nature communications

Article

https://doi.org/10.1038/s41467-024-54427-3

Genome evolution and diversity of wild and cultivated rice species

Received: 6 May 2024

Accepted: 11 November 2024

Published online: 18 November 2024

Check for updates

Weixiong Long[1,4] ✉, Qiang He[2,4], Yitao Wang[2,4], Yu Wang[2], Jie Wang[1], Zhengqing Yuan[3], Meijia Wang[2], Wei Chen[1], Lihua Luo[1], Laiyang Luo[1], Weibiao Xu[1], Yonghui Li[1], Wei Li[2], Longan Yan[1], Yaohui Cai[1] ✉, Huilong Du[2] ✉ & Hongwei Xie[1] ✉

Wild species of crops serve as a valuable germplasm resource for breeding of modern cultivars. Rice (*Oryza sativa* L.) is a vital global staple food. However, research on genome evolution and diversity of wild rice species remains limited. Here, we present nearly complete genomes of 13 representative wild rice species. By integrating with four previously published genomes for pangenome analysis, a total of 101,723 gene families are identified across the genus, including 9834 (9.67%) core gene families. Additionally, 63,881 gene families absent in cultivated rice species but present in wild rice species are discovered. Extensive structural rearrangements, sub-genomes exchanges, widespread allelic variations, and regulatory sequence variations are observed in wild rice species. Interestingly, expanded but less diverse disease resistance genes in the genomes of cultivated rice, likely due to the loss of some resistance genes and the fixing and amplification of genes encoding resistance genes to specific diseases during domestication and artificial selection. This study not only reveals natural variations valuable for gene-level studies and breeding selection but also enhances our understanding on rice evolution and domestication.

Rice (*Oryza sativa* L.) is a crucial crop globally and serves as a key model species for monocot and crop plant research[1]. Rice production will need to double by 2050 in order to meet the demand of an increasing world population[2]. To overcome the current production bottleneck that has stunted rice yields, the presence and absence of genes and interspecies allelic diversity should be exploited[3,4].

The *Oryza* genus comprises two cultivated rice species: Asian and African rice, along with 20 extant wild rice species. These wild rice are currently divided into 11 genome types: AA, BB, CC, EE, FF, GG, BBCC, CCDD, HHJJ, HHKK, and KKLL. The different genome-type rice species exhibit significant genetic and phenotypic diversity, which enables them to adapt to various ecological environments across Asia, Africa, the Americas, and Australia[5,6].

The availability of high-quality genome assemblies of cultivated rice has allowed for a thorough characterization of structural variation through comparative genomics analysis[7,8]. Various studies have contributed to the construction of pangenomes that integrate cultivars and the AA genome of wild rice[9,10]. A total of 20,045 gene families were generated from 111 cultivated rice, with 13,227 (65.7%) being core gene families[11]. The pangenome of 33 cultivated rice comprised 66,636 genes with 20,374 (30.57%) core genes[12]. While another pangenome of 251-accession comprising both cultivated and AA genome wild rice,

[1]Jiangxi Super -rice Research and Development Center, Jiangxi Provincial Key Laboratory of Rice Germplasm Innovation and Breeding, National Engineering Research Center for Rice, Jiangxi Academy of Agricultural Sciences, Nanchang, China. [2]School of Life Sciences, Institute of Life Sciences and Green Development, Hebei Basic Science Center for Biotic Interaction, Hebei University, Baoding, China. [3]College of Life Sciences, Wuhan University, Wuhan, China. [4]These authors contributed equally: Weixiong Long, Qiang He, Yitao Wang. ✉e-mail: longweixiong1219@163.com; caiyaohui@126.com; huilongdu@hbu.edu.cn; xhw206@jxaas.cn

51,359 non-redundant genes were identified with 21,888 (42.61%) core genes[10]. Currently, there is a lack of high-quality reference genomes for the majority of rice wild relatives and a pangenome of the rice genus. To date, a small number of non-A type *Oryza* genomes have been published to the chromosome level. *O. granulata* was assembled to the chromosome level using gradient resequencing and PacBio SMRT data. Analyses of the assembly revealed that positively selected genes related to photosynthesis and energy production may have facilitated the *O. granulata* tolerance to shade[13]. By using high fidelity (HiFi) and HiC data, the allotetraploid *O. coarctata* was reported to have a genome size of 573.4 Mb and an N50 of 23.1 Mb[14], Yu assembled a high-quality reference genome of *O. alta*, which enabled gene editing of important agricultural traits determining genes and the de novo domestication of allotetraploid wild rice[15].

In this work, we de novo assemble genomes of 13 wild rice accessions (Supplementary Table 1), along with two cultivated rice subspecies, *O. sativa* (cv. Nipponbare, referred to as NIP, representing *O. sativa* ssp. *japonica* and R498 of *O. sativa* ssp. *indica*)[16,17], *O. glaberrima*[18], and one common wild rice *O. rufipogon*[19]. We construct a non-redundant gene set for the *Oryza* genus and reconstruct the phylogenetic tree of this group. The expansion and contraction of long terminal repeat (LTR) subfamilies influence genome size variation within the *Oryza* genus. We identify satellite repeats and analyze the location, size, and repositioning of centromeres in rice. In addition, the assembled wild rice genomes facilitate the identification of candidate genes and genomic variations in rice. Furthermore, we identify genome-wide allelic variations based on collinear regions compared to the NIP[16] reference genome and analyze gene expression affected by structural variants (SVs) or copy number variations (CNVs). Pangenomic analyses assist in the identification of wild rice-specific resistance (*R*) genes, which hold the potential to breed disease-resistant rice cultivars. The *Oryza* genus pangenome provides breeders with a resource to improve cultivated rice and meet future food demands.

Results

High-quality assemblies of 13 representative wild rice species

We selected 13 representative wild rice species, comprising six allotetraploids and seven diploids, for de novo genome assembly. These accessions exhibit significant variation in geographical distribution and phenotype, as illustrated in Fig. 1a. We surveyed genome size and heterozygosity by sequencing to an average of 100-fold coverage at 150 bp insert size for each species using Novaseq 6000. The estimated heterozygosity ranged from 0.07% for *O. branchyantha* to 0.59% for *O. officinalis* in diploid wild rice and from 0.13% in *O. punctata* (BBCC) to 1.04% in *O. alta* in tetraploid rice (Supplementary Table 2 and Supplementary Fig. 1). Consequently, a total of 331.3 Gb of high fidelity (HiFi) reads were generated for the 13 wild rice accessions, representing approximately 60-fold coverage relative to the NIP[16] genome size of around 400 Mb (Supplementary Table 3). Through the integration of high-throughput chromosome conformation capture[20,21], all 13 wild rice accessions were assembled to a near-complete level, resulting in genome size ranging from 276.30 Mb to 1,059.66 Mb with an average N50 contig size ranging from 21.56 Mb to 45.47 Mb (Table 1 and Supplementary Figs. 1, 2). The mapping of Illumina short reads and HiFi reads to the 13 wild rice genome assemblies revealed high percentages of alignment, at more than 99.88% and 99.95%, respectively (Supplementary Fig. 3). All samples exhibited values above 95% towards the 1614 Benchmarking Universal Single-Copy Orthologs (BUSCO) identified in these assemblies, with an average BUSCO score of 95.66% and 99.25% in diploid and tetraploid rice, respectively. A low BUSCO value was observed in the subgenome, ranging from 89.7% for *O. grandiglumis* | D_t to 98.1% for *O. latifolia* | C_t[22] (Supplementary Table 4). Furthermore, an average LTR assembly index (LAI) of 23.64 was calculated for all assemblies[23], in addition to high consensus quality values, indicative of their superior quality, continuity, and completeness (Table 1 and Supplementary Table 5).

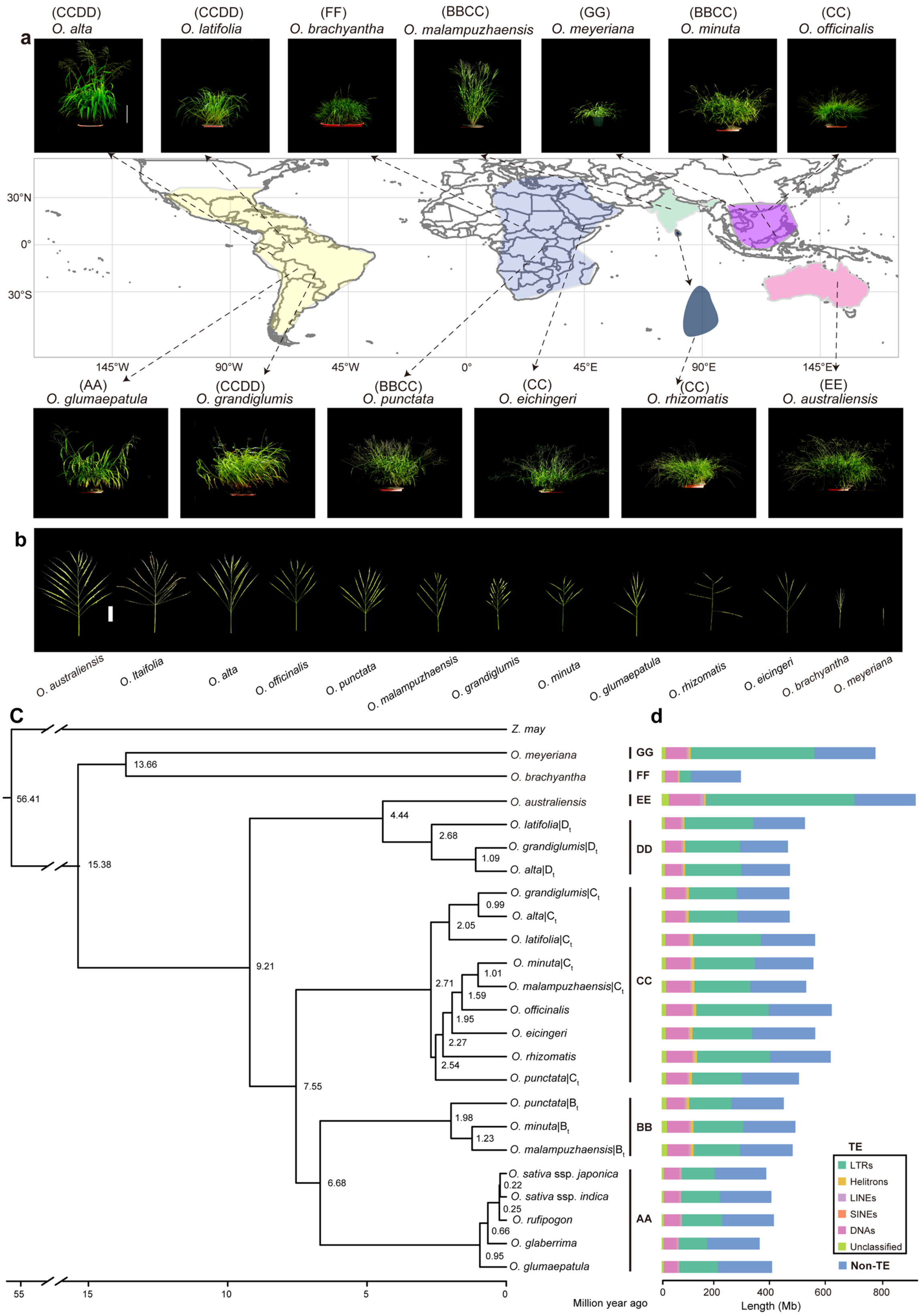
Gene structure annotation was performed using a combination of de novo, homologous, and transcript prediction methods based on the repeat-masked genome. The 13 wild rice protein-coding gene numbers exhibited a 2.28-fold variation, varying from 37,711 variants in *O. branchyatha* to 85,846 in *O. minuta* (Table 1). The transcript data supported a high percentage (ranging from 71.1% to 87.7%) of the predicted protein-coding genes, indicating the quality of the gene annotations (Supplementary Data 1). A phylogenetic tree based on 3,555 single-copy orthologs grouped the 17 rice species into 5 clusters-A, B, C, D/E, and F/G[24] (Fig. 1c). In addition, an Maximum Likelihood tree using chloroplast data (13 chloroplast genomes in this study and 5 chloroplast genomes from GenBank) constructed to trace the maternal progenitors of allopolyploid wild rice, revealed *O. eichingeri* as the female parent of *O. punctata* (BBCC) (Supplementary Fig. 4). Transposable elements (TEs) play a significant role in shaping large plant genomes and driving genome evolution through periodic bursts of amplification (Supplementary Fig. 5). The TE content in the 13 wild rice genomes ranged from 35.11% (*O. branchyantha*) to 76.35% (*O. australiensis*), with long terminal repeat retrotransposons (LTR-RTs) being the most abundant (Fig. 1d and Supplementary Data 2).

Unlocking untapped genes and hidden genomic variations using the pangenome of the *Oryza* genus

The rice pangenome was expanded to include 16 species (17 subspecies) in the *Oryza* genus, incorporating genomes from three AA-genome *Oryza* species. A pangenome, constructed using OrthoFinder analysis[25], clustered 808,478 predicted gene models from 13 wild rice species, along with three previously published AA genotype rice genomes and a reference genome of *O. sativa* (NIP), resulting in a pangenome cluster of 101,723 predicted gene family models (Fig. 2a, Supplementary Fig. 6 and Supplementary Table 6). A total of 9.66% of gene families (9834) shared across all 17 rice accessions, were classified as core gene families. Dispensable gene families, present in 2–15 individuals, made up 56.84% (57,822) of the *Oryza* genus pangenome, while 33.48% (34,067) were identified as species-specific gene families (Fig. 2a, b). Compared with cultivated rice, the pangenome constructed in this study provides an additional 63,881 gene families.

The 17 rice accessions, comprising 11 diploids and 6 allotetraploids, were categorized into diploid genome types labeled A–G (Supplementary Fig. 6). In addition, a syntenic pangenome was constructed to analyze variations among the 7 diploid genomes, which revealed that core gene families shared across these genomes accounted for 17.37% of the total gene sets. Dispensable families accounted for 29.73% of the total gene sets (Supplementary Fig. 6 and Supplementary Table 7), with the largest proportion represented by private gene sets, unique to individual genome types, making up 52.90% of the total gene sets (Supplementary Table 7).

Within the *Oryza* genus, 81.14% of the core genes could be assigned to protein domains available in the Pfam and InterPro databases. This number was almost double the percentage of dispensable genes (41.82%) and more than seven times higher than accession-specific genes (10.83%) (Supplementary Fig. 7). An average of 92.40% genes were expressed in the core genome, which is significantly higher than the 63.31% of genes expressed in the dispensable genome and 52.15% in the private genome (Fig. 2d). Core genes exhibited 6- to 20-fold higher expression levels compared to shell and private genes (Fig. 2e), and this tendency was also reflected in gene length. Core genes showed significantly lower (0.15-fold on average) pairwise non-synonymous substitution/synonymous substitution ratios (Ka/Ks) compared to the dispensable genes (Fig. 2f, g), indicating conservation of function among core genes in the *Oryza* genus, while variable genes evolved more rapidly. The average LTR insertion number per core gene accounted for 80.98% and 48.89% for shell- and species-specific



genes, respectively (Fig. 2i). This was likely due to the lower number of exons per gene in accession-specific genes compared to dispensable and core genes (Supplementary Fig. 8), suggesting that exon shuffling or loss contributes to the specificity of genes of a species.

Intriguingly, the core genes in the *Oryza* genus are primarily involved in fundamental functions such as transposition, iron ion binding, transport, and electron transport, indicating their role in maintaining essential activities of the *Oryza* genus (Fig. 2h). Additional genomic analysis of the different genome type private genes in rice was conducted to understand how they differ in term of function. The private genes of the A genome wild rice were enriched for traits associated with resistance to heavy metals and salt. In contrast,

Fig. 1 | Geographical distribution and phylogeny of wild and cultivated rice. **a** Geographic distribution of the 13 wild rice varieties and their diverse agronomic characteristics, such as plant height. The font colors of wild rice species correspond to the distribution on the world map. These 13 accessions covered 13 species in the rice genus. **b** Panicle architecture of the 13 wild rice species. **c** Phylogenetic tree based on the conserved single-copy gene illustrates evolutionary history in the *Oryza* genus. All the allotetraploid wild rice genome was split into subgenomes, the ASTRAL concatenation-based species tree estimated by 3555 single copy genes generated by Orthofinder. Numbers at nodes indicate the median value for the divergence time (Mya) estimates for each clade. **d** TE content in each subgenome/genome of the wild and cultivated rice. RNA transposons were presented as follows: blue for LTR, yellow represents Helitron, pink indicates LINEs, orange shows SINEs, and orange, blue stands for unclassified. DNA transposons were shown as wine. Source data are provided as a Source Data file.

Table 1 | Statics of the assembly and annotation of 13 wild rice species

Species	Genome type	Assembly size (Mb)	Contig N50 (Mb)	Anchoring rate (%)	LAI	Gene number	Repeat region % of assembly
O. glumaepatula	AA	393.8	14.18	99.48	27.94	44,527	53.48%
O. punctata	BBCC	925.4	32.88	99.92	21.79	85,475	57.48%
O. minuta	BBCC	1018.0	40.18	99.96	19.64	85,846	60.97%
O. malampuzhaensis	BBCC	982.3	32.72	99.93	19.18	85,302	60.38%
O. eichingeri	CC	547.0	49.78	99.58	23.06	44,719	59.48%
O. officinalis	CC	605.7	42.85	98.12	28.99	44,175	64.24%
O. rhizomatis	CC	602.0	45.97	99.85	24.13	44,596	65.70%
O.latifolia	CCDD	1059.0	45.47	99.88	24.51	85,302	64.02%
O. grandigumis	CCDD	906.1	21.73	99.87	21.87	81,960	60.12%
O. alta	CCDD	920.3	29.38	99.27	24.25	83,441	60.52%
O. australiensis	EE	903.3	71.18	99.93	26.39	43,435	76.35%
O. branchyantha	FF	276.3	21.56	99.32	19.92	37,711	35.71%
O. meyeriana	GG	761.1	30.46	97.56	25.67	41,989	73.04%

the B genome wild rice carried traits that enable the plant to respond to responses to oxidative stress. The C genome type wild rice private genes were significantly enriched in the regulation of salicylic acid and steroid hormones, contributing to their high resistance to various diseases. The D genome type wild rice private genes were significantly associated with arginine synthase, which enables the plant to sense and adapt to environmental changes. The E genome type wild rice private genes were significantly involved in cellulose synthase, which benefits large biomass production. The F genome type private genes were significantly associated with defense responses. Meanwhile, the G genome type rice-specific genes focused on electron transfer, which contributes to shade tolerance (Supplementary Fig. 9). The *Oryza* genus pangenome opens the door for non-AA genome wild rice resources to understand rice biology and improve breeding.

Transposons contribute to genomic variation in *Oryza*

The selective removal and retention of TEs significantly influences genome size, adaptation, and evolution of the *Oryza* genus[26]. Nonetheless, the specific TE subfamilies that influence *Oryza* genome size remain unresolved[27]. To address this, a classification of TEs within the *Oryza* genus was conducted, along with an in-depth analysis of the expansion profiles of LTR subfamilies (top 6 selected from the largest *Oryza* genome). Each LTR-RT sub-family displayed unique patterns of amplification across the species, impacting genome sizes (Fig. 3, Supplementary Fig. 10, and Supplementary Data 3). In addition to the amplification of the Ogre, Retand, and Tekay LTR of the Gypsy superfamily, the Angle LTR, which belongs to the Copia superfamily, has emerged as a significant contributor to the genome size of *O. australiensis* (EE), each accounting for 9.80%, 7.89%, 8.22%, and 6.64%., respectively This distinguishes *O. australiensis* from other rice species, whose genome size is primarily driven by the Gypsy superfamily (Fig. 3b, c and Supplementary Fig. 10). The genome sizes of B, C, and D type rice genomes were primarily influenced by the top three Gypsy subfamilies in descending order on average: Ogre, Retand and Tekay accounted for 7.49%, 3.15%, and 1.94% of the genome size, respectively (Fig. 3c). The genome size of the G genome type (*O. meyeriana*) was predominantly influenced by Retand LTR amplification, which accounted for 8.66% of the genome size (Fig. 3c). In contrast, the 10.17% abundance of Tekay in *O. glumaepatula* surpassed the 2.78% observed in *O. sativa*, contributing to an increase in genome size. All LTR subfamilies occupied only 4.05% of the genome size in *O. brachyantha* (FF), which is less than the 7.82% found in *O. sativa*, indicating a significant reduction of all LTR subfamilies (Fig. 3c). This finding aligns with the observed LTR density patterns across *Oryza* genomes (Supplementary Fig. 11). A correlation value of 0.86 was noted between genome size and the Retand LTR superfamily, which was higher than that of other subfamilies within *Oryza* (Fig. 3d). While the SIRE and CRM subfamilies occupied lengths of 3.41 Mb and 3.38 Mb, respectively, they corresponded to an average of 250 and 295 genes, respectively (Fig. 3c, Supplementary Fig. 10 and Supplementary Data 4).

The distribution of whole-genome intact LTRs indicated that the majority of LTR bursts occurred in the proximity of centromeric regions (Supplementary Fig. 12). Rice centromeres consist of organized satellite repeats (SRs), interrupted by centromere-specific retrotransposons (CRRs)[28]. It remains unclear if other lines or wild rice plants possess unique centromere satellites and whether centromere repositioning events occurred during centromere evolution. In cultivated rice (MH63 and ZS97), 155 bp and 165 bp CentO satellite repeats were categorized into seven distinct subsets across the 12 chromosomes[29]. Interestingly, only a few copies of satellite repeats were identified in *O. meyeriana* and *O. branchyantha*. Compared to the cultivated rice (NIP) genome, the C genome contains centromere-specific satellite repeats of 126 bp and 366 bp (Supplementary Data 5). Phylogenetic analysis results showed that the satellite repeats in *Oryza* can be classified into four groups (Fig. 3e), with chromosomes of the same type tending to cluster together, supporting models of repeated amplification events involving the central domain and local homogenization. Examination of the genetic characteristics of centromeres in the *Oryza* genus revealed a decrease in gene density closer to the centromere region, along with an increase in transposon density and the frequency of k-mers (Fig. 3f). While wild rice centromere sizes, with the exception of *O. brathyantha*, were significantly larger than those in cultivated rice, the opposite trend was observed in term of the number of genes (Fig. 3g, Supplementary Fig. 11 and Supplementary Data 6). A comparative sequence map of

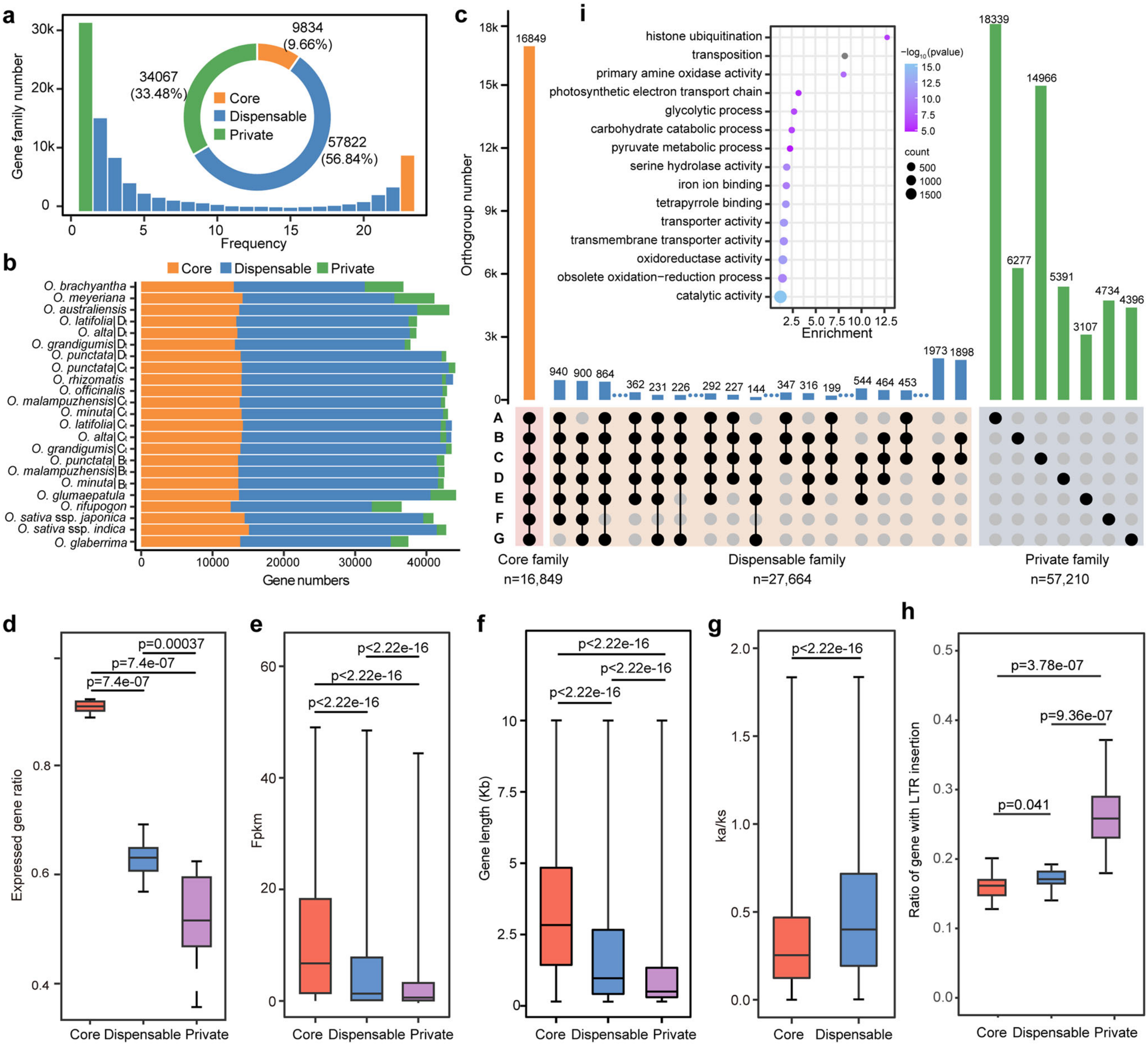
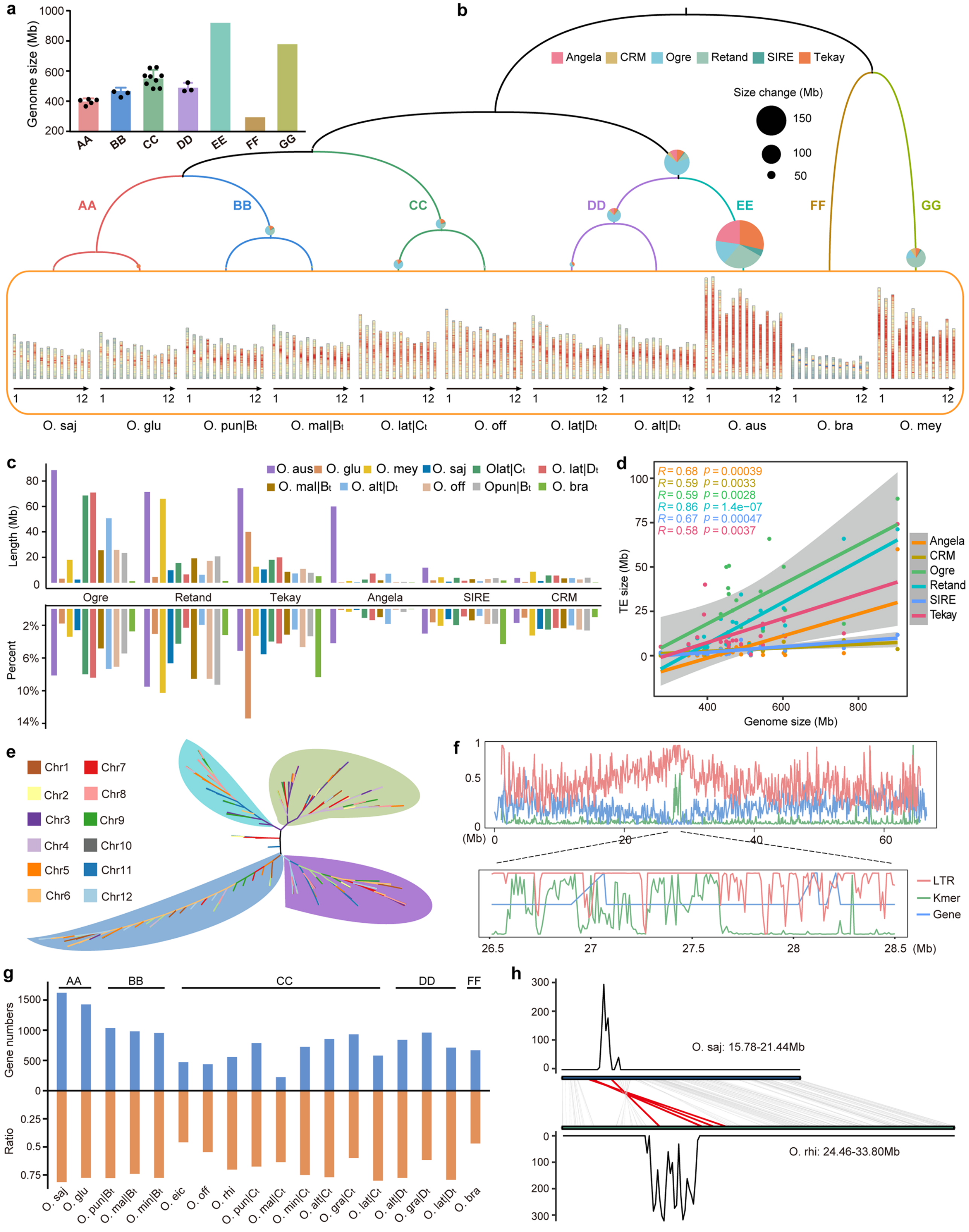


Fig. 2 | Pangenome of rice genus. a Compositions of the *Oryza* pangenome. The histogram displays the distribution of gene families in 23 subgenomes. The pie chart shows the proportion of gene families labeled by a component of pangenome, yellow: core gene families, blue: dispensable gene family, and green: private gene family. **b** Gene number of core gene families, dispensable gene families, and private gene families in each (sub)genome. **c** The upper diagram of the orthologous groups among the A, B, C, D, E, F, and G type group genomes. **d** The expression gene ratio in core, dispensable, and private genomes. **e**, **f** Gene expression and length in core, dispensable, and private genome. **g** Ka/Ks in core and dispensable genes. **h** The ratio of genes with LTR-RTs insertion in core, dispensable, and private genes. In **c**–**h**, *P*-values were calculated using two-tailed Student's *t*-tests. The middle bars show the median, and the bottle and top of each box indicate the 1/4 and 3/4 percentiles, respectively. the whiskers extend to 1.5 times the interquartile range. Sample sizes (n) represent the samples used for (**d**, **h**): *n* = 23, *n* = gene set numbers in (**e**, **f**, and **g**). **i** GO enrichment of the core genome. Significance was tested by a two-tailed Fisher's exact method. Source data are provided as a Source Data file.

centromere synteny between cultivated rice and wild rice highlighted extensive structural rearrangements in centromeric and pericentric regions across the *Oryza* genus (Fig. 3h and Supplementary Data 7). In addition, several centromere repositioning events were noted in the synteny analysis (Supplementary Fig. 13).

Large-scale chromosomal rearrangements and evolutionary history of *Oryza* genomes

To compare karyotype stability between cultivated rice and wild diploid genomes, we created a synteny map and conducted whole-genome pairwise alignments to identify large segment variations, such as translocations and inversions[5]. The large inversions (with more than 5 consecutive genes) shared by at least two species in the *Oryza* genus were prevalent in the genome alignments[8] (Supplementary Figs. 14 and 15). While reports on segregating inversions in wild rice are scarce, with none on natural polyploid wild rice, most of these events were observed in the low-recombining pericentromeric regions of the *Oryza* chromosomes, with a few inversions being species-specific (Supplementary Fig. 14). Through multiple species/genome comparisons, many large-scale genomic rearrangements were validated, including an inversion of an approximately 2.53 Mb segment comprising 166 genes on Chr6 in the modern cultivated rice NIP (Supplementary Fig. 16). The inversion occurred in all the rice species except for common wild rice.



Chromosomal rearrangements involving homoeologous groups 1, 3, and 6 of allotetraploid wild rice were initially identified in *Oryza* species (Supplementary Fig. 17a). Comparison between syntenic blocks revealed that chromosome $6D_t$ in *O. latifolia* displayed complete collinearity with the corresponding chromosome $3D_t$ in *O. alta* and *O. grandigumis*, However, a reciprocal translocation was observed between $3D_t$ and $1C_t$ in *O. alta* and *O. grandigumis* (Supplementary Fig. 17a), due to fragmental collinearity between $3C_t$ in *O. latifolia* and $1C_t$ in *O. alta* and *O. grandigumis*. Additionally, a translocation between $1C_t$ and $6C_t$ was identified, with the $1C_t$ segment translocated to the end of $6C_t$. Furthermore, a translocation was detected in homoeologous groups 4 and 7 in *Oryza* (Supplementary Fig. 17c–i).

Fig. 3 | Evolution of LTR Retrotransposons in genomes and centromeres. **a** Genome size of the *Oryza* genomes, including the A, B, C, D, E, F, and G genome types. The data was shown as the mean values ± SDs, *n* = sub/genome numbers. **b** LTR-RT subfamily expansion and contraction during the rice evolution, and the heatmap of TE density in wild rice species. Circle size indicates the length of the TE sequence. Color represents the LTR sub-family, light pink: Angela, orange: CRM, light blue: Ogre, light green: Retand, green: SIRE, light orange: Tekay. The TE density from low to high corresponds to the color from blue to red. **c** The size of LTR sub-families and the percentage of genes overlapped with LTR-RTs subfamilies in each rice species. **d** The correlation between genome size and LTR subfamily sequence length. Two-sides Pearson's correlations test with *P*-value = 0 < 0.001 was performed. **e** The cluster tree was generated based on the multiple sequence-structure alignment of centromeric repeat unit sequences. **f** Landscape of LTR density, repeat k-mer ratio, and gene number density in chromosome 1, 100 kb, and 10 kb were used as a window for the picture above and below. **g** The gene number in centromere region and the number of gene overlapped with LTR region account for the total gene number in A, B, C, D, E, and F genome. **h** Synteny map of centromere region of chromosome 1 between *O. rhizamatis* and *O. sativa* ssp. *japonica*. The red line indicates the collinearity of genes at both ends of the centromere. Source data are provided as a Source Data file.

By aligning allotetraploid wild rice resequencing data to the corresponding diploid BB and CC or CC and EE genomes, homeologous exchanges on each chromosome were identified based on the coverage depth calculated from unique reads. Several translocations of large segments between the subgenomes post-tetraploidization were discovered (Supplementary Fig. 17b, c). Chromosomes $1B_t$ and $1C_t$ exhibited high synteny, but the coverage depth of the reads to the BB and CC genomes indicated a significant homoeologous exchange between them (Supplementary Fig. 17b). The prospective history of homeologous exchange is depicted in Supplementary Fig. 17d.

Characterization of untapped, agronomically important SVs in the *Oryza* genomes

Despite extensive efforts to analyze genetic variants in cultivated rice and its ancestor species *O. rufipogon*[30], genetic diversity in distantly related wild rice species such as *O. punctata*, *O. rhizomatis*, and *O. meyeriana* remains poorly understood. We identified 2781-10,656 insertions, 2680-10,419 deletions, 4-52 translocations, and 7-22 inversions in the 16 rice accessions, with sizes ranging from 162.49-278.65 Mb, 182.13-705.17 Mb, 8.64-887.29 kb, and 41.51-11.33 Mb, respectively (Fig. 4a and Supplementary Data 8). Interestingly, the cultivated rice and the AA genome of wild rice showed a higher number of structural variations (SVs) compared to the non-AA genome of wild rice, although the size of the variations was smaller, likely due to more regions aligning with the reference genome (Fig. 4a). Wild rice species-specific SVs accounted for a significant portion of the total variation, indicating untapped genetic diversity in wild rice compared to cultivated varieties (Supplementary Fig. 18).

A total of 78.48% of SVs in cultivars and 39.10% in wild rice were shorter than 5 kb. As the length of SVs increased, there was a significant decrease in the number of SVs observed in both cultivars and wild rice. However, the wild rice variety exhibited 2.34%more SVs that were larger than 250 kb, which contributed to the presence of numerous private genes within the wild rice genome (Fig. 4b). The most common locations for SVs were intergenic regions as noted for 46.19% of *Oryza* genus followed by 16.74% in upstream gene regions (Supplementary Fig. 18b and Supplementary Data 9), in line with previous findings[12]. A total of 128,250 insertion sequences generated 30,945 specific genes identified in our pangenome recorded within the whole genome of Nipponbare (Fig. 4c, d). The cultivated rice displayed a higher number of SVs compared to wild rice, consistent with earlier observations (Fig. 4a). However, the size of structural variation in the wild rice *O. australiensis* and *O. meyeriana*, was 614.70 Mb and 711.90 Mb, respectively, was overlapped with repeat sequences. This is significantly higher than 274.43 Mb and 258.09 Mb associated with non-repeat sequences (Supplementary Fig. 17). Further examination of transposable elements in presence-absence variation (PAV) sequences revealed that DNA and other transposable elements were the primary components of both deletion and inversion variants (Supplementary Fig. 17d).

Recent findings suggested that gene loss could be linked to insertion/deletion events. For example, a 500 kb insertion corresponding to the NIP genome identified on chromosome 12 at 14.50 Mb only occurred in the *O. eichingeri* and C_t subgenome of *O. punctata* (BBCC), which led to the presence of protein-coding genes only in a few wild rice genomes (Supplementary Fig. 18).

Allelic and regulatory elements variations

Natural allelic variation of genes is essential for phenotypic diversity, environmental adaptation, and the process of domestication[31–33]. Our analysis focused on genome-wide allelic variations and their regulatory sequences (gene ± 10 kb) in the rice genome, as there are very few highly collinear blocks between non-AA genomic wild rice and cultivated rice (Supplementary Fig. 19 and Supplementary Table 8). As the divergence from cultivated rice increased, the number of colinear genes between wild rice diploids and cultivar rice decreased, ranging from 23,812 to 18,463, with an average of 20,288 (Supplementary Table 9). The inclusion of 19 published, high-quality, chromosome-level cultivated rice genomes (Supplementary Table 9) into our study allowed us to identify SV resources for both wild and cultivated rice[12]. By mapping collinear genes with a range of 10 kb adjacent regions onto the corresponding region of Nipponbare, we identified Single Nucleotide Polymorphisms (SNPs) and InDels of 50 bp or greater as PAV targets (Supplementary Fig. 20). We identified 125 variations per gene in wild rice, which resulted in a total number of $6*10^7$ variations in the wild rice pangenome. This number is significantly higher than that observed in cultivated rice, which had 7.28 variations per gene and a total of $5*10^6$ variations in the cultivar pangenome (Fig. 5a, b). To delve deeper into the functional impact of SVs on genes and proteins, combining variant alleles detected in each species into haplotypes and annotating each accession independently is essential. Wild rice (sub) genome displayed a higher number of alleles in collinear genes within the core genome compared to cultivated rice (Fig. 5c). The number of gene haplotypes (gHap) and gene-coding sequence haplotypes (gcHap) in wild rice was significantly greater than in cultivated rice (Fig. 5c). Analyses of protein diversity in collinear genes between wild and cultivated rice provided insights into their functional differentiation. A genome-wide protein cluster analysis was conducted based on domain similarity, revealing that wild rice contained an average of 7.71 clusters, while cultivated rice exhibited only 1.78 clusters (Fig. 5d). Furthermore, analysis of gene PAVs distinguished major species and highlighted significant differences between wild and cultivated rice (Supplementary Fig. 7b–d). The majority of group-unbalanced genes, accounting for 87.33%, were more prevalent in wild rice but less common in cultivated rice, underscoring the substantial legacy of mutations in wild rice (Fig. 5e). A well-documented gene that controls pericarp color (Rc) was selected as an example to display protein variation of protein between wild and cultivated rice (Fig. 5f).

Gene CNVs and *NLR* repertoire in rice

Recent studies have highlighted the significant role of copy number variations (gCNVs) in the evolution and domestication of crops[34,35]. However, the accurate identification of gCNVs in highly repetitive genome sequences within the rice genus is challenging. Leveraging our high-quality assemblies, we systematically investigated gCNVs by

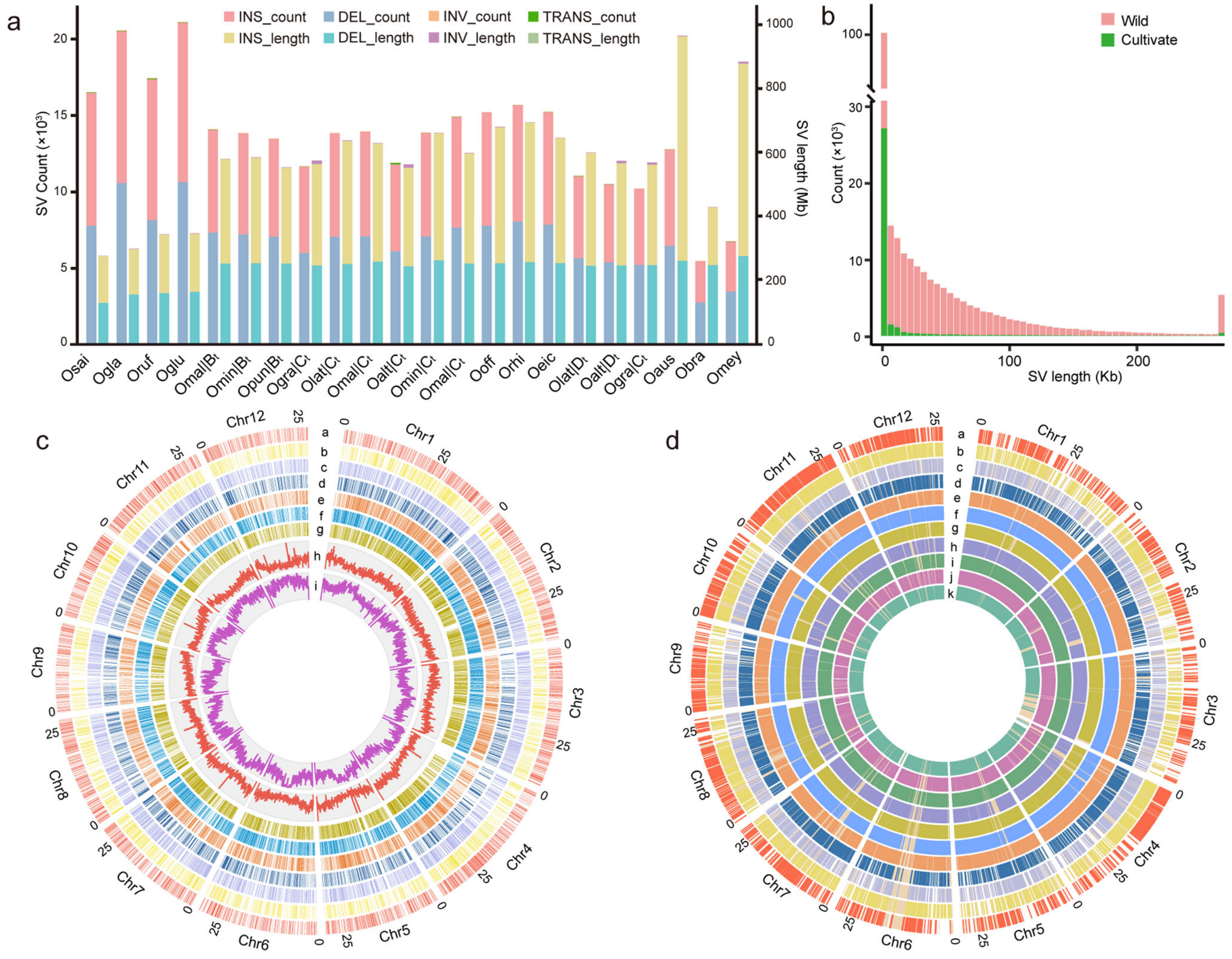


Fig. 4 | The landscape of structural variation among wild and cultivated rice. **a** Number and sequence length of different types of structural variation with each (sub)genome compared with the NIP reference genome. **b** The distribution of SV length in wild and cultivated rice. red color indicates wild species and blue color shows cultivated species. **c** The distribution of private genes of each rice diploid genome compared to NIP. From the outer-most track to the inner-most track: (**a**) the private gene position of R498 from A genome type rice. **b** *O. punctata* from B genome type. **c** *O. minuta* from C genome type. **d** *O. alta* from D genome. **e** *O. Australians* from E genome. **f** *O. brachyantha* from FF genome. (g) *O. meyeriana* from G genome type. **h** the gene density of NIP across the genome. **i** the PAV density, the sliding windows of (**h** and **i**)windows was 300 kb. **d** The distribution of PAV of each rice diploid genome. Diploid genomes from outermost circle to innermost circle: (**a**) R498(*O. sativa* ssp. indica); (**b**) CG14(*O. glaberrima*); (**c**) *O. rufipogon*; (**d**) *O. glumaepatula*; (**e**) *O. malampuzhaensis* | B_t; (**f**) *O. punctata* | C_t; (**g**) *O. latifolia*; (**h**) *O. australiensis*; (**i**) *O. brachyantha*; (**j**) *O. meyeriana*. Source data are provided as a Source Data file.

aligning collinear blocks of the rice accessions against the Nipponbare reference genome to assess their potential impact on important agronomic traits. Through whole-genome comparisons, we identified 207 genes with tandem repeats across the 14 wild rice assemblies, potentially influencing yield, resistance, grain quality, heading date, and biotic and abiotic resistance (Supplementary Data 10). To gain further insights into the functional roles of gCNVs in rice, we analyzed 4400 genes with known functions reported in a previous study[36]. Among these genes, 36 exhibited tandem repeats in the rice genus, impacting various agronomic traits related to yield, disease and pest resistance (e.g., blast, bacterial blight, rice brown planthoppers), biotic stress tolerance, element transport, and other important adaptation traits like heading date and hybrid sterility (Fig. 6a).

Expression levels of selected gCNVs were assessed to investigate potential changes in their expression profile. Several variations linked to the *Pi9* cluster, with gCNVs in the 10.38 Mb region of Nip genome chromosome 6, were also identified. *Pi9* is a well-known gene in rice that offers strong and long-lasting resistance to the fungus *Magnaporthe oryzae*[37]. Interestingly, *Pi9* is a typical NOD-like receptor encoding gene with copy number variation, which played a role in rice species' environmental adaptation (Supplementary Fig. 21).

Nucleotide-binding domain and leucine-rich repeat (NLR) proteins play a crucial role in plant immune systems[38]. Therefore, it is essential to have an accurate NLR dataset for rice genera. Plant *NLR* encoding genes often occur in clusters, making their identification challenging. To address this issue, we utilized RGAugury[39] and DupGen_finder[40] tools, resulting in a total of 7048 *NLR* encoding genes across the rice genus (Supplementary Table 10). The number of *NLR* encoding genes varied from 419 in *O. glabberima* to 511 in *O. sativa* ssp. indica (R498) in cultivated rice and from 159 in *O. australiensis* to 669 in *O. punctata* (BBCC) in wild rice (Fig. 6b and Supplementary Table 10), which revealed a significant expansion of *NLRs* in the cultivar rice.

Our study focused on identifying and categorizing *NLRs* in different rice species to establish a comprehensive understanding of their diversity within the rice genus. Interestingly, the diploid rice genome exhibited fewer *NLR* in wild rice compared that in cultivated rice, despite the larger genome size in wild rice (Fig. 6f). For instance, the genomes of *O. australiensis* and *O. meyeriana*, although twice the size

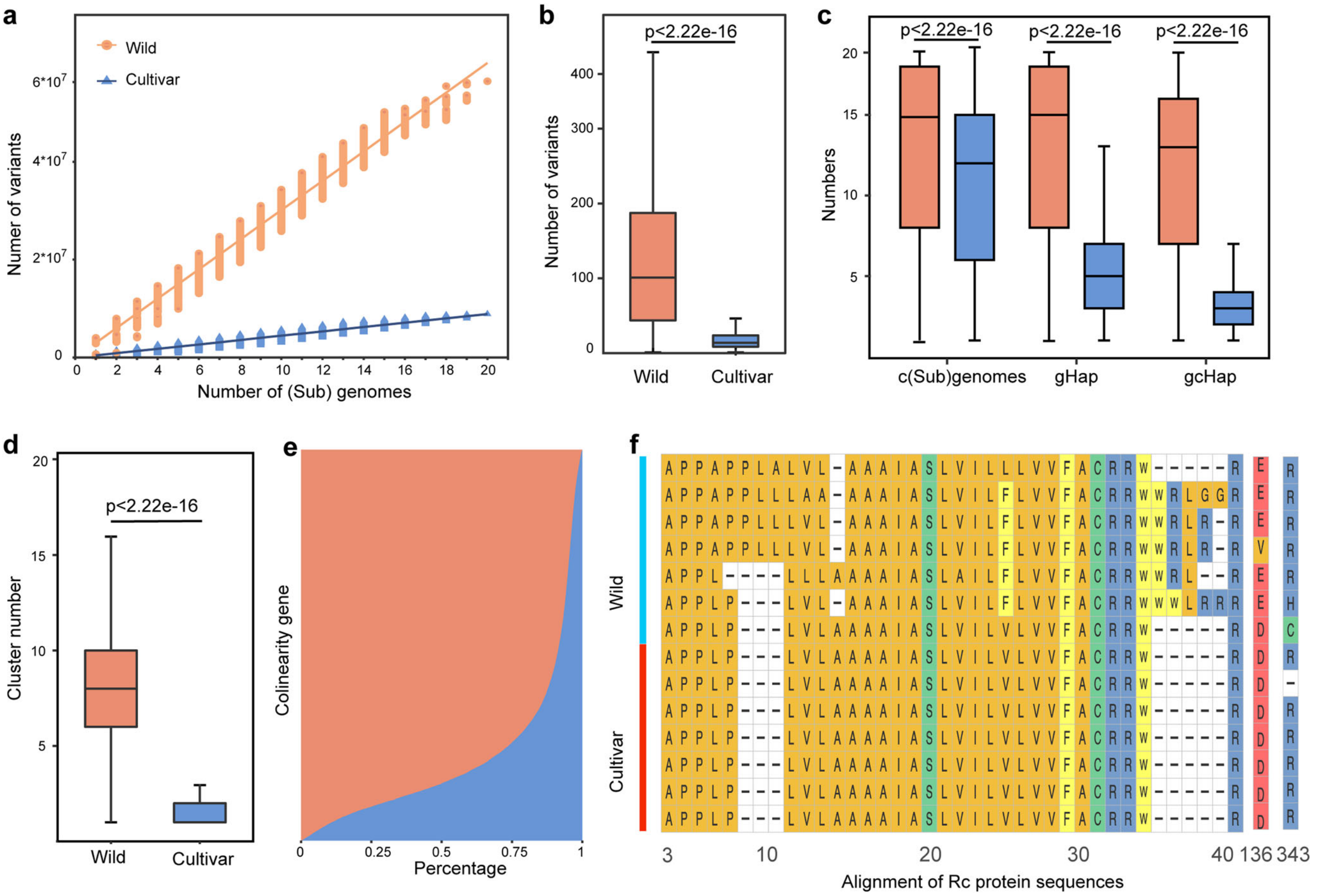


Fig. 5 | Alleles and their regulatory sequence variations in wild and cultivated rice accessions. a Genome-wide allelic variation and nearby 10 kb upstream/downstream sequences variation number in the wild and cultivated rice group along with equal amounts of rice diploid genomes. **b**–**d** Comparison of total SV (**b**), average collinear gene, gene haplotype, and CDS haplotype (**c**), orthologue protein cluster (**d**) between wild and cultivar rice group (*x*-axis). The y-axes represent the number of variants per gene, number of average collinear genes, gene haplotype, CDS haplotype, and number of protein clusters. The box plots show the medians (centerlines), interquartile ranges (boxes), and 1.5 times the interquartile ranges (whiskers), $n = 20$ for both cultivar rice and wild rice groups. Statistical significance (*P*-value) was determined using a two-sided Wilcoxon rank-sum test. **e** The unbalanced orthologue gene haplotype in wild and cultivated rice. **f** Alignment of an example Rc protein exhibited significantly more variation in wild than that in cultivated rice. Source data are provided as a Source Data file.

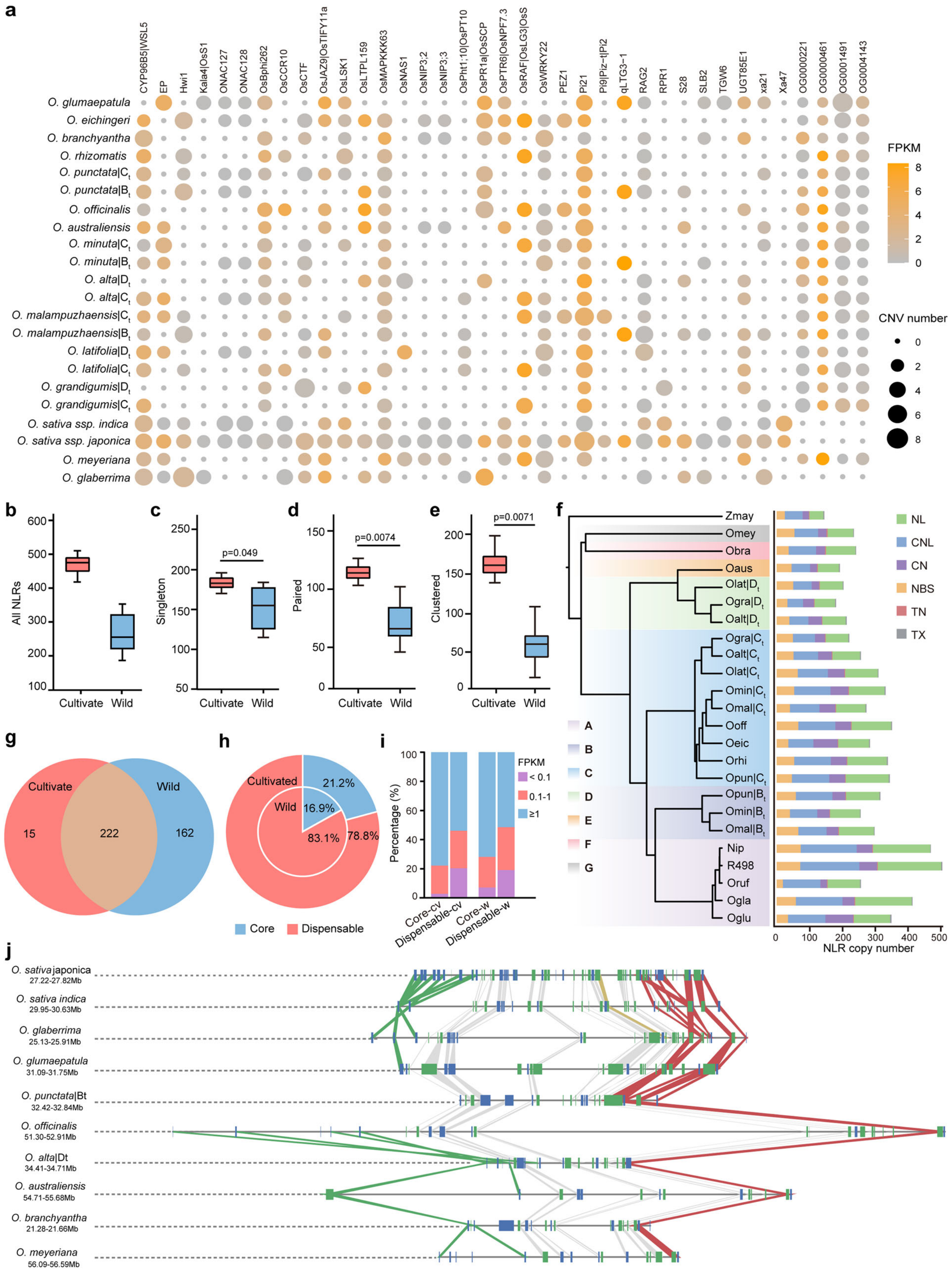
of NIP, contain only half the number of *NLRs* of that in cultivated rice (Fig. 6f). Analysis of *NLR* distribution showed that while *R* gene singletons were similar between wild and cultivated rice, cultivated rice tended to have a higher number of *R* genes in pairs or clusters compared to wild rice (Fig. 6b–d). Redundancy analysis revealed that 55.64% of *NLR* signatures were shared across all genomes, with 15 unique signatures in the cultivated group and 162 unique signatures in the wild rice group (Fig. 6g). The study found that as the number of cultivated rice accessions increased, the number of core *NLR* signatures also tended to increase. Analysis of *NLR* encoding genes in wild and cultivated rice revealed that 78.8% of the *NLR* genes in cultivated rice were dispensable, slightly lower than in wild rice (Fig. 6h). More than 90.0% of *NLR* genes in the core genome were expressed in both wild and cultivated rice, while around 20.0% of *NLR* genes in the dispensable genome exhibited low or no expression under normal conditions, suggesting specific expression upon encountering disease (Fig. 6i).

We classified *NLRs* of the rice genus into 359 clusters, of which 142 clusters were increased in cultivated rice (Supplementary Data 11), including well-studied rice *R* gene families, such as WRKY61, that provide resistance to rice blast disease caused by *Xanthomonas oryzae*. pv. *oryzae* (*Xoo*) (Fig. 6j). In addition, an *NLR* expansion event was observed in the wild rice pangenome (Fig. 6j), enabling these plants to adapt to various environments compared to cultivars. By leveraging lost *NLR* rice genes during domestication and artificial selection, we can enhance the resistance resources of cultivated rice and enrich the diversity of modern commercial rice. The total number of *NLRs* in cultivated rice species was higher compared to diploid wild rice species, despite some *NLR* gene losses, indicating that *NLR* expansion into cluster forms may be driven by breeding for specific pathogen resistance.

Discussion

In this study, we integrate genome assemblies of 13 wild rice species, three cultivated species[16–18], and one common wild rice[19] to construct a pangenome of the rice genus. Compared to cultivated rice, the pangenome generated in this study provides an additional 63,881 gene families. This dataset facilitates the functional investigation of novel sequence types, including large structural variations in non-coding regions, coding genes absent from the cultivated rice genome, private genes in wild rice, differentiation genes between wild and cultivated rice, and natural allelic variations. Furthermore, it aids in mapping candidate genes that control important traits in wild rice introgression populations. In addition, we reconstruct the phylogenetic tree for *Oryza* at the genome level and correct the evolutionary positions of the B, C, F, and G rice species.

Previous studies have revealed that TEs play a significant in driving of genome size evolution[26]. This study also enhances our understanding of the substantial variability in genome sizes during the evolution of *Oryza* and identifies which components of repeat



sequences predominantly contribute to rice genome size, serving as a model for similar analyses in other plant species. The genome size of the G genome type is significantly influenced by the Retand LTR subfamily, while the E genome type's genome size is increased by both the Gypsy (Ogre, Retand, and Tekay) and Copia (Ange) LTR superfamilies. In contrast, the genome sizes of B, C, and D genome types are primarily influenced by the top three Gypsy superfamilies in descending order: Ogre, Retand, and Tekay. Furthermore, this study reveals a strong correlation between genome size and the Retand LTR subfamily in rice, indicating that the varying chromosome size of *Oryza* appears to be driven by large-scale expansions and contractions of Retand elements in the intergenic regions.

Fig. 6 | Characteristic of gene CNVs associated with rice important agronomic traits and *R* gene in *Oryza* genus. a Feature of the CNV gene. Circle size indicates the gene copy number potentially results from a tandem duplicated mechanism. Color from gray to yellow shows the gene expression level from low to high. **b**–**e** Comparison of total *R* genes (**b**), singleton *NLRs* (**c**), paired *NLRs* (**d**), and clustered *NLRs* (**e**) number between wild and cultivar rice. The box plots show the medians (centerlines), interquartile ranges (boxes), and 1.5 times the interquartile ranges (whiskers). The sample size $n = 20$ for wild rice group, and $n = 3$ for cultivar rice group. *P*-values indicate the variation between wild and cultivar rice groups (Wilcoxon rank sum test). The red color represents wild rice species, the blue indicates cultivated rice. **f** The *NLR* gene number in each (sub)genome. The different colors in *NLR* copy numbers represent the subfamily of *NLRs*. The various color backgrounds indicate (sub) genomes from different genome types. **g** Venn diagram of orthologues *NLRs* gene in wild and cultivated rice. **h** The percentage of core and dispensable non-redundant *NLRs* in the wild and cultivated rice species. **i** Expression of core and dispensable non-redundant *NLRs* in wild and cultivated rice. pink shows the gene expression level is less than 0.1, red indicates the gene expression is larger than 0.1 and less than 1, and blue means the gene expression is larger than 1. **j** A representative region on chromosome 11 has less *R* gene in wild rice, but more *R* gene in cultivated rice. The same color means synteny *NLR* gene. Source data are provided as a Source Data file.

Our study also exemplifies how these new resources can enhance our understanding of the role of SVs, gCNV, and allelic variation in the processes of environmental adaptation and important agricultural traits in rice. A total of 21.52% of SVs in cultivars and 60.90% in wild rice are larger than 5 kb. Moreover, 2.34% of SVs larger than 250 kb were also found in wild rice, indicating that wild rice exhibits low collinearity with cultivated rice and contains large untapped private genes. These genetic variations likely play an important role in the local adaptation of wild rice to diverse habitats. In addition, a total of 36 CNVs in the rice genus, impacting various agronomic traits related to yield, disease, and pest resistance, have been displayed, which suggests that our CNV catalog is beneficial for investigating the hidden genomic variations underlying phenotypic diversity.

In addition, we observe that the number of *NLRs* in cultivated rice exceeds that in the wild rice diploid genome, yet exhibits lower disease resistance than that in wild rice (Fig. 6e–j); The number of clustered *NLRs* in cultivated rice is significantly higher than in wild rice, suggesting that some additional copies of *NLRs* may be redundant in providing resistance in cultivated rice. This finding aligns with the notion that multiple *NLRs* are necessary for the broad-spectrum resistance of Tetep to blast[41].

Understanding the genomes of various species within the genus *Oryza* enhances our ability to trace evolutionary history and develop strategies for crop improvement in the face of rapid climate change. The rice genus Pangenome represents a significant advancement in uncovering the underlying variations that influence traits, adaptation, and domestication in rice, while also facilitating the utilization of functional genes found in wild rice. Furthermore, when integrated with chromatin immunoprecipitation sequencing (ChIP-seq) and the Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq), the *Oryza* genomes provide valuable insights into the analysis of dynamic variations in centromere and gene regulation profiles. The next step in the *Oryza* genus pan-genomic will focus on increasing production, enhancing resistance to various diseases, and adaptation to changing environments through gene editing of private genes and alleles.

Methods

Sample selection

A total of 13 rice accessions derived from seven genome types, including AA, BBCC, CC, CCDD, EE, FF, and GG, were selected from the wild rice nursery in Hainan, China. These rice accessions included one AA genome type (*O. glumaepatula*), three BBCC allotetraploids (*O. punctata*, *O. minuta*, *O. malampuzhaensis*), three CC genome types (*O. eichingeri*, *O. officinalis*, *O. rhizomatis*), three CCDD allotetraploids (*O. latifolia*, *O. alta*, *O. grandiglumis*), one EE genome type (*O. australiensis*), one FF genome (*O. brachyantha*), and one GG genome type wild rice (*O. meyeriana*). The relevant sample information is provided in Supplementary Table 1.

Library preparation and resequencing

High-molecular-weight DNA was extracted from 13 wild rice pockets 20 days after transplant using the CTAB protocol. Sequencing libraries with an insert length of 300 bp were prepared employing the CLEANNGS DNA kit for Illumina sequencing. The sequencing was conducted on the Novaseq 6000 platform (Berry Genomic, Beijing, China) to generate 150 bp paired-end reads, following the manufacturer's guidelines. Short reads were filtered by removing adapter contamination, and low-quality reads, specifically, those containing base N's or exhibiting more than 10% of bases with a quality score below 20.

Tissue collection and RNA extraction

Plant material for RNA sequencing (RNA-seq) and Iso-Seq was grown in a greenhouse at the Jiangxi Academy of Agricultural Sciences, maintained under a 9-hour light and 15-hour dark cycle with daytime at temperatures of 28 °C and nighttime temperatures of 18 °C. Roots, leaves, stems, and panicles were collected, mixed, snap-frozen in liquid nitrogen, and stored at − 80 °C until RNA extraction was performed. RNA was extracted from the tissues using a Trizol extraction protocol. The integrity of the RNA was assessed using the Agilent 2100, and only samples with an RIN value greater than 8 were used for RNA-seq and Iso-Seq library construction.

RNA-seq library preparation and data generation

Sequencing libraries for the Illumina NovaSeq platform were prepared using the VAHTS mRNA-seq v2 Library Prep Kit, following the manufacturer's guidelines. This producess generated 150 bp paired-end reads with a 350 bp insert size. For tissue-specific data, please refer to Supplementary Data 1 for tissue-specific data. A custom pipeline was implemented to remove low-quality reads and adapters, while Hisat2[42] was employed to align high-quality bases to the reference genome using default settings. Subsequently, Cufflinks[43] was utilized to calculate transcription expression levels in FPKM. Differential expression analysis for homologous gene pairs on corresponding chromosomes of the subgenomes was performed using a student's *t*-test with Bonferroni correction.

De novo genome assembly of 13 wild rice accessions

Genome size and heterozygosity were estimated using a k-mer-based approach by using GCE[44]. The genomes of the 13 HiFi sequenced accessions were assembled into one primary and two haplotype draft contig genomes using HiFiasm v0.16.1[45], with parameters -h1 -h2 and -ul. Subsequently, the Hi-C reads were employed to anchor the assembled contigs onto chromosome pseudomolecules through sorting, orientation, and ordering utilizing 3D-DNA[46] and Juicer v1.5[47] to produce the final version of the wild rice genome assembly. Juicebox (https://github.com/aidenlab/Juicebox) was used to visualize the resulting Hi-C contact matrix, and manual corrections were performed based on neighboring interactions. The heatmap of genomic interactions was generated using HiCPlotter v2.7.0. Subgenome identification in this study was conducted based on the mapping depth of the diploid genome sequence[48], A reference-guided strategy based on subgenome sequence similarity, was employed to distinguish the subgenomes of CCDD allotetraploid wild rice (*O. alta*, *O. grandiglumis*, and *O. latifolia*) and BBCC wild rice (*O. malampuzhaensis*, *O. minuta* and *O. punctata*). The assembled genome of *O. officinalis* was divided into

100 bp non-overlapping fragments and aligned with the BBCC and CCDD genome assemblies using BWA[49] with the default settings. Only uniquely mapped fragments were retained. A syntenic block was defined by the presence of at least five syntenic fragments. Chromosomes exhibiting higher similarity to *O. officinalis* were designated as belonging to the C subgenome, whereas those with lower similarity were assigned to the B/D subgenome. Subgenome assignments were further validated by quantifying the depth of coverage of paired-end reads from *O. punctata* (BB, SRX15097569) in comparison to *O. officinalis* or *O. officinalis* and *O. australiensis*.

Evaluation of gene assemblies

The completeness of the assembled wild rice genomes was assessed using BUSCO[22] with the embryophyte_odb10 database. Genome continuity was evaluated using LAI[23] within the LTR_retriever[50] packages, focusing on the scaffold N50 parameter. We also examined the mapping rate of transcripts assembled with Trinity[51] to the corresponding genome assemblies through BLASTN[52], applying a minimum alignment length of 350 bp and a sequence identity threshold of 95%. Furthermore, the assemblies were assessed by mapping Illumina short reads and HiFi long reads using BWA and minimap2[53]. In addition, Merqury[54] was employed to evaluate the QV value and overall completeness of the genomes.

Identification, annotation, and classification of repetitive elements

De novo repeat identification and homology-based searches were conducted on the wild rice assemblies. For the ab initio-based search, results obtained from RepeatModeler (http://www.repeatmasker.org/RepeatModeler/), LTR_finder[55], LTR_retriever[50], and LTR_harvest[56] were integrated to construct a repetitive sequence library. This library was subsequently annotated and analyzed using RepeatMasker[57] with default parameters. The Repbase Transposable Element (TE) library and TE protein databases were utilized to mask TEs through the application of RepeatMasker[58] software and RepeatProteinMask programs. Full-length LTR retrotransposons (LTR-RTs) were identified using LTR_retriever[50] and were further classified using TEsorter[59]. In addition, solo-LTRs and intact LTRs were classified using LTRharvest (https://omictools.com/ltrharvest-tool).

Estimation of LTR insertion times

We estimated the insertion times of intact LTR retrotransposons. LTR_Finder[55] was utilized to de novo identify full-length LTR retrotransposons in the genomes of 13 wild rice species. Candidate intact LTRs were identified by locating matching LTR pairs within a specified distance. The full-length LTR sequences were aligned using MUSCLE[60]. Nucleotide variations (λ) within the intact LTR retrotransposons were calculated, and their divergence (K) was determined using the following formula[61]:

$$K = -3/4\ln(1-4\lambda/3) \tag{1}$$

The insertion time of these full-length LTRs was estimated based on the following formula[62]:

$$T = K/2r \tag{2}$$

where r in rice is estimated at 1.3×10^{-8} per bp per year[62].

Gene prediction and functional annotation

Gene models for wild rice were predicted by combining transcriptome data with both de novo and homology-based strategies. For the homology-based approach, protein sequences from six species (*Z. mays*, *O. sativa* ssp. indica, *O. sativa* ssp. japonica, *O. rufipogon*, *O. barthii*, and *L.perrieri*) were downloaded and aligned to the wild rice assemblies using Tblastn[63]. The resulting BLAST hits were subsequently utilized to predict the gene structures of the corresponding genomic regions with Genomethreader[64]. For predictions based on transcriptome data, RNA sequences from various tissues, including leaves, stem, panicles, and roots, were assembled using Trinity, and filtered with PASA[65]. The RNA data was then mapped onto the wild rice genomes using HISAT2[42], and the reads were assembled into transcripts using StringTie[66]. Transdecoder (https://github.com/TransDecoder/TransDecoder) was employed to predict open reading frames. For de novo predictions, three ab initio gene prediction programs, including GENSCAN[67], AUGUSTUS[68], and GlimmerHMM[69], were utilized to forecast coding regions from the repeat-masked genome. All gene model evidence obtained was subsequently amalgamated using EVidenceModeler[70] to generate non-redundant gene sets. Protein-coding genes were retrieved and annotated based on their best BLASTP[71] hits with an E-value of less than $1e^{-5}$ from various protein databases, including UniProtKB/Swiss-Prot (https://www.uniprot.org/downloads) and nonredundant protein database NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db). Gene Ontology (GO) terms and Pfam domains were assigned to each gene using InterProScan[72], and gene pathways were determined through homology searches against the KEGG database with an expected value of 10^{-5}.

Analyses of the protein-coding-gene-based *Oryza* pangenome

Combined the Asian cultivated rice (*O. sativa* indica, cv R498 and *O. sativa* japonica, cv Nipponbare)[12,17], African cultivated rice (cv. CG14), and common wild rice, along with 13 wild rice species, to represent the rice genus. To identify homologous relationships among the *Oryza* species, the longest transcript of each predicted gene in each genome was selected as a representative. An all-to-all comparison of 957,779 peptide sequences of protein-coding genes was conducted using BLASTP with an E-value of e^{-10}, followed by clustering the BLSAT results using OrthoFinder[25] with default parameters. The clusters were categorized into three groups: the core genome shared by all 23 genomes, the dispensable genome containing genes present in 2 to 22 genomes, and specific gene families unique to only one genome. Clade-specific gene families were defined as those found exclusively within one clade of the rice genus. ClusterProfiler[73] was employed for GO terms function enrichment analysis. Non-synonymous/synonymous substitution ratios (Ka/Ks) for core and soft-core clusters were calculated using ParaAT[74], with parameters set to '-m muscle -f axt -k'.

Oryza phylogenetic tree construction

Z. may be chosen as the outgroup for inferring the species phylogeny. A super-matrix tree was constructed using protein sequences from the longest transcripts of gene models selected from 14 wild rice and three cultivated rice accessions (Supplementary Table 1. BLASTP was employed for all-against-all alignments, followed by Orthofinder to infer orthologous genes. The 3555 single-copy orthologous genes were aligned using MAFFT[75], and the best model was determined with ProtTest3.0[76]. The phylogenetic tree of *Oryza* was constructed using RAxML software[77]. Bootstrap analyses with 1000 replicates were conducted for the concatenated sequence using the GTR + I model. Divergence times were estimated using the MCMCTree program in PAML (v4.9).

To mitigate the impact of incomplete lineage sorting, we utilized a multi-species concatenation-based approach implemented in ASTRAL[78] to construct a species tree. This estimation involved searching for the species tree that best aligned with the quartets obtained from the input gene trees.

Estimation of the divergence time

The divergence times of various species were calculated using MCMCTree from the PAML package[78] under the relaxed molecular clock hypothesis. To optimize computational resources, coding sequences (CDS) of single-copy genes from eight representative species (*O. sativa*, *O. glaberrima*, *O. rufipogon*, *O. punctata*, *O. alta*, *O. australiensis*, *O. brachyantha*, *O. meyeriana*) were selected for a preliminary assessment of substitution rates using BASEML with the model set to 4. Subsequently, MCMCTree was employed to estimate divergence times with the parameters 'model = 4, burnin = 50,000, sampfreq = 10, nsample = 20,000. Divergence times between *O. sativa*-*O. rufipogon* (0.7–1.7 Mya) and rice-maize (41.4–51.9 Mya) were utilized for calibration purposes.

Collinearity and chromosome rearrangement analysis

The protein sequences of annotated genes from wild rice assemblies and three previously sequenced genomes (*O. rufipogon*, *O. sativa* indica, and *O. glaberrima*) were aligned with the NIP[12] genome using BLASTP, applyingwith an e-value cut-off of $1e^{-5}$. The reciprocal best hit for each alignment was utilized to analyze the collinearity of the entire genomes between NIP and the other species within the Oryza genus, employing MCScanX[79] with default parameters. The resulting synteny outcomes were visualized using the jcvi software (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)), and significant chromosome rearrangements between NIP and the other species were identified.

The homoeologous chromosome sequences from groups 1, 3, and 6 of the C_t subgenome of *O. latifolia*, *O. alta,* and *O. grandiglumis* were aligned to analyze the syntenic regions using Mummer[80] with the parameter '-L 1000.' To assess the continuity of genome exchange in *O. latifolia*, Illumina reads were aligned back to the *O. latifolia* genome contigs using the default parameters of the Burrows-Wheeler Aligner software[49]. Sam tools was employed to calculate the sequencing depth, which was subsequently visualized by using R package ggplot2[81].

Centromere sequence identification

No results were obtained when the RCS2[82] family alignment was applied to the final 13 wild rice assemblies to identify centromere-related sequences. Subsequently, a search for tandem repeats was conducted across the rice genome using the Tandem Repeats Finder software[83]. The identified repeats were grouped into families based on an 80% similarity threshold, employing the SiLiX algorithm[84]. The relative abundance of each family was determined through k-mer analysis, utilizing Illumina short-read sequences and the Jellyfish software. To identify *O. brachyantha*, Chip-seq data (SRX4224850) were used and compared to known centromeric repeat sequences from *O. sativa* using the BLAST tool. The distribution of the five most prevalent repeats on each chromosome was manually examined to evaluate their characteristics as centromeric repeats. To further analyze the identified centromeric repeats, a neighbor-joining tree was constructed by randomly selecting 200 repeats from each family and using MEGA-X[85].

Identification of genome sequence structure variation

SV identification was performed using a modified approach that employed Minimap2 and SYRI software[86]. Pairwise genome alignment was executed between the genomes of 13 wild rice assemblies, using NIP[16] as the reference, through the Numcer program in Mummer[80] with default parameters. The alignment results were subsequently filtered to yield one-to-one alignment blocks. SVs in wild rice, relative to the NIP[16] genome, were identified using Minimap2[53] in conjunction with SYRI[86]. Initially, each assembled genome was aligned to the Nipponbare[16] genome using Minimap2[53]. The resulting alignments were then utilized for structural variation identification with SYRI[86] applying default parameters. To reduce false positives in SV calling, only SVs such as insertions, deletions, inversions (less than 1 Mb in size), and translocations (greater than50bp in length) were retained, while SVs containing 'N' sequences were excluded.

SV validation

Hi-C data were employed to validate SVs exceeding 10 kb in length. Paired-end reads from Hi-C were aligned to the corresponding genome assemblies, and the interaction heatmap for regions containing SVs was manually examined. For SVs shorter than 10 kb long reads were mapped to the genome assemblies, and the alignments at the boundaries of these SVs were manually verified. In addition, ten SVs were randomly selected for PCR amplification to evaluate fidelity in *O. glumaepatula*.

Distribution of PAV relative to the reference genome

To identify unique or absent sequences in comparison to the NIP[16] reference, the candidate PAV sequence was aligned to the NIP[16] genome using minimap2[53] with the parameter '-x asm10'. Sequences that covered more than 80% were filtered out to define the final PAV region. The genome wide distribution of PAV was visualized using Circos[87].

Distribution of private gene relative to Nipponbare genome

MCscanx was employed to identify collinear orthologs between the query genomes and the NIP[12] genome, requiring a minimum of five homologous genes to define a collinear block. Furthermore, certain translocations were not necessarily required to meet the threshold for synteny search. To characterize the private genes associated with NIP, it was essential to identify genes that shared a gene index pair with NIP in each genome. In addition, reciprocal best hits were employed as evidence to establish gene index pairs. Subsequently, collinear orthologs for each genome in relation to NIP were determined, while non-collinear orthologs were classified as private genes for each genome.

Collinear orthologues whose 10 kb upstream/downstream sequences variation

To better understand the evolutionary significance of variation in collinear orthologues within the rice genus, we conducted a sequence analysis of gene pairs along with their 10 kb upstream and downstream sequences. The genes and their respective sequences, obtained from GFF files, were aligned to orthologues on the NIP[16] reference genome to identify genomic variations, including SNPs, insertions, deletions, inversions, and PAVs (greater than 50 bp) using SYRI[86]with the show-snps parameter. Then we converted the aligned results to VCF format for each subgenome/genome, Bcftools[88] was used to split multiallelic variants into multiple rows and left-normalize INDELs before counting variants at the cohort level. To annotate the estimated effects of the identified genomic variations, we applied the software SnpEff[89]. We then annotated variants by haplotypes using CooVar[90] and ultimately combined the annotation results.

Haplotype analysis of genes, CDS, and proteins in the core genome

To capture the critical aspects of rice genomic diversity, it is essential to thoroughly characterize the functional haplotype diversity of nearly all rice genes, coding sequences (CDS), and proteins in wild rice. The gHAP dataset for the rice genus, which included 20 cultivated rice varieties and 14 wild rice species across 20 subgenomes, was constructed by concatenating the SNPs within the gene regions while ignoring synonymous SNPs. In instances where no SNPs or InDels are present in a gene region, we assign them the same gHap. For protein diversity, multiple sequence alignments for each orthologous protein cluster were performed using MAFFT[75], and the proteins were categorized based on their similarity.

Determination of group-unbalanced genes haplotype

Distributed genes (or gene families) haplotypes were further divided into three categories[3], including cultivated-predominant, wild-predominant, and group-balanced gene (or gene families) haplotypes. Group-unbalanced gene (or gene families) haplotypes are characterized by an unequal distribution among the cultivated and wild groups.

Characterization of copy number variant genes

Gene CNV identification was conducted with refinements based on previous reports[91]. For each rice variety, including NIP, DupGen_finder[40] was employed to identify tandem duplicated genes using the default parameter. A gene-CNV locus was defined as a duplicated locus, encompassing both tandem and proximal duplicates, that exhibited a differing copy number in at least one of the other rice varieties compared to NIP[12]. We obtained 4431 functionally known genes associated with traits in rice from a previous study (https://funricegenes.github.io/). Of these, we displayed the identified CNVs and expression levels according to each genome/subgenome. The expression levels were calculated as the average of three replicates across all tissues.

The analysis of pan-NLRome

The resistance genes were identified by integrating de novo genome annotation with RGAugury[39], using default parameters. For genome sequence annotation, the LRR domain was extracted from the de novo genome annotations. Subsequently, we verified the *NLRs* generated by these two methods using InterProScan[72] and hmm search to ensure the accuracy of the subsequent analyses. To determine how *NLRs* were regulated, we calculated the average expression levels (FPKM) of *NLRs* under standard growth conditions across all tissues.

To identify putatively expanded *NLR* clusters in rice, we classified the annotated *NLR* loci from 13 wild rice genomes and 3 cultivated rice genomes into clusters using a clustering method employed in pan-genome analysis. Based on the results from the identification of *NLRs* and OrthoFinder, we quantified the non-redundant *NLR* gene counts in both wild and cultivated rice.

The *NLRs* can be categorized into three categories types based on their positions[92]. If there were fewer than two *non-NLR* genes between any two *NLRs*, these two were defined as 'approaching'. Clusters with three or more approaching *NLRs* were classified as clusters, while two approaching *NLRs* were considered pairs. *NLRs* without approaching *NLRs* were labeled as singleton *NLRs*.

Collinearity and cluster of *NLR* genes

The accuracy of collinearity among *NLR* loci, as mentioned in the main text, was verified using the previously identified one-to-one alignment block (Supplementary Fig. 20). The *NLR* gene pairs and clustered genes were recognized as the same orthologous gene related to Nipponbare[12]. To better understand the expansion of *R* gene number in cultivated rice, we generated a cluster matrix of all *R* gene proteins using MAFFT. By sorting the number of genes in each group, we can observe the expansion of a specific *NLR* gene family.

Reporting summary

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

Data availability

The raw sequencing data and genome assembly have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession PRJNA1175549 and National Genomics Data Center under BioProject PRJCA024515. The genome assembly and annotation of 13 wild rice are available at Figshare [https://figshare.com/s/32d0ac68cee7f647e1e3]. Source data are provided in this paper.

References

1. Wing, R. A., Purugganan, M. D. & Zhang, Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet.* **19**, 505–517 (2018).
2. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
3. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
4. Khan, A. W. et al. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant Sci.* **25**, 148–158 (2020).
5. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat. Genet* **50**, 285–296 (2018).
6. Ge, S., Sang, T., Lu, B. R. & Hong, D. Y. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Natl. Acad. Sci. USA* **96**, 14400–14405 (1999).
7. Li, N. et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat. Genet.* **55**, 852–860 (2023).
8. Zhou, Y. et al. Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice. *Nat. Commun.* **14**, 1567 (2023).
9. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
10. Shang, L. G. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
11. Zhang, F. et al. Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.* **32**, 853–863 (2022).
12. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558.e3516 (2021).
13. Shi, C. et al. The draft genome sequence of an upland wild rice species, Oryza granulata. *Sci. Data* **7**, 131 (2020).
14. Zhao, H. et al. A high-quality chromosome-level wild rice genome of Oryza coarctata. *Sci. Data* **10**, 701 (2023).
15. Yu, H. et al. A route to de novo domestication of wild allotetraploid rice. *Cell* **184**, 1156–1170 (2021).
16. Kawahara, Y. et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).
17. Du, H. et al. Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* **8**, 15324 (2017).
18. Wang, M. et al. The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. *Nat. Genet.* **46**, 982–988 (2014).
19. Xie, X. et al. A chromosome-level genome assembly of the wild rice Oryza rufipogon facilitates tracing the origins of Asian cultivated rice. *Sci. China Life Sci.* **64**, 282–293 (2021).
20. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
21. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
22. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
23. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
24. Zou, X. H. et al. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol.* **9**, R49 (2008).
25. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
26. Pulido, M. & Casacuberta, J. M. Transposable element evolution in plant genome ecosystems. *Curr. Opin. Plant Biol.* **75**, 102418 (2023).
27. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).
28. Comai, L., Maheshwari, S. & Marimuthu, M. P. A. Plant centromeres. *Curr. Opin. Plant Biol.* **36**, 158–167 (2017).

29. 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).
30. Kou, Y. et al. Evolutionary genomics of structural variation in Asian Rice (Oryza sativa) domestication. *Mol. Biol. Evol.* **37**, 3507–3524 (2020).
31. Bai, F. et al. Natural allelic variation in GRAIN SIZE AND WEIGHT 3 of wild rice regulates the grain size and weight. *Plant Physiol.* **193**, 502–518 (2023).
32. Sun, X. et al. Natural variation of DROT1 confers drought adaptation in upland rice. *Nat. Commun.* **13**, 4265 (2022).
33. Huang, X. et al. Natural variation at the DEP1 locus enhances grain yield in rice. *Nat. Genet.* **41**, 494–497 (2009).
34. Wang, Y. et al. Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.* **47**, 944–948 (2015).
35. Deng, Y. et al. Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. *Science* **355**, 962–965 (2017).
36. Huang, F. et al. New data and new features of the FunRiceGenes (Functionally Characterized Rice Genes) database: 2021 update. *Rice* **15**, 23 (2022).
37. Qu, S. et al. The broad-spectrum blast resistance gene Pi9 encodes a nucleotide-binding site-leucine-rich repeat protein and is a member of a multigene family in rice. *Genetics* **172**, 1901–1914 (2006).
38. Feehan, J. M., Castel, B., Bentham, A. R. & Jones, J. D. Plant NLRs get by with a little help from their friends. *Curr. Opin. Plant Biol.* **56**, 99–108 (2020).
39. Li, P. et al. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).
40. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).
41. Wang, L. et al. Large-scale identification and functional analysis of NLR genes in blast resistance in the Tetep rice genome sequence. *Proc. Natl. Acad. Sci. USA* **116**, 18479–18487 (2019).
42. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
43. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
44. Liu, B. et al. Estimation of genomic characteristics by analyzing k mer frequency in de novo genome projects. *Preprint at* https://doi.org/10.48550/arXiv.1308.2012 (2020).
45. 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).
46. Dudchenko, O. et al. de novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
47. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
48. Peng, Y. et al. Reference genome assemblies reveal the origin and evolution of allohexaploid oat. *Nat. Genet.* **54**, 1248–1258 (2022).
49. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
50. 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).
51. Hu, Z. et al. Full-length transcriptome assembly of Italian ryegrass root integrated with RNA-seq to identify genes in response to plant cadmium stress. *Int J. Mol. Sci.* **21**, 1067 (2020).
52. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
54. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
55. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
56. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinforma.* **9**, 18 (2008).
57. Tarailo-Graovac M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* https://doi.org/10.1002/0471250953.bi0410s25 (2009).
58. Tempel, S. Using and understanding RepeatMasker. *Methods Mol. Biol.* **859**, 29–51 (2012).
59. Zhang, R. G. et al. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017 (2022).
60. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
61. Jukes, T. H. & CR, C. *Evolution of Protein Molecules*. (1969).
62. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **101**, 12404–12410 (2004).
63. Gertz, E. M., Yu, Y. K., Agarwala, R., Schaffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41 (2006).
64. 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).
65. Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
66. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
67. Aggarwal, G. & Ramaswamy, R. Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* **27**, 7–14 (2002).
68. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
69. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
70. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
71. Jacob, A., Lancaster, J., Buhler, J., Harris, B. & Chamberlain, R. D. Mercury BLASTP: Accelerating protein sequence alignment. *ACM Trans. Reconfigurable Technol. Syst.* **1**, 9 (2008).
72. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
73. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
74. Zhang, Z. et al. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **419**, 779–781 (2012).
75. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537**, 39–64 (2009).
76. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
77. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

78. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).
79. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
80. Marcais, G. et al. MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
81. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
82. Dong, F. et al. Rice (Oryza sativa) centromeric regions consist of complex DNA. *Proc. Natl. Acad. Sci. USA* **95**, 8135–8140 (1998).
83. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
84. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinform.* **12**, 116 (2011).
85. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
86. 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).
87. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
88. Narasimhan, V. et al. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).
89. Cingolani, P. Variant annotation and functional prediction: SnpEff. *Methods Mol. Biol.* **2493**, 289–314 (2022).
90. Vergara, I. A., Frech, C. & Chen, N. CooVar: co-occurring variant analyzer. *BMC Res. Notes* **5**, 615 (2012).
91. Wang, Y. et al. Time-ordering japonica/geng genomes analysis indicates the importance of large structural variants in rice breeding. *Plant Biotechnol. J.* **21**, 202–218 (2023).
92. Shang, L. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).

Acknowledgements

We thank all the members of the Longan Yan group at Jiangxi Academy of Agricultural Science for collecting and preserving the wild rice resource. We thank Dr. Zhilan Fan for providing the GG genome type wild rice *O. meyeriana* at the Guangdong Academy of Agricultural Science, and we thank Dr. Shengyi Liu at the Oil Crops Research Institute, Chinese Academy of Agricultural Science, for providing constructive suggestions. This work was supported by China Agriculture Research System (CARS-01-08 to H.X.), National Key Research and Development Program of China (2023YFD1201203 to Y.C.), National Natural Science Foundation of China (31960400 to H.X.), Major Discipline Academic and Technical Leaders Training Program of Jiangxi Province (20213BCJL22044 to H.X.), and Jiangxi Technology Innovation Guidance Program (20223AEI91010 to H.X.).

Author contributions

L.Y., Y.C., and H.X. supervised the work. L.H.L., L.L.L., W.X., and Y.L. collected samples for resequencing, HiC, and HiFi sequencing. M.W. and W.L. performed the genome assembly. Q.H. performed the genome annotation. W.X.L., Y.W., and Y.T.W. conducted the bioinformatic analysis. J.W., Z.Y., and W.C. collected samples for RNA-seq sequencing and conducted expression validation. W.X.L., H.D., and H.X. wrote the manuscript and designed the experiment.

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/s41467-024-54427-3.

Correspondence and requests for materials should be addressed to Weixiong Long, Yaohui Cai, Huilong Du or Hongwei Xie.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

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

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