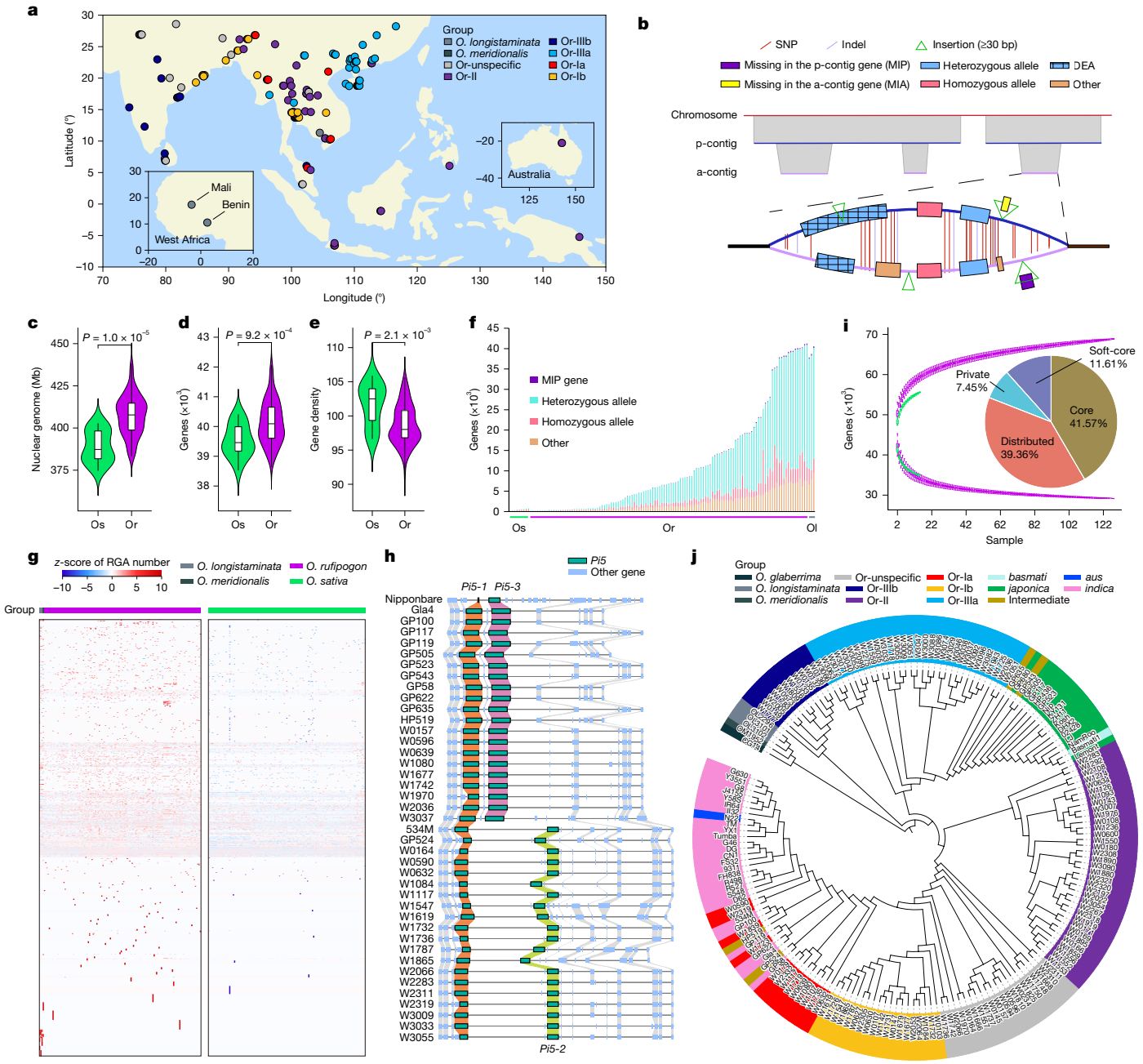


Fig. 1 | Pangenome analysis of 149 representative wild and cultivated rice accessions. a, Geographical distribution of 133 wild rice accessions, with dots coloured by group. Inset maps show enlarged views of West Africa and Australia. Positive values indicate east longitude (°E) and north latitude (°N). **b**, Schematic representation of p-contigs and a-contigs, showing gene models and variant information from HiFi genome assemblies. **c**–**e**, Violin plots comparing *O. sativa* (Os; $n = 16$) and *O. rufipogon* (Or; $n = 129$) varieties for nuclear genome size (**c**), gene number (**d**) and gene density (gene number per Mb) (**e**). The box edges mark the 25% and 75% quartiles and the central line marks the median. Whiskers extend to 1.5 times the interquartile range (IQR). Statistical significance was determined by two-sided Wilcoxon test; *P* values are shown at the top. **f**, Distribution of genes on a-contigs across 133 HiFi genomes, including *O. longistaminata* (Ol). **g**, Heat map showing RGA loci that have a higher average copy number in wild rice than in cultivated rice, normalized using a row-wise *z*-score method. Each row indicates an RGA locus, and each column indicates a variety. **h**, Local synteny plot of *Pi5* loci (chr. 9: 9.65–9.79 Mb in the Nipponbare reference genome) from 41 accessions. *Pi5* genes (*Pi5-1*, *Pi5-2* and *Pi5-3*) are shown in green with collinearity represented by lines of different colours; adjacent genes are shown in blue with grey collinearity lines. **i**, Accumulation curve of the number of pan-genes (uptrend lines) and core genes (downtrend lines) across *O. sativa* (green; $n = 16$) and *O. rufipogon* (purple; $n = 129$) varieties. The pie chart shows the gene proportions in the combined wild–cultivated pangenome ($n = 145$). **j**, Phylogenetic tree of the 180 *Oryza* varieties, on the basis of single-copy genes. Diverse groups are denoted by colour-coded labels.

negative regulatory factor of rice blast disease mediated by *OsMADS26*. There were many nonsynonymous variations between the resistant haplotype (Y476) and the non-resistant (Nipponbare) haplotype (Extended Data Fig. 1f). In our wild rice population, we discovered some accessions that had the same haplotype as Y476 (Extended Data Fig. 1g). Furthermore, we performed detailed analyses of another known blast-resistance locus, *Pi5*, which comprises two NBS–leucine-rich repeat (LRR) genes, *Pi5-1* and *Pi5-2* (ref. 33). Of note, the Nipponbare cultivar also contains two *Pi5* genes (*Pi5-1* and *Pi5-3*); however, *Pi5-3* lacks disease-resistance functionality. Through detailed homology comparisons and synteny analysis, we identified several accessions, particularly in *O. rufipogon*, that have this crucial resistance locus

(Fig. 1h and Supplementary Fig. 8). These results show that wild rice varieties are rich in disease-resistance resources and can be used as donors for developing highly resistant rice varieties.

Gene-based and graph-based pangenomes

Integrating genes from both a-contigs and p-contigs, we constructed pangenomes for 16 *O. sativa* accessions, 129 *O. rufipogon* accessions and a combined set of 145 taxa (Fig. 1i). The accumulation curves of pan-genes showed a pattern in which gene counts exhibited a notably steeper increase with increasing accession numbers in the wild rice pangenomes than in those of cultivated rice. This trend plateaued after the inclusion of 125 accessions. Owing to the near-saturated *O. rufipogon* accessions, the pan-gene number of 145 taxa revealed a total of 69,531 genes—more than several previously published rice pangenomes[15–17] (Supplementary Table 9). In the wild–cultivated rice pangenome, 41.57% (28,907) were core genes (present in all samples) and 7.45% (5,181) were private genes (present in only one sample), of which about 28.10% were expressed in at least one tissue. Notably, 13,728 genes (19.74%) were found exclusively in *O. rufipogon*. When compared with the same number of cultivated rice genomes ($n = 129$), 7,592 genes remained specific (Supplementary Table 10). These genes were related to defence responses and ADP-binding processes, consistent with our analysis of RGAs (Supplementary Fig. 9).

By incorporating gene data from a previously published pangenome of 33 cultivated rice accessions[16], we identified a set of 844 single-copy orthologues and constructed a phylogenetic tree (Fig. 1j). Building on our previous study[19], subsequent analysis further divided *O. rufipogon* into six distinct clades: Or-Ia, Or-Ib, Or-II, Or-IIIa, Or-IIIb and Or-unspecific. Clade Or-Ia was closely related to *indica*, and Or-IIIa to *japonica*, whereas Or-IIIb and Or-II showed more remote phylogenetic links to the domesticated cultivars. The group information for several accessions was redefined according to their positions in the phylogenetic tree (Supplementary Table 1). We further deduced the demographic history and estimated the timing of these events in *O. rufipogon* (Supplementary Fig. 10). These observations align well with the above phylogenetic relationships.

To further obtain scalable and comprehensive genetic diversity, we also constructed three base-level graph pangenomes for 15 *O. sativa* accessions, 129 *O. rufipogon* accessions and a combined set of 144 cultivated and wild rice species (Methods). The most expansive graph pangenome showed a total non-reference node length of 3.87 Gb (Supplementary Table 11). The isolated pangenomes of cultivated and wild rice revealed that each cultivated rice variety contributed an average of more than 17.47 Mb, whereas each *O. rufipogon* accession contributed up to 29.72 Mb to the non-reference node length. Research in both human[34] and plant[35] genetics has shown that graph pangenomes are more effective than linear reference genomes in calling all types of genetic variant.

Extensive variation and TE analyses

Despite advances in characterizing genetic variants among *O. sativa* and other wild species[9,13,15–19,22,24], large-scale and accurate analysis of the closely related *O. rufipogon* is still limited. Through alignment of sequencing reads and assembled contigs, we performed high-confidence calling of single-nucleotide polymorphisms (SNPs), small indels of fewer than 30 bp and SVs of more than 30 bp (including insertions, deletions, inversions and translocations) across each genome (Supplementary Table 12). Our cohort collectively encompassed 11.08 million SNPs, 10.97 million small indels, 362,194 deletions, 531,356 insertions, 4,495 inversions and 357,655 translocations, showing a marked enrichment in intergenic repetitive regions (Supplementary Fig. 11). Most deletions spanned 30–100 bp, and insertions ranged from 30 to 1,000 bp. Notably, translocations and inversions tended to be longer, especially inversions, with approximately 17% exceeding 50 kb in length (Fig. 2a). A previously reported[16,36] 4.5-Mb inversion on chromosome 6 was also observed in our study within 117 *O. rufipogon* accessions (92%) and 10 *O. sativa* cultivars (62%) (Supplementary Table 13). This was further verified by mapping the Hi-C paired reads to the Nipponbare genome and their corresponding genome assemblies (Supplementary Fig. 12a). It is worth mentioning that this inversion was also found in outgroups, which suggests that the reference allele possibly first appeared as the derived state in *O. rufipogon* (Supplementary Fig. 12b). Moreover, to capture a broader spectrum of wild rice diversity, we expanded our population by incorporating newly sequenced and non-redundant samples from a previous study[24] (Supplementary Table 14). We genotyped genetic variants by mapping short reads onto the cultivated and wild rice graph pangenome we constructed, yielding a dataset comprising 48,488,470 SNPs, 6,752,123 indels and 154,980 SVs (Methods).

We annotated TEs from all of the selected varieties in our study and from 28 non-redundant Asian cultivated rice assemblies from a previously published pangenome analysis of 33 rice accessions[16] in detail, and found that long terminal repeat retrotransposons (LTR-RTs), especially the Gypsy superfamily, were the most prevalent (Extended Data Fig. 2a and Supplementary Table 15), consistent with previous results[16]. Considering all intact LTR-RTs across different groups of *O. rufipogon*, the Copia superfamily exhibited a more modest but earlier expansion approximately 100,000 years ago, whereas the Gypsy superfamily underwent a significant expansion around 25,000 years ago, which was particularly notable in the Or-IIIa group (Extended Data Fig. 2b). Our comparative analysis validated the positive correlation between TE content and genome size (Extended Data Fig. 2c). *Oryza rufipogon* had an average TE content of 53.23%, which is slightly higher than that in *O. sativa* (52.32%), and could explain the observed lower gene density in *O. rufipogon* (Fig. 1e). The Or-IIIa group was distinguished by its significantly higher TE content, in contrast to *japonica*, which had the lowest TE content (Extended Data Fig. 2d). A considerable portion of the difference in genomic size between Or-IIIa and *japonica*, estimated at 78.90%, was attributed to TEs—predominantly the Gypsy superfamily (57.78%), most of which emerged within the past 200,000 years (Fig. 2b,c).

To construct a comprehensive pangenome TE library (Methods), we found that 97% of Gypsy families differed by less than 250 kb in length between Or-IIIa and *japonica* (Fig. 2d). However, 17 key Gypsy families accounted for a notable 25.85-Mb disparity, ranking among the most abundant in Or-IIIa (Extended Data Fig. 2e). In addition, we calculated the solo-to-intact ratio for these families, a metric indicative of TE removal through illegitimate recombination[37,38]. This ratio sheds light on the mechanisms behind the observed genomic disparities. Our analysis revealed that in 9 out of 17 families, about 9 Mb of the disparity in Gypsy elements in *japonica* was due to element removal, and the remaining difference of about 17 Mb was attributed to significant amplification of Gypsy elements in Or-IIIa (Fig. 2e). Overall, over the past 200,000 years, certain Gypsy families within Or-IIIa have undergone substantial amplification, yielding a multitude of young, intact elements that have significantly contributed to differences in genomic size. Moreover, *japonica* was subjected to severe genetic bottlenecks and a reduction in Gypsy elements owing to illegitimate recombination, further accentuating the genomic size difference between the two. We identified about 1,000 genes adjacent to these expanding Gypsy families in Or-IIIa; these were enriched in metabolic pathways—notably those governing carbohydrate and terpenoid metabolism—as well as cell-surface signal transduction processes (Extended Data Fig. 2f). This suggests that they have a positive role in environmental adaptation and disease-resistance processes[37].

Evolution of Asian cultivated rice

The origin of *O. sativa* has been actively debated in the scientific community[19,21–24]. To further address this complexity, our dataset has been

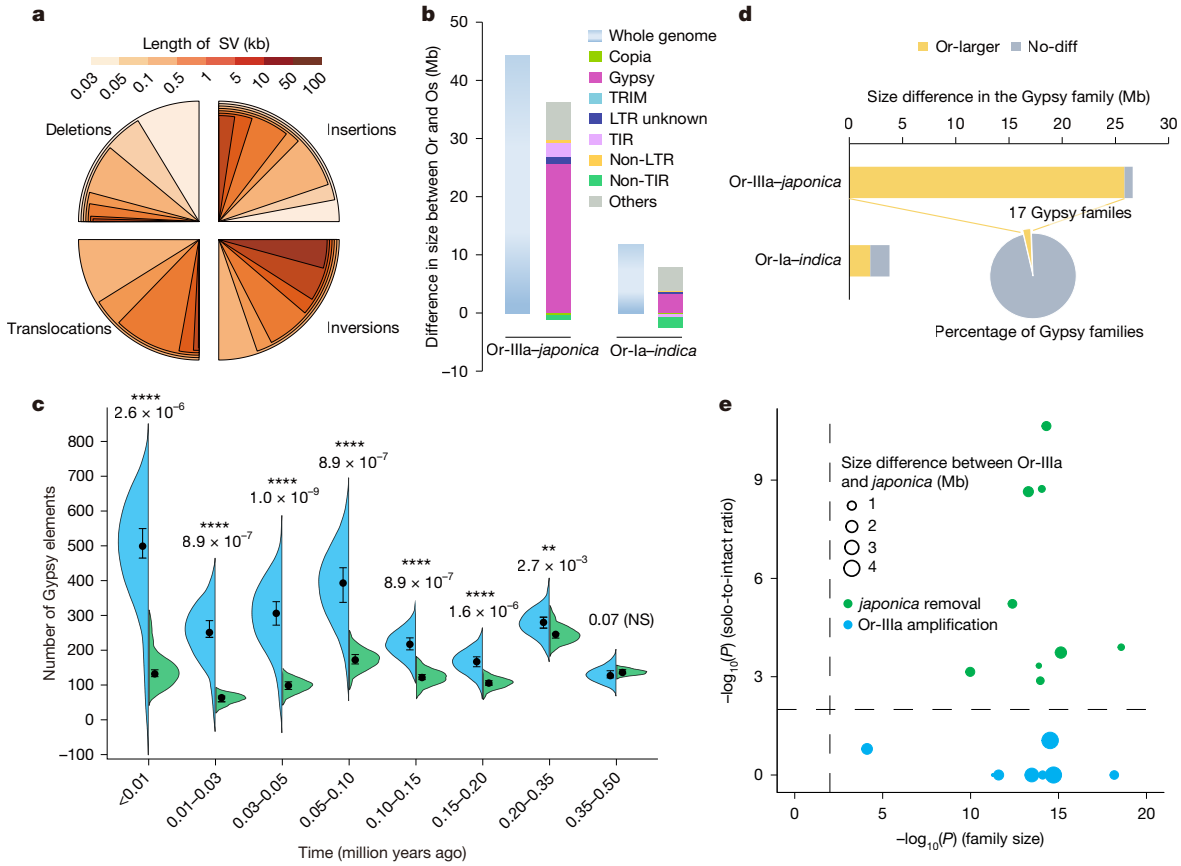


Fig. 2 | Genomic variation and TE composition in wild and cultivated rice. **a**, Fan charts showing frequency distribution by length for various SVs. **b**, Bar chart illustrating the difference in genome size and TE family size between *O. rufipogon* (Or-IIIa or Or-Ia) and *O. sativa* (*japonica* or *indica*). Positive values indicate larger sizes in *O. rufipogon* and negative values indicate larger sizes in *O. sativa*. TIR, terminal inverted repeats; TRIM, terminal-repeat retrotransposon in miniature. **c**, Expansion dynamics of intact Gypsy elements in Or-IIIa (blue; $n = 31$) and *japonica* (green; $n = 19$) over time. The bars represent the 75% and 25% quartiles; and the central circles indicate the median. *P* values (two-sided Wilcoxon test) are shown at the top (NS, not significant; $^{**}P < 0.01$; $^{****}P < 0.0001$). **d**, Size differences in the Gypsy family between *O. rufipogon* (Or-IIIa or Or-Ia) and *O. sativa* (*japonica* or *indica*). 'Or-larger' families categorize those in which Or-IIIa or Or-Ia has a size difference greater than 250 kb relative to *japonica* or *indica*; 'no-diff' encompasses families with size differences between −250 kb and 250 kb. **e**, Scatter plot showing the classification of Gypsy families that are larger in Or-IIIa genomes than in *japonica* genomes. Each dot denotes a distinct Gypsy family, with its size proportional to the family size difference between Or-IIIa and *japonica*. Both the *x* and *y* axes are scaled by $-\log_{10}$. Horizontal and vertical black dashed lines indicate the threshold for the two-sided solo-to-intact ratio Student's *t*-test and the family size Student's *t*-test, respectively ($P = 0.01$).

augmented with published high-quality genomic data[15,16,28,39], resulting in a comprehensive cohort of 280 individual varieties (Supplementary Table 16). Using whole-genome SNP data, we constructed a neighbour-joining phylogenetic tree and performed archetypal analysis (Methods), for advanced evolutionary research (Fig. 3a,b). Population structure analysis indicated that the origins of *japonica* and *basmati* can be traced to Or-IIIa; those of *indica* to Or-Ia; and those of *aus* to Or-Ib, mainly from South Asia. These evolutionary relationships were further confirmed through comprehensive evolutionary analyses of our two expanded populations (Supplementary Figs. 13 and 14 and Supplementary Tables 14 and 17). This enabled us to examine the relationships among *O. rufipogon* classified by different methods[24] and to address the ongoing debate about the origins of cultivated rice (Supplementary Note 1).

basmati and *aus*, which are genetically distinct from *japonica* and *indica*, are two subpopulations concentrated mainly in the Indian subcontinent[40]. *basmati* is known for its exceptionally fine grain and pleasant aroma, whereas *aus* is highly tolerant to environmental stresses such as drought and heat[41,42]. Wild rice exhibited a higher nucleotide diversity (often referred to using the symbol π) and a faster decay of linkage disequilibrium (LD) than did cultivated rice (Fig. 3c and Supplementary Fig. 15a). Analyses of both fixation index (F_{ST}) and individual-level pairwise genomic distance (DST) revealed that the genomic distance between Or-Ia and *indica* is notably lower than that between other domesticated rice groups and their wild progenitors (Fig. 3c and Supplementary Fig. 15b). We found that Or-Ia had about 40 Mb of introgression fragments from *indica* (3,951 10-kb windows with similarity greater than 99.99%), which also exceeded that in the comparison of other cultivated rice with its ancestral parents (Supplementary Fig. 16a), suggesting that there might be subsequent gene flow from *indica* into Or-Ia (Supplementary Note 2). Gene flow from *indica* was evident in other wild rice clades; Or-unspecific is likely to have emerged from a hybridization event between Or-II and *indica* (Fig. 3b, Supplementary Fig. 16b and Supplementary Table 18). Subsequently, we used the F_{ST} index to identify and exclude introgression regions from *indica* in Or-I (Or-Ia and Or-Ib). Our analysis revealed significant divergence within the top 20% of F_{ST} regions between Or-I and *indica*, in which Or-Ia and Or-Ib persisted as separate branches (Supplementary Fig. 16c), indicating pre-domestication divergence and independent origins of *indica* and *aus*.

The phylogenetic tree revealed a specific *indica* lineage from South Asia clusters with Or-Ib and *aus* (Fig. 3a). Admixture analysis showed that, from $K = 4$ to $K = 7$, this lineage diverged from the rest of *indica* owing to shared ancestral components with *aus* (Fig. 3b), suggesting

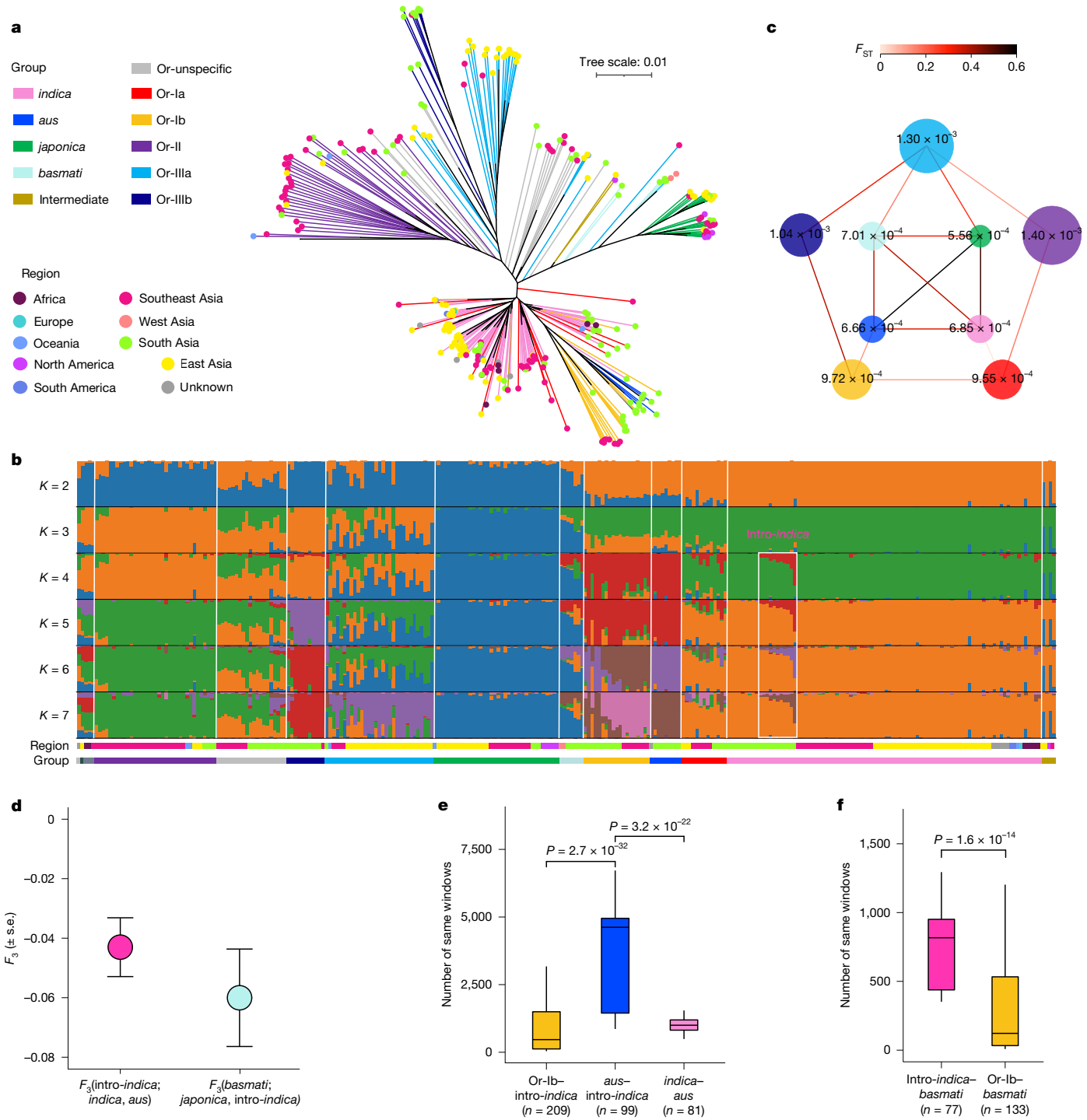


Fig. 3 | Population structure and introgression of wild and cultivated rice. **a**, Phylogenetic tree of 280 *Oryza* accessions, based on whole-genome SNPs. The scale bar represents 0.01 nucleotide substitutions per site. The tree branches are colour-coded by group, and the external circles denote the geographical distribution of each accession. **b**, ADMIXTURE clustering of 280 *Oryza* accessions from $K = 2$ to $K = 7$. The white box indicates the intro-*indica* group from $K = 4$ to $K = 7$. The colours corresponding to population groups and regions are defined in **a**. **c**, Mean nucleotide diversity (π) in each group (nodes) and mean population differentiation (F_{ST}) between groups (edges). The size of the nodes and values within the nodes represent the nucleotide diversity, and the colour of the edge represents F_{ST}. The colour of the node represents the group, corresponding to the colours in **a**. **d**, Intro-*indica* and *basmati* originated from admixture between cultivated rice in South Asia, as revealed by F_3-admixture tests. Error bars represent the standard error (s.e.). **e**,**f**, Differences in the number of same windows (10-kb windows with ≥99.99% similarity) when intro-*indica* (**e**) and *basmati* (**f**) are compared with other groups. The box edges mark the 25% and 75% quartiles and the central line marks the median. Whiskers extend to 1.5 × IQR. Statistical significance was determined by two-sided Wilcoxon test; *P* values are shown at the top.

crossbreeding among cultivated rice varieties in South Asia. By combining introgression and pairwise differentiation analyses, this *indica* lineage, dubbed 'intro-*indica*', was confirmed to be a hybrid of *indica* and *aus* (Fig. 3d,e, Supplementary Fig. 17a,b and Supplementary Table 18). Moreover, *basmati* also exhibited admixture between intro-*indica* and *japonica* (Fig. 3d,f and Supplementary Table 18).

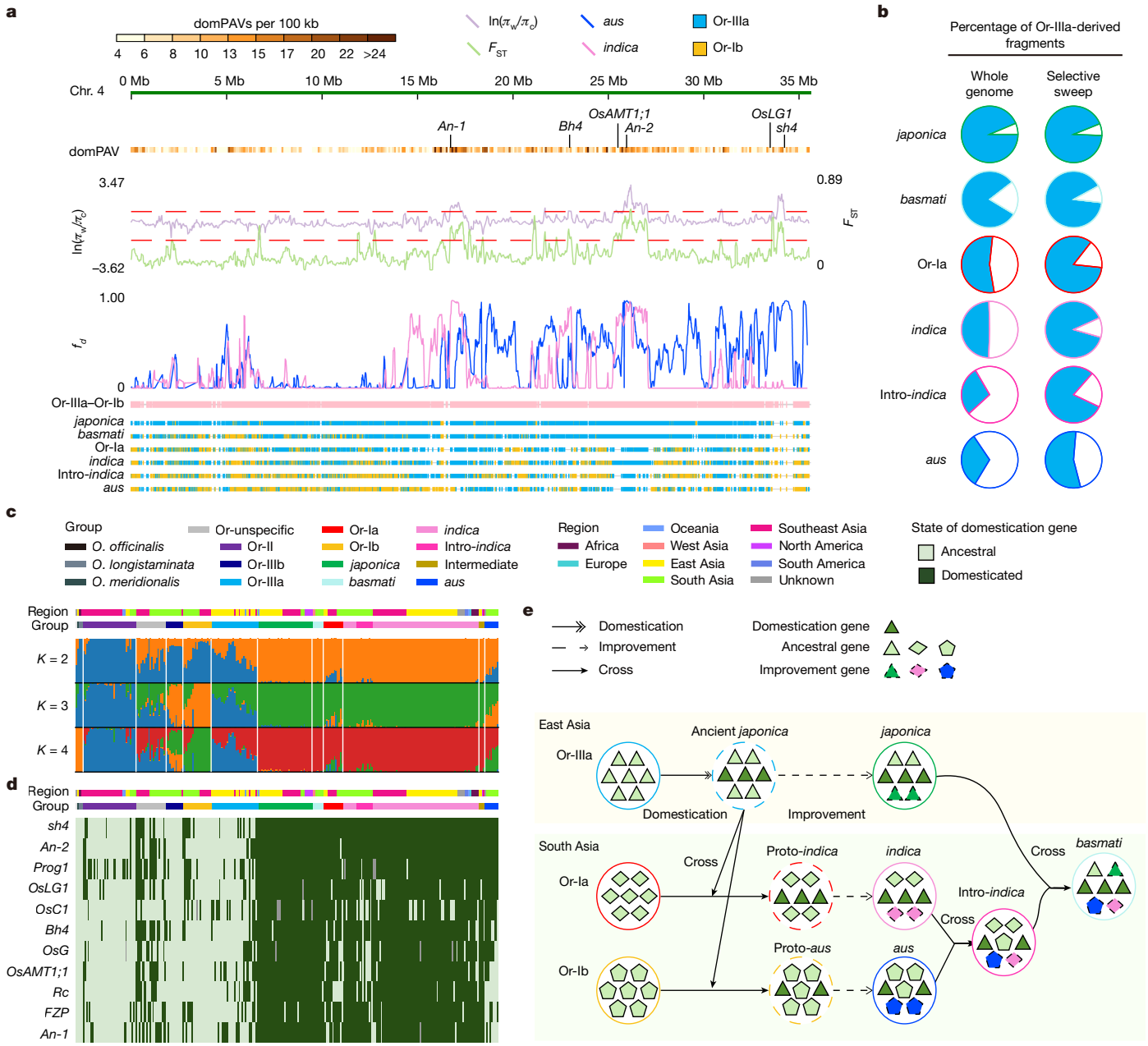


Fig. 4 | Selective sweep and evolutionary routes of Asian cultivated rice. **a**, Nucleotide diversity ratio (π_w/π_c) and fixation index (F_{ST}) between wild rice (Or-IIIa and Or-Ib) and cultivated rice (*japonica*, *basmati*, *indica*, intro-*indica* and *aus*) along chromosome 4. Horizontal red dashed lines indicate genome-wide thresholds of selection signals ($\ln(\pi_w/\pi_c) = 1.09$ and $F_{ST} = 0.3$). In the line chart of f_d value distribution, the populations Or-Ib, *aus*, *japonica* and *O. longistaminata* are represented by the blue line, and the populations Or-Ia, *indica*, *japonica* and *O. longistaminata* by the pink line. Highly differentiated SNP sites ($n = 566{,}513$) between Or-IIIa ($n = 31$) and Or-Ib ($n = 19$) are indicated in pink. Within six populations (*japonica*, *basmati*, Or-Ia, *indica*, intro-*indica* and *aus*), the major alleles matching Or-IIIa-specific alleles are blue, and those matching Or-Ib-specific alleles are orange. **b**, Percentage of Or-IIIa-derived fragments in the whole genome and in selective sweep regions across major groups. **c**, ADMIXTURE clustering of 280 *Oryza* accessions, based on the SNPs within the selective sweep from $K = 2$ to $K = 4$. **d**, Domestication state of 11 known domestication genes in each accession. Each row indicates a domestication gene, and each column indicates an accession. Colours correspond to population groups and regions as defined in **c**. **e**, Overview of the evolutionary history of Asian cultivated rice. Schematic outlining the key milestones in the domestication and improvement of various groups of Asian cultivated rice, highlighting notable crossbreeding events. Dark green triangles represent domestication genes. Ancestral genes are depicted as light green triangles (from Or-IIIa), diamonds (from Or-Ia) and pentagons (from Or-Ib). Improvement genes are shown with dashed-line emerald green triangles (in *japonica*), pink diamonds (in *indica*), and blue pentagons (in *aus*).

Domestication of Asian cultivated rice

In the evolutionary narrative of Asian cultivated rice, another widely debated topic is whether different groups of cultivated rice shared a common initial domestication event or underwent independent domestications[19–24]. Or-Ia is known to have experienced extensive gene flow from *indica*, which could confound the identification of domestication regions. Therefore, to accurately identify selective sweeps, we excluded Or-Ia and retained only the progenitors (Or-IIIa and Or-Ib) of other cultivated rice groups as the ancestral groups (Supplementary Figs. 18 and 19 and Supplementary Note 3). By genome-wide comparison of genetic diversity ratios (π_w/π_c) and F_{ST} values between ancestral groups and cultivated groups, we identified 50 loci with strong selection signals, covering a total region of 12.35 Mb (Methods). In addition, beyond

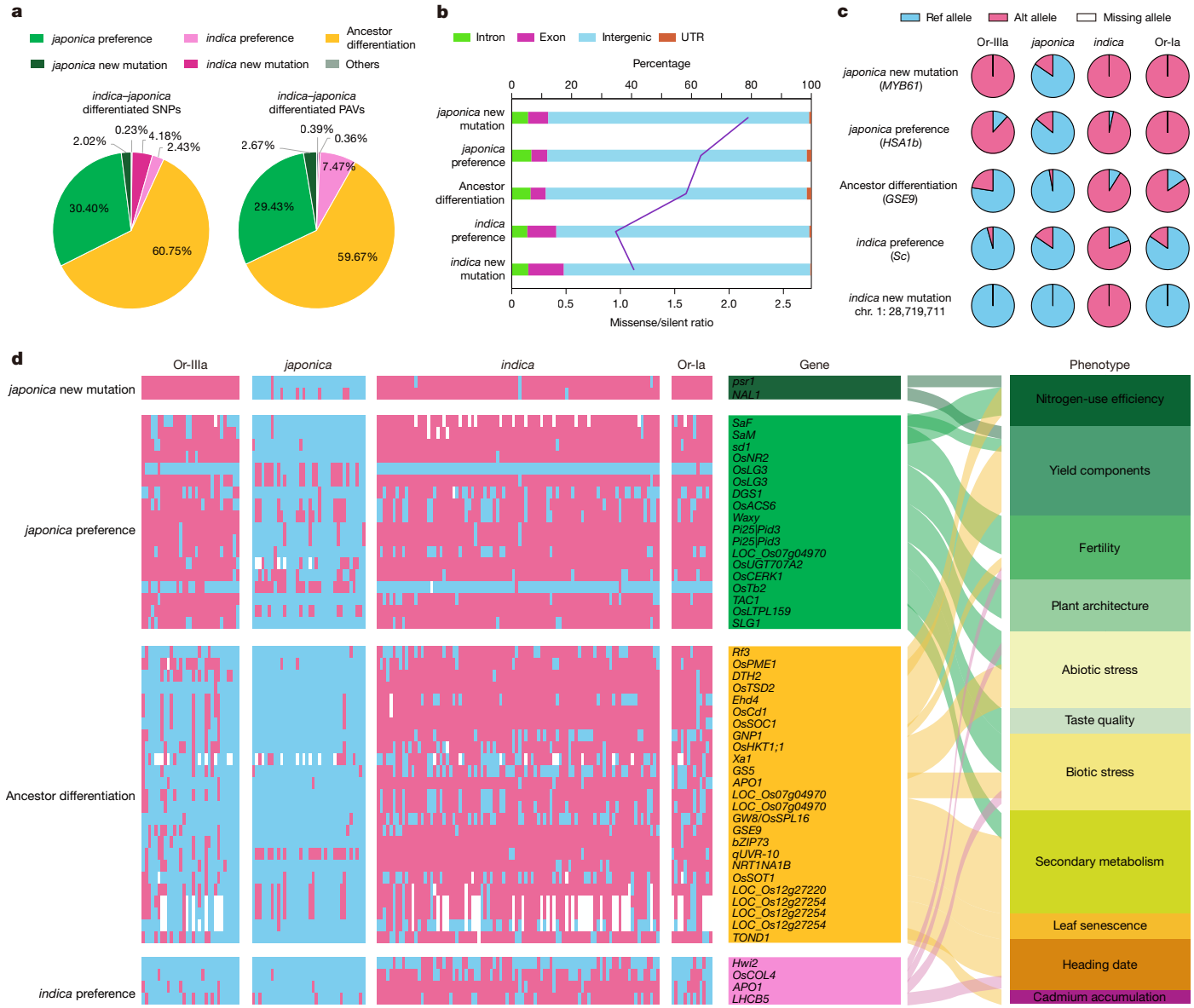


Fig. 5 | Genetic divergence between *indica* and *japonica*. **a**, Pie charts showing the origins of *indica*–*japonica* differentiated SNPs and PAVs, coloured by their composition. **b**, Genomic locations of different origins of *indica*–*japonica* differentiated SNPs, with the percentage indicated on the top *x* axis. Concurrently, the missense/silent ratio for these SNPs is plotted on the bottom *x* axis. UTR, untranslated region. **c**, Allele frequencies of the causal *indica*–*japonica* differentiated SNPs or PAVs in the *MYB61*, *HSA1b*, *GSE9* and *Sc* genes and in chr. 1: 28,719,711 across Or-IIIa, *japonica*, *indica* and Or-Ia. Blue, reference (ref) allele; pink, alternative (alt) allele. **d**, Left, allele state of the 49 *indica*–*japonica* differentiated QTNs in Or-IIIa, *japonica*, *indica* and Or-Ia. Each row indicates a QTN site, and each column indicates an accession. The colours of the allele types are as in **c**. Right, Sankey diagram showing the association between a specific gene (left) and the phenotype it regulates (right). The colour of each left strand corresponds to different origins of *indica*–*japonica* differentiated QTNs, and the colour of each right strand corresponds to different phenotypic categories.

SNP-based analysis, we identified domestication-related presence–absence variations (domPAVs) using Fisher's exact tests to compare PAV frequencies between both populations. This method effectively captured most well-known domestication genes in the first stages of rice domestication within identified selective sweep peaks or domPAV hotspots, including *Bh4* (hull colour)[43], *PROG1* (tiller angle)[44,45], *sh4* (seed shattering)[46], *OsC1* (leaf sheath colour and apiculus colour)[47], *OsAMT1;1* (nitrogen-use efficiency)[48], *OsG* (seed dormancy)[49], *FZP* (panicle architecture)[50], *OsLG1* (panicle shape and ligule development)[51], *An-1* (awn length)[52], *An-2* (awn length)[53,54] and *Rc* (pericarp colour)[55] (Fig. 4a and Supplementary Fig. 20).

We used differentiated SNPs between Or-IIIa and Or-Ib as markers to trace the origin of domestication regions, revealing that many domestication genes are situated within segments derived from Or-IIIa, such as *An-1*, *OsAMT1;1* and *An-2* (Fig. 4a). Selective regions in cultivated rice showed a higher proportion of Or-IIIa-derived segments, compared with the whole genome, particularly in *indica*, *aus* and intro-*indica*, indicating a monophyletic origin from Or-IIIa (Fig. 4b). In contrast to the genome-wide patterns, archetypal and phylogenetic analyses based on SNPs within these selective sweeps clearly showed that all cultivars clustered together, and that the Or-IIIa population tended to be the direct origin of them[19,22] (Fig. 4c and Supplementary Figs. 21–24). Specifically, approximately 55.3% of the selective sweep regions in *aus* were derived from Or-IIIa, the lowest proportion among all cultivars (Fig. 4b), aligning with the idea that *aus* underwent a milder or differently directed domestication process[9,55]. Owing to introgression from *aus*, the proportion of Or-IIIa-derived segments in intro-*indica* was lower than that in *indica*. Similarly, introgression from *indica* resulted

in Or-Ia showing higher similarity to cultivated rice within selective sweeps.

To investigate whether the introgression events of *indica* and *aus* derived from *japonica* were independent, we conducted an ABBA–BABA analysis. The results confirmed gene flow from *japonica* into both *indica* and *aus*, with no evidence of gene flow from *indica* to *aus* (Supplementary Fig. 25a and Supplementary Table 19). We compared the top 1-Mb, 3-Mb and 5-Mb introgression segments, as defined by the f_d value between the two groups. The overlap was only 20%, 10% and 14%, respectively, confirming distinct introgression events from *japonica* to both *indica* and *aus* (Supplementary Fig. 20 and Supplementary Fig. 25b). Consistent results emerged from pairwise comparisons of representative varieties, in which *indica* and *aus* both showed higher similarity to Or-IIIa within selective sweep regions, and *aus* possessed unique segments from Or-IIIa (Supplementary Fig. 25c,d). Collectively, these findings reveal differences in the degree of introgression and selection patterns between *indica* and *aus*, demonstrating that the domestication regions in *aus* were not directly derived from *indica*.

The 11 well-characterized domestication genes mentioned above provided crucial clues for tracing the early domestication history of rice (Supplementary Note 4). Haplotype analysis revealed that all domestication genes in cultivated rice shared a major haplotype, which was either directly linked to the haplotype of Or-IIIa (such as *sh4* and *PROG1*) or linked through a haplotype composed mainly of *japonica* (such as *An-2*) (Supplementary Fig. 26). This finding reinforced the existence of a shared domestication event in cultivated rice originating from Or-IIIa. Variations between the major haplotype and the nearest haplotypes of Or-IIIa were also observed, which might be sites that were selected for or linked to these functional sites during initial domestication (Supplementary Fig. 27). Although most of these sites have been confirmed in previous studies, there are still some unknown potential mutation sites that need further functional verification. Aligning haplotype data with genealogies (Supplementary Fig. 28), we delineated each domestication gene into domesticated or ancestral states. Their distribution across varieties, shown in Fig. 4d, reflected domestication patterns consistent with our deduced pathways. In detail, *japonica* was first domesticated from Or-IIIa in southern China[19], and then spread to South Asia, where it crossed with the local wild rice varieties Or-Ia and Or-Ib, giving rise to *indica* and *aus*, respectively. Extensive hybridization occurred among cultivated rice varieties in South Asia. Intro-*indica* emerged from crosses between *indica* and *aus*, and intro-*indica* further hybridized with *japonica*, leading to the formation of *basmati* (Fig. 4e).

Divergence between *indica* and *japonica*

indica and *japonica*, the two main subspecies of Asian cultivated rice, exhibited significant genetic differentiation (Fig. 3c). Our analysis identified a total of 855,122 highly differentiated SNPs between 90 *indica* cultivars and 36 *japonica* cultivars. We also found 13,853 highly differentiated PAVs between 26 *indica* cultivars and 13 *japonica* cultivars. We further analysed the profiles of these differentiated variations in their respective ancestors, focusing on loci classified as major alleles (with a frequency of 60% or higher). On the basis of this criterion, approximately 77% of differentiated SNPs and 83% of differentiated PAVs between *indica* and *japonica* in the ancestral populations were used for further ancestral tracing analysis (Supplementary Table 20). Notably, of the SNPs meeting the above conditions, 60.75% had already differentiated between Or-IIIa and Or-Ia; 30.40% and 2.43% were preferred in *japonica* and *indica*, respectively, with an additional 2.02% and 4.18% representing novel mutations in *japonica* and *indica* (Fig. 5a). Applying the same method, we observed a similar distribution pattern in PAVs, which also showed a stronger preference in *japonica*. This suggested that a larger genetic bottleneck existed during the domestication of *japonica*. In *indica*, a notable proportion of preferred and novel SNPs was found in exons, albeit with fewer nonsynonymous mutations (Fig. 5b). Conversely, *japonica* exhibited a higher frequency of nonsynonymous mutations in both preferred and new mutations, exemplified by the sterility gene *HSA1b* (ref. 56) (Fig. 5c).

An analysis of 49 differentiated quantitative trait nucleotides (QTNs) in *indica* and *japonica*[57] showed that 51% and 37% of QTNs were derived from ancestral differentiation and *japonica* preference, respectively. These QTNs affected phenotypes that differentiate *indica* and *japonica*, such as plant stature, grain size, heading date and fertility[58] (Fig. 5d). However, we discovered that ancestors exhibited minimal differences in certain traits, such as stem height and grain shape (Supplementary Fig. 29). Consequently, the ancestors of *indica* and *japonica* significantly diverged about 0.1 million years ago (Supplementary Fig. 10b). Their divergence was further amplified by a particularly pronounced genetic bottleneck in *japonica*, which had a crucial role in their distinct evolutionary paths.

Discussion

Wild rice faces critical threats from habitat destruction and limited global dispersal, which hinders research and conservation efforts. To address these challenges and preserve these resources, we constructed a reference-level wild–cultivated rice pangenome, mainly using PacBio HiFi sequencing technology. This near-saturation dataset encompassed 129 *O. rufipogon* accessions, capturing extensive genetic diversity across a broad geographical distribution. Through precise classification of *O. rufipogon* populations, we mapped the evolutionary and domestication pathway leading to major cultivated rice groups, including a newly defined subpopulation, intro-*indica*. Our analysis contributes substantially to researchers' understanding of rice origins and domestication. Furthermore, we investigated the origin of key functional genes of rice, particularly those involved in *indica–japonica* differentiation, laying the groundwork for future studies that combine beneficial genes from different rice subspecies. The development of cross populations between *O. sativa* and *O. rufipogon*, combined with investigations of domestication-related phenotypes in wild rice varieties, enables the identification of new domestication genes, promoting the de novo domestication of wild rice in an accelerated manner[59].

Another key finding of this study is the identification of a rice gene pool containing 68,901 genes, of which about 20% are specific to wild rice. These genetic resources can improve our understanding of rice environmental adaptation, phenotypic plasticity and regeneration potential[60] (Supplementary Fig. 30 and Supplementary Note 5). By bridging the gap between wild and cultivated rice genetics, our study opens new avenues and provides useful wild rice resources for developing superior and more productive rice varieties. These improved varieties could incorporate valuable traits from wild rice species, potentially enhancing their resilience to rapid environmental changes.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-08883-6.

1. Chen, E., Huang, X., Tian, Z., Wing, R. A. & Han, B. The genomics of *Oryza* species provides insights into rice domestication and heterosis. *Annu. Rev. Plant Biol.* **70**, 639–665 (2019).
2. Gnanamanickam, S. S. in *Biological Control of Rice Diseases* (ed. Gnanamanickam, S. S.) 1–11 (Springer, 2009).
3. Godfray, H. C. J. et al. Food security: the challenge of feeding 9 billion people. *Science* **327**, 812–818 (2010).
4. Goff, S. A. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
5. Yu, J. et al. A draft sequence of the rice genome (Oryza *sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).

6. International Rice Genome Sequencing Project & Sasaki, T. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
7. Huang, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
8. Huang, X. et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **44**, 32–39 (2012).
9. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
10. Wang, X. et al. Selective and comparative genome architecture of Asian cultivated rice (*Oryza sativa* L.) attributed to domestication and modern breeding. *J. Adv. Res.* **42**, 1–16 (2022).
11. Tao, Y., Jordan, D. R. & Mace, E. S. Crop genomics goes beyond a single reference genome. *Trends Plant Sci.* **24**, 1072–1074 (2019).
12. Lei, L. et al. Plant pan-genomics comes of age. *Annu. Rev. Plant Biol.* **72**, 411–435 (2021).
13. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
14. Zhou, Y. et al. Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice. *Nat. Commun.* **14**, 1567 (2023).
15. Zhang, F. et al. Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.* **32**, 853–863 (2022).
16. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558 (2021).
17. Shang, L. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
18. Wang, J. et al. A pangenome analysis pipeline provides insights into functional gene identification in rice. *Genome Biol.* **24**, 19 (2023).
19. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
20. Huang, X. & Han, B. Rice domestication occurred through single origin and multiple introgressions. *Nat. Plants* **2**, 15207 (2016).
21. Choi, J. Y. & Purugganan, M. D. Multiple origin but single domestication led to *Oryza sativa*. *G3* **8**, 797–803 (2018).
22. Wu, D. et al. A syntelog-based pan-genome provides insights into rice domestication and de-domestication. *Genome Biol.* **24**, 179 (2023).
23. Civan, P., Craig, H., Cox, C. J. & Brown, T. A. Three geographically separate domestications of Asian rice. *Nat. Plants* **1**, 15164 (2015).
24. Jing, C.-Y. et al. Multiple domestications of Asian rice. *Nat. Plants* **9**, 1221–1235 (2023).
25. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
26. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).
27. Shang, L. et al. A complete assembly of the rice Nipponbare reference genome. *Mol. Plant* **16**, 1232–1236 (2023).
28. Song, J.-M. et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* **14**, 1757–1767 (2021).
29. Wang, C. et al. High-resolution genetic mapping of rice bacterial blight resistance gene *Xa23*. *Mol. Genet. Genomics* **289**, 745–753 (2014).
30. Li, R. et al. The evaluation and utilization of new genes for brown planthopper resistance in common wild rice (*Oryza rufipogon* Griff.). *Mol. Entomol.* **4**, 365–371 (2010).
31. Yang, M. et al. Identification of a novel planthopper resistance gene from wild rice (*Oryza rufipogon* Griff.). *Crop J.* **8**, 1057–1070 (2020).
32. Huang, J. et al. Haplotype-resolved gapless genome and chromosome segment substitution lines facilitate gene identification in wild rice. *Nat. Commun.* **15**, 4573 (2024).
33. Lee, S.-K. et al. Rice *Pi5*-mediated resistance to *Magnaporthe oryzae* requires the presence of two coiled-coil–nucleotide-binding–leucine-rich repeat genes. *Genetics* **181**, 1627–1638 (2009).
34. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
35. Zhou, Y. et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
36. Du, H. et al. Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* **8**, 15324 (2017).
37. Ou, S. et al. Differences in activity and stability drive transposable element variation in tropical and temperate maize. *Genome Res.* **34**, 1140–1152 (2024).
38. Tian, Z. et al. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* **19**, 2221–2230 (2009).
39. Shenton, M. et al. Evolution and diversity of the wild rice *Oryza officinalis* complex, across continents, genome types, and ploidy levels. *Genome Biol. Evol.* **12**, 413–428 (2020).
40. Choi, J. Y. et al. Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biol.* **21**, 21 (2020).
41. Casartelli, A. et al. Exploring traditional aus-type rice for metabolites conferring drought tolerance. *Rice* **11**, 9 (2018).
42. Zhou, J. et al. Interspecific hybridization is an important driving force for origin and diversification of Asian cultivated rice *Oryza sativa* L. *Front. Plant Sci.* **13**, 932737 (2022).
43. Zhu, B.-F. et al. Genetic control of a transition from black to straw-white seed hull in rice domestication. *Plant Physiol.* **155**, 1301–1311 (2011).
44. Jin, J. et al. Genetic control of rice plant architecture under domestication. *Nat. Genet.* **40**, 1365–1369 (2008).
45. Tan, L. et al. Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**, 1360–1364 (2008).
46. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
47. Zheng, J. et al. Determining factors, regulation system, and domestication of anthocyanin biosynthesis in rice leaves. *New Phytol.* **223**, 705–721 (2019).
48. Ding, Z., Wang, C., Chen, S. & Yu, S. Diversity and selective sweep in the *OsAMT1;1* genomic region of rice. *BMC Evol. Biol.* **11**, 61 (2011).
49. Wang, M. et al. Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.* **50**, 1435–1441 (2018).
50. Huang, Y. et al. Variation in the regulatory region of *FZP* causes increases in secondary inflorescence branching and grain yield in rice domestication. *Plant J.* **96**, 716–733 (2018).
51. Ishii, T. et al. *OsLG1* regulates a closed panicle trait in domesticated rice. *Nat. Genet.* **45**, 462–465 (2013).
52. Luo, J. et al. *An-1* encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. *Plant Cell* **25**, 3360–3376 (2013).
53. Gu, B. et al. *An-2* encodes a cytokinin synthesis enzyme that regulates awn length and grain production in rice. *Mol. Plant* **8**, 1635–1650 (2015).
54. Hua, L. et al. *LABA1*, a domestication gene associated with long, barbed awns in wild rice. *Plant Cell* **27**, 1875–1888 (2015).
55. Sweeney, M. T. et al. Global dissemination of a single mutation conferring white pericarp in rice. *PLoS Genet.* **3**, e133 (2007).
56. Kubo, T., Takashi, T., Ashikari, M., Yoshimura, A. & Kurata, N. Two tightly linked genes at the *hsa1* locus cause both F_1 and F_2 hybrid sterility in rice. *Mol. Plant* **9**, 221–232 (2016).
57. Wei, X. et al. A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat. Genet.* **53**, 243–253 (2021).
58. Jiang, L. et al. Research progress on the divergence and genetic basis of agronomic traits in *xian* and *geng* rice. *Crop J.* **10**, 924–931 (2022).
59. Yu, H. et al. A route to de novo domestication of wild allotetraploid rice. *Cell* **184**, 1156–1170 (2021).
60. Wang, C. & Han, B. Twenty years of rice genomics research: from sequencing and functional genomics to quantitative genomics. *Mol. Plant* **15**, 593–619 (2022).

Methods

Plant material and sequencing

We selected a total of 149 rice accessions according to the phylogenetic relationships of 1,529 accessions reported previously[19], which were preserved at the China National Rice Research Institute in Hangzhou, China, and the National Institute of Genetics in Mishima, Japan. For sampling, all accessions were cultivated in Lingshui County, Hainan.

We extracted genomic DNA from fresh leaves, which were then immediately flash-frozen in liquid nitrogen and stored at −80 °C. HiFi sequencing was performed on the Pacific Biosciences Sequel II platform for 133 accessions, and 16 accessions were subjected to nanopore sequencing on the PromethION platform, following their respective standard protocols. After removing low-quality reads, the average N50 lengths of pass reads were 16 kb for the HiFi-sequenced accessions and 26 kb for the ONT-sequenced accessions. The sequencing yielded 6.8–17.4 Gb of HiFi reads for the 133 rice accessions and 28.4–55.3 Gb of nanopore reads for the 16 rice accessions. The leaves of 142 rice accessions were also used to extract genomic DNA and construct Illumina libraries with a 350-bp insert size, which were then sequenced on the HiSeq 4000 platform (Illumina). Detailed information on the geographical origins and sequencing coverage of these accessions can be found in Supplementary Tables 1 and 2.

We selected leaves at the seedling stage from 30 representative varieties for Hi-C sequencing. Following the standard protocol, Hi-C libraries were constructed from these samples, which were subsequently sequenced on the NovaSeq platform (Illumina).

Tissues were collected from various plant parts, including leaves during vegetative growth from 149 accessions, roots from 146 accessions, seedlings (about 15 days old) from 137 accessions and panicles (3–5 cm) from 122 accessions. RNA was then extracted using TRIzol (Invitrogen). RNA sequencing (RNA-seq) libraries with an insert size of 350 bp were prepared and sequenced on the HiSeq 4000 platform (Illumina) according to the instructions provided.

Genome assemblies

We assembled the genomes of 133 accessions sequenced with HiFi technology using Hifiasm[61] (v.0.16.0), with default parameters. This process yielded both p-contigs and a-contigs. For seven accessions sequenced with nanopore technology, we used NECAT[62] (v.20200803) with default parameters for assembly. The remaining nine accessions were assembled using NextDenovo[63] (v.2.1). To improve the consistency and single-base accuracy of nanopore genomes, we used Racon[64] (v.1.0.0) for one round of polishing using nanopore reads, followed by two rounds of polishing with NextPolish[65] (v.1.0.2) using Illumina reads. To exclude organelle genomes, we aligned the chloroplast (NC 001320.1) and mitochondrion (NC 011033.1 and NC 001751.1) reference sequences from IRGSP-1.0 against the assembly contigs of each accession using MUMmer[66] (v.4.0.0beta2). Contigs showing more than 50% coverage and measuring less than 500 kb were specifically targeted and removed. The remaining contigs, attributed to the nuclear genome, were subsequently evaluated and anchored to pseudo-chromosomes using ALLMAPS[67].

To achieve chromosome-level genome assembly, the obtained Hi-C reads were aligned to contigs using Chromap[68] (v.0.2.6). These alignments were then converted into bed or bam format by SAMtools[69] (v.1.20), and processed with YaHS[70] (v.1.1). Subsequently, Hi-C contact maps for visualization were generated by juice_tools (v.1.19.02), followed by manual optimization with JuiceBox Assembly Tools[71] (v.1.9.8).

Evaluation of genome assemblies

To assess the quality of genome assemblies, we implemented several indexes. First, we evaluated gene completeness using the embryophyta_odb10 database, using BUSCO[72] (v.5.2.2), and repeat completeness on the basis of the LTR assembly index (LAI)[25], using LTR_retriever (v.2.9.0) with parameters '-maxlenltr 7000'. Furthermore, the overall assembly quality (QV) was measured using Inspector[26] (v.1.0.1), a reference-free assembly evaluator. To round off our evaluation, the number of mismatches between the Nipponbare genome assembled in this study and the reference genomes IRGSP-1.0 and T2T-NIP was assessed using QUAST[73] (v.5.0.1).

For the syntenic analysis at the genome level between the Nipponbare (IRGSP-1.0) genome and four representative genomes in *O. sativa* (Gla4), *O. rufipogon* (W1943), *O. longistaminata* (OL3101) and *O. meridionalis* (OM1952), we used the nucmer program from MUMmer[66] (v.4.0.0beta2). Syntenic blocks identified were filtered using the delta-filter program with parameters '-i 85 -l 5000 -o 85' and then visualized with the mumplot program.

We performed a comparative genomic analysis between chromosomes constructed using Hi-C technology and pseudo-chromosomes obtained with the ALLMAPS method. At first, we aligned the genomes using the nucmer program from MUMmer[66]. Syntenic blocks identified were filtered using the delta-filter program with parameters '-m'. Subsequently, we compared alignments between two chromosome-level assemblies and identified synteny blocks and structural rearrangements using SyRI[74] (v.1.4).

Structure and functional annotation of protein-coding genes

Three distinct strategies, comprising ab initio, homology-based and transcriptome-based predictions, were integrated to generate the predicted gene models. For ab initio prediction, four different programs were used: FGENESH+[75] (v.3.1.1), SNAP[76] (v.2006-07-28), GeneMark-ES[77] (v.4.68_lic) and AUGUSTUS[78] (v.3.3.2). In the homology-based prediction, homologous protein sequences, sourced from the Rice Genome Annotation Project database (v.7.0, http://rice.plantbiology.msu.edu), were aligned to the assembled genomes using GenomeThreader[79] (v.1.7.1). RNA-seq reads from four different tissues of each accession were mapped to their respective assembled genomes using HISAT2 (ref. 80) (v.2.0.5) and then assembled into transcripts with StringTie[81] (v.2.0). We performed both de novo and genome-guided RNA-seq assemblies using Trinity[82] (v.2.12.0), subsequently aligning them to the genome assemblies with PASA[83] (v.2.0.1). By assigning appropriate weights to every predicting method, we synthesized all predicted gene structures into consensus gene models using EVidenceModeler[84] (v.1.1.1). The protein-coding annotations were assessed using BUSCO[72] (v.5.2.2).

For functional annotation of genes, InterProScan[85] (v.5.56-89.0) was used to predict potential protein domains. For calculating gene-expression levels, low-quality RNA-seq reads were first removed using fastp[86] (v.0.23.0) with parameters '-l 30'. Then, the filtered paired-end reads were mapped against the index of decoy sequences, which concatenated the genome to the end of the annotated transcripts, using Salmon[87] (v.1.6.0) in mapping-based mode with parameters '-l A --validateMappings --gcBias'. Finally, gene-expression levels were quantified by counting the number of reads mapping to each transcript and calculating the transcripts per million (TPM) values.

Identification of allelic and differentially expressed allelic genes

We developed a pipeline to discern allelic genes between p-contigs and a-contigs within HiFi genome assemblies (Supplementary Fig. 31). First, allelic gene pairs were determined to be reciprocal best hits using GeneTribe[88] (v.1.2.0) with the default parameters. Subsequently, the full-length sequences of genes from a-contigs underwent a BLASTN (v.2.9.0+) search against those from p-contigs. Genes with both identity and coverage values equating to 100% were classified as homozygous alleles, and all others as heterozygous alleles. In addition, genes present in a-contigs but missing in p-contigs (MIP genes) were isolated on the basis of two stringent conditions: (1) no hit when gene sequences from a-contigs were aligned to p-contigs using BLASTN (v.2.9.0+) and (2) no path found when the coding sequences (CDSs) from a-contigs were

aligned to p-contigs using GMAP[89] (v.2021-05-27). The other genes on a-contigs that are not classified were defined as 'others'.

The identification of differentially expressed allelic genes was based on three standards that must be simultaneously satisfied:
(1) $TPM_{p\text{-}gene}/TPM_{a\text{-}gene} > 2$ or $TPM_{p\text{-}gene}/TPM_{a\text{-}gene} < 1/2$;
(2) $NumReads_{p\text{-}gene}/(NumReads_{p\text{-}gene} + NumReads_{a\text{-}gene}) > 0.75$ or $NumReads_{p\text{-}gene}/(NumReads_{p\text{-}gene} + NumReads_{a\text{-}gene}) < 0.25$;
(3) $NumReads_{p\text{-}gene} + NumReads_{a\text{-}gene} \geq 10$.

Identification of RGAs

The RGAugury[90] pipeline was used to predict the RGAs in 133 wild rice accessions (including 129 *O. rufipogon*, 3 *O. longistaminata* and 1 *O. meridionalis*) and 129 cultivated rice accessions (Supplementary Table 8). RGA candidates were identified and classified into four major families on the basis of the presence of combinations of these RGA domains and motifs: RLKs, NBSs, TM-CC proteins and RLPs. Considering that RGAs tend to cluster together in the genome, tandem duplications were identified using the 'jcvi.para.catalog' module of the MCscan[91] (Python version) pipeline, with the default parameters based on their location on the chromosomes. RGAs in each locus were categorized as singletons, pairs and clusters if they numbered 1, 2 and more than 2, respectively.

Collinear blocks for each accession relative to all others were constructed using the 'ortholog' tool of the 'jcvi.compara.catalog' module in the MCscan (Python version) pipeline. Tandem genes were integrated using the 'mcscan' tool of the 'jcvi.compara.synteny' module with the parameter '--mergetandem'. Then, all of the collinear blocks for each accession with all others were joined to a matrix using the 'join' tool of the 'jcvi.formats.base' module. Finally, a comprehensive RGA matrix was created by merging, sorting and deduplicating all collinear matrices using a custom script.

Identification of centromeres and telomeres

The sequences of CentO satellite repeats in rice, which had been reported previously[92], were aligned against nuclear genomes using ClustalW[93] (v.2.1). The generated alignment file was converted to Stockholm format using an online sequence conversion tool (http://sequenceconversion.bugaco.com/converter/biology/sequences/), which then served as the input for constructing an HMM file using the 'hmmbuild' function in HMMER[94] (v.3.1b2). Next, a homology search was performed to identify the CentO repeats for each chromosome across all accessions using nhmmer, with an E-value threshold set to 1×10^{-5}. We adopted the strategy of extracting one CentO repeat unit at every fiftieth interval on each chromosome to select some CentO repeats for a subsequent similarity comparison across genomes using MAFFT[95] (v.7.490). Subsequent phylogenetic analysis was performed using IQ-TREE[96] (v.1.6.12), incorporating a bootstrap value of 1,000 for robustness.

To identify the telomere sequences in each chromosome, the telomere sequence 5′-CCCTAAA-3′ and the reverse complement of the seven bases were searched directly by the custom script.

TE annotation

TEs of 149 chromosome-level genomes were identified by the EDTA[97] (v.2.1.0) pipeline, using both a manually curated rice TE library (rice6.9.5.liban) and the annotated CDSs of each genome. Then, individual non-redundant TE libraries, generated from each genome, were combined with the curated TE library (rice6.9.5.liban) by panEDTA[37], leading to the formation of a comprehensive pangenome TE library. This pangenome TE library was then used to reannotate whole-genome TEs in our study's 149 assemblies, as well as in 28 rice assemblies from a previously published pangenome of 33 cultivated rice accessions[16] (excluding *Oryza barthii*, *Oryza glaberrima*, *aus*, *basmati* and WSSM) using RepeatMasker (http://repeatmasker.org) (v.4.1.2). The insertion time of each intact Gypsy and Copia element was estimated using LTR_retriever[98] (v.2.9.0) with default parameters.

In a previous study[37], the dynamics of the LTR family were determined by comparing family size and the ratio of solo-to-intact LTRs within each family between two groups. Families showing significant differences in both sizes and the solo-to-intact LTR ratio were categorized as 'removal families'. We classified cases in which only the family size was significantly different as 'amplification families'. On the other hand, if there was a notable change in the solo-to-intact LTR ratio without a corresponding shift in family size, these were classified as 'balanced families'. Families that did not exhibit a significant difference in either dimension were termed 'drifting families'. To classify the dynamics of Or-IIIa-larger Gypsy families, we first identified solo Gypsy elements using the 'solo_finder.pl' script from the LTR_retriever package, and obtained family information for the intact Gypsy elements from each genome's final annotation results. Then, the solo-to-intact ratio was calculated by dividing the number of solo elements by the number of intact elements within the Gypsy families. Finally, we applied Student's *t*-tests to compare the family size and solo-to-intact ratio between Or-IIIa and *japonica* groups, with $P < 0.01$ as the cut-off for significance.

We mapped the sequences of identified insertions and deletions to the comprehensive pangenome TE library using BLASTN (v.2.12.0+). If both the identity and the coverage reached 80%, the PAV was defined as a TE insertion polymorphism (TIP)[99]. To identify the genes adjacent to the Or-IIIa-larger Gypsy families, we mapped the gene sequences against the TIP sequences containing the Or-IIIa-larger Gypsy families in each Or-IIIa genome. Genes with an identity of at least 95% and a coverage of at least 50% were classified as adjacent to these families. Gene Ontology (GO) enrichment analysis of these genes was performed in the R package clusterProfiler (v.4.6.2), with $P \leq 0.05$ as the threshold for significance.

SV calling

We adopted three strategies to detect PAVs (large insertions and deletions) in the 133 Hifi genomes. (1) We mapped HiFi reads to the IRGSP-1.0 reference genome using pbmm2 (https://github.com/PacificBiosciences/pbmm2/) (v.1.4.0) with the '--preset CCS' parameters. Then, pbsv (https://github.com/PacificBiosciences/pbsv) (v.2.6.2) was used for variant calling of each accession with the parameters '--min-sv-length 30 --max-ins-length 100K --max-dup-length 100K'. (2) We mapped HiFi reads to the IRGSP-1.0 reference genome using minimap2 (ref. 100) (v.2.21-r1071) with the '-x map-hifi' parameters. CuteSV[101] (v.1.0.11) was then used for variant calling of each accession with the parameters '--min_support 3 --min_size 30 --max_size 100000 --max_cluster_bias_INS 1000 --diff_ratio_merging_INS 0.9 --max_cluster_bias_DEL 1000 --diff_ratio_merging_DEL 0.5'. (3) We mapped assembled contigs to the IRGSP-1.0 reference genome using minimap2 (ref. 100) (version 2.22-r1071) with the parameters '-x asm5 –cs -r 2k'. For variant calling, SVIM-asm[102] (v.1.0.7) was operated in 'haploid' mode, with the parameters '–min_sv_size 30 –max_sv_size 100000'. For obtaining high-confidence variations, we merged all insertions within a 50-bp range and deletions within a range of 50% of their length using SURVIVOR[103] (v.1.0.6). We reported only those variants that were corroborated by at least two of the calling methods and for which there was a consensus on the variant type. To detect PAVs in 16 nanopore genomes, we tailored our approach by using the second and third strategies of the above method, with some modifications to suit the characteristics of nanopore data: (1) the minimap2 (ref. 100) (v.2.21-r1071) parameters for mapping nanopore reads were adjusted to '-x map-ont'; (2) the parameters for CuteSV[101] (v.1.0.11) were modified to increase the minimum support threshold to 10; and (3) the minimum supporting caller number in SURVIVOR[103] (v.1.0.6) was adjusted to one.

To detect translocations and inversions, we first aligned each pseudochromosome to the reference genome across 149 genomes and then used the SyRI[74] (v.1.4) pipeline for variation calling. Per the classifications provided by SyRI, INV variants were categorized as inversions,

in comparison with the Nipponbare reference. Both TRANS and INVTR variants were categorized as translocations. To detect small indels, we extracted INS and DEL variants consisting of fewer than 30 bp.

SNP calling and annotation

We implemented both read-mapping-based and assembly-based approaches to identify SNPs using Nipponbare as the reference genome. For read mapping, we called SNPs through Longshot[104] (v.0.4.1), using the alignment results of reads by minimap2 (ref. 100) during the process of PAV calling. Parameters were set to '-c 3 -D 3:10:50' for HiFi genomes and '-c 10 -D 3:10:50' for nanopore genomes. For assembly-based calling, we first aligned each contig to the reference genome using the 'nucmer' program and refined the alignments to one-to-one matches using the 'delta-filter' program. SNPs were then identified using the 'show-snps' program with the '-C -I' parameters, all from the MUMMER package[66] (v.4.0.0beta2). To minimize false positives, we only considered SNPs detected by both methodologies.

From a population of 280 accessions, including 132 long-read sequencing genomes sourced from published studies (Supplementary Table 16), SNP calling was performed by mapping the assembled contigs using MUMMER[66] alone. We merged the resulting SNP datasets for each sample using a custom Perl script. To compile a high-confidence SNP dataset, we used the 'VariantFiltration' function in the Genome Analysis Toolkit[105] (v.4.1.4.0) with the '--cluster-window-size 10 --cluster-size 3' parameters. This dataset served as the basis for further evolutionary analysis. Finally, we annotated and predicted the effects of our identified SNPs using SnpEff[106] (v.55.0), to ensure a comprehensive understanding of their potential impact. The same method was also applied to the population of 510 samples (Supplementary Table 17).

Pangenome construction

We performed all-versus-all CDS alignment in the pangenomes for 16 *O. sativa* accessions, 129 *O. rufipogon* accessions and a combined set of 145 wild–cultivated rice accessions (16 *O. sativa* and 129 *O. rufipogon*) using BLASTN (v.2.2.18). If a gene was aligned with at least 95% identity and at least 50% coverage, it was considered present in the corresponding genome. On the basis of their frequency, we classified genes into the following four categories: core (those present in all individuals), soft-core (those present in more than 90% of samples but not all), dispensable (those present in more than one but less than 90% of samples) and private (present in only one accession). To achieve a balanced comparison, we incorporated 113 non-redundant cultivated rice genomes from 3 previously published pangenomic datasets to match the size of the wild rice population (Supplementary Table 10). Gene annotation was performed uniformly across all samples using a consistent methodology. In the dataset comprising 129 *O. rufipogon* and 129 *O. sativa* accessions, genes present exclusively in wild rice and absent in all cultivated rice were defined as wild-rice-specific genes.

To construct three distinct pangenome graphs for our study, we applied the Minigraph-Cactus pipeline[107] to the assembled genomes of 16 *O. sativa*, 129 *O. rufipogon* and a combined set of 145 accessions. The first step involved using minigraph[108] (v.0.19-r551) to develop a primary pangenome graph, capturing the SVs within the input assemblies. Subsequently, these assemblies were remapped onto the primary graph using minigraph[108]. The mapping results were then used as the input for Cactus[109] (v.2.2.1), which facilitated the generation of the final graphs. We defined the graph size as the total length of all nodes, and nodes that were not included in the reference genome (non-ref) were defined as novel sequences. To call variants for 142 accessions from our study (the remaining 6 samples lacked next-generation data) and 407 newly sequenced samples (33 *O. sativa* and 374 *O. rufipogon* or *Oryza nivara*) from another study[24] (Supplementary Table 14), the Illumina short paired-end reads from each accession were mapped against the graph-based cultivated–wild pangenome using vg giraffe[34] (v.1.43.0). The variations were then called using DeepVariants[110] (v.1.6.1) with the NGS model, and all individual variants were merged using GLnexus[111] (v. 1.4.1-0-g68e25e5).

Phylogenetic tree construction

Chloroplast genomes are very conserved across different species and are frequently used to construct phylogenetic evolutionary trees, which can be instrumental in studying species classification and understanding their evolutionary relationships[112]. To assemble chloroplast genomes of wild rice in our study, the HiFi reads were first aligned to the reference chloroplast genome of Nipponbare (Gene Bank ID: GU592207.1) using minimap2 (ref. 100) (v.2.21-r1071) with the '-x map-hifi' parameters. Chloroplast-derived reads with higher than 70% coverage were then extracted using a custom Perl script. The final assembly of the chloroplast genome was then performed using hifiasm[61] (v.0.16.0) with default parameters. Locally collinear blocks among 72 assembled chloroplast genomes, along with published chloroplast genomes of *O. barthii* (KF359904.1), *Oryza glumipatula* (NC_027461.1), *O. longistaminata* (NC_027462.1), *O. meridionalis* (OV049999.1), *O. rufipogon* (NC_017835.1) and Nipponbare, were identified for constructing multi-sequence alignments using HomBlocks[113]. The phylogenetic tree was then constructed using IQ-TREE[96] (v.2.2.0.3) with 1,000 bootstraps. On the basis of the results, we re-identified three accessions of *O. longistaminata* and one of *O. meridionalis*.

We performed an all-versus-all comparison of the amino acid sequences of protein-coding genes using DIAMOND[114] (v.2.0.15). These genes were from 149 genome assemblies and 31 assemblies (excluding the same species NIP and WSSM) from a cultivated rice pangenome[16]. The alignment results were then input into OrthoFinder[115] (v.2.5.4) to find orthogroups and orthologues. Using 844 identified single-copy orthologues, we constructed a gene-based maximum-likelihood phylogenetic tree using IQ-TREE[96] (v.2.2.0.3) with 1,000 bootstraps.

To determine the phylogenetic relationships of three populations, including wild rice and cultivated rice (Supplementary Tables 14, 16 and 17), we first converted the SNP VCF files into tfam format using PLINK[116] (v.1.90b6.9 64-bit). After this, a kinship matrix was generated using EMMAX[117] (v.beta-07Mar2010) with the '-v -h -s -d 10' parameters. The neighbour-joining phylogenetic tree was then constructed using the PHYLIP package (https://phylipweb.github.io/phylip/) (v.3.66). For visualizing the resulting phylogenetic trees, the interactive tool iTOL[118] was used.

Archetypal analysis

We first identified core SNP subsets of three populations, including wild rice and cultivated rice (Supplementary Tables 14, 16 and 17), each exhibiting a minor allele frequency of more than 0.05 and a missing rate of less than 0.8, using VCFTools[119] (v.0.1.16). Further refinement was done using PLINK[116] (v.1.07) to exclude SNPs with substantial LD ($r^2 \geq 0.5$) in each sliding window (in windows of 100 SNPs within steps of 10 SNPs). Archetypal analysis[120] of these SNP sets was performed with the parameters '-tolerance 0.0001 --max_iter 400'.

Calculations of π, F_{ST}, LD and DST

The nucleotide diversity (π) of each group and the fixation index (F_{ST}) between different groups were both estimated using VCFTools[119] (v.0.1.16) with a window size of 100 kb and a step size of 10 kb. The genome-wide LD decay pattern for each group was calculated using PopLDdecay[121] (v.3.42) and plotted using the Plot_MultiPop.pl script in the PopLDdecay package with parameters '-bin1 500 -bin2 7000 -break 5000'. DST was calculated using PLINK[116] (v.1.07) with the '--genome' and '--genome-full' options. Heat plots of 1-DST matrices were made with the ggplot2 package in R (v.4.1.3).

Demographic history inference

We used MSMC2 (ref. 122) (v.2.1.4) to infer the population separation history. Our analysis began with the preparation of a negative mask

file for the coding region of IRGSP-1.0 (MSU7.0) and a mappability mask file using seqbility (http://lh3lh3.users.sourceforge.net/snpable.shtml) (v.20091110) and makeMappabilityMask.py. The phased SNP sites with uniquely mapped reads and mean coverage depths greater than threefold were acquired using Longshot[104] (v.0.4.1) and the high-quality regions of each genome were acquired using the filtered results of show-snps from MUMmer[66] (v.4.0.0beta2). The MSMC2 input files were constructed by merging VCF and mask files using the 'generate_multihetsep.py' script. Because *O. rufipogon* naturally uses both cross-pollination and self-pollination, we followed an established approach of constructing pseudodiploids, which has been widely used in similar studies of inbreeding species such as *Caenorhabditis*[123], *Arabidopsis thaliana*[124], soybean[125] and African wild rice[126,127]. We randomly selected four samples from each population and treated each sample as a single haplotype. We then paired chromosomes from haplotypes within the same population to construct pseudodiploids. The population split inference focused on 2 individuals (4 haplotypes) per group, calculating median population split times based on 50 random combinations for each comparative analysis. A mutation rate of 8.09×10^{-9} per site per generation[128] and a generation time of one year were applied to estimate demographic history.

Introgression analysis

We used TreeMix[129] (v.1.13) to infer population admixture graphs for major groups of *O. rufipogon* (Or-IIIa, Or-Ia and Or-Ib) and *O. sativa* (*japonica*, *basmati*, *indica*, intro-*indica* and *aus*) from East Asia and South Asia. *Oryza officinalis* (CC genome), *O. longistaminata* and *O. meridionalis* were set as outgroups to construct the phylogenetic tree. We systematically varied the number of migrations from 0 to 10, performing 10 iterations. For each migration event, TreeMix was executed by randomly sampling approximately 80% of the SNP loci using a random seed, applying the '-global -k 500' parameters for global allele frequency estimation. The optimal number of migration edges ($m = 4$) was determined using the R package OptM[130] (v.0.1.6).

To detect potential admixture events of the form (target; source 1, source 2), we performed an F_3-admixture test using the qp3Pop program in ADMIXTOOLS[131] (v.7.0.2). Under the null hypothesis that the target population is not a mixture of populations related to source 1 and source 2, the expected F_3 statistic would yield a non-negative mean. A negative mean of the F_3 statistic, on the other hand, would suggest admixture in the target population, with genetic contributions from groups related to source 1 and source 2. A z-score below −3 was considered indicative of significant admixture in population C.

Using a four-taxon model (((P1, P2), P3), PO), we calculated the D-statistic to perform the ABBA–BABA test, using the script calculate_abba_baba.r (https://github.com/palc/tutorials-1/tree/master/analysis_of_introgression_with_snp_data/src). With *O. longistaminata* designated as the outgroup, our analysis revealed a significantly positive D-statistic ($P < 0.01$), suggesting introgression between P3 and P2. To delve deeper into introgression segments between *indica* and *aus* from *japonica*, we computed the f_d statistic across the genome in 100-kb sliding windows with a step size of 10 kb, using the script ABBABABAwindows.py from genomics_general toolkit (https://github.com/simonhmartin/genomics_general). The minimum number of SNPs per window was set to 250, and the minimum proportion of samples genotyped per site was set to 0.4. The $f_d < 0$ values are converted to zero, and $f_d > 1$ values are converted to 1. Finally, to assess the congruence of introgression regions between *indica* and *aus* from *japonica*, we catalogued the putative introgression segments within the top 10, 30 and 50 100-kb windows.

Pairwise differentiation comparison

To quantify the genetic similarities between the two groups, we performed a comprehensive analysis of all possible pairwise combinations of varieties. We focused on identifying 'same windows', defined as those with a similarity exceeding 99.99%. The similarity index for each 10-kb window was calculated using the following formula:

$$\text{Similarity} = (10{,}000 - \text{Num}_{\text{diff}} - \text{Num}_{\text{nan}})/10{,}000.$$

Here, Num_{diff} is the number of differing SNPs, and Num_{nan} is the number of sites with missing data within each window. This methodology was also applied to predict potential introgression fragments between two taxa. For this purpose, we selected a representative variety from each group and mapped out the count of different SNPs (that was termed as pairwise differentiation) across 10-kb non-overlapping windows within specified regions.

Identification of genomic selective sweep

To detect selective sweeps associated with artificial selection during domestication, we calculated $\pi_{\text{wild}}/\pi_{\text{cultivated}}$ and F_{ST} using VCFtools[119] (v.0.1.16) with a 100-kb sliding window and a 10-kb step. After this, we used BEDTools[132] (v.2.30.0) with the parameter '-d 30000' to merge overlap regions that were identified within the top 5% of two values. It is worth highlighting that our analysis was restricted to Or-IIIa and Or-Ib as representatives of the wild rice groups, given the extensive gene flow observed in Or-Ia from *indica*. The cultivated rice category encompassed *indica*, *japonica*, *aus* and *basmati*. Phylogenetic tree construction and archetypal analysis based on the SNPs within identified selective sweeps were in keeping with the above methods for whole genomes. To identify domPAVs, we performed a two-sided Fisher's test comparing wild and cultivated rice[133], considering PAVs with a false discovery rate (FDR)-adjusted $P < 0.05$ as significant.

To trace the origins of domestication regions in rice, we first identified 566,513 differentiated SNPs between 31 Or-IIIa accessions and 19 Or-Ib accessions, exhibiting an allele frequency greater than 0.8. We then assessed the major alleles, which have a frequency of 90% or higher, at these SNP sites within the Asian cultivated rice groups and Or-Ia. If the major allele matched that of Or-IIIa, the SNP was classified as originating from Or-IIIa; otherwise, it was classified as deriving from Or-Ib. The whole-genome distribution was visualized using RectChr (https://github.com/BGI-shenzhen/RectChr) (v.1.36).

Haplotype analysis and construction of phylogenetic trees of domestication genes

To construct phylogenetic trees of domestication genes, we selected 11 representative genes known for their roles in the early stages of rice domestication, as documented in published literature. Using GMAP[89] (v.2021-05-27), we extracted gene region sequences or gene regions along with their specified upstream and downstream regions from the 280 rice accessions in this study. These sequences were then aligned using MAFFT[95] (v.7.490), applying the parameter '--maxiterate 1000' to optimize alignment accuracy. For phylogenetic tree construction, we extended model selection followed by tree inference with 1,000 bootstrap replications using IQ-TREE[96] (v.2.2.0.3), and set *O. officinalis* (CC genome), *O. longistaminata*, *O. meridionalis* and *O. glaberrima* as outgroups. In the haplotype analysis phase, we filtered the intron sequences without candidate functional sites or QTNs for haplotype analysis and visualized haplotype networks using the R package geneHapR[134] (v.1.1.9) after manually trimming the alignment sequences. Our analysis retained haplotypes with a frequency greater than two and those closely related to the major haplotype for clarity.

Identification of *indica–japonica* differentiated variations

We identified SNP sites with highly differentiated alleles between 90 *indica* cultivars and 36 *japonica* cultivars, requiring that they must have an allele frequency higher than 0.9 in both groups. We identified a total of 855,121 *indica–japonica* differentiated SNPs. To ascertain the ancestral origin of these SNPs, we analysed the states of major alleles (frequency ≥ 0.6) in their respective ancestral groups and divided

them into six categories (Supplementary Table 20). We also applied the same standard to identify 13,853 *indica–japonica* differentiated PAVs between 26 *indica* cultivars and 13 *japonica* cultivars.

Distribution maps

The geographical records of all wild rice in the study were obtained by collecting field samples. Approximate latitude and longitude information on their distribution ranges was used for spatial mapping, and this can be found in Supplementary Table 2. The distribution map was generated using the open-source Python tool GeoPandas[135] (v.0.14.4) (BSD-3-Clause licence), with base map layers derived from a public-domain Natural Earth dataset (https://www.naturalearthdata.com/).

Reporting summary

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

Data availability

All raw sequencing data, transcriptome data and Hi-C data generated in this study have been deposited in the European Nucleotide Archive under the BioProject accession number PRJEB73710. The whole-genome-sequencing data reported in this paper have been deposited in the Genome Warehouse in the National Genomics Data Center (NGDC)[136,137], Beijing Institute of Genomics, Chinese Academy of Sciences–China National Center for Bioinformation with the BioProject accession number PRJCA024131. All assemblies with annotations, variant VCF files and graph pangenome files are available at Figshare[138] (https://doi.org/10.25452/figshare.plus.25697817) and the RicePandb database (http://ricepandb.ncgr.ac.cn). The embryophyta_odb10 database, used for genome completeness assessment, was downloaded from https://busco-data.ezlab.org/v4/data/lineages/. Source data are provided with this paper.

Code availability

All of the analysis scripts used in this study are available at GitHub (https://github.com/dongling-hub/Wild-rice-Pangenome-Project) and Zenodo[139] (https://doi.org/10.5281/zenodo.14881729).

61. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
62. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **12**, 60 (2021).
63. Hu, J. et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
64. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
65. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
66. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
67. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
68. Zhang, H. et al. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat. Commun.* **12**, 6566 (2021).
69. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
70. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
71. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at *bioRxiv* https://doi.org/10.1101/254797 (2018).
72. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
73. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
74. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
75. Solovyev, V. in *Handbook of Statistical Genetics* 3rd edn (eds. Balding, J. et al.) 97–159 (Wiley, 2007).
76. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
77. Lomsadze, A. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
78. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
79. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
80. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
81. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
82. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
83. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
84. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
85. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
86. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
87. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
88. Chen, Y. et al. A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the Triticeae tribe as a pilot practice in the plant pangenomic era. *Mol. Plant* **13**, 1694–1708 (2020).
89. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
90. Li, P. et al. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).
91. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
92. Cheng, Z. et al. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**, 1691–1704 (2002).
93. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
94. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
95. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
96. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
97. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
98. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
99. Cai, X. et al. Transposable element insertion: a hidden major source of domesticated phenotypic variation in *Brassica rapa*. *Plant Biotechnol. J.* **20**, 1298–1310 (2022).
100. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
101. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
102. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
103. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
104. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).
105. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
106. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
107. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
108. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
109. Armstrong, J. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
110. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
111. Yun, T. et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2021).
112. Li, H.-T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* **5**, 461–470 (2019).
113. Bi, G., Mao, Y., Xing, Q. & Cao, M. HomBlocks: a multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genomics* **110**, 18–22 (2018).
114. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
115. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
116. 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).
117. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

118. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
119. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
120. Gimbernat-Mayol, J., Dominguez Mantes, A., Bustamante, C. D., Mas Montserrat, D. & Ioannidis, A. G. Archetypal analysis for population genetics. *PLoS Comput. Biol.* **18**, e1010301 (2022).
121. 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).
122. Schiffels, S. & Wang, K. in *Statistical Population Genomics* (ed. Dutheil, J. Y.) 147–166 (Springer, 2020).
123. Thomas, C. G. et al. Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. *Genome Res.* **25**, 667–678 (2015).
124. Durvasula, A. et al. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **114**, 5213–5218 (2017).
125. Kim, M.-S. et al. The patterns of deleterious mutations during the domestication of soybean. *Nat. Commun.* **12**, 97 (2021).
126. Meyer, R. S. et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* **48**, 1083–1088 (2016).
127. 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).
128. Wang, Y. & Obbard, D. J. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. *Evol. Lett.* **7**, 216–226 (2023).
129. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
130. Fitak, R. R. OptM: estimating the optimal number of migration edges on population trees using Treemix. *Biol. Methods Protoc.* **6**, bpab017 (2021).
131. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
132. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
133. He, Q. et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
134. Zhang, R., Jia, G. & Diao, X. geneHapR: an R package for gene haplotypic statistics and visualization. *BMC Bioinformatics* **24**, 199 (2023).
135. Van den Bossche, J. et al. geopandas/geopandas: v0.14.4. *Zenodo* https://doi.org/10.5281/zenodo.11080352 (2024).
136. Chen, M. et al. Genome Warehouse: a public repository housing genome-scale data. *Genomics Proteomics Bioinformatics* **19**, 584–589 (2021).
137. CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res.* **52**, D18–D32 (2024).
138. Zhao, Y. A pangenome reference of wild and cultivated rice. *Figshare* https://doi.org/10.25452/figshare.plus.25697817.v1 (2025).
139. Guo, D., Li, Y. & Lu, H. All codes and scripts used in the Wild Rice Pangenome Project. *Zenodo* https://doi.org/10.5281/zenodo.14881729 (2025).

Acknowledgements This work was supported by grants from the National Natural Science Foundation of China (32388201), the Strategic Priority Research Program of the Chinese Academy of Sciences (Precision Seed Design and Breeding, XDA24020205) and the Ministry of Agriculture and Rural Affairs (2022YFF1003301). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the Rice Collection of the National Institute of Genetics and Systems Functional Genetics Project of the Transdisciplinary Research Integration Center, ROIS, Japan for most of the wild rice accessions. All of the cultivated rice germplasm and the rest of the wild rice accessions were distributed from China National Rice Research Institute.

Author contributions B.H. and Q. Zhao conceived and designed the project. D.G. and Y. Li participated in the bioinformatics analyses and visualization. X.W. and N.K. contributed to the collection of rice accessions. Q. Zhao and Z.G. contributed to the selection of rice accessions. A.W., Y.W., Q. Zhan, Ziqun Wang and Zixuan Wang performed field management and phenotype investigation. D.F., C.Z., Y. Lu, Q.T., Q.W. and Q.F. performed sampling, library construction and sequencing. H.L. performed de novo genome assembly. D.G., Y. Li and Y.Z. performed genome annotation and evolutionary study. C.W. contributed to functional analyses. T.H. and L.Z. were responsible for managing the dataset. D.G. and Y. Li wrote the manuscript draft. Y. Li, Q. Zhao, X.H. and B.H. reviewed and revised the manuscript. All authors contributed to manuscript preparation and read, commented on and approved the manuscript.

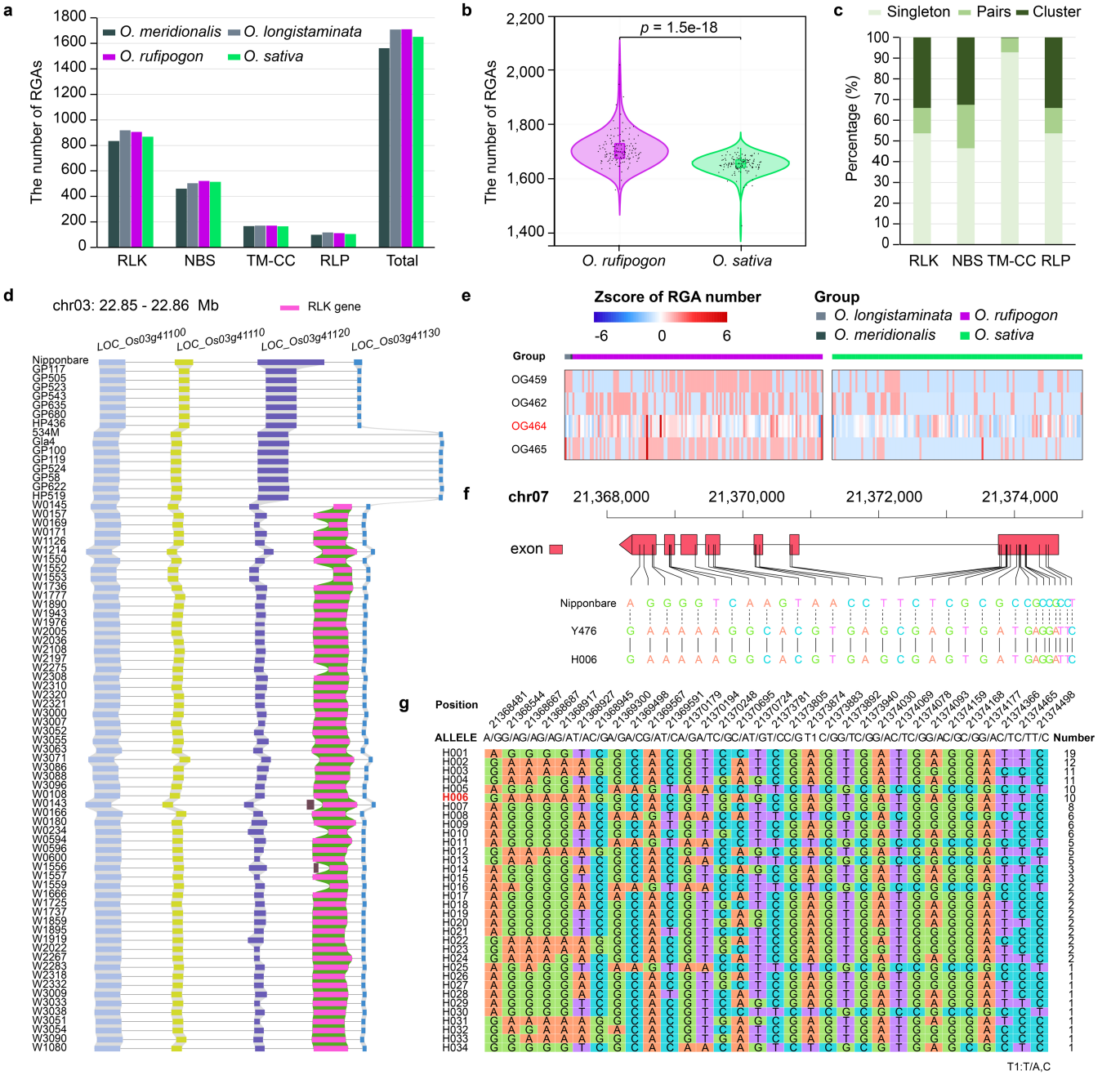
Competing interests The authors declare no competing interests.

Additional information

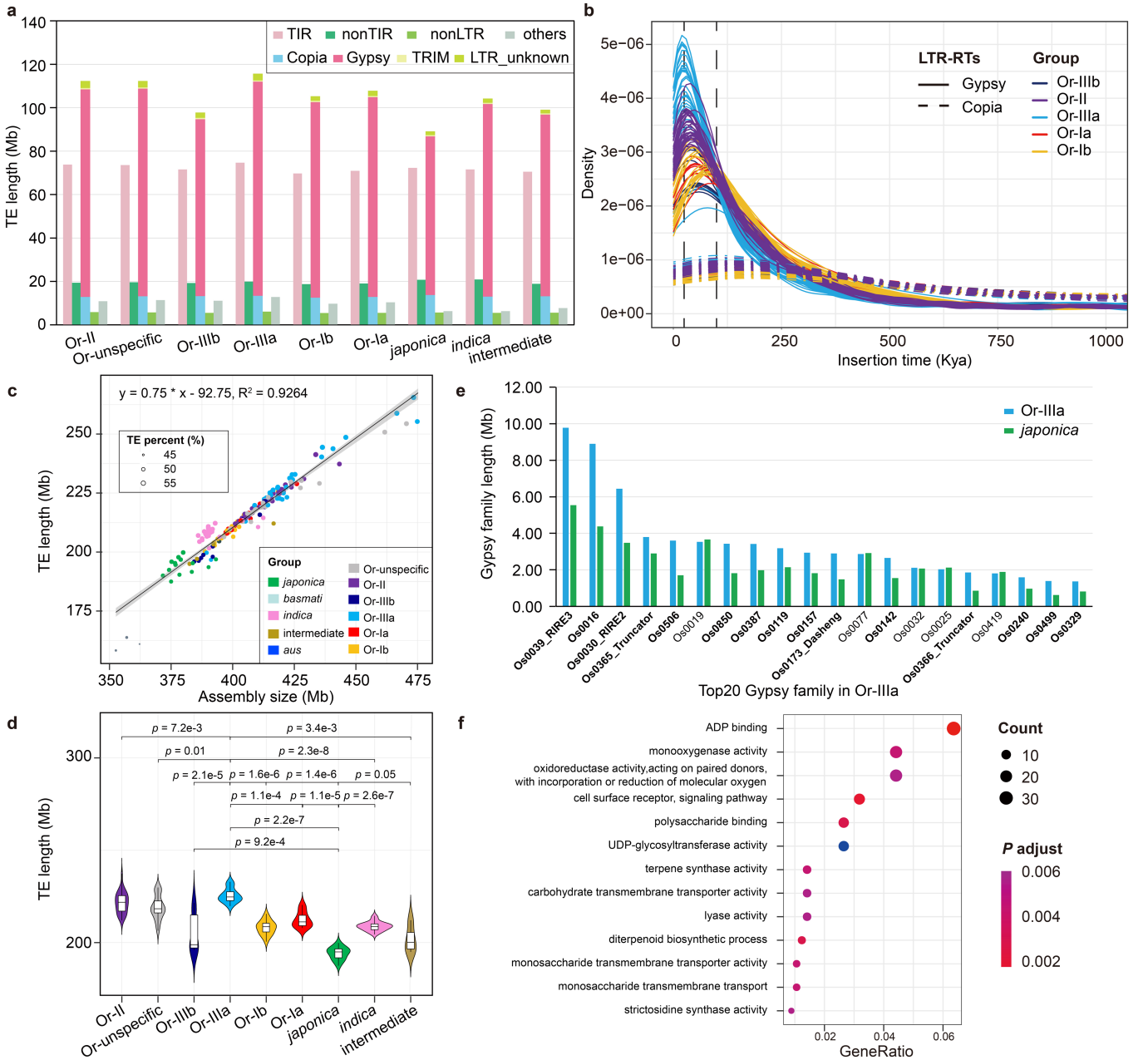
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-08883-6.

Correspondence and requests for materials should be addressed to Qiang Zhao or Bin Han.

Peer review information *Nature* thanks Shigui Li, Xianran Li, J. Chris Pires and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Characterization of RGAs in wild and cultivated rice. **a**, Numbers of each RGA type across different populations. **b**, Differences in RGA number between *O. sativa* ($n = 129$) and *O. rufipogon* ($n = 129$). The 25% and 75% quartiles are marked by the lower and upper box edges in box plots, with the median represented by the central line. Whiskers extend to 1.5 × IQR. Statistical significance was assessed using two-sided *t*-tests, with *P* values labelled above the plots. **c**, Percentage of singleton, pairs and cluster genes across RGA types. **d**, Local synteny plot in the 22.85–22.86 Mb on chromosome 3 in Nipponbare from 77 accessions. Pink boxes indicate RLK orthologues (*LOC_Os03g41130*) in each accession and green lines show their collinearity. Boxes in other colours indicate other adjacent genes and grey lines indicate their respective collinearity. **e**, The landscape of several RGA loci exhibits a higher average copy number in wild rice compared with cultivated rice. The locus containing the disease-resistance gene *LOC_Os07g35680*, designated as OG464, is marked in red. **f**, Distribution of nonsynonymous mutations between disease-resistant (the haplotype of Y476) and non-resistant (the haplotype of Nipponbare) variants in *LOC_Os07g35680*. **g**, Haplotype analysis of nonsynonymous mutations in *LOC_Os07g35680* among our wild rice population, with the disease-resistant haplotype (H006) marked in red.



Extended Data Fig. 2 | Composition of TEs in wild and cultivated rice. **a**, Length of each TE type across different populations. **b**, Distribution of insertion times for the intact Gypsy superfamily (solid lines) and Copia (dotted lines) superfamily in major groups of *O. rufipogon*. The colour of the line in the graph represents different groups of wild rice. Two vertical dashed lines represent the peak of amplification times of Gypsy (about 25,000 years ago) and Copia (about 100,000 years ago) in the Or-IIIa. **c**, Correlation between the length of TE and assembly size. Various groups are represented by coloured dots, with the dot size proportional to the percentage of TE content. A linear regression line along with a 95% confidence interval is also depicted, illustrating the relationship between these two parameters. **d**, Comparative TE lengths across major *O. rufipogon* and *O. sativa* groups. *P* values between populations using two-sided Wilcoxon tests are labelled above the plot. **e**, Length comparison of top 20 Gypsy families between Or-IIIa and *japonica*. Families with names in bold indicate those exceeding 250 kb in length in Or-IIIa as compared with *japonica*. **f**, GO enrichment analysis for genes adjacent to expanding Gypsy families in Or-IIIa. The *x* axis quantifies the gene ratio, and the *y* axis lists the GO terms. Bubble size reflects the count of genes associated with the function, and its colour indicates the significance level (the adjusted *P* values).

nature portfolio

Corresponding author(s): Bin Han

Last updated by author(s): Nov 3, 2025

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 to collect data. Sequencing platforms used to generate the raw data are listed as followed: Pacific Biosciences Sequel II, PromethION, Illumina HiSeq 4000, Illumina NovaSeq.
Data analysis	We used publicly available and appropriately cited software in the Methods. No commercial software and code were used in this study. Software are listed as follows: 1. raw reads assembly: Hifiasm (version 0.16.0), Necat (version 20200803), NextDenovo (version 2.1), Racon (version 1.0.0), NextPolish (version 1.0.2), BWA (version 0.7.17), chromap (version 0.2.6), samtools (version 1.20), YaHS (version 1.1), juice_tools (version 1.19.02), JBAT (version 1.9.8). 2. pseudo-chromosomes construction: MUMMER (version 4.0.0beta2), ALLMAPS. 3. genome assessment: BUSCO (version 5.2.2), LTR_retriever (version 2.9.0), Inspector (version 1.0.1), QUAST (version 5.0.1). 4. identification of centromeres and telomeres: ClustalW (version 2.1), HMMER (version 3.1b2), MAFFT (version 7.490), IQ-TREE (version 1.6.12). 5. gene annotation and expression levels calculation: FGENESH+ (version 3.1.1), SNAP (version 2006-07-28), GeneMark-ES (version 4.68_lic) and AUGUSTUS (version 3.3.2), GenomeThreader (version 1.7.1), HISAT2 (version 2.0.5), StringTie (version 2.0), Trinity (version 2.12.0), PASA (version 2.0.1), EVidenceModeler (version 1.1.1), InterProScan (version 5.56-89.0), fastp (version 0.23.0), Salmon (version 1.6.0). 6. identification of allelic genes and differential expression allelic genes:GeneTribe(version 1.2.0),BLASTN (version 2.9.0+), GMAP (version 2021-05-27) 7. identification of resistance gene analogs: RGAugury, MCscan (Python version). 8. annotation of transposable elements: EDTA (version 2.1.0), panEDTA, RepeatMasker (version 4.1.2), LTR_retriever (version 2.9.0), R package clusterProfiler (version 4.6.2).

9. SV calling: pbmm2 (version 1.4.0), PBSV (version 2.6.2), minimap2 (version 2.21-r1071), CuteSV (version 1.0.11), SVIM-asm (version 1.0.7), SURVIVOR (version 1.0.6), SyRI (version 1.4).
10. SNP calling: longshot (version 0.4.1), minimap2 (version 2.21-r1071), MUMMER package (version 4.0.0beta2), GATK (version 4.1.4.0), SnpEff (version 55.0).
10. pan-genome construction: BLASTN (version 2.2.18), minigraph (version 0.19-r551), Cactus (version 2.2.1), vg giraffe (version 1.43.0), DeepVariants (version 1.6.1), GLnexus (version 1.4.1-0-g68e25e5).
11. evolutionary analysis: HomBlocks, DIAMOND (version 2.0.15), OrthoFinder (version 2.5.4), IQ-TREE (version 2.2.0.3), PLINK (version 1.90b6.9 64-bit), EMMAX (version beta-07Mar2010), PHYLIP (version 3.66), iTOL, VCFTools (version 0.1.16), archetypal-analysis, PopLDdecay (version 3.42), R (version 4.1.3), Python (veresion 3.8.10), BEDTools (version 2.30.0), RectChr (version 1.36).
12. demographic history inference: MSMC2 (version 2.1.4), seqbility (version 20091110).
13. gene flow analysis: TreeMix (version 1.13), R package OptM (version 0.1.6), AdmixTools (version 7.0.2), genomics_general toolkit.
14. haplotype analysis: GMAP (version 2021-05-27), MAFFT (version 7.490), R package geneHapR (version 1.1.9).
15. Distribution map: GeoPandas (version 0.14.4), numpy (version 1.26.4), pandas (version 2.2.2), matplotlib (3.8.4)
16. All the analysis scripts used in this study are available at Github (https://github.com/dongling-hub/Wild-rice-Pangenome-Project) and Zenodo (https://doi.org/10.5281/zenodo.14881729) repository.

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 raw sequencing data, transcriptome data and Hi-C data generated in this study have been deposited in the European Nucleotide Archive (ENA) under the BioProject accession number PRJEB73710. The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center (NGDC), Beijing Institute of Genomics, Chinese Academy of Sciences/ China National Center for Bioinformation with the BioProject accession number PRJCA024131. All assemblies with annotations, variant VCF files and graph pangenome files are available at Figshare (https://doi.org/10.25452/figshare.plus.25697817) and the RicePandb database (http://ricepandb.ncgr.ac.cn). The embryophyta_odb10 database, used for genome completeness assessment, was download from the https://busco-data.ezlab.org/v4/data/lineages/. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	not applicable
Reporting on race, ethnicity, or other socially relevant groupings	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

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	PacBio HiFi: 133 samples; Nanopore: 16 samples; Hi-C: 30 samples; Illumina: 696 samples. We built the pan-genome based on 145 representative accessions in our study. We selected this number of populations because they sufficiently cover the genetic and geographic distribution of both Oryza rufipogon and Oryza sativa. These materials are evenly distributed within an evolutionary tree that includes a larger set of approximately 1500 samples.

Data exclusions	No data was excluded.
Replication	The study primarily focuses on genome assembly and evolutionary analyses, relying on genetic data and computational models rather than experimental protocols that would necessitate replication.
Randomization	A randomized complete block design was used in planting.
Blinding	All accessions were only labeled by numbers when planting and data collection.

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
☐	☒ Plants

Methods

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

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☐	☒	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Hazards	*Please describe the agents/technologies/information that may pose a threat, including any agents subject to oversight for dual use research of concern.*

For examples of agents subject to oversight, see the United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern.

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Precautions and benefits

Biosecurity precautions	*Describe the precautions that were taken during the design and conduct of this research, or will be required in the communication and application of the research, to minimise biosecurity risks. These may include bio-containment facilities, changes to the study design/ methodology or redaction of details from the manuscript.*
Biosecurity oversight	*Describe any evaluations and oversight of biosecurity risks of this work that you have received from people or organizations outside of your immediate team.*
Benefits	*Describe the benefits that application or use of this work could bring, including benefits that may mitigate risks to public health, national security, or the health of crops, livestock or the environment.*
Communication benefits	*Describe whether the benefits of communicating this information outweigh the risks, and if so, how.*

Plants

Seed stocks	We selected a total of 149 rice varieties according to the phylogenetic relationships and geographic distribution from a previously reported population, which were cultivated in Lingshui County, Hainan.
Novel plant genotypes	We only collected wild accessions and cultivars in this study. No novel plant were used.
Authentication	All samples were from the China National Rice Research Institute in Hangzhou, China, and the National Institute of Genetics in Mishima, Japan.