nature genetics

Article

https://doi.org/10.1038/s41588-025-02183-5

Oryza genome evolution through a tetraploid lens

Received: 26 May 2024

Accepted: 26 March 2025

Alice Fornasiero[1], **Tao Feng**[2], **Noor Al-Bader**[1], **Aseel Alsantely**[1,10], **Saule Mussurova**[1], **Nam V. Hoang**[2], **Gopal Misra**[1], **Yong Zhou**[1], **Leonardo Fabbian**[1], **Nahed Mohammed**[1], **Luis Rivera Serna**[1], **Manjula Thimma**[1], **Victor Llaca**[3], **Praveena Parakkal**[3], **David Kudrna**[4], **Dario Copetti**[4], **Shanmugam Rajasekar**[4], **Seunghee Lee**[4], **Jayson Talag**[4], **Chandler Sobel-Sorenson**[4], **Marie-Christine Carpentier**[5,6], **Olivier Panaud**[5,6], **Kenneth L. McNally**[7], **Jianwei Zhang**[8], **Andrea Zuccolo**[1,9], **M. Eric Schranz**[2] & **Rod A. Wing**[1,4]

Oryza is a remarkable genus comprising 27 species and 11 genome types, with ~3.4-fold genome size variation, that possesses a virtually untapped reservoir of genes that can be used for crop improvement and neodomestication. Here we present 11 chromosome-level assemblies (nine tetraploid, two diploid) in the context of ~15 million years of evolution and show that the core *Oryza* (sub)genome is only ~200 Mb and largely syntenic, whereas the remaining nuclear fractions (~80–600 Mb) are intermingled, plastic and rapidly evolving. For the halophyte *Oryza coarctata*, we found that despite detection of gene fractionation in the subgenomes, homoeologous genes were expressed at higher levels in one subgenome over the other in a mosaic form, demonstrating subgenome equivalence. The integration of these 11 new reference genomes with previously published genome datasets provides a nearly complete view of the consequences of evolution for genome diversification across the genus.

The genetic bottleneck imposed by thousands of years of domestication has inevitably impoverished the rice crop genetic diversity needed to adapt to the changing environment[1–3]. New solutions are needed to overcome current and future challenges in rice production and sustainability, in light of a predicted expansion of the human population to 10 billion by 2050 (ref. 4). To help reduce this bottleneck, we are exploring and exploiting the standing genetic diversity of the genus *Oryza*. The genus includes Asian and African cultivated rice and 25 wild species (that is, 15 diploid genomes with $2n = 2x = 24$ chromosomes and ten allotetraploid genomes with $2n = 4x = 48$ chromosomes) spanning 11 extant genome types (AA, BB, CC, BBCC, CCDD, EE, FF, GG, HHJJ, HHKK and KKLL) defined on the basis of cytogenetic[5] and/or molecular and phylogenetic evidence[6]. The *Oryza* species collectively encompass ~15 million years (Myr) of evolutionary history[6–8] and represent a crucial resource for tolerance and resistance traits that could be harnessed for crop improvement and/or serve as the raw material for neodomestication[9,10].

Since the late 1990s, genetic and genomic characterizations of the genus *Oryza* have advanced from single gene trait discovery and

---

[1]Biological and Environmental Sciences and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [2]Biosystematics Group, Wageningen University and Research, Wageningen, the Netherlands. [3]Research and Development, Corteva Agriscience, Johnston, IA, USA. [4]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ, USA. [5]Laboratoire Génome et Développement des Plantes, UMR 5096 CNRS/IRD 52, Université de Perpignan, Perpignan, France. [6]EMR MANGO Université de Perpignan/CNRS/IRD, Perpignan, France. [7]Rice Breeding Innovations Department, International Rice Research Institute (IRRI), Los Baños, Philippines. [8]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan, China. [9]Institute of Crop Science, Scuola Superiore Sant'Anna, Pisa, Italy. [10]Present address: National Center for Vegetation Development and Combating Desertification (NCVC), Riyadh, Saudi Arabia. e-mail: alice.fornasiero@kaust.edu.sa; andrea.zuccolo@kaust.edu.sa; eric.schranz@wur.nl; rod.wing@kaust.edu.sa

cloning[11–13] to restriction fragment length polymorphism[14] and physical mapping[15], the first reference genome sequences of rice[16–18] and a set of chromosome-level reference genomes representing the 15 distinct subpopulations of Asian rice[19]. The release of a chromosome-level reference genome of *Oryza alta* and an optimized protocol for neodomestication of this species through editing of key domestication genes has led to a new era in improvement of polyploid cereal crops[20]. Next steps include the generation of a complete digital GeneBank for cultivated rice as well as a set of ultra-high-quality reference genomes of the wild relatives of rice, including all *Oryza* tetraploid species[21].

Here, we report the generation and interrogation of 11 chromosome-level reference genomes, from nine underutilized wild tetraploid *Oryza* species[22,23] (*Oryza malampuzhaensis* (BBCC), *Oryza minuta* (BBCC), *O. alta* (CCDD), *Oryza grandiglumis* (CCDD), *Oryza latifolia* (CCDD), *Oryza coarctata* (KKLL), *Oryza schlechteri* (KKLL[HHKK]), *Oryza longiglumis* (HHJJ) and *Oryza ridleyi* (HHJJ)) and two wild diploid species (*Oryza australiensis* (EE) and *Oryza meyeriana* (GG)), using PacBio long-read sequencing technology and Bionano optical validation mapping. This dataset, combined with ten previously published diploid reference genomes, was used to describe how genome size and composition have evolved across the species in the genus, showing that some species are more malleable than others. The role of transposable elements (TEs) in shaping genome size was particularly evident in the *ridleyi* complex (*O. ridleyi* and *O. longiglumis*, both HHJJ), in which the differential expansion of a collection of TEs produced a striking size variation of the homoeologous subgenomes. We revisited a previous phylogenetic analysis of the *Oryza* genus[24] by adding new evidence of the relationships among subgenomes. Based on our molecular and phylogenetic evidence, which demonstrates high similarity between the *O. schlechteri* and *O. coarctata* genomes, we propose changing the genome type designation of *O. schlechteri* from HHKK to KKLL. We performed synteny analysis at both macro and micro scales to define major chromosomal rearrangements and gene presence/absence variation in the wild species with respect to the AA genomes, explored the extent of gene fractionation[25] in the subgenomes after polyploidization, and investigated the phenomena of subgenome dominance and equivalence in *O. coarctata*.

The release of 11 new chromosome-level reference genomes, in combination with previously published genome assemblies, represents a comprehensive and unique opportunity for future research in the fields of evolutionary biology, functional genomics, population genetics and biodiversity conservation of *Oryza* and could provide a robust instrument for neodomestication of climate-adapted rice crops[26–28].

Results

The wild *Oryza* genome data package

We generated chromosome-level reference assemblies for nine tetraploid and two diploid wild *Oryza* species (Extended Data Table 1 and Supplementary Note 1) and extensively assessed them for quality and contiguity (with Bionano optical maps, genome assembly evaluation using BUSCO[29] and the genome mapping rate) as described in Supplementary Note 2 and Supplementary Tables 1–3. These assemblies were annotated for genes and TEs using a uniform annotation pipeline (Extended Data Table 2, Supplementary Note 2 and Supplementary Tables 4–7). BioSample, BioProject, and Sequence Read Archive (SRA) accession numbers are provided in Supplementary Data 1, and code is available at https://github.com/alicefornasiero/IOMAP-3/tree/main.

TE dynamics in the *ridleyi* complex

Genome size across the *Oryza* genus varied ~3.4-fold, with the polyploids in the *ridleyi* complex being the largest (that is, the HHJJ genome species *O. longiglumis* (1,147 Mb) and *O. ridleyi* (1,203 Mb)) (Fig. 1 and Extended Data Table 2). Subgenome size was strikingly different in these species, with the HH subgenome showing ~1.5-fold variation with respect to the JJ subgenome (Supplementary Table 5). Analysis of TEs showed that size variation in the subgenomes of *O. longiglumis* and *O. ridleyi* could be attributed to a difference in TE abundance (Fig. 2). The ratios of TE content over non-TE content in HH subgenomes were 1.87 and 1.90 for *O. longiglumis* and *O. ridleyi*, respectively, whereas the same ratios for the JJ subgenomes were 1.21 and 1.25 for *O. longiglumis* and *O. ridleyi*, respectively (Supplementary Table 5). To investigate the preferential expansion of TEs in the HH subgenomes, we investigated the distribution of TEs belonging to six *Oryza*-specific superfamilies (CACTA, Ty1/Copia, Ty3/Gypsy, MuDR, hAT and LINE) in each subgenome and generated neighbor-joining trees. This analysis did not reveal any evidence of preferential expansion of a TE superfamily over others (Supplementary Figs. 1 and 2). In these species, the majority of long terminal repeat retrotransposons (LTR-RTs) were amplified after polyploidization, estimated ~2.25 million years ago (Ma)[24]: 83.4% and 83.7% in the HH and JJ subgenomes of *O. ridleyi*, respectively; and 71.9% and 76.1% in the HH and JJ subgenomes of *O. longiglumis*, respectively (Supplementary Fig. 3). We then determined whether the variation in TE content in the subgenomes was due to a differential rate of either TE accumulation or TE removal in one of the two subgenomes. Unequal recombination and illegitimate recombination serve as mechanisms for LTR-RT elimination, the former generating solo LTRs by recombining LTRs within or between different LTR-RTs, and the latter acting on dissimilar DNA sequences to remove sections of TE sequences and occasionally leaving incomplete elements[30,31]. To assess TE removal efficacy in the subgenomes of *O. ridleyi* and *O. longiglumis*, we calculated the ratio of solo LTRs to complete LTR-RTs; we found no relevant difference between the HH and JJ subgenomes. In *O. ridleyi*, the ratios of solo LTRs to complete LTR-RTs in the HH and JJ subgenomes were 1.2 and 1.5, respectively. In *O. longiglumis*, the ratios were 1.1 and 1.2 in the HH and JJ subgenomes, respectively. These values were similar to those found for *Oryza sativa*[31,32]. This evidence confirms previous findings of El Baidouri and Panaud, who showed that the ratio of solo LTRs to complete LTR-RTs does not depend on genome type[32]. In summary, our results indicate that the difference in subgenome size in the HHJJ genome species is primarily due to preferential accumulation of LTR-RT-related sequences in the HH subgenome. The distribution of the main *Oryza*-specific TE families showed no preferential expansion of specific families, and the solo LTR to complete LTR-RT ratios showed no evidence of differential efficiency in TE removal, thereby favoring TE accumulation as the primary mechanism contributing to subgenome size disparity.

Macro-synteny and large-scale chromosomal rearrangements

To understand and visualize the syntenic relationships across the genus, we built a syntenic map of the entire *Oryza* genus (Fig. 1) that included 21 species (nine tetraploid and two diploid species from this study, plus ten publicly available diploid species, including the International Rice Genome Sequencing Project (IRGSP) reference sequence) and the outgroup species *Leersia perrieri* (Supplementary Table 8). The synteny map (shown as a riparian plot) tracks the syntenic homologous blocks across the 11 genome types (AA, BB, CC, DD, EE, KK, LL, HH, JJ, FF and GG), showing ~15 Myr of evolution in inversions, duplications and translocations across the genus (Fig. 1). The riparian plot in Supplementary Fig. 4 shows collinear syntenic blocks inverted in consecutive (sub)genome pairs (shown as blue ribbons). The small-scale segmental duplication[33] on chromosome 11 and 12 was shared by all *Oryza* species and *L. perrieri*. The *O. alta* (CCDD) and *O. grandiglumis* (CCDD) genomes shared five unbalanced translocations relative to the *O. sativa* genome (for example, t(Chr1CC; Chr3), t(Chr6CC; Chr1), t(Chr7CC; Chr4); t(Chr3DD; Chr6); t(Chr4DD; Chr7)) (Supplementary Fig. 5), adding molecular support to the hypothesis of their conspecific nature based on previous morphological, cytogenetic and distribution data[5]. None of these translocations was found in CCDD species *O. latifolia* or in any other species included in this study (Fig. 1). Large chromosomal rearrangements are often associated with

https://doi.org/10.1038/s41588-025-02183-5

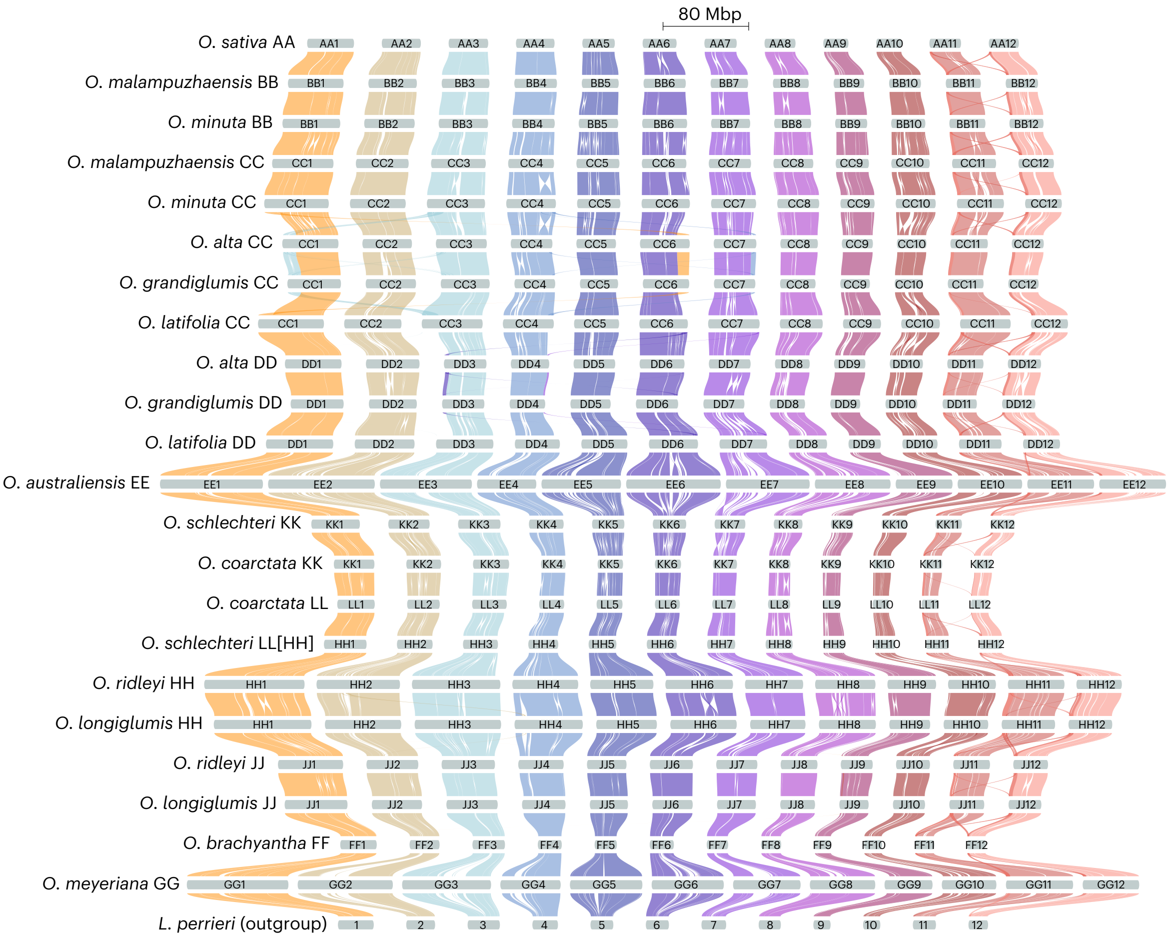


Fig. 1 | Overview of the syntenic landscape and large-scale structural rearrangements of 12 *Oryza* species (21 (sub)genomes) with the outgroup *L. perrieri*. Riparian plot showing macro-syntenic regions and large-scale structural rearrangements (large duplications and translocations) across the chromosomes of 12 *Oryza* species (21 (sub)genomes) and outgroup species *L. perrieri*. Genome types are shown according to the phylogenetic order in the genus, from the top (*O. sativa* (AA)) to the bottom (*O. meyeriana* (GG)). Each chromosome is colored as follows: Chr1, orange; Chr2, beige; Chr3, celeste; Chr4, steel blue; Chr5, navy blue; Chr6, deep purple; Chr7, plum; Chr8, magenta; Chr9, raspberry; Chr10, ruby; Chr11, coral; Chr12, salmon. Chromosomes are scaled by assembly length.

heterochromatic regions composed of repetitive DNA, such as satellites, minisatellites and simple sequence repeats[34–36]. When comparing the *O. alta* and *O. grandiglumis* genomes with an *O. sativa* reference genome, we found that nine of ten putative translocation breakpoints showed the presence of nearby simple sequence repeats, that is, AT repeats (Supplementary Figs. 6 and 7), suggesting that these sequences may have facilitated the occurrence of these chromosomal rearrangements. Reciprocal translocations between homoeologous chromosomes in polyploid genomes can be found by aligning subgenomes with each diploid relative genome species. When aligning the BBCC genome species with their diploid relative genome species (*Oryza punctata* (BB) and *Oryza officinalis* (CC)), we found a reciprocal translocation between Chr1BB and Chr1CC (that is, t(Chr1BB; Chr1CC)), as confirmed with optical maps, in both *O. minuta* (~9 Mb translocation size) and *O. malampuzhaensis* (~8 Mb translocation size) (Supplementary Fig. 8).

To provide further support for the structural description of the *Oryza* genus provided above, we generated genome-type-level pangenomes for the AA, BB, CC and DD genome types ('Data availability'). For each pangenome, we calculated the core pangenome size, which ranged from 180 Mbp to 270 Mbp (Extended Data Fig. 1a), providing strong support for our estimates obtained by calculating the non-TE-related content in individual genomes (Supplementary Note 2). Regarding large chromosomal rearrangements, the pangenomes for the CC and DD genome types showed (as expected) evidence of the large translocations identified by the macro-synteny analysis, further supporting the presence of these rearrangements (Extended Data Fig. 1b,c).

The syntenic pangenome

To identify core gene sets conserved during *Oryza* evolution, as well as accessory gene sets that underwent duplication, translocation and/or gene loss, we performed a micro-synteny analysis at the (sub) genome level (Fig. 3). A total of 832,658 gene sequences identified from the gene prediction of 30 *Oryza* (sub)genomes (Supplementary Table 8) and the outgroup *L. perrieri* were grouped into 77,482 syntenic gene clusters (groupings of syntenic homologous genes) and used to build a syntenic pangenome (Fig. 3a). For *O. alta* and *O. grandiglumis*, owing to chromosomal duplications and unbalanced translocations (as described above), underlying genes were also duplicated and translocated, replacing genes on the chromosomal portions that were lost. In Fig. 3a, the yellow (duplicated genes) and gray (depleted genes) tracks belonging to the same clusters of the dendrogram in either subgenome of *O. alta* and *O. grandiglumis* correspond to these genes (Supplementary Table 9). Congruent with the random occurrence of translocations, we could not detect overrepresented gene ontology (GO)-slim terms when comparing *O. sativa* homologs of either *O. alta* or *O. grandiglumis* genes duplicated in the CC (DD) subgenome and depleted in the DD (CC) subgenome with *O. sativa* homologs of genes in the CC (DD) subgenome.

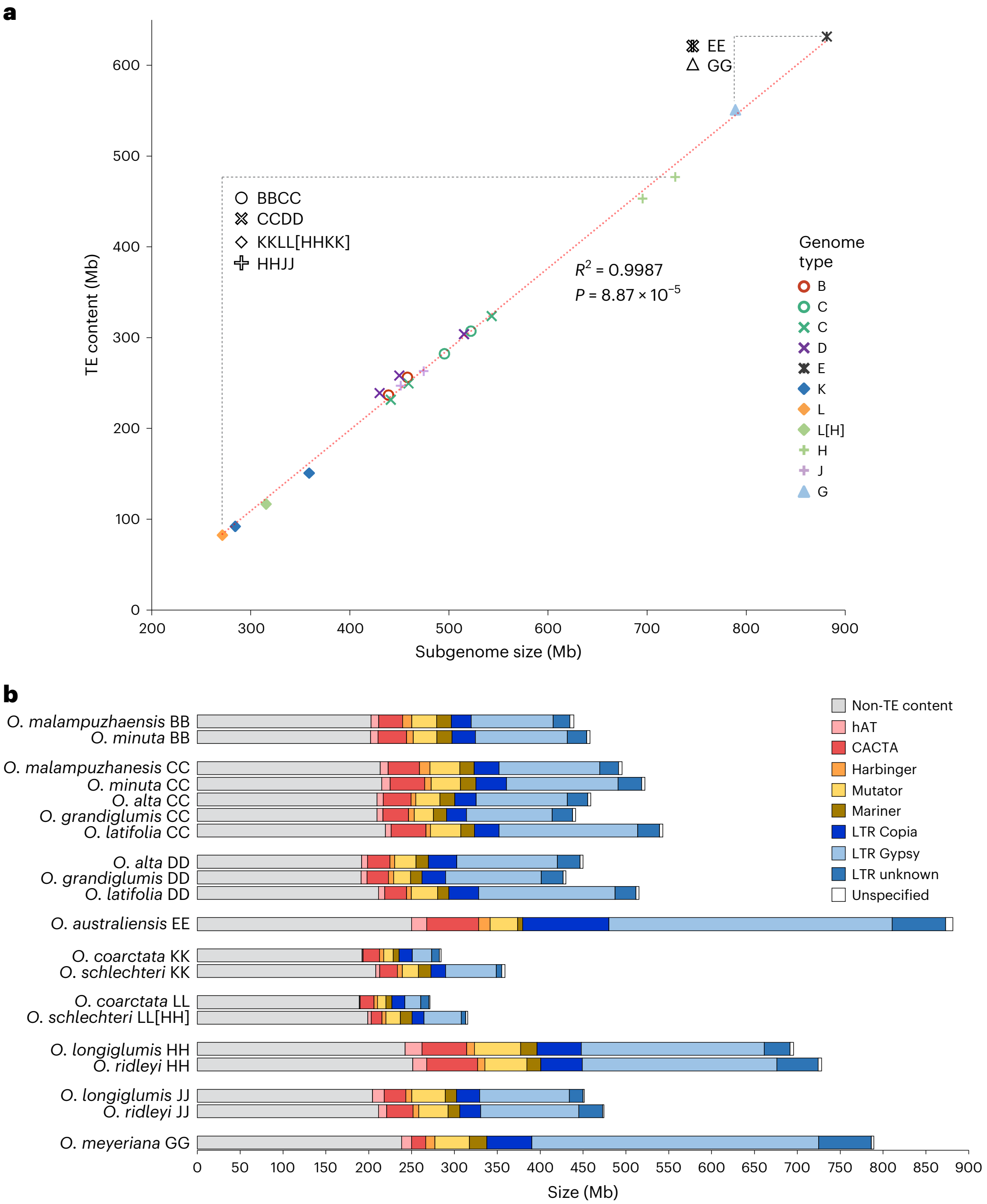


Fig. 2 | (Sub)genome size and TE content of the wild *Oryza* species. **a**, Correlation between (sub)genome size (Mb) and TE content (Mb) in the nine tetraploid and two diploid wild *Oryza* species. The significance of the linear correlation (Pearson's correlation coefficient, R^2) was ascertained by two-sided *t*-test. **b**, Abundance of the main classes of TEs (Mb). DNA transposons are shown as follows: hAT (DTA), pink; CACTA (DTC), red; Harbinger (DTH), orange; Mutator (DTM), yellow; Mariner (DTT), ochre. LTR retrotransposons are shown as: LTR Copia, dark blue; LTR Gypsy, light blue; LTR unknown, steel blue. Unspecified TEs are shown in white; non-TE content is shown in gray. (Sub)genomes of the species are ordered by genome type (BB, CC, DD, EE, KK, LL, HH, JJ, GG).

Syntenic core, softcore, dispensable and private clusters were defined as those present in all 30 (sub)genomes, in 27–29 (sub) genomes (≥90%), in 2–26 (sub)genomes and in one (sub)genome, respectively (Fig. 3b,c). We found 6,256 (8.1%), 9,865 (12.7%), 61,130 (78.9%) and 231 (0.3%) syntenic core, softcore, dispensable and private clusters, respectively (Fig. 3b). Functional enrichment analysis of GO-annotated genes showed that the most significantly overrepresented and underrepresented GO terms referred to expected functions for different portions of the syntenic pangenome (that is, the core portion contained genes with 'essential' functions, shared by all species, whereas the softcore and dispensable portions contained genes with 'malleable' functions that were partially shared) (Supplementary Table 10).

Reconstruction of the evolutionary history of *Oryza*

To gain insight into the evolutionary history and maternal origins of the genus *Oryza*, we first reconstructed a chloroplast-genome-based phylogenetic tree using the chloroplast sequences of 26 *Oryza* species (the ten chloroplast genomes assembled in this study and 16 chloroplast genomes obtained from NCBI) and the outgroup *Leersia japonica* (Supplementary Table 11). The chloroplast-genome-based phylogenetic tree showed that after divergence from *Leersia*, *Oryza* split into two main clades: a basal clade that includes the GG, FF and HHJJ genome types and a core clade consisting of all other genome types (Fig. 4a). Our tree was highly consistent with previously reported chloroplast-genome-based trees[37,38]. Of note, the former study did not include *O. coarctata* and *O. schlechteri* genomes, whereas the

https://doi.org/10.1038/s41588-025-02183-5

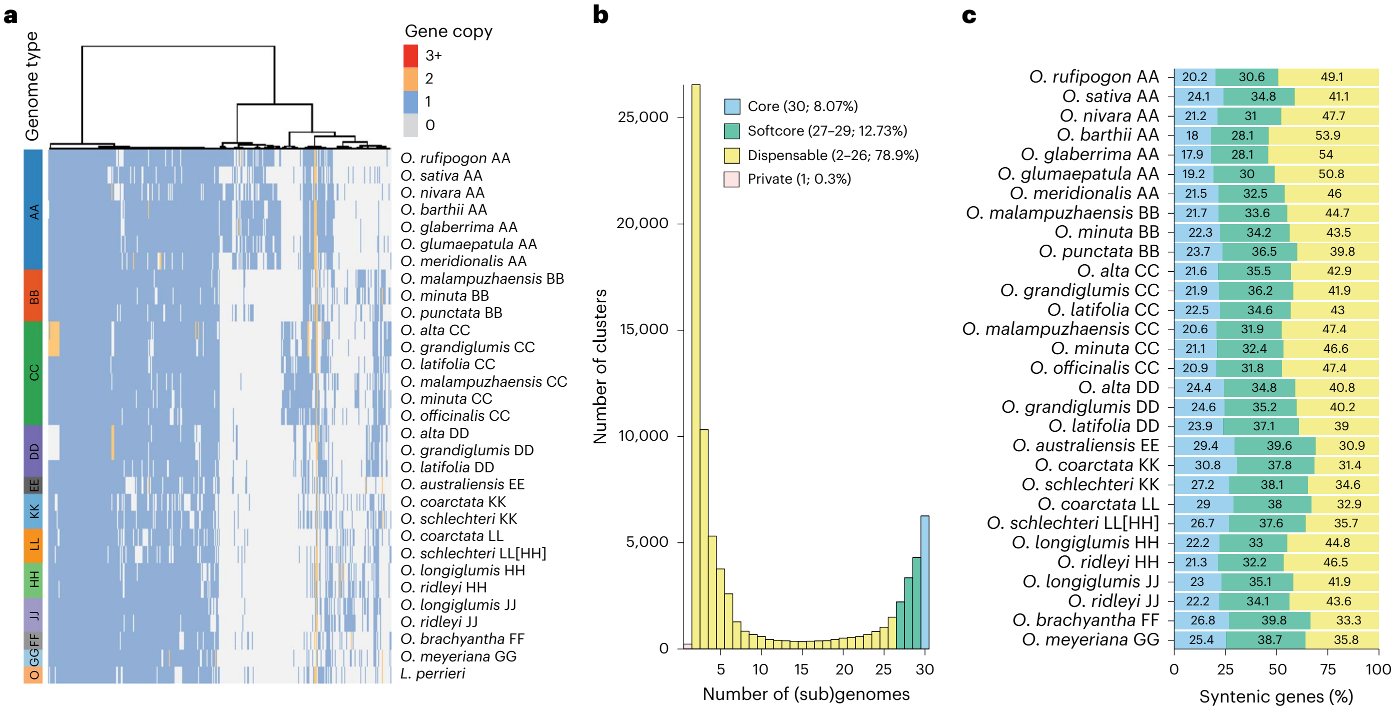


Fig. 3 | The syntenic pangenome. a, Phylogenomic profiling of clusters of syntenic genes across 30 *Oryza* (sub)genomes (the nine tetraploid and two diploid species presented here and ten additional diploid species listed in Supplementary Table 8) and outgroup species *L. perrieri*. In the heat map, each row represents a (sub)genome, and each column shows a syntenic cluster (that is, a grouping of syntenic homologous genes across two or more *Oryza* species and/or between *Oryza* and *L. perrieri*). The minimum number of genes in a cluster is two homologous genes in two different species or genera. Clustering of the *Oryza* (sub)genomes was based on presence and absence patterns of syntenic clusters using Euclidean distance and is shown as a dendrogram on top of the figure. For each syntenic cluster, gene copy-number variation is represented as follows: gene absence, light gray; one gene copy, blue; two gene copies, yellow; three or more gene copies, red. On the left side, vertical bars represent the genome type of the (sub)genomes: AA, dark blue; BB, red; CC, dark green; DD, purple; EE, black; KK, sky blue; LL, orange; HH, light green; JJ, lilac; FF, light gray; GG, light blue; O (outgroup), peach. **b**, Histogram showing the frequency distribution of syntenic clusters in the 30 (sub)genomes shown in **a** and shared by increasing numbers of *Oryza* (sub)genomes (*x* axis). The legend shows the percentage of core (found in all 30 subgenomes), softcore (found in 27–29 subgenomes), dispensable (found in 2–26 subgenomes) and private syntenic clusters. **c**, Percentages of genes classified in the different syntenic cluster categories (core (blue), softcore (green) and dispensable (yellow)) in the 30 *Oryza* (sub)genomes. Percentages of genes classified in private syntenic clusters are not shown.

latter merged the KKLL and HHKK genome types into a single HHKK type. Analogous to the tree presented by Zhang and colleagues[38], our chloroplast-based tree showed that *O. longiglumis* and *O. ridleyi* (both HHJJ genome types) and *O. schlechteri* (KKLL[HHKK] genome type) do not form a monophyletic group, supporting different maternal donors (Fig. 4a).

To better understand the evolutionary relationships of 21 of the 27 *Oryza* species for which a chromosome-level assembly was available, we performed coalescent phylogenetic analyses using 3,728 single-copy syntenic genes present in all 30 *Oryza* (sub)genomes and the outgroup. This gene set enabled grouping of the *Oryza* (sub)genomes into six highly supported monophyletic groups: AA, BB, CC, DD/EE, KK and LL/HH/JJ/FF/GG (Supplementary Fig. 9). The diploid CC genome (*O. officinalis*) clustered with the CC subgenomes of BBCC species *O. minuta* and *O. malampuzhaensis* but not with the CC subgenomes of CCDD species *O. alta*, *O. grandiglumis* and *O. latifolia*. This was consistent with previous studies[39,40] suggesting that a CC genome species (probably *O. officinalis* (CC)) is the paternal donor of the BBCC tetraploid species, and the maternal donor is a BB genome species (probably *O. punctata* (BB); Fig. 4a); whereas the CC genome in the CCDD tetraploid species served as the maternal parent and might be different from *O. officinalis* (Supplementary Fig. 9).

Coalescent phylogenetic analysis of the species in the unclassified group[41] (*O. coarctata* (KKLL) and *O. schlechteri* (KKLL[HHKK])) showed that the LL genome of *O. coarctata* (KKLL) clustered with the LL[HH] genome of *O. schlechteri* (KKLL[HHKK]), and the latter did not form a monophyletic group with the HH genomes of the HHJJ species (*O. ridleyi* and *O. longiglumis*) (Supplementary Fig. 9). This incongruity was investigated in greater detail at both the phylogenetic and molecular levels to explore the relationships of the HH, JJ, KK and LL genome types (Fig. 4b,c). The results of this analysis, described in Supplementary Note 3 and Supplementary Figs. 10–13, led us to conclude that the 'HH' subgenome of *O. schlechteri* should be renamed 'LL' (Fig. 5).

To date the *Oryza* phylogeny, we used the least square dating method[42] on a concatenated alignment of 528 single-copy genes and obtained a phylogeny that was highly consistent with that obtained with the coalescent approach (Fig. 4b). Using the molecular calibration of 14.5 Ma for the crown age of *Oryza* and 6 Ma for the divergence of the CC and AA-BB species[39,43], we estimated the divergence times of the genome types and inferred the hybridization times of the tetraploid genomes (Fig. 4b and Supplementary Table 12). Although Zou and colleagues suggested in 2015 that two independent polyploidization events represented the origins of *O. malampuzhaensis* and *O. minuta* (both BBCC species)[39], the grouping of the BB and CC subgenomes of these two species first, followed by grouping with either diploid relative genome (Fig. 4b), suggests a common origin of the BBCC species. Therefore, we considered the average between 1.53 Ma (the estimated divergence time of the BB subgenomes) and 1.92 Ma (the estimated divergence time of the CC subgenomes) to infer a polyploidization time of 1.73 Myr for the BBCC species. The three CCDD species (*O. alta*, *O. grandiglumis* and *O. latifolia*) are assumed to have originated from a single hybridization event[44,45]. We therefore considered

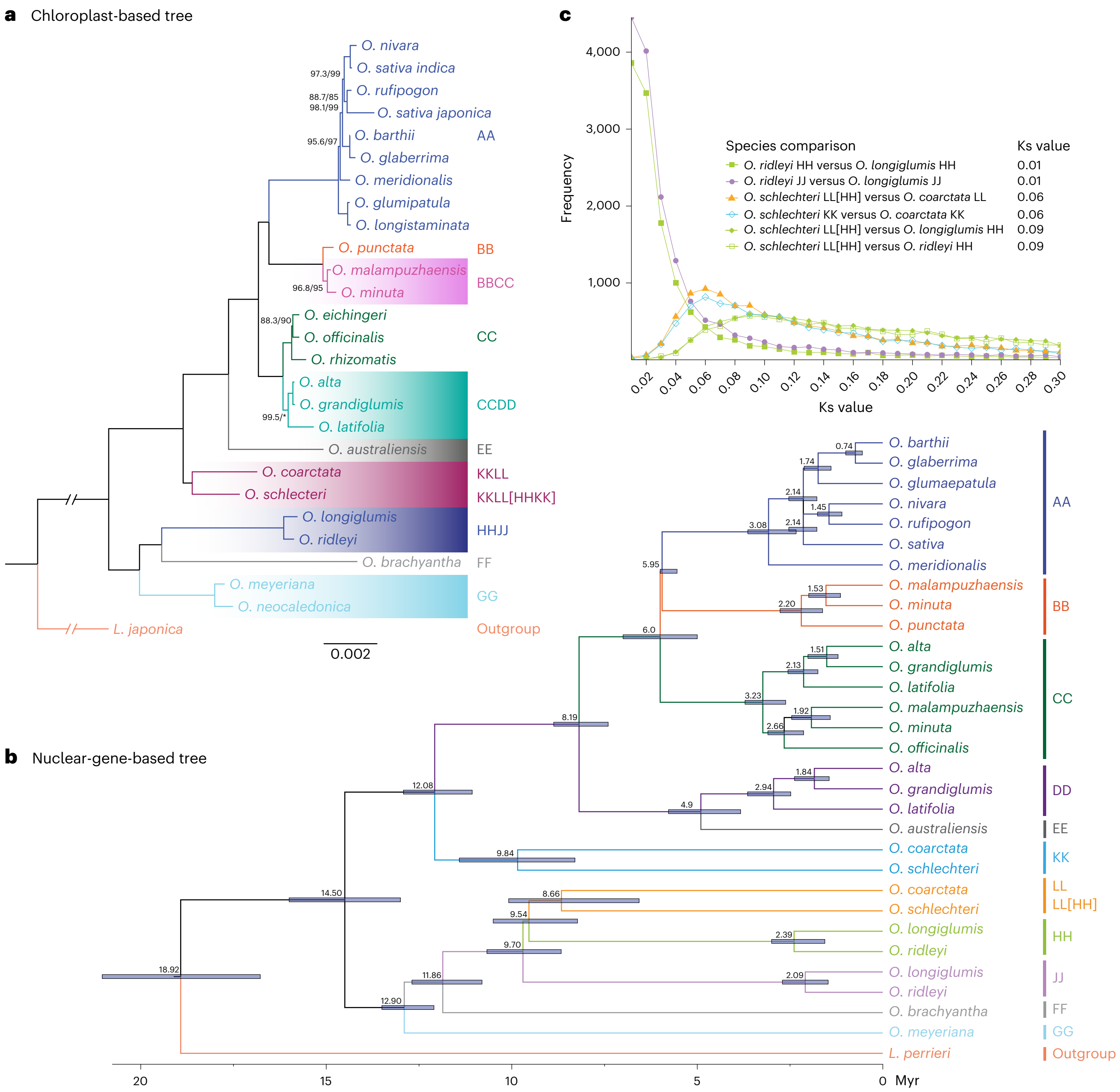


Fig. 4 | Phylogenetic analysis of the *Oryza* (sub)genomes. a, Phylogenetic tree based on chloroplast genome sequences. IQ-TREE was used to reconstruct a maximum likelihood phylogeny using the large-single-copy regions of the chloroplast genomes of 26 *Oryza* species (the chloroplast genomes of the ten species presented here and 16 additional chloroplast genomes of diploid species; Supplementary Table 11). Supporting values next to each branch are SH-aLRT (Shimodaira–Hasegawa-like approximate likelihood ratio) support (%)/ultrafast bootstrap support (%). 100%/100% support values are not shown. Branch length indicates substitutions per site. Trees were rooted using *L. japonica* as outgroup. **b**, Time-dated phylogenetic tree based on nuclear gene sequences. The phylogeny was inferred using the maximum likelihood method with a concatenated alignment of 528 single-copy genes. Phylogenetic dating was obtained using the molecular calibration for the crown age of *Oryza* (14.5 Ma) and the divergence of CC and AA-BB (6 Ma)[39,43]. **c**, Ks value distribution plot for HH, JJ, KK and LL genome types (*O. ridleyi* JJ versus *O. longiglumis* JJ, closed purple circle; *O. ridleyi* HH versus *O. longiglumis* HH, closed green square; *O. schlechteri* LL[HH] versus *O. coarctata* LL, closed orange triangle; *O. schlechteri* KK versus *O. coarctata* KK, open blue diamond; *O. schlechteri* LL[HH] versus *O. longiglumis* HH, green closed diamond; *O. schlechteri* LL[HH] versus *O. ridleyi* HH, open green square). The genome types used in the phylogenetic trees refer to definitions based on cytogenetic and hybridization experiments[5] and the molecular evidence provided here for the renaming of *O. schlechteri* from HHKK to KKLL genome type.

the average between 2.13 Ma (the estimated divergence time of the CC subgenomes) and 2.94 Ma (the estimated divergence time of the DD subgenomes) to infer a polyploidization time of 2.54 Myr for the CCDD species. *O. ridleyi* and *O. longiglumis* (both HHJJ species) are closely related and diverged quite recently[46]. For these species, we considered the average between 2.39 Ma (the estimated divergence time of the HH subgenomes) and 2.09 Ma (the estimated divergence time of the JJ subgenomes) to infer a polyploidization time of 2.24 Myr. The divergence times of the HH, JJ, KK and LL subgenomes suggested that the formation of the tetraploid genomes of *O. coarctata* (KKLL) and

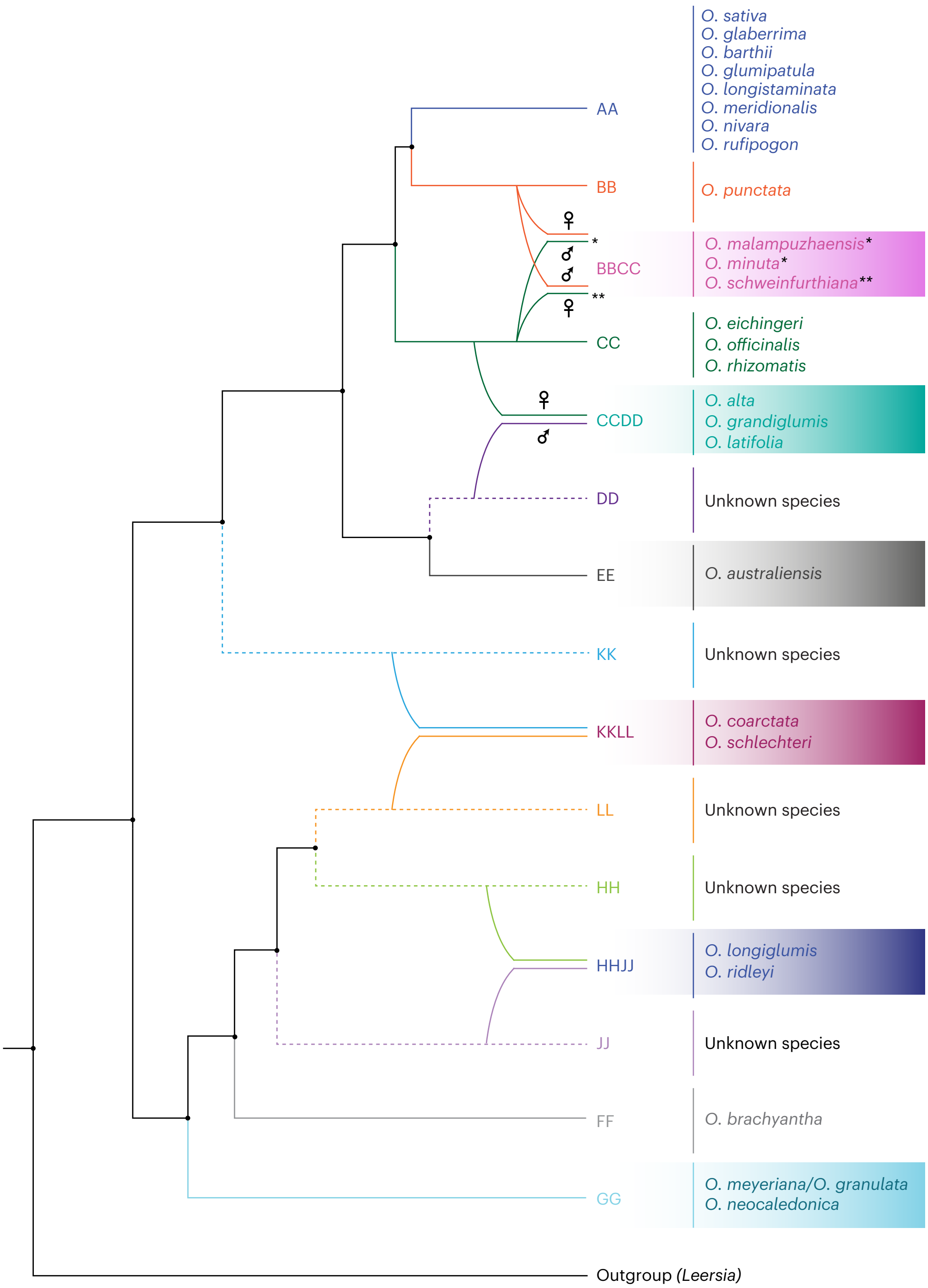


Fig. 5 | Consensus tree summarizing origins and evolutionary history of diploid and tetraploid *Oryza* species. Single-line branches denote diploid species, whereas double-line branches denote tetraploid species. Single dashed lines represent unknown diploid wild relative species. Forward slash (/) indicates that the species names—*O. meyeriana* and *O. granulata*—are considered synonyms (https://powo.science.kew.org). One asterisk (*) denotes that the maternal donor is a BB genome species. Two asterisks (**) denote that the maternal donor is a CC genome species. Genome types and known representative species are shown next to the terminal nodes. The relative times of hybridization events are based on the current study. The tree includes the new designation of *O. schlechteri* as KKLL genome type (the same as *O. coarctata*) proposed in this work. *L. perrieri* and *L. japonica* (here collectively referred to as *Leersia*) are the outgroups.

O. schlechteri (KKLL[HHKK]) occurred much earlier than that of the HHJJ genomes (Fig. 4b).

According to estimates of LTR-RT insertion times (Supplementary Fig. 3), amplification of these retroelements occurred mostly after the polyploidization event in the allotetraploids, with the majority of LTR-RTs (ranging from 77.7% in the LL subgenome of *O. coarctata* to 98.7% in the DD subgenome of *O. latifolia*) having inserted in the past 3 Myr.

Homoeologous gene fractionation

Following whole-genome duplication (via allopolyploidization, the hybridization of two or more distinct species; or autopolyploidization, the multiplication of a complete chromosome set within a species), gene copies can be lost from one homoeologous chromosome or the other(s), resulting in gene fractionation. Over evolutionary time, gene fractionation leads to reduction of a polyploid genome back to a diploid state in which the overall genomic structure has changed

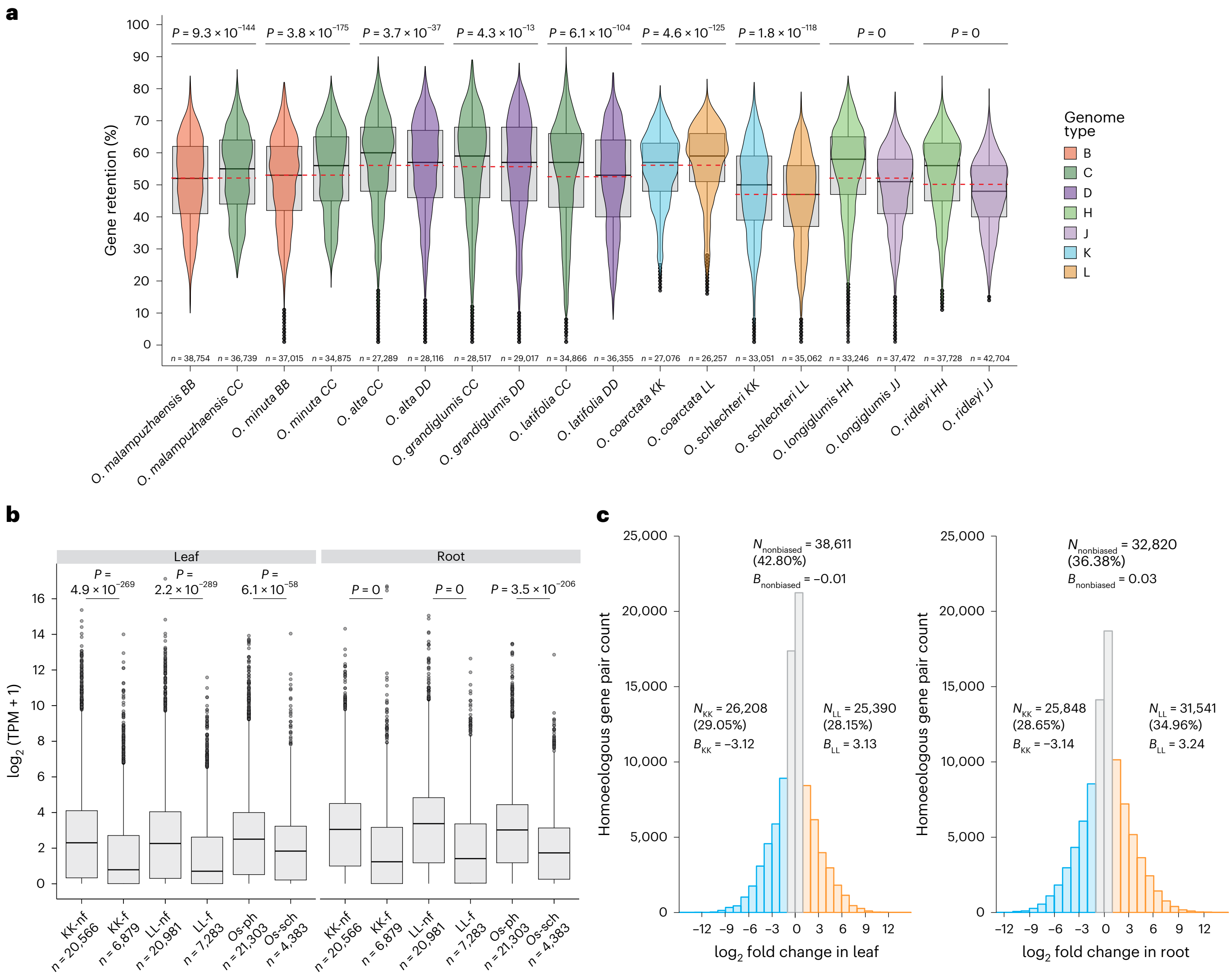


Fig. 6 | Homoeologous gene retention in *Oryza* and subgenome equivalence in *O. coarctata*. a, Distribution of gene retention (percentage, *y* axis) in the subgenomes of the tetraploid species (*x* axis). Each genome type is colored as in Figs. 3a and 4b. The red dashed line indicates the average percentage of gene retention calculated genome-wide for each species. *P* values from two-sided Wilcoxon rank-sum tests and numbers of sliding windows (*n*) are shown. **b**, Transcript abundance of homoeologous genes in *O. coarctata* and their homologs in *O. sativa*. Gene expression as $\log_2$ (TPM + 1) was measured in the leaf and in the root considering the replicates together. *P* values from two-sided Wilcoxon rank-sum tests are shown. In **a** and **b**, the 50th percentiles are defined by middle lines; lower and upper hinges correspond to the 25th and 75th percentiles. The upper whisker extends from the hinge to the largest value at most 1.5 times the interquartile range from the hinge. The lower whisker extends from the hinge to the smallest value at most 1.5 times the interquartile range of the hinge. Data beyond the end of the whiskers were considered to be outliers and plotted as individual points. **c**, Homoeologous gene pair expression bias (*B*) in the leaf (left) and the root (right) of *O. coarctata*. Blue and orange bars represent the expression of homoeologs biased toward KK ($B < -1$) and LL ($B > 1$) subgenomes, respectively. Homoeolog pairs with $-1 \leq B \leq 1$ (gray bars) are defined as nondominantly expressed. *N* represents the number of homoeologous gene pairs in the three categories (N_{KK}, homoeologous gene dominantly expressed in KK subgenome; N_{LL}, homoeologous gene dominantly expressed in LL subgenome; $N_{nonbiased}$, homoeologous gene not dominantly expressed). B_{KK}, B_{LL} and $B_{nonbiased}$ represent average expression bias for the homoeologous pairs in the respective categories. nf, nonfractionated (homoeologous gene pairs); f, fractionated genes; Os-ph, *O. sativa* genes homologous to nonfractionated *O. coarctata* genes (paired homologous); Os-sch, *O. sativa* genes homologous to fractionated *O. coarctata* genes (single-copy homologous).

substantially[47]. Gene fractionation (Fig. 6a) was measured as a percentage of homoeologous gene retention in the subgenomes of the tetraploid *Oryza* genomes. A lower percentage of gene retention with respect to the genome-wide average indicates overfractionation (greater gene loss), whereas a higher percentage indicates underfractionation (greater gene retention). Statistical comparisons (using two-sided Wilcoxon rank-sum tests) of gene retention between subgenomes showed the most pronounced difference within *O. longiglumis* and *O. ridleyi*, with higher gene retention in the HH subgenome, followed by the species in the *officinalis* complex, with higher gene retention in the CC subgenome ($P < 0.001$; Fig. 6a).

We used the total average percentage of gene retention and the inferred polyploidization time (Fig. 6a and Supplementary Table 12) of the BBCC, CCDD and HHJJ species to estimate the average gene loss per million years for these genome types (Supplementary Fig. 14). Gene loss was faster in younger species (the BBCC species that originated ~1.73 Ma) and progressively slowed in older species (the HHJJ and CCDD species that originated ~2.24 Ma and ~2.54 Ma, respectively).

Subgenome dominance and equivalence in *O. coarctata*

Subgenome dominance is a widely observed phenomenon in polyploid plant genomes (for example, in *Brassica rapa*[48], monkeyflower[49]

(*Mimulus peregrinus*), maize[50] (*Zea mays*), cotton[51] (*Gossypium*) and switchgrass[52] (*Panicum virgatum*)) in which genes from one subgenome tend to be expressed at higher levels than those from the homoeologous subgenome. Over evolutionary time, the less expressed subgenome (that is, the submissive subgenome) tends to lose more homoeolog copies than the more highly expressed subgenome, generating biased fractionation. Alternatively, subgenome equivalence means neither genome is 'dominant' over the other, and gene loss occurs at an approximately equal extent in the subgenomes[47]. Absence of biased fractionation and subgenome dominance has been observed in *Capsella bursa-pastoris*[53], pear[54] (*Pyrus bretschneideri*) and Chinese sprangletop[55] (*Leptochloa chinensis*). *Cucurbita* ssp. *maxima* and *moschata* (pumpkin and gourd) exhibit biased gene-loss patterns in some chromosomes and lack of subgenome dominance[56]. Similarly, the allotetraploid broomcorn millet (*Panicum miliaceum*) shows subtle biased gene fractionation and no significant dominance of either subgenome[57]. To investigate these phenomena in the *Oryza* tetraploids, we analyzed gene expression patterns in *O. coarctata* (KKLL), owing to its importance as a halophytic species[58] and the availability of transcriptome data. The average gene retention in *O. coarctata* was 56.1%, meaning that, on average, ~56% of genes were retained in duplicate and were syntenic in the homoeologous chromosomes (Supplementary Fig. 15). Overall, the KK subgenome showed overfractionation (that is, greater gene loss) with respect to the LL subgenome ($P < 0.001$, two-sided Wilcoxon rank-sum test; Fig. 6a).

To investigate the correlation between gene expression and gene loss, we analyzed the transcript abundance of nonfractionated genes (homoeologous gene pairs) and fractionated genes in *O. coarctata* and compared it with that of paired homologous and single-copy homologous genes in *O. sativa*, in leaf and root tissues. The median gene expression in *O. coarctata* was significantly higher ($P < 0.001$, two-sided Wilcoxon rank-sum test) for homoeologous gene pairs than for fractionated genes (Fig. 6b and Supplementary Table 13). *O. sativa* genes homologous to nonfractionated *O. coarctata* genes (*O. sativa* paired homologous genes) showed significantly higher expression levels ($P < 0.001$, two-sided Wilcoxon rank-sum test) than *O. sativa* genes homologous to fractionated *O. coarctata* genes (*O. sativa* single-copy homologous genes) (Fig. 6b and Supplementary Table 13). This evidence suggests that in *O. coarctata*, genes with two homoeologous copies tend to be expressed at higher levels than genes that have undergone fractionation. Moreover, *O. sativa* genes homologous to *O. coarctata* nonfractionated genes tended to be expressed at higher levels than *O. sativa* genes homologous to *O. coarctata* fractionated genes. These results suggest that in *O. coarctata*, genes with low expression tend to be preferentially lost compared with highly expressed genes during the process of rediploidization over evolutionary time. Similar analyses and results have been reported for broomcorn millet[57], for which the expression of homoeologous genes was compared with that of homologous genes in the diploid wild relative *Panicum hallii*.

We then compared expression in homoeologous gene pairs to investigate whether there was subgenome dominance in the tissue-specific transcriptome data of *O. coarctata*. Among the dominantly expressed homoeologous gene pairs, a higher fraction was dominantly expressed in the KK subgenome in leaf and in the LL subgenome in root (one-sided binomial test, $P < 0.001$) (Fig. 6c and Supplementary Table 14).

Taken together, the results of our analyses show that despite greater gene retention in the LL subgenome of *O. coarctata* (suggesting biased gene fractionation), homoeologous genes are expressed at higher levels in one subgenome over the other in a mosaic form, suggesting subgenome expression equivalence[59]. Additional tissues need to be analyzed to provide further evidence of subgenome equivalence in *O. coarctata*.

Discussion

We have generated a comprehensive resource of publicly available wild diploid and tetraploid *Oryza* reference genomes spanning all tetraploid genome types and the EE and GG diploid genome types, using PacBio long-read sequencing, optical maps and long/short read back-mapping validation. Previous efforts in this direction were made within the framework of the International *Oryza* Map Alignment Project (IOMAP[60]).

Analysis of our dataset resulted in insights into genome evolution within the genus. We generated a macro-synteny description of the *Oryza* (sub)genomes and built and characterized a syntenic pangenome. We characterized large chromosomal rearrangements that resulted in our present-day inventory of living *Oryza* species. For example, the description of five large nonreciprocal translocations shared between *O. alta* (CCDD) and *O. grandiglumis* (CCDD) but not present in *O. latifolia* (CCDD) added robust evidence to a previous hypothesis of their conspecific nature[38,45]. Given that *O. alta* is considered to be a synonym for *O. latifolia* in the Plants of the World Online (https://powo.science.kew.org), and intermediate forms between *O. alta* and *O. grandiglumis* have been observed[61,62], genetic studies at the population level are needed to investigate the history and composition of these species and potentially resolve their taxonomic classification[63]. According to estimation of LTR-RT insertion time in the allotetraploids, TE amplification occurred mostly after polyploidization. The distribution of LTR-RT insertion events in *O. coarctata* (KKLL) and *O. longiglumis* (HHJJ) showed a higher fraction of older LTR-RT elements in these two species compared with the other *Oryza* species. This could be due to a slower mutation rate and/or a less efficient TE removal mechanism in these species. This analysis corroborates previous findings[32] showing that LTR-RT insertion time and accumulation are not dependent on genome type. We analyzed an interesting example of size variation between homoeologous genomes in the species of the *ridleyi* complex and identified TEs as the driving force of genome size change, thereby clarifying the mode of amplification in these species—that is, the involvement of the entire TE complement, as opposed to the rapid explosion of a few TE families as shown previously in *O. australiensis*[64] (EE) and *Oryza granulata*[65] (GG).

We retested the phylogenetic placement of the genomes, confirming previous findings[7,37,40,44,66]. We also estimated divergence times of the subgenomes and inferred polyploidization timing for the BBCC, CCDD and HHJJ species. Our whole-genome-scale phylogenetic analysis provided robust confirmation of previous inferences (based on the analysis of a few genes) regarding the origins of the HH, KK and JJ genome types, according to which *O. schlechteri* and *O. coarctata* are more closely related to each other and share the same genome type[44]. In addition, genome sequence similarity, TE abundance and distribution, and presence/absence patterns of syntenic genes clearly showed greater similarity at the molecular level between the LL[HH] subgenome of *O. schlechteri* and the LL subgenome of *O. coarctata* compared with the other HH subgenomes. Our phylogenetic and molecular evidence is consistent with a previous report on *O. coarctata*[6], the authors of which proposed the renaming of this species (previously designated an HHKK genome type) as a KKLL genome type. Thus, we recommend the renaming of *O. schlechteri* from an HHKK to a KKLL genome type, as shown in the phylogenetic tree in Fig. 5.

Our investigation of gene fractionation in the tetraploids found variable patterns of gene fractionation among subgenomes, as recently described in other plant systems[47]. Investigation of subgenome dominance/equivalence in leaf and root tissue of *O. coarctata* (KKLL) did not reveal evidence of expression dominance of one subgenome over another, even though gene fractionation was higher in subgenome KK. Additional transcriptome data for *O. coarctata* and the other tetraploid species will be needed to investigate the phenomenon of subgenome dominance/equivalence in *Oryza*.

This dataset provides a valuable resource for future investigations, which could include the discovery of adaptive genes and/or traits to improve cultivated rice, the neodomestication of the wild *Oryza* species, and population genetics studies of wild *Oryza* species across their species ranges for conservation and enhancement of their genetic diversity for the planet's future[63].

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/s41588-025-02183-5.

References

1. Tanksley, S. D. & McCouch, S. R. Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**, 1063–1066 (1997).
2. Ross-Ibarra, J., Morrell, P. L. & Gaut, B. S. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl Acad. Sci. USA* **104**, 8641–8648 (2007).
3. Purugganan, M. D. An evolutionary genomic tale of two rice species. *Nat. Genet.* **46**, 931–932 (2014).
4. *World Population Prospects: The 2017 Revision* (United Nations, 2017); https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2017_world_population_prospects-2017_revision_databooklet.pdf
5. Nayar, N. M. Origin and cytogenetics of rice. *Adv. Genet.* **17**, 153–292 (1973).
6. Lu, F. et al. Comparative sequence analysis of MONOCULM1-orthologous regions in 14 *Oryza* genomes. *Proc. Natl Acad. Sci. USA* **106**, 2171–2076 (2009).
7. Ammiraju, J. S. S. et al. Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell* **20**, 3191–3209 (2008).
8. Ammiraju, J. et al. Spatio-temporal patterns of genome evolution in allotetraploid species of the genus *Oryza*. *Plant J.* **63**, 430–442 (2010).
9. 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).
10. Zhu, X. G. & Zhu, J. K. Precision genome editing heralds rapid de novo domestication for new crops. *Cell* **184**, 1133–1134 (2021).
11. Song, W.-Y. et al. A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* **270**, 1804–1806 (1995).
12. Shomura, A. et al. Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**, 1023–1028 (2008).
13. Ashikawa, I. et al. Two adjacent nucleotide-binding site–leucine-rich repeat class genes are required to confer *Pikm*-specific rice blast resistance. *Genetics* **180**, 2267–2276 (2008).
14. Xiao, J. et al. Identification of trait-improving quantitative trait loci alleles from a wild rice relative, *Oryza rufipogon*. *Genetics* **150**, 899–909 (1998).
15. Chen, M. et al. An integrated physical and genetic map of the rice genome. *Plant Cell* **14**, 537–545 (2002).
16. International Rice Genome Sequencing Project & Sasaki, T. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
17. Yu, J. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
18. Goff, S. A. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
19. Zhou, Y. et al. A platinum standard pan-genome resource that represents the population structure of Asian rice. *Sci. Data* **7**, 113 (2020).
20. Yu, H. et al. A route to de novo domestication of wild allotetraploid rice. *Cell* **184**, 1156–1170.e14 (2021).
21. Wing, R. A. Harvesting rice's dispensable genome. *Genome Biol.* **16**, 217 (2015).
22. Vaughan, D. A., Morishima, H. & Kadowaki, K. Diversity in the *Oryza* genus. *Curr. Opin. Plant Biol.* **6**, 139–146 (2003).
23. Khush, G. S. in *Oryza: From Molecule to Plant* (eds Sasaki, T. & Moore, G.) 25–34 (Springer, 1997); https://doi.org/10.1007/978-94-011-5794-0_3
24. Jacquemin, J., Bhatia, D., Singh, K. & Wing, R. A. The International *Oryza* Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Curr. Opin. Plant Biol.* **16**, 147–156 (2013).
25. Langham, R. J. et al. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* **166**, 935–945 (2004).
26. Mussurova, S., Al-Bader, N., Zuccolo, A. & Wing, R. A. Potential of platinum standard reference genomes to exploit natural variation in the wild relatives of rice. *Front. Plant Sci.* **11**, 579980 (2020).
27. Fornasiero, A., Wing, R. A. & Ronald, P. Rice domestication. *Curr. Biol.* **32**, R20–R24 (2022).
28. Gutaker, R. M. et al. Scaling up neodomestication for climate-ready crops. *Curr. Opin. Plant Biol.* **66**, 102169 (2022).
29. Simão, 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).
30. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075–1079 (2002).
31. Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869 (2004).
32. El Baidouri, M. & Panaud, O. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol. Evol.* **5**, 954–965 (2013).
33. Yu, J. et al. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38 (2005).
34. Badaeva, E. D. et al. Chromosomal rearrangements in wheat: their types and distribution. *Genome* **50**, 907–926 (2007).
35. Raskina, O., Barber, J. C., Nevo, E. & Belyayev, A. Repetitive DNA and chromosomal rearrangements: speciation-related events in plant genomes. *Cytogenet. Genome Res.* **120**, 351–357 (2008).
36. Molnár, I., Cifuentes, M., Schneider, A., Benavente, E. & Molnár-Láng, M. Association between simple sequence repeat-rich chromosome regions and intergenomic translocation breakpoints in natural populations of allopolyploid wild wheats. *Ann. Bot.* **107**, 65–76 (2011).
37. Gao, L.-Z. et al. Evolution of *Oryza* chloroplast genomes promoted adaptation to diverse ecological habitats. *Commun. Biol.* **2**, 278 (2019).
38. Zhang, W. et al. DNA barcoding of *Oryza*: conventional, specific, and super barcodes. *Plant Mol. Biol.* **105**, 215–228 (2021).
39. Zou, X.-H. et al. Multiple origins of BBCC allopolyploid species in the rice genus (*Oryza*). *Sci. Rep.* **5**, 14876 (2015).
40. 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).
41. Brar, D. S. & Khush, G. S. in *The Wild Oryza Genomes* (eds Mondal, T. K. & Henry, R. J.) 1–25 (Springer International Publishing, 2018).
42. To, T.-H., Jung, M., Lycett, S. & Gascuel, O. Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**, 82–97 (2016).
43. Zou, X. H., Yang, Z., Doyle, J. J. & Ge, S. Multilocus estimation of divergence times and ancestral effective population sizes of *Oryza* species and implications for the rapid diversification of the genus. *New Phytol.* **198**, 1155–1164 (2013).

44. 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).
45. Bao, Y. & Ge, S. Origin and phylogeny of *Oryza* species with the CD genome based on multiple-gene sequence data. *Plant Syst. Evol.* **249**, 55–66 (2004).
46. Nishikawa, T., Vaughan, D. A. & Kadowaki, K. Phylogenetic analysis of *Oryza* species, based on simple sequence repeats and their flanking nucleotide sequences from the mitochondrial and chloroplast genomes. *Theor. Appl. Genet.* **110**, 696–705 (2005).
47. Cheng, F. et al. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* **4**, 258–268 (2018).
48. Cheng, F. et al. Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS ONE* **7**, e36442 (2012).
49. Edger, P. P. et al. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* **29**, 2150–2167 (2017).
50. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci. USA* **108**, 4069–4074 (2011).
51. Yoo, M.-J., Szadkowski, E. & Wendel, J. F. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171–180 (2013).
52. Lovell, J. T. et al. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* **590**, 438–444 (2021).
53. Douglas, G. M. et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proc. Natl Acad. Sci. USA* **112**, 2806–2811 (2015).
54. Li, Q. et al. Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri* Rehd.). *Hortic. Res.* **6**, 34 (2019).
55. Wang, L. et al. Genomic insights into the origin, adaptive evolution, and herbicide resistance of *Leptochloa chinensis*, a devastating tetraploid weedy grass in rice fields. *Mol. Plant* **15**, 1045–1058 (2022).
56. Sun, H. et al. Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Mol. Plant* **10**, 1293–1306 (2017).
57. Sun, Y. et al. Biased mutations and gene losses underlying diploidization of the tetraploid broomcorn millet genome. *Plant J.* https://doi.org/10.1111/tpj.16085 (2023).
58. Mondal, T. K. et al. Draft genome sequence of first monocot-halophytic species *Oryza* coarctata reveals stress-specific genes. *Sci. Rep.* **8**, 13698 (2018).
59. Garsmeur, O. et al. Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.* **31**, 448–454 (2014).
60. Wing, R. A. et al. The Oryza Map Alignment Project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol. Biol.* **59**, 53–62 (2005).
61. Oliveira, G., Morishima, H. & Martins, P. *Investigations of Plant Genetic Resources in the Amazon Basin with the Emphasis on the Genus Oryza* Report of Study Tour 1992/93 (National Institute of Genetics, 1994).
62. Morishima, H. *Reports of the Study-Tours for Investigation of Wild and Cultivated Rice Species. Part II* (NBRP, 2002); https://shigen.nig.ac.jp/rice/oryzabase/ricereport/
63. Alsantely, A. et al. The International *Oryza* Map Alignment Project (IOMAP): the Americas—past achievements and future directions. *J. Exp. Bot.* **74**, 1331–1342 (2023).
64. Piegu, B. et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269 (2006).
65. Ammiraju, J. S. S. et al. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *Plant J.* **52**, 342–351 (2007).
66. 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).

Methods

Sample collection

Single seed descent germplasm for *O. alta* (IRGC 105143), *O. australiensis* (IRGC 100882), *O. grandiglumis* (IRGC 105669), *O. latifolia* (IRGC 100890), *O. longiglumis* (IRGC 106525), *O. malampuzhaensis* (IRGC 80765), *O. meyeriana* (IRGC 106473), *O. minuta* (IRGC 101141), *O. ridleyi* (IRGC 100821) and *O. schlechteri* (IRGC 82047) were obtained from the International Rice Research Institute (IRRI, Philippines) under the Standard Material Transfer Agreement. Seeds were sown in potting soil and grown at an air temperature of 24–29 °C with 15–25% humidity in a greenhouse. *O. coarctata* (IRGC 104502) leaf tissue was obtained from a vegetative voucher plant imported from IRRI through the USDA and grown at an air temperature of 24–29 °C with 15–25% humidity in the greenhouse at the University of Arizona.

Construction of genome-type-level pangenomes

Given the large sequence variation among distant *Oryza* species, pangenomes for the AA, BB, CC and DD genome types were generated separately (that is, genome-type-level pangenomes). The pangenomes were built using PanGenome Graph Builder[67], and each chromosome was considered individually. The settings used were as follows: for AA genomes, the similarity threshold was set to 90%, the *k*-mer length to 7 bp and the segment length to 15,000 bp; for the other genome types, the similarity threshold was lowered to 80%, with the other settings remaining the same. The core pangenome size statistics were calculated by applying the Panacus[68] tool to the GFA files (https://github.com/GFA-spec/GFA-spec). The AA pangenome included the species *Oryza barthii*, *Oryza glaberrima*, *Oryza glumaepatula*, *Oryza meridionalis*, *Oryza nivara*, *Oryza rufipogon* and *O. sativa*. The BB pangenome included the species *O. punctata*, and *O. malampuzhaensis* and *O. minuta* BB subgenomes. The CC pangenome included the species *O. officinalis*, and *O. malampuzhaensis*, *O. minuta*, *O. alta*, *O. grandiglumis* and *O. latifolia* CC subgenomes. The DD pangenome included *O. alta*, *O. grandiglumis* and *O. latifolia* DD subgenomes.

TE annotation

Extensive De-novo TE Annotator (EDTA[69] v.1.9.0) was used to generate a de novo nonredundant TE library for each genome, using default settings. De novo TE libraries were then used to annotate TEs in each respective genome using EDTA and to soft-mask genome sequences using RepeatMasker[70] (v.4.1.0) for the subsequent step of gene prediction.

Gene prediction

After evaluating read quality with FastQC (v.0.11.8) and removing adapters with Trimmomatic[71] (v.0.38), we aligned RNA sequencing (RNA-seq) reads to their respective genomes using the STAR[72] (v.2.7.8a) aligner embedded in the OmicsBox tool (Bioinformatics Made Easy, BioBam Bioinformatics, https://www.biobam.com/omicsbox, v.2.0.36). Minimap2 (ref. 73) aligner implemented in OmicsBox was used to align high-quality transcripts from full-length isoform sequencing (Iso-seq) reads. Gene prediction using extrinsic evidence was carried out on soft-masked genomes using Augustus[74] (v.3.4.0) software embedded in OmicsBox with a model training set derived for *O. sativa*, for all species except *O. australiensis*. For *O. australiensis*, MAKER-P[75] (v.3.01.03) was used to carry out gene prediction using RNA-seq data from leaf tissue. MAKER-P was run on the soft-masked genome of *O. australiensis* with the Augustus[74] (v.3.2.1), SNAP[76] (v.2013-11-29) and Fgenesh+[77] gene predictors. Genes and transcripts were retained if the annotation edit distance was less than 1.

For the macro-synteny and phylogenetic analyses, gene models in each genome species were filtered to retain the longest isoform using the AGAT tool suite (https://agat.readthedocs.io/en/latest/index.html). The resulting gene prediction for each genome species was evaluated for completeness using the poales_odb10 database in BUSCO[29] (v.5.1.2) with assessment mode 'proteins'.

Functional gene annotation

Functional gene annotations were generated for the 11-genome dataset presented in this work and ten additional diploid species obtained from NCBI and used to build the syntenic pangenome. Diamond blastp (v.2.1.8) was used to search the protein sequences in the NCBI nonredundant database (https://ftp.ncbi.nlm.nih.gov/blast/db/v5/) using an *e*-value threshold of 1×10^{-6}, a minimum sequence identity of 80% and a minimum subject cover of 40% to report an alignment. The best blast hit for each predicted protein sequence was retained. Functional information was then translated into GO terms using the Blast2GO[78] (v.6.0) suite. GO mapping of the protein blast hits onto the Gene Ontology Annotation database (v.2023.08) was performed (GO terms refer to the UniProt ID-mapping). GO annotation of the GO term candidates was performed using an annotation score cut-off of 75% to select the GO term for a given GO branch.

GO enrichment analysis

O. sativa homologs of either *O. alta* and *O. grandiglumis* genes duplicated in the CC(DD) subgenome and depleted in the DD(CC) subgenome were used for GO enrichment analysis in PANTHER (v.18.0) (https://www.pantherdb.org/). The reference gene set consisted of *O. sativa* homologs of either *O. alta* or *O. grandiglumis* genes in the CC(DD) subgenome. PANTHER GO-Slim annotations for each ontology (molecular function, biological process and cellular component) were assigned to test and reference gene sets. Fisher's exact test was used with a *P* value threshold of 0.001, and *P* values were corrected using the false discovery rate. For functional enrichment analysis of the syntenic pangenome, two-tailed Fisher's exact test with a *P* value threshold of 0.001 was used on the set of GO-annotated genes in each pangenome category (core, softcore and dispensable) and a reference set containing all GO-annotated genes in the pangenome. *P* values were corrected using the false discovery rate. The top ten overrepresented and underrepresented GO terms were then selected for each GO ontology.

Insertion time of LTR-RTs

The insertion time of complete LTR-RTs in the 11 genome species was estimated using the method proposed by SanMiguel and colleagues[79], which is based on the nucleotide distance between the two flanking LTR sequences of a complete LTR-RT. For each genome species, LTR sequences flanking complete LTR-RTs were aligned pairwise using the global aligner STRETCHER[80] from the EMBOSS tool suite (v.6.6.0). The nucleotide distance between the two LTR sequences was quantified using the Kimura 2p method[81] as implemented in the DISTMAT software[80] from EMBOSS (v.6.6.0). The nucleotide distance (D) was then converted to insertion time (T) in Myr using the formula: $T = \frac{D}{2\mu} \times 10^{6}$, where μ is the substitution rate of 1.3×10^{-8} per site per year[82]. Distributions were represented as percentages of complete LTR-RT insertions in bins of width 0.5 (Myr).

TE amplification in the homoeologous subgenomes of *O. ridleyi* and *O. longiglumis*

Conserved tracts of 100 amino acids from transposase and reverse transcriptase enzymes were used as queries in tBlastn[83] (v.2.14.0) to identify DNA TEs and retroelements, respectively, in the subgenomes of *O. ridleyi* and *O. longiglumis*. Five-hundred paralogs covering at least 80% of the query length were randomly selected among the tBlastn output hits for six TE superfamilies (Ty1/Copia and Ty3/Gypsy LTR retroelements, LINE and CACTA DNA TEs, MuDR and hAT) and aligned with their respective subgenomes using MUSCLE[84] (v.3.8.425). The multiple sequence alignments for each TE superfamily were then used to build a neighbor-joining tree using MegaX[85] (v.10.1). Bootstrap values were calculated in MegaX for 1,000 replicates using the pairwise deletion option and shown on the tree when greater than 50. Evolutionary distances were estimated using the Poisson correction distance[86].

Macro- and micro-synteny analysis

To track the genomic dynamics across the genus *Oryza*, we performed a macro-synteny analysis using GENESPACE[87] (v.1.3.1). This software implements and improves features from OrthoFinder[88] (v.2.5.4) and MCScanX_h[89] to link gene sequence homology with gene coordinates and traces the processes of genome polyploidization, reduction, rearrangement and translocation across a set of genomes. The protein sequences of the nine tetraploid and two diploid genomes reported in this study, as well as the publicly available protein-coding genes of ten diploid *Oryza* species and the outgroup *L. perrieri* (Supplementary Table 8), were analyzed in GENESPACE using default settings. The macro-synteny analysis was run with each tetraploid genome separated into two individual subgenomes (that is, the analysis was performed considering each genome type separately). The macro-synteny results were visualized as a riparian plot using the embedded plot_riparian function, and the homology and collinearity information at the subgenome level was used to build a synteny-constrained phylogenomic framework of the *Oryza* (sub)genomes.

Gene micro-synteny analysis across the *Oryza* genomes was performed using a modified version of a previously developed pipeline[90] in which pairwise synteny was inferred by GENESPACE and syntenet clusters were generated using custom developed scripts (https://github.com/xiaoyezao/Rice-Phylogenomics). Copy-number variation of syntenic homologs (a group of genes derived from a single common ancestor and retained in syntenic relationships) was profiled using the SYNTENET[91] (v.1.5.2) package, and visualized using heat maps. The syntenic homologs were then used for multilocus phylogenomic analysis to infer the phylogeny of the (sub)genomes, as described in the next section.

Phylogenetic relationship analysis of the chloroplast and nuclear genome across the *Oryza* genus

We performed a phylogenomic analysis using chloroplast-based and multinuclear-gene-based approaches to infer the evolutionary history of the *Oryza* species/genomes. For chloroplast-based phylogeny (https://github.com/nam-hoang/rice_cp_phylo_analysis), we used the chloroplast genome sequences of ten *Oryza* species (*O. malampuzhaensis*, *O. minuta*, *O. alta*, *O. grandiglumis*, *O. latifolia*, *O. coarctata*, *O. schlechteri*, *O. ridleyi*, *O. longiglumis* and *O. meyeriana*) assembled from whole-genome PacBio sequencing data (in this study), plus 17 publicly available chloroplast genome sequences from 16 diploid *Oryza* species and the outgroup *L. japonica* (Supplementary Table 11). To construct the maximum likelihood phylogenetic tree, we aligned the large-single-copy regions of the chloroplast genomes using MAFFT[92] (v.7.480). Poorly aligned regions were trimmed using trimAL[93] (v.1.4) with the option '-automated1'. The alignment files were subjected to IQ-TREE[94] (v.1.6.12) with default settings (1,000 bootstrap iterations) and with the best-fit substitution model identified by ModelFinder[95]. The resulting tree was visualized in FigTree (v.1.4.3) (http://tree.bio.ed.ac.uk/software/figtree/) and rooted using *L. japonica* as the outgroup. For nuclear-gene-based phylogeny, a nuclear gene dataset was obtained from the GENESPACE[87] output as described above. A total of 3,728 single-copy syntenic genes found in the 30 *Oryza* (sub)genomes and the outgroup genome were used for the phylogenomic analysis. Syntenic homologous genes were aligned using MAFFT[92] (v.7.520) (-genafpair; -maxiterate 1,000), and the alignments were cleaned using trimAl[93] (v.1.4.1) (-gt 0.6; -st 0.001). The gene trees were inferred using RAxML-NG[96] (v.1.2.0) with the Q.plants model[97]. For all gene trees, a coalescent algorithm in Astral-Pro2 (ref. 98) (v.1.16.1.3) was used to infer the phylogeny at the (sub)genome level. Branch supports of the Astral tree were estimated as local posterior probabilities[99].

To date the *Oryza* phylogeny, the least square dating method[42] as implemented in IQ-TREE[100] v.2.3.6 was used on a concatenated alignment of 528 single-copy genes. The substitution model was determined by ModelFinder implemented in IQ-TREE v.2.3.6, and the best maximum likelihood tree was obtained by maximum likelihood inference with 20 initial parsimony trees. The resulting phylogeny was highly consistent with the phylogeny estimated using the coalescent approach described above, and it was used for phylogenetic dating with a molecular calibration of 14.5 Ma for the crown age of *Oryza* and 6 Ma for the divergence of CC and AA-BB[39,43]. The confidence interval of the estimated dates was obtained by resampling branch lengths 100 times. Machine-readable phylogenetic trees are provided in Supplementary Data 2.

Synteny-based phylogenetic analysis

The synteny-based phylogenetic inference was performed following the Syn-MRL approach[101] as implemented in the R package syntenet[91] v.1.7.1. Specifically, the pangene table of the 31 (sub)genomes generated from the GENESPACE macro-synteny analysis was phased using a custom script (https://github.com/xiaoyezao/Rice-Phylogenomics) to identify syntenic homologous genes and syntenic clusters. Then, phylogenomic profiling of the syntenic clusters to determine which (sub)genomes contained which clusters was performed using the cluster_network function as implemented in syntenet. The phylogenomic profile was converted into a binary matrix in which the presence or absence of a syntenic gene from a given (sub)genome was coded as 1 or 0, respectively. Finally, phylogenetic analysis of the binary matrix was conducted using IQ-TREE v.2.3.6 with 1,000 Shimodaira–Hasegawa-like approximate likelihood ratio bootstrap replicates.

Ks analysis

The SynMap[102] and CodeML[103] tools, implemented in the Comparative Genomics (CoGe[104]) platform v.7 (https://genomevolution.org/coge), were used to identify collinear blocks of homologous genes between subgenomes of the same type (for example, between the KK subgenomes of *O. coarctata* and *O. schlechteri*) and calculate the fraction of synonymous substitutions per synonymous site (Ks). The SynMap analysis was run using default parameters (comparison algorithm: Last; window size: 100 genes; minimum number of aligned pairs: 5 genes; maximum distance between two matches: 20 genes). The distribution of Ks values ≤0.3 for each pair of subgenomes was plotted using a bin width of 0.01.

Gene fractionation

Gene fractionation was run on each tetraploid genome using SynMap[102] and FractBias[105], both implemented in CoGe[104] (v.7). SynMap was used to define gene homoeology and collinearity between subgenomes in each tetraploid genome as described above. FractBias was used to calculate and plot gene retention by setting a quota align ratio of 1:1.

Expression of homoeologous genes in *O. coarctata*

To identify pairs of homoeologous genes (that is, orthologous genes between the KK and LL subgenomes) among the predicted genes in the *O. coarctata* genome, Blastp[106] (v.2.14.0) was used with an *e*-value cut-off of 1×10^{-5}, and reciprocal hits were filtered using a bits score threshold of 50 and a minimum alignment identity of 50% for reciprocally aligning regions. The same method was used to identify genes in *O. sativa* homologous to genes in *O. coarctata*. In *O. coarctata*, genes were divided into two main categories: nonfractionated genes, homoeologous gene pairs found in both the KK and LL genomes; and fractionated genes, those found in one subgenome with no homoeologous pair in the other subgenome. In *O. sativa*, genes were also divided into two main categories: paired homologous genes, *O. sativa* genes that had homology to nonfractionated gene pairs in *O. coarctata*; and single-copy homologous genes, *O. sativa* genes that had homology to fractionated genes in *O. coarctata*. As the density distribution of

$\log_2$-transformed raw read counts was homogenous in all samples across the three plant growth tanks (Supplementary Fig. 16), we considered the individual plants as technical replicates. Paired-end reads from all the plants were mapped on each individual subgenome sequence of *O. coarctata* using TopHat2 (ref. 107) (v2.1.1). For reads that mapped to both subgenome sequences, we used EAGLE-RC[108] (v.1.1.1) to determine the likelihood of read alignment against each subgenome without knowing the genotype differences explicitly (-ngi) and chose the best alignment. Expression levels were determined using transcripts per kilobase of exon model per million mapped reads (TPM) with TPMCalculator[109] (v.0.0.3). To compare expression of genes in *O. coarctata* homologous to genes in *O. sativa*, we obtained publicly available datasets of Illumina paired-end RNA-seq of flag leaf (SAMN22452874, SRR16526865, SRR16526866), leaf (SRR4017523, SRR4017527) and root (SRR25078452, SRR25078455, SRR25078456) of *O. sativa*. Sequencing data were mapped to the *O. sativa* IRGSP reference sequence using TopHat2 (ref. 107) (v.2.1.1). TPM values were calculated as described above (the code is available at https://github.com/gopal-misra/Oryza_coarctata_expression). Differences in expression levels in the six gene categories were visualized as $\log_2(\mathrm{TPM}+1)$, and P values between the nonfractionated and fractionated and between the paired homologous and single-copy homologous categories were obtained using a two-sided Wilcoxon rank-sum test.

Expression bias (B) was quantified in homoeologous gene pairs over leaf and root tissue using the $\log_2$ fold change:

$$B = \log_2\left(\frac{\mathrm{TPM}_{\mathrm{LL}}+1}{\mathrm{TPM}_{\mathrm{KK}}+1}\right),$$

where $\mathrm{TPM}_{\mathrm{KK}}$ and $\mathrm{TPM}_{\mathrm{LL}}$ indicate the expression levels in the KK and LL subgenomes, respectively. Expression bias values greater than 1 ($B > 1$) indicate that the homoeologous copy on the LL subgenome is dominantly expressed, whereas expression bias values less than −1 ($B < -1$) indicate that the homoeologous copy on the KK subgenome is dominantly expressed[49]. Homoeologous gene pairs showing less than a twofold change ($-1 \leq B \leq 1$) were classified as nondominantly expressed. One-sided binomial test and a P value threshold of 0.001 were used to assess differences in numbers of dominantly expressed genes in the subgenomes for each tissue.

Statistics and reproducibility

To assess the significance of the correlations between (sub)genome size (Mb) and TE content (Mb), we calculated Pearson's linear correlation coefficient (R^2) with a two-sided t-test. For functional enrichment analysis of the syntenic pangenome, two-tailed Fisher's exact test with a P value threshold of 0.001 was used on the set of GO-annotated genes in each pangenome category (that is core, softcore and dispensable) and a reference set containing all GO-annotated genes in the pangenome. P values were corrected using the false discovery rate. To test for differences in percentages of gene retention between subgenomes in each tetraploid species, we used a two-sided Wilcoxon rank-sum test with a P value threshold of 0.001 (the wilcox.test function in R v.4.3.2 with continuity correction).

To test for differences in the expression levels of genes of six categories (four gene categories in *O. coarctata*: nonfractionated genes (nf), homoeologous gene pairs found in both the KK and LL genomes; and fractionated genes (f), those found in one subgenome with no homoeologous pair in the other subgenome; and two gene categories in *O. sativa*: paired homologous genes (ph), *O. sativa* genes that had homology to nonfractionated gene pairs in *O. coarctata*; and single-copy homologous genes (sch), *O. sativa* genes that had homology to fractionated genes in *O. coarctata*), we used a two-sided Wilcoxon rank-sum test and a P value threshold of 0.001 (the wilcox.test function in R v.4.3.2 with continuity correction). To assess differences in the number of genes dominantly expressed in the subgenomes of *O. coarctata*, in the leaf and root, we used a one-sided binomial test and P value threshold of 0.001.

Reporting summary

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

Data availability

Genome assemblies and DNA sequencing and RNA-seq data were deposited in GenBank and the NCBI SRA, respectively, under the following BioProjects: *O. alta*, PRJNA1039467; *O. australiensis*, PRJNA591699; *O. coarctata*, PRJNA439330; *O. grandiglumis*, PRJNA737282; *O. latifolia*, PRJNA737486; *O. longiglumis*, PRJNA1016142; *O. malampuzhaensis*, PRJNA757598; *O. meyeriana*, PRJNA1039468; *O. minuta*, PRJNA757599; *O. ridleyi*, PRJNA687623; *O. schlechteri*, PRJNA1115974. BioSamples, SRAs and GenBank genome IDs are specified in Supplementary Data 1. The GFA files for the AA, BB, CC and DD genome-type-level pangenomes are available on figshare (https://doi.org/10.6084/m9.figshare.28622279 (ref. 111). Publicly available data used in this paper are as follows. GenBank accessions of reference genome sequences: *O. sativa japonica* Nipponbare IRGSP (GCF_001433935.1), *O. punctata* (GCA_000573905.2), *O. officinalis* (GCA_008326285.1), *Oryza brachyantha* (GCF_000231095.2). Illumina paired-end RNA-seq of flag leaf (SAMN22452874, SRR16526865, SRR16526866), leaf (SRR4017523, SRR4017527) and root (SRR25078452, SRR25078455, SRR25078456) of *O. sativa*. GenBank accessions of chloroplast genomes: *O. barthii* (NC_027460.1), *O. glaberrima* (NC_024175.1), *O. nivara* (OL912836), *O. sativa japonica* (NC_001320.1), *O. rufipogon* (NC_017835.1), *O. sativa indica* (NC_008155.1), *O. meridionalis* (NC_016927.1), *Oryza glumipatula* (NC_027461.1), *Oryza longistaminata* (NC_027462.1), *O. punctata* (KF359908), *Oryza eichingeri* (NC_034759.1), *O. officinalis* (KF359910), *O. rhizomatis* (NC_034758.1), *O. australiensis* (KF359916), *O. brachyantha* (KF359917), *Oryza neocaledonica* (NC_053276.1), *L. japonica* (KF359922.1). Proteome sequences in Gramene *Oryza* release 7: *O. sativa japonica* Nipponbare IRGSP (https://ftp.gramene.org/oryza/release-7/fasta/oryza_sativa/pep), *O. glaberrima* (https://ftp.gramene.org/oryza/release-7/fasta/oryza_glaberrima/pep), *O. barthii* (https://ftp.gramene.org/oryza/release-7/fasta/oryza_barthii/pep), *O. glumipatula* (https://ftp.gramene.org/oryza/release-7/fasta/oryza_glumaepatula/pep), *O. meridionalis* (https://ftp.gramene.org/oryza/release-7/fasta/oryza_meridionalis/pep), *O. nivara* (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-57/fasta/oryza_nivara/pep), *O. rufipogon* (https://ftp.gramene.org/oryza/release-7/fasta/oryza_rufipogon/pep), *O. punctata* (https://ftp.gramene.org/oryza/release-7/fasta/oryza_punctata/pep), *O. brachyantha* (https://ftp.gramene.org/oryza/release-7/fasta/oryza_brachyantha/pep), *L. perrieri* (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-57/fasta/leersia_perrieri/pep). Public databases for genome annotation: BUSCO Poales database v.5 (https://busco-data.ezlab.org/v5/data/lineages/poales_odb10.2024-01-08.tar.gz), RepBase *Oryza* LINEs (https://www.girinst.org/repbase/update/browse.php?type=All&format=FASTA&autonomous=on&division=Oryza&letter=L), NCBI nonredundant database v.5 (https://ftp.ncbi.nlm.nih.gov/blast/db/v5/), NCBI reference sequence protein database (https://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref100/), InterPro (https://www.ebi.ac.uk/interpro/download/InterPro/), GO (https://geneontology.org/docs/download-ontology/) and PANTHER v.18.0 (https://www.pantherdb.org/).

Code availability

The custom scripts used in this study are available via GitHub (https://github.com/alicefornasiero/IOMAP-3, https://github.com/xiaoyezao/Rice-Phylogenomics, https://github.com/nam-hoang/rice_cp_phylo_analysis, https://github.com/gopal-misra/Oryza_coarctata_expression) and Zenodo (https://doi.org/10.5281/zenodo.14936239)[110].

References

67. Garrison, E. et al. Building pangenome graphs. *Nat Methods*. **21**, 2008–2012 (2024).
68. Parmigiani, L. et al. Panacus: fast and exact pangenome growth and core size estimation. *Bioinformatics* **40**, btae720 (2024).
69. Su, W., Ou, S., Hufford, M. B. & Peterson, T. in *Plant Transposable Elements* Vol. 2250 (ed. Cho, J.) 55–67 (Springer, 2021).
70. Tempel, S. in *Mobile Genetic Elements* Vol. 859 (ed. Bigot, Y.) 29–51 (Humana Press, 2012).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
73. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
74. Hoff, K. J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics* **65**, e57 (2019).
75. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics* **2014**, 4.11.1–4.11.39 (2014).
76. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
77. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
78. Conesa, A. & Götz, S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008**, 619832 (2008).
79. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
80. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
81. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
82. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101**, 12404–12410 (2004).
83. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2012).
84. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
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. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, 2000).
87. Lovell, J. T. et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* **11**, e78526 (2022).
88. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
89. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
90. Zhao, T. & Schranz, M. E. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl Acad. Sci. USA* **116**, 2165–2174 (2019).
91. Almeida-Silva, F., Zhao, T., Ullrich, K. K., Schranz, M. E. & Van De Peer, Y. syntenet: an R/Bioconductor package for the inference and analysis of synteny networks. *Bioinformatics* **39**, btac806 (2023).
92. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
93. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
94. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–W235 (2016).
95. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
96. 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).
97. Minh, B. Q., Dang, C. C., Vinh, L. S. & Lanfear, R. QMaker: fast and accurate method to estimate empirical models of protein evolution. *Syst. Biol.* **70**, 1046–1060 (2021).
98. Zhang, C. & Mirarab, S. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* **38**, 4949–4950 (2022).
99. Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
100. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
101. Zhao, T. et al. Whole-genome microsynteny-based phylogeny of angiosperms. *Nat. Commun.* **12**, 3498 (2021).
102. Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E. & Lyons, E. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* **33**, 2197–2198 (2017).
103. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
104. Albert, V. A. & Krabbenhoft, T. J. in *Polyploidy* Vol. 2545 (ed. Van De Peer, Y.) 19–45 (Springer, 2023).
105. Joyce, B. L. et al. FractBias: a graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics* **33**, 552–554 (2017).
106. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
107. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
108. Kuo, T., Frith, M. C., Sese, J. & Horton, P. EAGLE: explicit alternative genome likelihood evaluator. *BMC Med. Genomics* **11**, 28 (2018).
109. Vera Alvarez, R., Pongor, L. S., Mariño-Ramírez, L. & Landsman, D. TPMCalculator: one-step software to quantify mRNA abundance of genomic features. *Bioinformatics* **35**, 1960–1962 (2019).
110. Fornasiero, A. alicefornasiero/IOMAP-3: v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.14936239 (2025).
111. Fornasiero, A., Zuccolo, A., Schranz, M. E. & Wing, R. A. *Oryza* genome evolution through a tetraploid lens. *figshare* https://doi.org/10.6084/m9.figshare.28622279 (2025).

Acknowledgements

This publication is based on work supported by the King Abdullah University of Science and Technology (KAUST) with Baseline Funding and the Bud Antle Endowed Chair in Excellence in Agriculture (University of Arizona) to R.A.W. T.F. was supported by the Dutch Research Council (NWO) (grant OCENW.KLEIN.419) and the NWO TTW Project (grant P17-19). M.E.S. was supported by the Dutch Research Council (NWO) (grant OCENW.KLEIN.419). We acknowledge contributions from M. Miculan and V. Melino to the experimental design for transcriptome analysis and subgenome dominance in *O. coarctata*, and support from the Ibex computational scientists and system administrators at KAUST Supercomputing Laboratory. We are also indebted to past (R. Sackville Hamilton and D. Brar) and present (V. Ramaiah) members of IRRI's International Rice Genebank for their >20 years of continuous support and long-term maintenance of the *Oryza* voucher specimens used in this study.

Author contributions

R.A.W. conceived and designed the project. R.A.W., A.Z. and A.F. coordinated the project. N.M. prepared the tissue samples for genome and transcriptome sequencing with assistance from D.K., D.C., S.R., S.L. and J.T. (at the Arizona Genomics Institute). C.S.-S. provided computational support for data generation and transfer. N.A.-B. grew and prepared the tissue samples for subgenome dominance analysis in *O. coarctata*. P.P. and V.L. generated optical maps for genome assembly validation. A.F., N.A.-B., A.A., S.M., G.M., L.F., M.-C. C. and A.Z. performed genome assembly and validation and gene and TE annotation of the *Oryza* species with substantial contributions from Y.Z., L.R.S., V.L. and M.T. S.M. and A.Z. performed TE analysis in *O. ridleyi* and *O. longiglumis*. T.F. and N.V.H. performed phylogenomic analysis. N.A.-B. performed Ks analysis. A.F. and T.F. performed macro- and micro-synteny analysis. A.F. performed gene fractionation analysis with substantial contributions from M.E.S. G.M. performed analysis of subgenome dominance in *O. coarctata*. O.P. contributed transcriptome data for *O. australiensis*. K.L.M. grew, maintained and provided the germplasm collection of the *Oryza* tetraploids. M.E.S. and J.Z. contributed valuable suggestions to analyses and interpretation of results. A.F. wrote the paper with substantial contributions from T.F., N.A.-B., N.V.H., A.Z., M.E.S. and R.A.W.

Competing interests

V.L. and P.P. are affiliated with Corteva Agriscience. The other authors declare no competing interests.

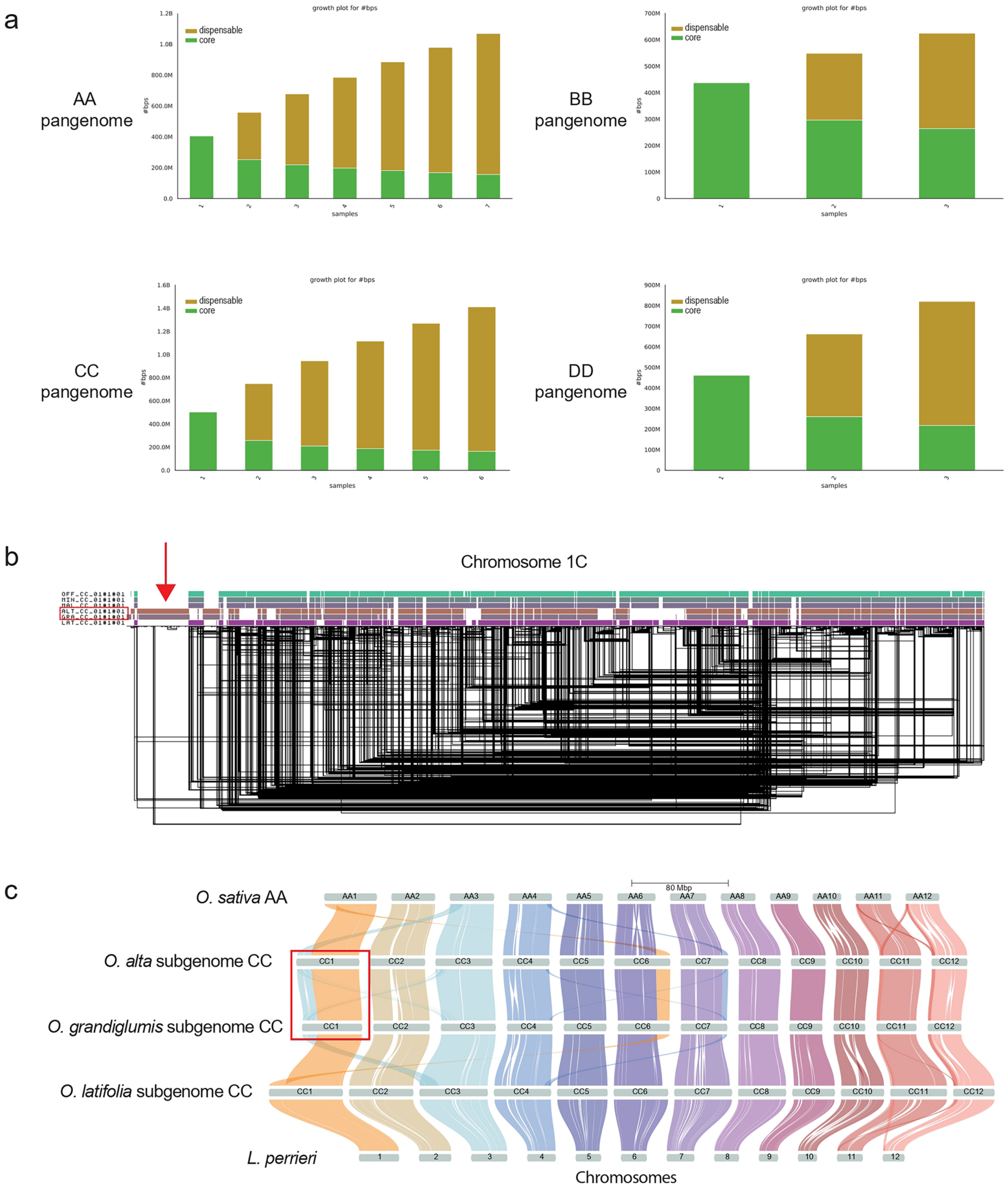
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02183-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02183-5.

Correspondence and requests for materials should be addressed to Alice Fornasiero, Andrea Zuccolo, M. Eric Schranz or Rod A. Wing.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Genome type-level pangenomes for the AA, BB, CC, and DD types. a) A growth histogram relative to the cumulative number of (sub)genomes on the x-axis is shown for the AA, BB, CC, and DD pangenomes, respectively. Sequences shared by all (sub)genomes represent the core pangenome (green), partially shared sequences represent the dispensable pangenome (ochre). **b**) Pangenome visualization of the translocation on chromosome 1 C in *O. alta* and *O. grandiglumis*. The red arrow points to the chromosomal region corresponding to an unbalanced translocation of a portion of Chr3 on Chr1 (see also panel **c**). This chromosomal rearrangement is present in *O. alta* and *O. grandiglumis* (red box) and absent from the other CC types used to build the CC pangenome (that is *O. officinalis*, *O. minuta*, *O. malampuzhaensis* and *O. latifolia*). c) The riparian plot shows the synteny between *O. sativa* (AA), *O. alta* (CC), *O. grandiglumis* (CC), *O. latifolia* (CC) and *L. perrieri* chromosomal regions. The red box highlights the duplication and translocation of a portion of Chr3 on Chr1 in *O. alta* and *O. grandiglumis*.

https://doi.org/10.1038/s41588-025-02183-5

Extended Data Table 1 | The wild *Oryza* genome data package

Species	IRRI Accession	Country of origin	Sequencing reads	Raw data (Gbp)	Number of subreads	Mean subread length (bp)	Total Assembly length (Mbp)	Coverage (fold)
O. alta	IRGC 105143	Guyana	PacBio CLR	125	4,940,816	25,381	909.02	138
O. australiensis	IRGC 100882	Australia	PacBio CLR	125	8,092,633	15,446	881.42	142
O. coarctata	IRGC 104502	Bangladesh	PacBio CCS	107	5,458,678	19,600	555.71	193
O. grandiglumis	IRGC 105669	Brazil	PacBio CLR	300	16,455,569	18,228	871.18	344
O. latifolia	IRGC 100890	Cuba	PacBio CLR	304	15,365,458	19,812	1058.40	288
O. longiglumis	IRGC 106525	Papua New Guinea	PacBio CCS	47	2,675,801	17,710	1146.78	41
O. malampuzhaensis	IRGC 80765	India	PacBio CLR	310	15,554,278	19,960	934.73	332
O. meyeriana	IRGC 106473	Philippines	PacBio CCS	37	2,262,199	16,205	789.12	46
O. minuta	IRGC 101141	Philippines	PacBio CLR	441	24,643,407	17,899	980.38	450
O. ridleyi	IRGC 100821	Thailand	PacBio CLR	169	6,801,463	24,825	1202.81	140
O. schlechteri	IRGC 82047	Papua New Guinea	PacBio CLR	135	5,376,255	25,041	674.35	200

IRGC accession, country of origin, and sequencing data type and statistics of the 11 wild *Oryza* genomes presented in this study. Passport data of the *Oryza* species are available in the IRRI GrinGlobal genebank at: https://gringlobal.irri.org/gringlobal/search

Extended Data Table 2 | Genome summary statistics

	O. malampuzhaensis	*O. minuta*	*O. alta*	*O. grandiglumis*	*O. latifolia*	*O. australiensis*	*O. coarctata*	*O. schlechteri*	*O. longiglumis*	*O. ridleyi*	*O. meyeriana*
Genomic feature											
Genome type	BBCC	BBCC	CCDD	CCDD	CCDD	EE	KKLL	KKLL[HHKK]	HHJJ	HHJJ	GG
Chromosomes (#)	24	24	24	24	24	12	24	24	24	24	12
Assembly											
Total assembly size (Mbp)	**934.73**	**980.38**	**909.02**	**871.18**	**1,058.40**	**881.42**	**555.71**	**674.35**	**1,146.78**	**1,202.81**	**789.12**
Contig N50 (Mbp)	19.09	20.60	18.70	19.30	35.13	44.75	22.97	22.26	51.17	53.84	66.88
Contig N90 (Mbp)	8.99	11.25	10.73	12.01	15.86	7.29	16.51	9.86	31.97	33.86	49.09
Contig L50 count	19	20	19	19	13	7	10	12	9	9	5
Contig L90 count	44	44	42	41	31	24	21	28	20	21	11
Scaffold N50 (Mbp)	39.29	41.47	38.39	37.53	45.31	77.73	22.97	28.20	51.17	53.84	66.88
Scaffold N90 (Mbp)	30.49	31.40	29.35	27.76	34.13	54.84	16.51	20.81	31.97	33.86	50.49
Scaffold L50 count	11	11	11	11	11	5	10	10	9	9	5
Scaffold L90 count	21	21	21	21	21	11	21	21	21	21	11
GC content (%)	43.36	43.43	43.85	43.87	44.22	44.98	41.86	43.12	43.40	43.48	46.05
Number of gaps	**40**	**45**	**34**	**31**	**17**	**67**	**0**	**17**	**2**	**1**	**1**
Complete BUSCO genome assembly (%)	**99.20**	**99.20**	**99.30**	**99.20**	**99.10**	**97.90**	**98.40**	**99.00**	**99.30**	**99.40**	**98.30**
Annotation											
TE size (Mbp)	518.60	563.15	507.89	470.35	627.55	631.54	174.42	267.45	700.05	740.09	550.85
TE fraction (%)	55.49	57.44	55.86	53.99	59.28	71.63	31.38	39.67	61.06	61.54	69.81
Predicted genes (#)	77,869	74,266	66,387	66,356	73,734	39,392	55,709	70,857	73,324	82,808	45,092
Complete BUSCO gene prediction (%)	**94.30**	**89.70**	**93.40**	**93.40**	**93.80**	**88.10**	**86.60**	**81.10**	**90.10**	**95.40**	**90.70**

Genomic features and summary statistics for the contig- and scaffold-level assembly, gene prediction and TE annotation of the 11 wild *Oryza* genomes presented in this study.

nature portfolio

Corresponding author(s): Alice Fornasiero
Andrea Zuccolo
M. Eric Schranz
Rod A. Wing

Last updated by author(s): Feb 27, 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/codes were used for data collection.
Data analysis	Optical mapping molecule assembly and hybrid scaffolding: Bionano Access and Tools software (v1.4.1 , v1.5.2 , v1.6.1). De novo genome assembly of PacBio CLR reads: Mecat2 (v20193.14), Canu (v2.0), Flye (v2.8.1). De novo genome assembly of PacBio HiFi reads: Hifiasm (v0.19.5), HiCanu (v2.2). Scaffolding, contig assignment, ordering and orientation, and gap fixing: GCPP (https://github.com/PacificBiosciences/gcpp), Pilon (v1.23), Mashmap (v2.0), HAKMER, Phylogenetic Analysis using PAUP (v4.0a), Genome Puzzle Master, RagTag (v2.1.0), BWAmem (v0.7.17), SAMtools (v1.16.1). Assembly quality evaluation: BUSCO (v5.1.2). Chloroplast genome assembly: Blastn (v2.14.0), Minimap2 (v2.19). Transposable element annotation: EDTA (v1.9.0), RepeatMasker (v4.1.0). Gene prediction: OmicsBox (v2.0.36) (https://www.biobam.com/omicsbox) and embedded tools (FastQC (v0.11.8), Trimmomatic (v0.38), STAR (v2.7.8a), Minimap2 (v2.19), Augustus (v3.4.0)). Gene prediction in O. australiensis: MAKER-P (v3.01.03) and embedded tools (Augustus (v3.2.1), SNAP (v2013-11-29), Fgenesh+), AGAT tool suite (https://agat.readthedocs.io/en/latest/index.html). Functional gene annotation: non-redundant (nr) and NCBI Reference Sequence (RefSeq) protein databases v5 (http://www.ncbi.nlm.nih.gov/RefSeq/), Diamond Blastp (v0.9). InterProScan (v5.39), Blast2GO (v6.0). Gene function annotation in O. australiensis: MAKER-P (v3.01.03), BUSCO (v5.1.2). GO enrichment analysis: PANTHER (v18.0). TE amplification and insertion time of long terminal repeat-retrotransposons: tBlastn (v2.14.0), MUSCLE (v3.8.425), MegaX (v10.1), EMBOSS (V6.6.0). Macro-synteny analysis: GENESPACE (v1.3.1) and embedded software (OrthoFinder (v2.5.4) and MCScanX_h). Gene micro-synteny analysis: custom script (https://github.com/xiaoyezao/Rice-Phylogenomics), SYNTENET (v1.5.2). Phylogenomic analysis using chloroplast-based approach: MAFFT (v7.480), trimAL (v1.4), IQ-TREE (v1.6.12), FigTree (v1.4.3). Phylogenomic analysis using multi nuclear gene-based approach: MAFFT (v7.520), trimAl (v1.4.1), RAxML-NG (v1.2.0), Astral-Pro2 (v1.16.1.3). Ks analysis: SynMap and CodeML tools in Comparative Genomics (CoGe) platform v7 (https://genomevolution.org/coge). Gene fractionation: SynMap and FractBias implemented in CoGe platform v7 (https://genomevolution.org/coge). Expression of homoeologous genes in O. coarctata: Blastp (v2.14.0), Tophat2 (v 2.1.1), Eagle-RC (V1.1.1), TPMCalculator (v0.0.3).

The custom scripts used in this study are available in GitHub (https://github.com/alicefornasiero/IOMAP-3, https://github.com/xiaoyezao/Rice-Phylogenomics, https://github.com/nam-hoang/rice_cp_phylo_analysis, https://github.com/gopal-misra/Oryza_coarctata_expression) and deposited in Zenodo (https://doi.org/10.5281/zenodo.14936239) (ref. 110).

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

Genome assemblies and DNA and RNA sequencing data were deposited in GenBank and NCBI Sequence Read Archive (SRA), respectively, under the following BioProjects: O. alta, PRJNA1039467; O. australiensis, PRJNA591699; O. coarctata, PRJNA439330; O. grandiglumis, PRJNA737282; O. latifolia, PRJNA737486; O. longiglumis, PRJNA1016142; O. malampuzhaensis, PRJNA757598; O. meyeriana, PRJNA1039468; O. minuta, PRJNA757599; O. ridleyi, PRJNA687623; O. schlechteri, PRJNA1115974. BioSamples, SRAs and GenBank Genome IDs are specified in Supplementary Data 1. The GFA files for the AA, BB, CC, and DD genome type-level pangenomes are available on Figshare (https://doi.org/10.6084/m9.figshare.28622279).

Publicly available data used in this paper - GenBank accessions of reference genome sequences: O. sativa japonica Nipponbare IRGSP (GCF_001433935.1), O. punctata (GCA_000573905.2), O. officinalis (GCA_008326285.1), O. brachyantha (GCF_000231095.2). Illumina paired-end RNA-Seq of flag leaf (SAMN22452874, SRR16526865, SRR16526866), leaf (SRR4017523, SRR4017527) and root (SRR25078452, SRR25078455, SRR25078456) of O. sativa. GenBank accessions of chloroplast genomes: O. barthii (NC_027460.1), O. glaberrima (NC_024175.1), O. nivara OL912836), O. sativa japonica (NC_001320.1), O. rufipogon (NC_017835.1), O. sativa indica (NC_008155.1), O. meridionalis (NC_016927.1), O. glumipatula (NC_027461.1), O. longistaminata (NC_027462.1), O. punctata (KF359908), O. eichingeri (NC_034759.1), O. officinalis (KF359910), O. rhizomatis (NC_034758.1), O. australiensis (KF359916), O. brachyantha (KF359917), O. neocaledonica (NC_053276.1), L. japonica (KF359922.1). Proteome sequences in Gramene Oryza release 7: O. sativa japonica Nipponbare IRGSP (https://ftp.gramene.org/oryza/release-7/fasta/oryza_sativa/pep), O. glaberrima (https://ftp.gramene.org/oryza/release-7/fasta/oryza_glaberrima/pep), O. barthii (https://ftp.gramene.org/oryza/release-7/fasta/oryza_barthii/pep), O. glumipatula (https://ftp.gramene.org/oryza/release-7/fasta/oryza_glumaepatula/pep), O. meridionalis (https://ftp.gramene.org/oryza/release-7/fasta/oryza_meridionalis/pep), O. nivara (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-57/fasta/oryza_nivara/pep), O. rufipogon (https://ftp.gramene.org/oryza/release-7/fasta/oryza_rufipogon/pep), O. punctata (https://ftp.gramene.org/oryza/release-7/fasta/oryza_punctata/pep), O. brachyantha (https://ftp.gramene.org/oryza/release-7/fasta/oryza_brachyantha/pep), L. perrieri (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-57/fasta/leersia_perrieri/pep). Public databases for genome annotation: BUSCO Poales database v5 (https://busco-data.ezlab.org/v5/data/lineages/poales_odb10.2024-01-08.tar.gz), RepBase Oryza LINEs (https://www.girinst.org/repbase/update/browse.php?type=All&format=FASTA&autonomous=on&division=Oryza&letter=L), NCBI non-redundant (nr) database v5 (https://ftp.ncbi.nlm.nih.gov/blast/db/v5/), NCBI Reference Sequence (RefSeq) protein database (https://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref100/), InterPro (https://www.ebi.ac.uk/interpro/download/InterPro/), Gene Ontology (https://geneontology.org/docs/download-ontology/), PANTHER v18.0 (https://www.pantherdb.org/).

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	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No statistical methods were required to establish sample size in this study. The germplasm used to generate the reference genomes of the 11 wild Oryza species in this study represents the collection of accessions that, historically, has been chosen for the International Oryza Map Alignment Project (IOMAP). The selection for the present work aimed at covering all genome types of the genus Oryza.
Data exclusions	No data were excluded from the analysis. Filtering parameters for raw sequencing data, sequence alignments, and gene models were described in the methods.
Replication	O. coarctata cuttings were grown in three separated plant growth tanks with six replicates each, in hydroponic conditions with freshwater irrigation at the KAUST greenhouse facility. As density distribution of log2-transformed raw read counts was homogenous in all samples across the three plant growth tanks (Supplementary Figure 11), we considered the 18 individual plants as technical replicates. For each phylogenetic analysis, we performed 1,000 bootstrap replicates to assess the robustness of the inferred tree topology. Each replicate represents an independent reconstruction of the same dataset, and the analysis was performed once with 1,000 replicates. All bootstrap attempts were successfully completed and the final bootstrap support values are shown in the trees.
Randomization	Group allocation was not relevant to this study and therefore randomization was not necessary. Analysis of gene expression (i.e. subgenome dominance/equivalence in O. coarctata) was performed at the subgenome level within the nucleus of each single genotype.
Blinding	Group allocation was not relevant to this study and therefore blinding was not necessary.

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

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

Plants

Seed stocks	Single seed descent (SSD) germplasm for O. alta (IRGC 105143), O. australiensis (IRGC 100882), O. grandiglumis (IRGC 105669), O. latifolia (IRGC 100890), O. longiglumis (IRGC 106525), O. malampuzhaensis (IRGC 80765), O. meyeriana (IRGC 106473), O. minuta (IRGC 101141), O. ridleyi (IRGC 100821) and O. schlechteri (IRGC 82047) were obtained from the International Rice Research Institute (IRRI, Philippines) under the Standard Material Transfer Agreement.
Novel plant genotypes	No novel plant genotypes were generated in this study.
Authentication	Authentication procedure for the germplasm accessions used in this study were performed by the International Rice Genebank curators, maintained by IRRI.