



OPEN The *Oryza australiensis* genome reveals potential as a source of genes for rice improvement

Sabrina Morrison^{1,2}, Ardashir K. Masouleh², Lena Constantin³, Olivier Panaud⁴, Rod A. Wing⁵ & Robert J. Henry^{1,2}✉

The wild relatives of cultivated rice (*Oryza sativa* L.) represent largely untapped sources of novel genetic material for crop improvement. Here, we present a chromosome-level, haplotype-resolved genome for *O. australiensis*, an Australian wild rice relative, assembled de novo using PacBio HiFi long reads (80× coverage). The final assembly had a 99.3% BUSCO completeness score and a total length of 909 Mb. The assembly was highly contiguous (contig N50 of 72.6 Mb) despite its extremely high repeat element content, which was estimated to comprise 76% of the genome. There were 26,529 transcripts and 23,929 potential genes annotated from the assembly, of which 1,431 were species-specific to *O. australiensis* and not found in cultivated rice. Annotation of these revealed enrichment for functions associated with DNA integration, as well as those linked to abiotic and biotic response signalling. Structural analyses highlighted considerable synteny between *O. australiensis* and cultivated rice, despite a long period of evolutionary divergence between the two species. Despite the presence of little structural variation between the haplotypes, there was considerable variation observed between the current assembly and a previously published *O. australiensis* genome, suggesting local divergence that may be associated with adaptation of populations to heterogeneous environments.

Keywords Wild rice, Disease resistance, Genome assembly, Haplotype, *Oryza*, Chromosome-level

Abbreviations

BLAST	Basic local alignment search tool
BPH	Brown planthopper
BUSCO	Benchmarking Universal Single-Copy Orthologues
CC	Coiled-coil
CDRH	Cloned disease resistance homologue
CDS	Coding sequence
CPAT	Coding potential analysis test
CTAB	Cetyltrimethyl ammonium bromide
DNA	Deoxyribonucleic acid
GO	Gene ontology
LAI	LTR Assembly Index
LRR	Leucine-rich repeat
LTR	Long terminal repeat
NBS	Nucleotide-binding site
NCBI	National Center for Biotechnology Information
NLR	Nucleotide-binding leucine-rich repeat
RAP	Rice Annotation Project
RGA	Resistance gene analogue
RLK	Receptor-like kinases
RLP	Receptor-like proteins
RNA	Ribonucleic acid
RPW8	Resistance to Powdery Mildew 8

¹Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St Lucia 4072, Australia.

²ARC Centre of Excellence for Plant Success in Nature and Agriculture, University of Queensland, St Lucia 4072, Australia. ³Frazer Institute, University of Queensland, St Lucia 4072, Australia. ⁴Université de Perpignan, 66100 Perpignan, France. ⁵King Abdullah University of Science and Technology, 23955 Thuwal, Saudi Arabia. ✉email: robert.henry@uq.edu.au

TDG	Tandem duplicated gene
TIR	Terminal inverted repeat
TIR	Toll/Interleukin-1 receptor
TM	Transmembrane
UQ	University of Queensland

Rice (*Oryza sativa* and *O. glaberrima*) is grown in over 100 countries as a staple food for an estimated 3.5 billion people¹. The wild *Oryza* species represent largely untapped reservoirs of novel genetic material that can be used to improve agronomically-important traits in cultivated rice^{2,3}. There are four major wild relatives of rice existing in Australia; *Oryza australiensis* (EE genome), *O. meridionalis* (AA genome), *O. rufipogon* (AA genome), and *O. officinalis* (CC genome)⁴. Of these, the diploid *O. australiensis* (2n = 24) is the most distantly related, having diverged from the progenitor of cultivated rice more than 7 million years ago⁵. This species possesses a number of unique phenotypic characteristics that make it a potential candidate for rice improvement⁴. Genotypic and morphological tolerance traits related to drought⁶, heat⁷, and salinity⁸ have enabled the species to survive perennially within the coastal and inland regions of northern Australia. *Oryza australiensis* also possesses resistance to a number of globally important pests and diseases of rice, including the brown planthopper insect⁹, bacterial blight (BB)¹⁰, and rice blast disease¹¹.

The assembly and annotation of wild *Oryza* reference genomes have revealed insights into the evolutionary history of the genus and enabled the discovery of novel gene targets for rice breeding programs^{2,12}. A number of gene variants identified in wild rices have been found to confer advantageous traits, such as variations in the BB resistance gene *Xa27* from *O. minuta*¹³ and in the heat tolerance gene *HTH5* from *O. rufipogon*¹⁴. The *O. australiensis* genome is more than twice the size (~900 Mb) of cultivated rice due to a rapid expansion of LTR retrotransposon families, and this has impeded the generation of a reference genome assembly until recently^{12,15–17}. As genome sequencing and assembly technologies continue to advance, an updated, high-quality de novo reference for the *O. australiensis* genome will further elucidate genetic insights from this species. Here, we assemble a chromosome-scale, haplotype-resolved de novo assembly for *O. australiensis* using PacBio long reads and HiC-integrated sequencing data. We aimed to explore the potential of *O. australiensis* as a candidate source of novel genetic material by identifying species-specific genes and protein clusters that are absent in cultivated rice. We also investigated the structural variations across haplotypes, accessions, and species to understand intra- and inter-species diversity. We identified some of the genomic regions comprising potential disease resistance gene loci and located gene homologues to cloned resistance genes in rice, highlighting *O. australiensis* as a source for novel disease resistance.

Results

Generation of a de novo, chromosome-level *O. australiensis* genome assembly

The two PacBio Sequel II SMRT cells used for long-read sequencing of *Oryza australiensis* produced HiFi yields of 36.3 Gb (40X) and 36.5 Gb (40X) with a Q35 median read quality and a total combined genome coverage of 80X. Assembly of the HiFi reads was conducted with the Hifiasm assembler¹⁸ using paired-end Hi-C read integration to produce a primary contig assembly. The primary assembly yielded a total of 1452 contigs, with an N50 of 72.6 Mb and a total length of 975 Mb. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis with universal single copy genes within the viridiplantae lineage identified a genome coverage of 99.3%.

The primary assembly was scaffolded using SALSA2¹⁹ with Hi-C proximity ligation libraries. D-Genies²⁰ was used to assess alignment of the scaffolded *O. australiensis* pseudomolecules with the *Oryza sativa* reference genome. There were 14 major (>0.5 Mb) *O. australiensis* scaffolds that aligned with the 12 *O. sativa* chromosomes. Ten of these scaffolds aligned wholly with a chromosome in *O. sativa* and possessed telomeric repeats on both ends. Ragtag²¹ joined three scaffolds corresponding to *O. sativa* chromosome 5, resulting in a telomere-to-telomere pseudochromosome 5 of length 79.2 Mb. RagTag did not join the two scaffolds corresponding to chromosome 9.

Eleven out of 12 scaffolds in the finalised assembly possessed telomeres at both ends, representing complete ‘telomere-to-telomere’ (T2T) pseudochromosomes. Pseudochromosome lengths ranged from 49.4 Mb (Oaus_09) to 99.3 Mb (Oaus_02) (Supplementary table 1). The final length of this assembly was 909 Mb with a BUSCO completeness of 99.3% (Table 1). The genome lengths estimated by flow cytometry for *O. sativa* ssp. *japonica* cv. Nipponbare and *O. australiensis* were 391 Mb at 1C ($\pm 1.69\%$ CV) and 909 Mb at 1C ($\pm 0.80\%$ CV), respectively.

Resolution of *O. australiensis* haplotype assemblies

Assembly of the HiFi reads using the Hifiasm assembler¹⁸ generated a pair of haplotype-resolved assemblies, designated haplotype 1 (hap1) and haplotype 2 (hap2). The hap1 contig assembly had a total of 1,254 contigs, an N50 of 72.4 Mb and a total length of 962 Mb. The hap2 assembly produced 678 contigs, an N50 of 69.6 Mb and a total length of 939 Mb. Scaffolding with SALSA2 and RagTag joined two and four contigs in the hap1 and hap2 assemblies, respectively. There were two major contigs in each of the haplotype assemblies corresponding with chromosome 5 in the collapsed assembly, however these were not joined during scaffolding (Supplementary Fig. 1). Scaffolds from the hap1 assembly that corresponded to pseudochromosomes in the primary genome were designated OausHap1_01 to OausHap1_12. Remaining scaffolds with telomeric regions or scaffolds that appeared to align with the primary assembly but were not assigned as chromosomes were also included in the finalised hap1 assembly, which consisted of 16 total scaffolds (Table 1). This was repeated for the hap2 assembly, resulting in the selection of 18 final scaffolds. The BUSCO analyses for hap1 and hap2 assemblies reported a gene coverage of 99.0% and 99.2%, respectively. There were 10 T2T pseudochromosomes in the finalised hap1 assembly and nine T2T pseudochromosomes generated in the hap2 assembly (Supplementary table 1). The two

	Primary	Haplotype 1	Haplotype 2
Total assembly size (bp)	908,685,645	906,642,021	903,514,669
Complete BUSCOs (%)	99.3	99.0	99.2
Scaffold number	13	16	18
Largest scaffold (bp)	99,334,057	99,281,607	98,984,626
scaffold N50 (bp)	76,570,987	72,406,816	69,627,118
scaffold L50	6	6	6
GC content (%)	45.0	45.0	44.9
Total telomere number	23	24	23
Number of genes	23,929	23,578	23,653
Number of transcripts	26,529	26,125	26,182
Repeat element content (%)	75.9	76.5	76.3

Table 1. Primary and haplotype genome assembly and annotation statistics.

haplotype assemblies were observed to be highly syntenic with each other (Fig. 1). Around 873.6 Mb (99.9%) of the hap1 chromosome-level assembly was predicted to consist of syntenic regions with the hap2 assembly. Unaligned regions accounted for less than 0.04% (396 kb) of the haplotype assembly lengths, whilst inversions and translocations corresponded to ~ 57 kb and ~ 33 kb of the total assembly lengths, respectively. Heterozygosity of the *O. australiensis* genome was predicted to be ~ 0.04% using k-mer-based statistical approaches.

Structural annotation and repeat element content analyses of *O. australiensis*

Approximately 689 Mb (75.9%) of the genome was designated as interspersed repeat content, of which 43.7% were LTR retrotransposons. The majority of repeat element content was designated into two classes of LTR retrotransposon—*Ty1/Copia* (10.6%) and *Gypsy/DIRS1* (32.3%) retroelements. Around a quarter (24.3%) of transposable elements were unclassified. Further analysis with EDTA²² designated these as unknown LTR retrotransposons. The final LTR retrotransposon content estimated by EDTA was 58.2%. A further 11.3% of repeat content was designated as terminal inverted repeat (TIR) DNA transposons and 4.4% was classed as non-TIR rolling-circle (Helitron) transposons. Analysis of LTR elements using LTR_retriever^{23,24} demonstrated a high LTR Assembly Index (LAI) score of 21.51, indicating high assembly contiguity.

The primary *O. australiensis* assembly was annotated using viridiplantae protein data and RNA-seq data as input files. There were 23,929 putative genes and 26,529 transcripts predicted by Braker3²⁵. Of these 26,529 predicted transcripts, 23,388 (88%) were supported by the RNA-seq data. Analyses of the annotated protein sequences identified 98.2% of complete BUSCOs within the viridiplantae lineage. Annotation of the haplotype assemblies predicted 23,578 potential coding genes and 26,125 transcripts in hap1 and 23,653 potential genes and 26,182 transcripts in hap2 (Table 1). The gene densities and distributions of transposable elements across the *O. australiensis* chromosomes were visualised using shinyCircos V2.0²⁶ (Fig. 2). Gene density appeared to be highest towards the ends of chromosomes and seemed to coincide with DNA transposon distribution but was inversely distributed with retroelements.

Intra- and inter-species structural variations in *O. australiensis*

There was extensive structural variation observed between the primary assembly and a previously published *O. australiensis* genome¹² (Fig. 3). There were 4,607 translocations, 1,084 duplications, and 163 inversion events between the two assemblies, corresponding to 98.8 Mb or 10.9% of the primary *O. australiensis* genome. More than half of this structural variation was attributed to large inversions (> 1 Mb and > 1.25% total chromosome length), spanning a cumulative length of 47.7 Mb (5.2% total genome length) (Supplementary table 2). The largest inversion was 6.3 Mb (8.0% of total chromosome length) and found on chromosome 7, followed by an inversion of 5.6 Mb on chromosome 11 (9.0% of chromosome length). The largest translocation region in the primary assembly was 584 kb in length and on chromosome 12 in both assemblies. Chromosome lengths were relatively consistent between the current and previously published assemblies. The difference in chromosome lengths was < 2 Mb in nine of the twelve assembled chromosomes. The largest difference in length occurred on chromosome 11 (5.9 Mb length difference).

Structural variation was additionally compared between the *O. australiensis* chromosomes and a cultivated *O. sativa* ssp. *japonica* rice genome (Fig. 4). Despite being considerably evolutionarily diverged, the two genome assemblies demonstrated a high level of structural synteny. Around 526.9 Mb (58.0% total genome length) of the *O. australiensis* genome was predicted to be syntenic with cultivated rice. Syntenic regions between the two genomes were concentrated on the distal end of chromosomes, whilst large sections of unaligned or inverted regions tended to be distributed centrally. There were 159 inversions and 354 translocations between the primary *O. australiensis* genome and *O. sativa*. Thirteen inversions were larger than 1 Mb and greater than 1.25% of the chromosome length, comprising a total of 48.1 Mb (5.3%) of the *O. australiensis* genome. Nine of these inversions were common (position of inversion within 200 bp and < 3 kb inversion size difference in *O. sativa*) in both the current and previously published *O. australiensis* genomes, whilst the remaining four were unique to the current assembly (Tables 2 and 3). The largest inversion between *O. australiensis* and cultivated rice was 20.1 Mb (22.5% total chromosome length) on chromosome 6.

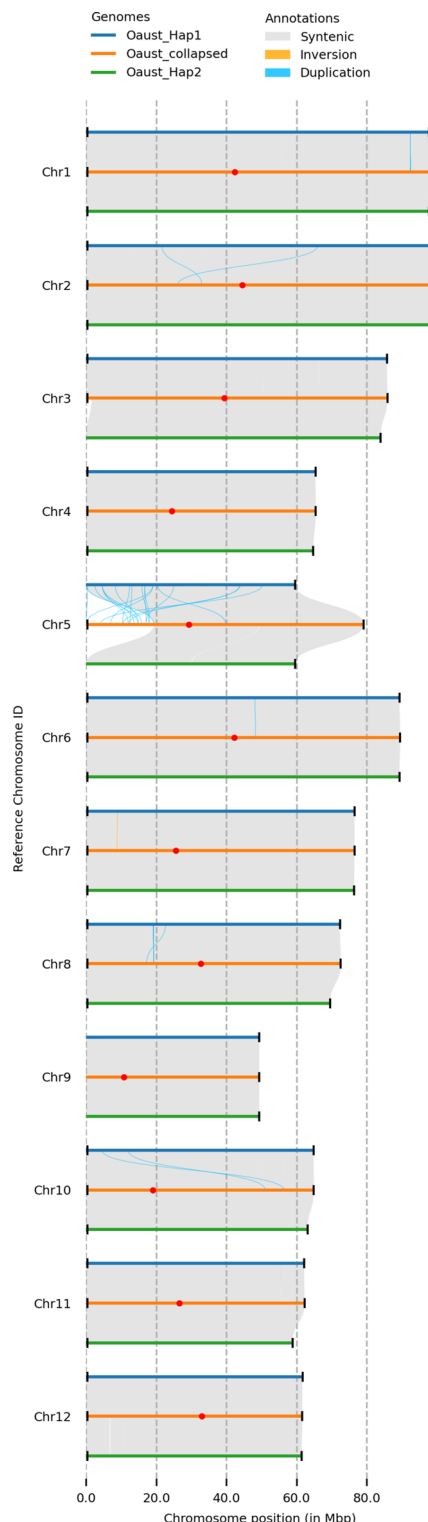


Fig. 1. Structural variation between the primary *O. australiensis* assembly (collapsed) and haplotypes (Hap1 & Hap2). Black vertical lines (|) indicate predicted telomeres across each assembly. Red circles (o) indicate predicted centromeres in the collapsed assembly.

Functional annotation of collapsed *O. australiensis* genome assembly

The primary *O. australiensis* coding sequences were functionally annotated. Of the 26,529 transcript sequences, 3,388 sequences had BLAST hits, 280 were further mapped to Gene Ontology (GO) terms, and 22,495 sequences further underwent GO annotation (Supplementary data 1). There were 366 (1.4%) sequences that did not yield any BLAST hits and were not GO annotated, designated as ‘no-BLAST’ sequences. The coding potentials of these

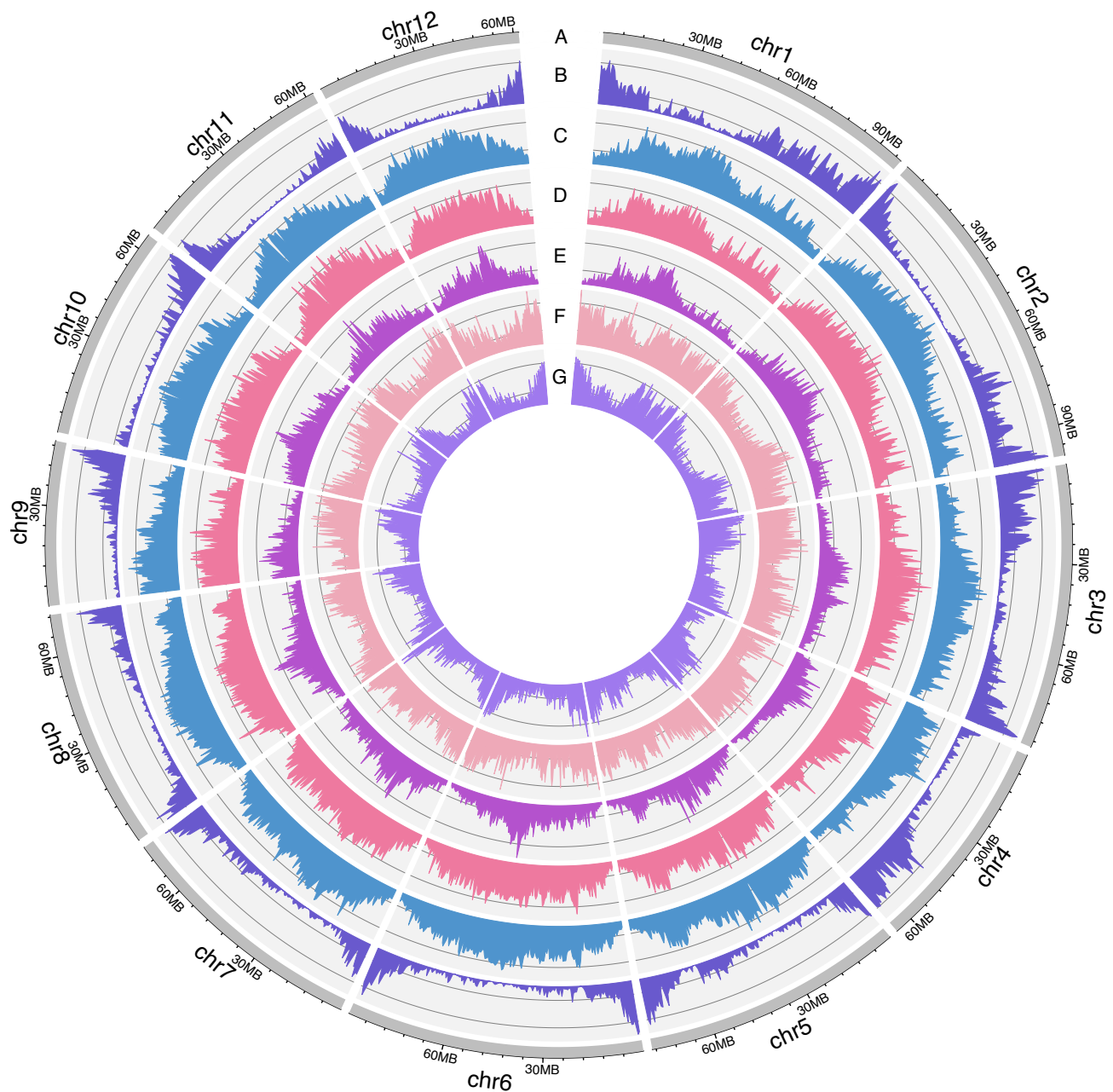


Fig. 2. Gene density and repeat element content in *O. australiensis*. (A) chromosome number, (B) Gene density, (C) Total retroelement density, (D) *Gypsy* retrotransposon density, (E) *Copia* retrotransposon density, (F) terminal inverted repeat (TIR) DNA transposon density, (G) Helitron density.

no-BLAST sequences were assessed using *Arabidopsis* (0984) and *Oryza* (45,270) as reference models. There were 314 (85.8%) and 331 (90.4%) *O. australiensis* no-BLAST sequences predicted to have coding potential when analysed against the *Arabidopsis* and *Oryza* references, respectively.

There were 11,496 GO terms assigned to genes within the *O. australiensis* genome. GO terms associated with abiotic and biotic stress response were investigated in OmicsBox (OmicsBox—Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019, <https://www.biobam.com>). There were 1,597 coding sequences that possessed GO annotations associated with abiotic stresses or factors. A large number of genes possessed GO functions associated with salt stress response (449), followed by response to cold (282). There were 1,665 coding sequences with GO annotations specifically associated with biotic interactions. The most common GO annotation function related to biotic stress was 'defense response to bacteria' (296) followed by 'defense response to fungus' (280).

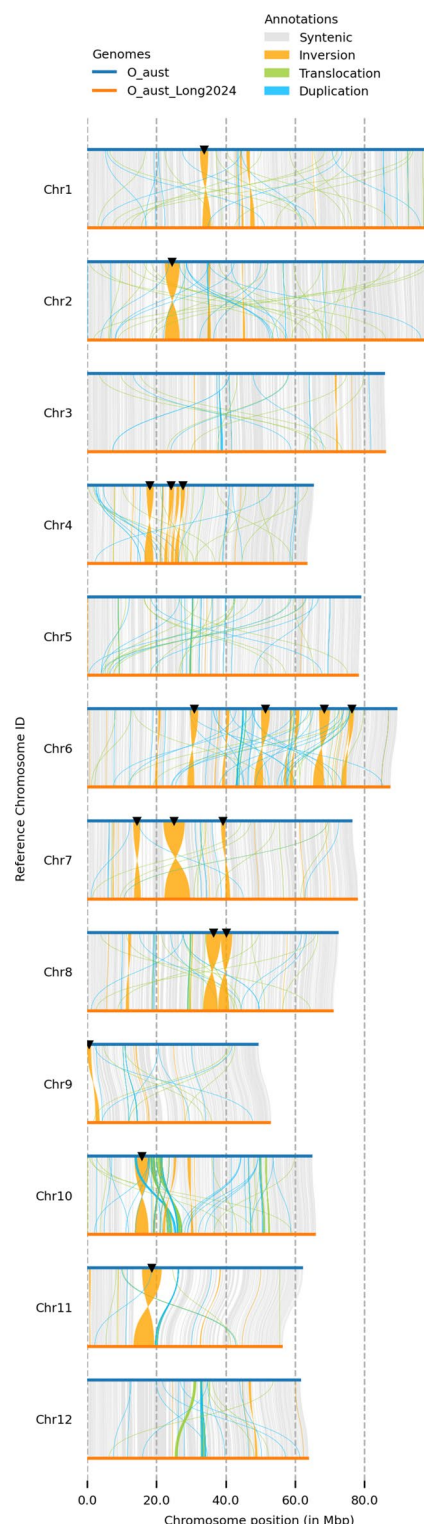


Fig. 3. Structural variation between *O. australiensis* assembly (O_aust) and previously published genome¹² (O_aust_Long2024). Arrows indicate large inversions (> 1.25% of chromosome length).

Oryza australiensis possesses species-specific genes and protein clusters

Species-specific protein clusters in *O. australiensis* were identified using OrthoVenn3²⁷, which grouped the 23,929 *O. australiensis* protein sequences from the primary assembly into 18,241 homologous clusters. Of these, 18,070 (99.1%) clusters were orthologous with *O. sativa* and 171 were unique to *O. australiensis*. There were 87 unique, *O. australiensis*-specific clusters that contained three or more protein sequences. These were annotated and assessed for common GO annotation terms using OrthoVenn3 (Supplementary table 3). Forty-six of these

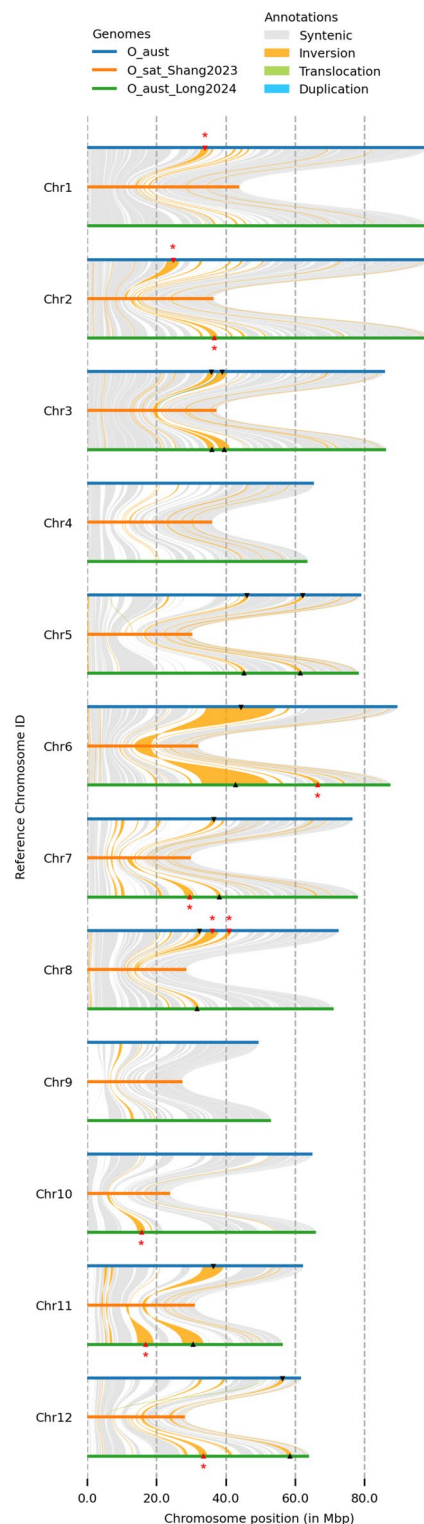


Fig. 4. Structural variations between cultivated rice (*O. sativa*) (O_sat_Shang2023) and *O. australiensis*. Black arrows indicate large inversions (> 1.25% of chromosome length) that are common to both the current (O_aust) and previously published¹² (O_aust_Long2024) assemblies. Red asterisks indicate large inversions unique to that assembly.

87 clusters remained unannotated and were not designated a GO function. The largest cluster, with 41 proteins, was annotated with the GO term 'DNA recombination' and matched to the Swiss-Prot hit 'retrovirus-related Pol polyprotein from transposon' (Q94HW2). Of the remaining nine largest protein clusters, two (both containing 12 proteins) were annotated with the GO term 'DNA integration' and similarly matched to the Swiss-Prot hit

Chromosome	Inversion starting position*	Inversion size (Mb)*	Inversion size difference** (Mb)	% of total chromosome length*	Unique ID (<i>O. australiensis</i> / Long et al. (2024) assembly)
Chr3	37,466,163	2.99	0.2039	3.48	INV690/INV692
Chr3	34,766,007	1.89	0.0391	2.20	INV689/INV691
Chr5	61,479,811	1.27	0.0028	1.60	INV720/INV719
Chr5	45,472,961	1.08	0.0784	1.37	INV718/INV717
Chr6	34,290,333	20.12	0.7676	22.46	INV732/INV734
Chr7	35,816,779	1.27	0.0369	1.66	INV751/INV751
Chr8	31,729,251	1.19	0.0009	1.64	INV770/INV769
Chr11	33,361,604	5.91	0.1385	9.49	INV651/INV646
Chr12	55,659,685	1.41	0.0968	2.29	INV660/INV661

Table 2. Common inversions with *O. sativa* found in *O. australiensis* assemblies. *Inversion sizes in current *O. australiensis* genome. **Inversion size difference between current and previous *O. australiensis* assemblies.

Chromosome	Inversion start position*	Inversion end position	Inversion size* (Mb)	% of total chromosome length	Unique ID
<i>O. australiensis</i> assembly					
Chr1	32,833,933	35,053,497	2.22	2.27	INV622
Chr2	22,957,943	26,598,750	3.64	3.67	INV669
Chr8	34,255,951	37,914,225	3.66	5.04	INV772
Chr8	40,185,784	41,638,446	1.45	2.00	INV622
Long et al. (2024) <i>O. australiensis</i> assembly					
Chr2	35,618,687	37,629,065	2.01	2.02	INV673
Chr6	65,599,041	67,350,432	1.75	2.00	INV737
Chr7	28,250,849	30,873,145	2.62	3.36	INV749
Chr10	14,483,181	16,709,441	2.23	3.37	INV637
Chr11	14,533,216	19,153,179	4.62	8.19	INV645
Chr12	32,406,185	34,609,095	2.20	3.44	INV655

Table 3. Unique inversions with *O. sativa* found in *O. australiensis* assemblies. *Inversion position and size in reference (*O. australiensis* assemblies).

‘retrovirus-related Pol polyprotein from transposon’ (P10978, P04323). One cluster, containing 19 proteins, was annotated with the GO term ‘nuclease activity’ (Swiss-Prot hit: Q9M2U3—Protein ALP1-like), and the remaining six largest clusters remained unannotated.

O. sativa ssp. *japonica* cv. Nipponbare Illumina reads were mapped to the *O. australiensis* predicted gene set to identify potential species-specific genes. There were 1,431 (6.0%) *O. australiensis* genes that could not be mapped with *O. sativa* reads. These were designated as ‘unmapped’, species-specific genes. Of these genes, 968 (67.7%) were GO-annotated in OmicsBox, 374 (26.1%) and 46 (3.2%) received BLAST hits or GO mapping, respectively, but no GO annotation, and 43 (3%) remained without BLAST hits, GO mapping, or annotation. These 43 genes underwent coding potential analyses, of which 34 (79.1%) and 41 (95.3%) genes were classified as coding genes when compared against the *Arabidopsis* and *Oryza* models, respectively. There were 102 unmapped gene sequences found in *O. australiensis* that possessed functions related to abiotic and biotic stress response. From these sequences, the most frequently occurring GO term was ‘response to other organism’ (GO:0051707), which was found in 60 of the unmapped coding sequences, followed by ‘defense response to other organism’ (43 genes) (GO:0098542). Gene enrichment analysis of the unmapped gene subset was conducted using a two-tailed Fisher’s Exact Test. There were twelve GO functions identified to be overrepresented in the unmapped *O. australiensis* genes (Table 4). Enriched GO functions related to biological processes included DNA integration (GO:0015074) ($p < .0001$), nucleosome assembly (GO:0006334) ($p < .0001$), calcium-mediated signalling (GO:0019722) ($p = .0003$), response to auxin (GO:0009733) ($p = .0004$), and cellular response to potassium ion starvation (GO:0051365) ($p = .0006$). Molecular functions that were overrepresented were related to protein heterodimerization activity (GO:0046982) ($p < .0001$), molecular tag activity (GO:0141047) ($p = .0009$), protein tag activity (GO:0031386) ($p = .0009$), and structural constituent of chromatin (GO:0030527) ($p < .0001$). There were also three enriched GO terms related to cellular components: nucleosome (GO:0000786) ($p < .0001$), protein-DNA complex (GO:0032993) ($p < .0001$), and extracellular regions (GO:0005576) ($p < .0001$).

The *O. australiensis* genome possesses homologues to cloned disease resistance genes
Two previously published datasets, one containing experimentally-verified plant nucleotide-binding leucine-rich repeat (NLR) disease resistance genes²⁸ (RefPlantNLR dataset) and one containing rice-specific cloned disease

GO Term	GO Name	GO Category	No. genes in unmapped subset	Total no. genes in genome	Unadjusted P-value (two-tailed Fisher's Exact Test)
GO:0015074	DNA integration	Biological process	34	93	<.0001
GO:0030527	Structural constituent of chromatin	Molecular function	19	32	<.0001
GO:0000786	Nucleosome	Cellular component	19	44	<.0001
GO:0046982	Protein heterodimerization activity	Molecular function	23	117	<.0001
GO:0032993	Protein-DNA complex	Cellular component	20	102	<.0001
GO:0005576	Extracellular region	Cellular component	87	923	<.0001
GO:0006334	Nucleosome assembly	Biological process	12	46	<.0001
GO:0019722	Calcium-mediated signalling	Biological process	9	33	.0003
GO:0009733	Response to auxin	Biological process	44	426	.0004
GO:0051365	Cellular response to potassium ion starvation	Biological process	5	9	.0006
GO:0141047	Molecular tag activity	Molecular function	7	23	.0009
GO:0031386	Protein tag activity	Molecular function	7	23	.0009

Table 4. Overrepresented GO functions in the unmapped, species-specific *O. australiensis* gene set.

resistance genes²⁹ (CDRH dataset) were used for identifying disease resistance homologues in *O. australiensis*. Seventy-nine (79) genes from the *O. australiensis* genome shared homology with 52 of the 481 protein sequences in the RefPlantNLR database, which were clustered into a total of 45 shared orthologous groups (Supplementary data 2). Within the 52 RefPlantNLR resistance genes identified to have homologues in *O. australiensis*, there were two (4%) leaf rust (*Lr*) resistance genes (*Lr13*—*g4425*, *Lr21*—*g22628*), three (6%) stripe rust (*Yr*) resistance genes (*Yr10*—*g22403*, *YrAS2388R*—*g19116*, *YrU1-RGA4*—*g13087*), and one bacterial blight (*Xo1*) resistance gene (*g11308*).

In comparisons against the rice-specific disease resistance gene set, there were 41 *O. australiensis* genes identified as homologues to 50 of the CDRH sequences, resulting in 39 orthologous clusters between the two datasets (Table 5).

Of these 41 genes, 11 were homologues to multiple cloned disease resistance genes, four were homologues to only *bph* resistance genes, 14 were homologues to *Pi* resistance genes, four were homologues to virus resistance genes, and eight were homologues to *Xa* resistance genes. One gene, *g18778*, located on *O. australiensis* chromosome 8, was clustered into an orthologue group with five cloned disease resistance genes (*Pi56*, *Pi5*, *Pish*, *Pi37*, *BPH14*). There were 10 genes in *O. australiensis* that were distributed on different chromosomes to their homologous counterpart in *O. sativa* (*g18778*, *g5963*, *g5428*, *g18763*, *g22620*, *g20366*, *g22575*, *g23592*, *g267*, *g23146*) (Fig. 5). *O. australiensis* chromosome 11 possessed the highest number of CDRHs (15 genes), followed by chromosome 4 (6 genes), chromosome 12 (5 genes), and chromosome 8 (3 genes). Gene homologues were predominantly concentrated around the distal chromosomal regions.

Prediction of disease resistance gene candidates in *O. australiensis*

The *O. australiensis* unmasked genome assembly and coding sequence files were processed through the NLR-Annotator gene prediction tool to identify potential NLR gene candidates (Table 6). Within the *O. australiensis* genome assembly there were 271 regions predicted to contain NLR motifs. In the *O. sativa* genome assembly, 517 sequences with NLR motifs were identified. In the coding sequence files, 132 and 464 potential NLR genes were predicted in *O. australiensis* and *O. sativa*, respectively. Chromosome 11 in *O. australiensis* possessed the greatest number of potential NLR loci (33 gene candidates) whilst chromosome two possessed the least (five candidates).

Potential RGAs were identified from annotated *O. australiensis* amino acid sequences using RGAugury (Table 6). There were 161 non-redundant NBS-encoding disease resistance gene candidates. Of these, 61 candidates possessed coiled-coil (CC), NBS, and LRR domains together (CNL-type), 56 possessed only NBS and LRR domains (NL-type), 18 possessed a CC and NBS domain, but no LRR (CN-type), and 19 consisted of solely an NBS domain (NBS-type). One NBS-LRR gene possessed a Resistance to Powdery Mildew 8 (RPW8) on the N-terminal domain (RNL-type). There were no genes of the TNL type (containing Toll/Interleukin-1 receptor (TIR), NBS, and LRR domains) identified in *O. australiensis*, however there were three genes containing a TIR and NBS domain (TN-type) and three containing a TIR and unknown domain (TX-type). In *O. sativa*, RGAugury predicted 556 NBS-encoding disease resistance gene candidates, of which 447 contained both NBS and LRR domains. There were 647 non-NBS-encoding disease resistance gene candidates identified in *O. australiensis*, of which 499 were RLKs, 39 were RLPs, 107 possessed transmembrane (TM)-CC domains, and two possessed RPW8-like domains (Table 2). In *O. sativa*, RGAugury identified 1042 RLKs, 119 RLPs, 190 TM-CCs, and one RPW8-like gene candidate.

Duplicated genes were identified within the *O. australiensis* genome using MCScanX. Across the entire genome, there were 1,870 genes classified as tandem duplicated genes (TDGs), 671 were proximal duplications, and 7,453 were dispersed duplications. Of the RGA candidates identified by RGAugury, 131, 83, and 337 genes were predicted to be tandem, proximal, and dispersed duplicates, respectively. The 131 RGAs classified as TDGs formed 53 clusters of two or more genes. The most commonly occurring TDG RGA class was RLKs (80 TDGs in

Cloned resistance gene	Chromosome in <i>O. sativa</i>	Homolog ID	Chromosome in <i>O. australiensis</i>	Position (bp)
BPH14	3	g18778 ^e	8	70,725,643
BPH17 = BPH15	4	g9829 g5963 ^a	4 2	12,989,050 94,387,918
BPH30	4	g9773 ^b g9771 ^c	4 4	6,502,789 6,493,820
BPH6	4	g9773 ^b g9771 ^c	4 4	6,502,789 6,493,820
BPH29	6	g13789	6	332,651
BPH32 = BPH3	6	g13853 g5428	6 2	811,139 80,727,292
BPH1 = BPH10 = BPH18 = BPH21/BPH2 = BPH26/BPH9/BPH7	12	g18763 ^d	8	70,497,397
Pi37	1	g18778 ^e	8	70,725,643
Pi64	1	g2494	1	80,963,823
Pish/Pi35	1	g18778 ^e	8	70,725,643
Pit	1	g22620 ^f	11	60,928,299
Pib	2	g18763 ^d	8	70,497,397
bsr-d1	3	g8179	3	35,610,037
pi21	4	g10138	4	41,854,052
Pi63	4	g11308 ^g	4	60,566,963
Pi9/Pi2 = Piz-5/Piz-t/PigmR/Pizh/Pi50	6	g14583 ^h	6	24,135,238
Pid2	6	g5963 ^a	2	94,387,918
Pi25/Pid3/Pid3-i1	6	g20366	10	9,726,536
Pid4	6	g14583 ^h	6	24,135,238
Pi36	8	g22575 ⁱ	11	60,094,608
Pi5 = Pi3 = Pi5-1	9	g18778 ^e	8	70,725,643
Pi56	9	g18778 ^e	8	70,725,643
NLS-1	11	g23592	12	56,614,115
Pb1	11	g22564	11	59,775,819
Pb2	11	g22628	11	61,122,843
Pb3	11	g22620 ^f	11	60,928,299
Pi65	11	g22647 g22649	11 11	61,521,424 61,710,179
Pia/Pi-CO39	11	g21945 g21947	11 11	8,906,283 8,914,911
Pik/Pik-m/Pik-p/Pi1/Pi7	11	g22457 ^j	11	57,177,540
Pik-h = Pi54	11	g22526	11	58,855,054
Pita	12	g23241	12	22,287,743
Pita2 = Ptr	12	g23527	12	54,874,956
RYMV2	1	g3213	1	94,012,729
RYMV1	4	g10662	4	49,739,246
RYMV3	11	g22575 ⁱ	11	60,094,608
STV11	11	g22199	11	48,878,289
tsv1	7	g16664	7	60,578,998
Xa1/Xa2 = Xa31(t)/Xa14/ Xa45(t)/CGS-Xo1	4	g11308 ^g	4	60,566,963
xa5	5	g11815	5	646,045
xa13/OsSWEET11/Os8N3	8	g18737	8	70,114,010
Xa10	11	g22393 ^k	11	55,921,351
Xa21	11	g267	1	2,980,453
Xa23	11	g22393 ^k	11	55,921,351
Xa3 = Xa26	11	g23146	12	8,680,012
Xa4	11	g22648	11	61,692,971
xa41(t)/OsSWEET14/Os11N3	11	g22206	11	49,408,893
Xa46 (t)	11	g22390	11	55,889,500
Xa47(t)	11	g22457 ^j	11	57,177,540
xa25/OsSWEET13/OsMtN3	12	g23351	12	45,463,545

Table 5. Cloned disease resistance gene homologues (CDRHs) identified in *O. australiensis*. Letters that are the same indicate *O. australiensis* genes that correspond to multiple cloned disease resistance genes. "=" indicates identical alleles, "/" indicates different alleles at the same locus.

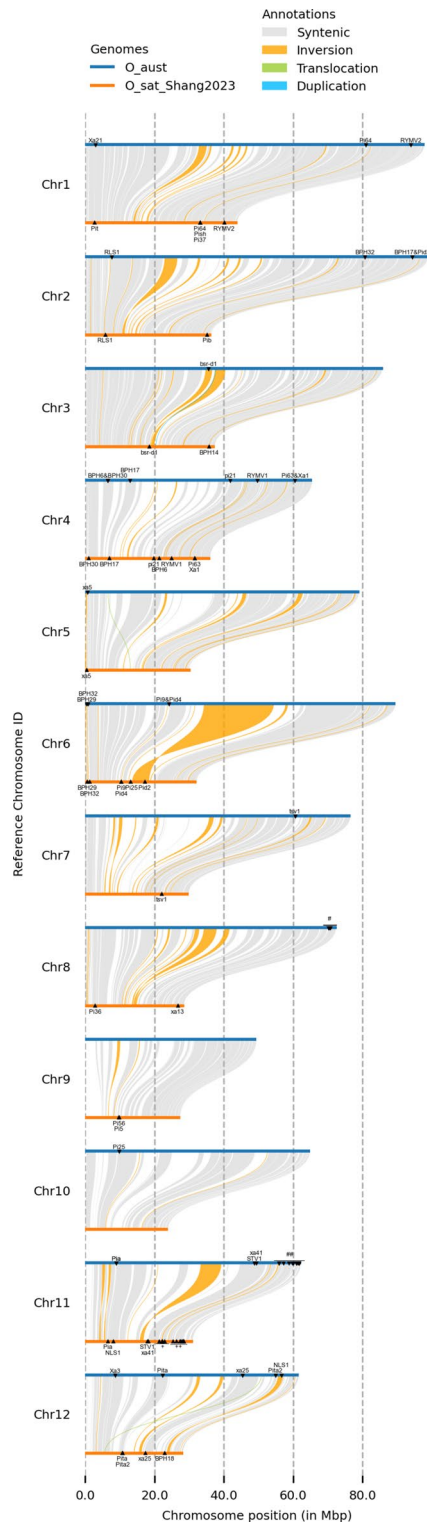


Fig. 5. Distribution of CDRHs in *O. australiensis* (O_aust) and corresponding position in *O. sativa* (O_sat_Shang2023). &= indicates that one gene in *O. australiensis* is homologous to more than one cloned disease resistance gene. # = gene positions from left to right: *xa13*, *BPH18&Pib*, *BPH14&Pi37&Pish&Pi5&Pi56*. ## = gene positions from left to right: *Xa46*, *Xa10&Xa23*, *Pik&Xa47*, *Pi54*, *Pb1*, *Pi36&RYMV3*, *Pit&Pb3*, *Pb2*, *Xa4*, *Pi65*. + = gene positions from left to right: *Xa21*, *Xa46*, *Xa10*, *Xa23*, *Pb1*. ++ = gene positions from left to right: *Pi54*, *RYMV3*, *Pb3*, *Pb2*, *Pik*, *Xa4*, *Pi65*, *Xa3*.

	NBS-encoding										
	NBS	CNL	CN	TN	NL	TX	RNL	RLP	RLK	TM-CC	RPW8
RGAugury	19	61	18	3	56	3	1	39	500	107	2
NLR-Annotator	7	60	7	0	53	0	1	–	–	–	–

Table 6. Summaries of number of disease resistance gene candidates identified in *O. australiensis* by NLR-Annotator and RGAugury. NBS = nucleotide binding site domain, C = coiled-coil domain, N = NBS, L = leucine rich-repeat (LRR) domain, T = Toll/interleukin-1/resistance (TIR) domain, X = other domain, RPW8 = Resistance to Powdery Mildew 8, R = RPW8-like domain, TM = transmembrane, RLP = receptor-like protein, RLK = receptor-like kinase.

31 clusters), followed by NBS-encoding (39 TDGs in 16 clusters), RLPs (6 TDGs in 3 clusters), TM-CCs (4 TDGs in 2 clusters), and RPW8 (2 TDGs in 1 cluster) (Supplementary table 4). Tandem duplications were typically distributed similarly to CDRHs, along the distal ends of chromosomes (Supplementary Fig. 2). Chromosomes 7 and 11 both had the highest number of clusters (seven clusters), however the clusters in chromosome 7 were larger (25 TDGs) compared to chromosome 11 (15 TDGs). Chromosomes 3 and 10 had the smallest number of TDGs (2 TDGs, 1 cluster).

Discussion

The wild relatives of crops are potential sources of novel genes that can improve agronomic traits in cultivated species. *Oryza australiensis* possesses characteristics that make it a promising source of abiotic and biotic stress resistance for rice improvement. *O. australiensis* has the largest genome within the diploid *Oryza* group owing to a rapid expansion of LTR retrotransposable elements, which have hindered the assembly and exploration of its genome until recent years^{12,15–17}. This paper reports a high-quality de novo assembly of a near-complete, chromosome-level genome for *O. australiensis*. Genome sequencing coverage was 80X, resulting in a highly contiguous assembly with a contig N50 of 72.6 Mb, which is higher than previously published assemblies^{12,15,17}. Eleven of the scaffolds from this assembly possessed telomeric repeats on both the forward and reverse ends of the scaffold, representing complete, ‘telomere-to-telomere’ pseudochromosomes. Ten of these *O. australiensis* pseudochromosomes were complete at the contig-level after initial assembly with the Hifiasm assembler. Overall, eleven of the twelve *O. australiensis* pseudochromosomes were complete post-scaffolding, all 24 telomeric ends of the pseudochromosomes were identified, and the final BUSCO score of the collapsed assembly was 99.3%. This represents a remarkably high-quality assembly of the *O. australiensis* genome. The final estimated size of the *O. australiensis* genome assembly was 909 Mb, which aligned exactly with the genome size estimated by flow cytometry. Previous length estimates of the *O. australiensis* genome have ranged from 858¹⁵ to 1056 Mb³⁰.

In addition to a collapsed genome assembly, two phased de novo haplotype assemblies were also assembled to chromosome-level. The haplotype assemblies exhibited similar BUSCO completeness, number of genes, repeat element content, and genome length to the collapsed assembly, with the exception of chromosome 5, where scaffolding tools joined two contigs to create a longer chromosome in the collapsed assembly, but not in the haplotype assemblies. Analyses of the assemblies revealed high structural synteny and low heterozygosity (~0.04%) between the haplotypes, suggesting genetic homogeneity within *O. australiensis* populations. Interestingly, comparisons against a previously published *O. australiensis* genome¹² identified a large number of inversions and translocations between the previous and current assemblies, suggesting a high level of genetic diversity across populations. These findings corroborate previous suggestions that genetic distance between *O. australiensis* populations is linked to geographic distance¹⁵. Extensive structural variation has also been reported between accessions of wild relatives of other crops, such as barley and cucumber^{31–35}. These variations, and specifically chromosomal inversions, are suggested to facilitate adaptive trait selection to abiotic and biotic stresses^{32,36,37}. Inversions between plant populations additionally play a role in suppressing recombination, resulting in the preservation of favourable allele combinations³⁸. It is therefore possible that *O. australiensis*—a widely distributed species found across a large geographic region in Australia—possesses extensive intra-species variation to enable and maintain local adaptation of populations to heterogenous environments. Additional work is needed to confirm that these variations are indeed biological and not a result of assembly artifacts. However, the high assembly quality and contiguity in both the current and previously published genomes strongly support the suggestion that these differences are real. As previously discussed in other studies, these findings also demonstrate the difficulty of capturing the extent of diversity within wild crop relatives using single reference genomes^{32,35}.

Despite an estimated 7 million years of evolutionary divergence, the *O. australiensis* assembly showed remarkable synteny with the *O. sativa* genome^{5,15}. Several large inversions (> 1.25% of total chromosome length) were observed to comprise 7.3% of the *O. australiensis* genome. The largest inversion between *O. australiensis* and *O. sativa* was found on chromosome 6 and comprised 22.5% (20.1 Mb) of the total chromosome length in *O. australiensis*. This inversion was common to both the current and previously published *O. australiensis* genomes. Similar large inversions occurring in the centric region of chromosome 6 have also previously been reported in other wild *Oryza* species^{12,39}.

Analysis of the *O. australiensis* repeat content revealed that approximately 76% of the genome consists of repeat elements, of which around 58% are LTR retrotransposons. In other studies, the *O. australiensis* total repeat and LTR retrotransposon content have similarly been estimated to be around 74–76% and ~60%, respectively^{12,15,16,40}. The *O. australiensis* repeat element content exceeds that of the other wild Australian rice

species, which are estimated to be 42% for *O. rufipogon*, 27% for *O. meridionalis*, and 65% for *O. officinalis*^{2,40}. The most commonly occurring transposable elements belonged to the *Gypsy* (32.3%) and *Copia* (10.6%) retrotransposon families. Previous studies suggest that these repeat elements are likely to be primarily comprised of the *Kangourou* and *Wallabi* (*Gypsy*-type) and *RIRE1* (*Copia*-type) LTR retrotransposons^{16,41}.

There were 23,929 genes predicted in the *O. australiensis* genome, supported by a BUSCO completeness of 98.2%. Whilst this higher BUSCO score supports our gene predictions, this gene number was lower than previous estimates for *O. australiensis*^{12,15}, and was also lower than those for *O. sativa* cultivated varieties, which range between 35,500 and 38,500 genes². However, our findings support previous suggestions that the large size of the *O. australiensis* genome is attributed to the expansion of retroelements, as opposed to gene duplications^{15,16}. Moreso, these findings suggest that the expansion of transposable elements and movement of retrotransposons within or near genes may have contributed to gene disruption, resulting in a decrease in functional genes within *O. australiensis*⁴². This is plausible given that transposable element activity is often deleterious and can induce strong loss-of-function mutations^{43,44}.

Mapping of *O. sativa* data to the *O. australiensis* genome identified 1,431 coding sequences and 171 protein clusters that were specific to *O. australiensis* and not found in cultivated rice. Of the 1,431 species-specific transcripts, 43 did not yield any BLAST or GO annotation results. Coding potential analyses showed that between 79% and 95% of these transcripts possessed coding potential, indicating that these could encode for novel genes that have not yet been identified in cultivated rice or other species. Genes and protein clusters specific to *O. australiensis* were found to be enriched in functions associated with DNA integration and retrotransposon activity, which coincides with the high transposable element content observed in this species¹⁶. Species-specific *O. australiensis* genes were also enriched in functions related to calcium-mediated signalling, which plays an important role in both abiotic and immune response signalling^{45,46}, as well as auxin response, which plays a role in plant growth but has also been linked to abiotic salt stress response in *Arabidopsis*⁴⁷.

Homologues to genes related to rice blast, bacterial blight, and BPH resistance were identified in the *O. australiensis* genome. Similar to other *Oryza* species², a large proportion of the *O. australiensis* disease resistance genes were positionally clustered, with the highest number of genes observed on the distal end of chromosome 11. In cultivated rice, major clusters of leaf blast (*Pi*) resistance genes have been identified on chromosomes 6, 11, and 12, whilst chromosomes 3 and 7 possess the least number of *Pi* resistance genes^{48,49}. Similarly in *O. australiensis*, several homologues to known *Pi* genes were located on chromosomes 6, 12, and the distal ends of chromosome 11, and no *Pi*-like genes were identified on chromosomes 3 and 7. In cultivated rice, the majority of brown planthopper (*bph*) resistance genes are located on chromosomes 3, 4, 6, and 12⁵⁰. In *O. australiensis*, *bph* homologues were found on all of these chromosomes. There were bacterial blight (*Xa*) resistance homologues observed to cluster in similar regions to the *Pi* genes on chromosome 11, similar to observations made in cultivated rice⁵¹.

There were between 118 and 132 nucleotide-binding leucine-rich repeat (NLR) disease resistance gene candidates identified within the *O. australiensis* genome, similar to previous estimates of 159 NLRs¹². There were no NLR genes of the TIR-NBS-LRR (TNL) type found in *O. australiensis*, supporting previous suggestions that the TNL gene type was discarded during evolution of the monocot lineage^{52–56}. Cultivated rice *O. sativa* possesses a higher number of NLRs, with previous estimates ranging between 374 and 535 genes, depending on cultivar and sub-species^{2,55,57–59}. Similarly, we predicted between 447 and 464 NLR genes within the *O. sativa* genome. Our results support previous suggestions that wild rices possess fewer NLR genes compared to cultivated rices as a result of artificial selection for pathogen resistance during domestication^{2,12}. Despite having a lower total number of genes, *O. australiensis* possessed homologues to almost all of the current cloned genes conferring resistance to the most globally important rice diseases. Allelic variations within known resistance loci can be sources of novel or additional disease resistance, for example, several variants of the *Pi9* blast resistance locus have been utilised across a number of rice cultivars^{29,60–62}. Moreso, in *O. australiensis*, a major QTL *Pi40* was also identified at the *Pi9* locus and introgressed into cultivated rice to confer broad spectrum resistance to rice blast in Korea and Türkiye^{11,63}. Consequently, the homologues identified in *O. australiensis* represent real candidates for disease resistance improvement in cultivated rice.

Conclusion

In this study we explored the genome of *O. australiensis*, a crop wild relative of rice. Despite 7 million years of evolutionary divergence and possessing a markedly larger genome, *O. australiensis* remains highly syntenic with cultivated rice, which has likely facilitated the introgression of agronomically valuable genes from the species previously^{11,64}. There is substantial structural variation between *O. australiensis* accessions, alluding to local environmental adaptation between populations, yet very low heterozygosity between haplotypes. This study provides new insights into the genetic diversity within this species, and supports the potential of *O. australiensis* to be a candidate source for novel abiotic and biotic stress resistance genes.

Materials and methods

Sample collection, DNA and RNA extractions, and sequencing

Oryza australiensis germplasm was sourced from the Australian Grains Genebank (AusTCRF#300,134, seed pack #593). Young fresh leaves from an *Oryza australiensis* plant (ID: Oaus_01) were collected from a glasshouse at The University of Queensland (UQ) Gatton campus. Total genomic DNA was extracted from pulverized leaf tissues using a CTAB (Cetyltrimethyl ammonium bromide) DNA extraction protocol⁶⁵ with the following amendments. Experimental steps involving phenol were replaced with 100% chloroform. Extracted DNA was dissolved and stored in 10 mM TrisHCl buffer. RNase A (Qiagen, 19,101) was added to a final concentration of 4 ng/uL after extraction and dissolving of DNA. Quality analyses of DNA extraction were conducted using

spectrophotometer and gel electrophoresis methods as described in the above protocol⁶⁵. High-fidelity (HiFi) reads were generated at the Australian Genome Research Facility (AGRF), UQ, using Pacific Biosciences (PacBio) circular consensus sequencing methods with two PacBio Sequel II SMRT cells. For the RNA-seq data, RNA was extracted from *O. australiensis* leaf tissue and preparation of RNA libraries was conducted by the Ramaciotti Centre for Genomics using the TruSeq Stranded Total RNA with Ribo-Zero plant kit. Sequencing was conducted at the Ramaciotti Centre for Genomics, University of New South Wales, Australia, using a NextSeq 500 system and 300-cycle MID output kit. For Hi-C sequencing, fresh young leaves were collected from an *O. australiensis* plant grown in glasshouse conditions at the Arizona Genomics Institute, the University of Arizona, USA. Hi-C library preparation and sequencing was performed by Dovetail Genomics, California, USA. The Dovetail Hi-C libraries were QC'd by sequencing ~1–2 M PE75bp reads using a MiSeq instrument and mapping the data back to the de novo assembly.

Genome length estimation by flow cytometry

Mature *Oryza sativa* ssp. *japonica* cv. Nipponbare, *Oryza australiensis*, and *Macadamia tetraphylla* were used for cytometric analyses. Mechanical dissociation was used to prepare nuclear suspensions as previously described⁶⁶, with modifications for woody plant species. Briefly, young leaves were freshly harvested and then chopped into ice-cold 0.5 mL of Arumuganathan and Earle⁶⁷ nuclear isolation buffer in a 5 cm polystyrene Petri dish. To estimate nuclear DNA content for rice, 40 mg of the internal standard *M. tetraphylla* was co-chopped with 15 mg of rice for approximately 7–8 min. Resulting homogenates were gently filtered through a pre-soaked 40-µm nylon mesh into a 5 mL round bottom polystyrene tube. Homogenates were then stained with 50 µg/mL of propidium iodide (PI) (Sigma, P4864-10ML) and 50 µg/mL of RNase A (Qiagen, 19,101) for 10 min on ice.

The BD Biosciences LSR II Flow Cytometer and FlowJo software package was used to analyse the homogenates. Fluorescence was collected using a 488 nm excitation laser tuned to 514.4 nm and a 610/20 nm bandpass filter. Instrument settings were kept constant across experiments: forward scatter voltage at 200, side scatter voltage at 350, fluorescence intensity voltage at 500, with a slow flow rate (20–50 events/s). Three biological replicates were performed on three different days. For each biological replicate, a minimum of 1,500 PI-stained events were collected per PI-stained peak. The %CV for the PI-stained peaks ranged between 2.2 and 4.3, with a mean of 3.0. Nuclear DNA content was calculated as previously described⁶⁸ using 796 Mb at 1C for the assumed size of *M. tetraphylla*.

Genome assembly

Hi-C sequencing data from Dovetail Genomics, California, USA, was integrated with PacBio HiFi reads using the Hifiasm Denovo assembler v0.19.4 (r575)¹⁸ to create two de novo contig haplotype assemblies and a collapsed, primary assembly. Genome assembly completeness was assessed against the 425 Benchmarking Universal Single-Copy Orthologues (BUSCO v5.4.6) within the viridiplantae lineage, using the Galaxy AUGUSTUS program workflow^{69,70}. Genome contiguity was assessed with the quality assessment tool for genome assemblies (QUAST v5.2.0)⁷¹. Contigs were mapped with Hi-C proximity ligation libraries using Chromap v0.2.5⁷² and scaffolded using the SALSA2 scaffolding tool¹⁹. Additional scaffolds were joined with the RagTag scaffolding tool v2.1.0²¹, using *Oryza sativa* (Osativa323v7 Phytozome) as a reference (<https://phytozome-next.jgi.doe.gov/>). Resulting scaffolds were assigned to corresponding *O. sativa* chromosomes. Lengths of pseudochromosomes were estimated using the seqtk toolkit⁷³. Telomeric repeat regions were characterised in pseudochromosomes using the Telomere Identification Toolkit (tidk) (<https://github.com/tolkkit/telomeric-identifier>) and the plant telomeric repeat 'AAACCCT'. Centromeric regions were predicted using CentroMiner within quarTeT⁷⁴. K-mer analysis was conducted on the *O. australiensis* genome using Jellyfish (v2.3.1)⁷⁵ and GenomeScope2.0⁷⁶ using a k-mer size of 21.

Structural annotation

Structural annotation was performed on the finalised *O. australiensis* collapsed and haplotype assemblies. Transposable element repeats within the *O. australiensis* genome were identified and masked using RepeatModeler2 v2.0.5⁷⁷ and RepeatMasker v4.1.5⁷⁸. Repeat element content within the *O. australiensis* genome was assessed with RepeatMasker v4.1.5 using the soft masking option. Paired-end RNA-seq files were adaptor-trimmed using TrimGalore! v0.6.10⁷⁹ and aligned to genomes using HISAT2⁸⁰. Gene prediction was performed with Braker3 v3.0.3²⁵ using RNA-seq data and viridiplantae protein databases as support inputs. Subread⁸¹ and featureCounts⁸² were used to assess the proportion of genes with RNA-seq support. BUSCO analysis using the viridiplantae lineage was used to predict genome completeness of the predicted protein sequences. The LTR Assembly Index (LAI) score for assessing assembly contiguity was calculated using LTR_retriever^{23,24}.

Density distributions of predicted genes and repeat element content was visualised with shinyCircos V2.0²⁶. The repeat element data used for density visualisations were generated using EDTA v2.2.0 (parameters: -annoc 1)²².

Structural synteny and genome visualisation

O. australiensis chromosomes were compared with those from *O. sativa* (NCBI: PRJNA953663⁸³ and Osativa323v7 from Phytozome), as well as those from a previously published *O. australiensis* assembly¹² using D-Genies v1.5²⁰ and Syri⁸⁴ with Minimap2 v2.28⁸⁵ to investigate intra- and inter-species structural synteny. Inversions between assemblies were designated as 'large' if inversion length exceeded 1.25% of the total chromosome length³⁴.

Functional annotation

Functional annotation of the *O. australiensis* CDS annotation file was performed in OmicsBox 3.1.11 (OmicsBox—Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019, <https://www.biobam.com>)

using the following workflow. Sequences were searched with the Basic Local Alignment Search Tool (BLAST) against a non-redundant protein sequence database using BLASTX with viridiplantae taxonomy and an e-value threshold of 1.0×10^{-10} . Sequences were run through InterProScan⁸⁶ and Gene Ontology (GO) terms were retrieved from BLAST hits using Gene Ontology mapping with Blast2GO⁸⁷ annotation. Blast2GO annotations were then combined with GO terms produced from InterProScan. Sequences with no BLAST hits were run through the Coding Potential Analysis Test (CPAT)⁸⁸ in OmicsBox 3.1.11 to search for potential coding genes. CPAT was run using the prebuilt *Arabidopsis thaliana* model and with a *Oryza* specific model (organism: 45270 *Oryza*).

Prediction of candidate unique coding sequences and protein clusters in *O. australiensis*

O. sativa ssp. *japonica* Illumina read data was extracted from the National Center for Biotechnology Information (NCBI) (SRR15967546 & SRR15967547) and mapped against the *O. australiensis* gene list to identify candidate unique coding sequences. Data was mapped using the Map Reads to Reference tool (length fraction: 0.9 and similarity fraction: 0.9) on the QIAGEN CLC Genomics Workbench 24.0.2 (QIAGEN, Aarhus, Denmark).

O. australiensis genes that were not mapped with any *O. sativa* Illumina reads were designated as ‘unmapped’ genes and were functionally annotated with OmicsBox 3.1.11. Sequences that had no BLAST hits were then assessed for coding potential using the CPAT analysis described above. OmicsBox was used with FatiGO⁸⁹ to conduct a two-tailed Fisher’s Exact test to assess for enriched GO functions in the unmapped genes compared to the overall *O. australiensis* gene set. Adjusted *p*-values were calculated within the FatiGO pipeline using the Benjamini and Hochberg false discovery rate method^{89,90}. The threshold for determining differentially represented GO terms was determined using a *p*-value cut-off of .05.

The protein sequence data for both *O. australiensis* and *O. sativa* (Osativa323v7 protein file Phytozome) were filtered for the longest isomer and then analysed for orthologous and unique protein clusters within the *O. australiensis* genome using OrthoVenn3 (parameters: OrthoFinder algorithm, E-value: 1×10^{-2} , Inflation value: 1.50)^{27,91}. GO enrichment analyses and annotation for unique *O. australiensis* clusters were automatically run through the OrthoVenn3 platform.

Identification of cloned disease resistance gene orthologues in *O. australiensis*

Cloned disease resistance gene sequences in rice compiled in a previous paper²⁹ were searched and downloaded from the Rice Annotation Project (RAP)⁹² and MSU Rice Genome Annotation Project databases^{93,94} (Supplementary table 5). This compiled list of genes was designated as the cloned disease resistance homologue (CDRH) list. An additional database, RefPlantNLR, containing 481 amino acid sequences corresponding to experimentally-verified plant NLRs within a number of important plant species was also retrieved²⁸. The RefPlantNLR database and compiled CDRH gene lists were searched for orthologues within the *O. australiensis* genome using OrthoFinder v2.5.5 with default parameters⁹¹.

Identification of potential resistance gene analogues in *O. australiensis*

The collapsed *O. australiensis* genome was run through two tools, NLR-Annotator⁹⁵ and RGAugury⁵⁵, to identify disease resistance gene candidates. NLR-Annotator annotates potential NLR loci by identifying amino acid motifs associated with NLRs. NLR-Annotator was run against both the *O. australiensis* unmasked assembly and the annotated CDS sequence file, as well as against the *O. sativa* (Osativa323v7 protein file Phytozome) unmasked genome assembly and CDS file. The annotated *O. australiensis* amino acid sequence file was run through RGAugury, which identifies resistance gene analogues (RGAs), including NLRs, receptor-like kinases (RLKs), and receptor-like proteins (RLPs) using the programs BLAST⁹⁶, nCoils⁹⁷, pfam_scan⁹⁸, Phobius⁹⁹, and InterProScan⁸⁶.

Identification of tandem duplicated resistance genes

Tandem duplicated genes were identified with MCScanX¹⁰⁰ using the *O. australiensis* longest isomer protein sequence file and corresponding GFF file. The *O. australiensis* protein sequence file was blasted against itself to identify homologous genes using BLASTp⁹⁶ within DIAMOND (v2.1.7.161)¹⁰¹. The following parameters were specified: –more-sensitive, –max-target-seqs 5, –evalue 1×10^{-45} , –query-cover 70. The BLAST output and the *O. australiensis* GFF file were run through duplicate_gene_classifier within MCScanX¹⁰⁰. RGA candidates identified by RGAugury that were classified as duplicated gene pairs were extracted and manually designated as tandem duplicates if they belonged to the same RGA gene family and satisfied one or both of the following criteria: 1) were within 100 kb of each other, or 2) were adjacent with no intervening non-homologous genes in between them.

Data availability

The primary *O. australiensis* genome assembly and annotation is available at the China National Center for Bio-information Genome Warehouse repository under BioProject PRJCA030604, Accession GWHFIGT00000000.1. Primary and haplotype assemblies are submitted under the National Center for Biotechnology Information (NCBI) repository (primary BioProject: PRJNA1144493, haplotype 1 BioProject: PRJNA1162258, haplotype 2 BioProject: PRJNA1162260). HiFi read sequencing data was submitted to the NCBI Sequencing Read Archive (SRR31008054) and *O. australiensis* sample details were submitted under BioSample SAMN43753486. Additional data related to haplotype annotation (<https://doi.org/10.6084/m9.figshare.28245260.v1>), species-specific genes and protein clusters (<https://doi.org/10.6084/m9.figshare.28252145.v1>), structural variations (<https://doi.org/10.6084/m9.figshare.28605827.v1>), and disease resistance gene identification (<https://doi.org/10.6084/m9.figshare.28366184.v1>) is published in FigShare (https://figshare.com/projects/Oryza_australiensis_as_a_source_for_novel_genes/234,941).

Received: 30 May 2025; Accepted: 23 July 2025

Published online: 30 July 2025

References

- Muthayya, S., Sugimoto, J. D., Montgomery, S. & Maberly, G. F. An overview of global rice production, supply, trade, and consumption. *Ann. N. Acad. Sci.* **1324**, 7–14. <https://doi.org/10.1111/nyas.12540> (2014).
- 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. <https://doi.org/10.1038/s41588-018-0040-0> (2018).
- Tirnaz, S. et al. Application of crop wild relatives in modern breeding: An overview of resources, experimental and computational methodologies. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2022.1008904> (2022).
- Henry, R. J. et al. Australian *Oryza*: Utility and conservation. *Rice* **3**, 235–241. <https://doi.org/10.1007/s12284-009-9034-y> (2010).
- 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. <https://doi.org/10.1016/j.pbi.2013.02.014> (2013).
- Hamzelou, S. et al. Wild and cultivated species of rice have distinctive proteomic responses to drought. *Int. J. Mol. Sci.* <https://doi.org/10.3390/ijms21175980> (2020).
- Scafaro, A. P. et al. Heat tolerance in a wild *Oryza* species is attributed to maintenance of Rubisco activation by a thermally stable Rubisco activase ortholog. *New Phytol.* **211**, 899–911. <https://doi.org/10.1111/nph.13963> (2016).
- Yichie, Y., Brien, C., Berger, B., Roberts, T. H. & Atwell, B. J. Salinity tolerance in Australian wild *Oryza* species varies widely and matches that observed in *O. sativa*. *Rice (N Y)* **11**, 66. <https://doi.org/10.1186/s12284-018-0257-7> (2018).
- Jena, K. K., Jeung, J. U., Lee, J. H., Choi, H. C. & Brar, D. S. High-resolution mapping of a new brown planthopper (BPH) resistance gene, *Bph18(t)*, and marker-assisted selection for BPH resistance in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* **112**, 288–297. <https://doi.org/10.1007/s00122-005-0127-8> (2006).
- Multani, D. S. et al. Development of monosomic alien addition lines and introgression of genes from *Oryza australiensis* Domin. to cultivated rice *O. sativa* L. *Theor. Appl. Genet.* **88**, 102–109. <https://doi.org/10.1007/BF00222401> (1994).
- Jeung, J. U. et al. A novel gene, *Pi40(t)*, linked to the DNA markers derived from NBS-LRR motifs confers broad spectrum of blast resistance in rice. *Theor. Appl. Genet.* **115**, 1163–1177. <https://doi.org/10.1007/s00122-007-0642-x> (2007).
- Long, W. et al. Genome evolution and diversity of wild and cultivated rice species. *Nat. Portfolio* <https://doi.org/10.21203/rs.3.rs-4350570/v1> (2024).
- Gu, K. et al. R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature* **435**, 1122–1125. <https://doi.org/10.1038/nature03630> (2005).
- Cao, Z. et al. Natural variation of *HTH5* from wild rice, *Oryza rufipogon* Griff., is involved in conferring high-temperature tolerance at the heading stage. *Plant Biotechnol. J.* **20**, 1591–1605. <https://doi.org/10.1111/pbi.13835> (2022).
- Phillips, A. L. et al. The first long-read nuclear genome assembly of *Oryza australiensis*, a wild rice from northern Australia. *Sci. Rep.* **12**, 10823. <https://doi.org/10.1038/s41598-022-14893-5> (2022).
- Piegu, B. et al. Doubling genome size without polyploidization: Dynamics of retrotransposon-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* **16**, 1262–1269. <https://doi.org/10.1101/gr.5290206> (2006).
- Fornasiero, A. et al. *Oryza* genome evolution through a tetraploid lens. *Nat. Genet.* <https://doi.org/10.1038/s41588-025-02183-5> (2025).
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175. <https://doi.org/10.1038/s41592-020-01056-5> (2021).
- Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genom.* **18**, 527. <https://doi.org/10.1186/s12864-017-3879-z> (2017).
- Cabanettes, F. & Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958. <https://doi.org/10.7717/peerj.4958> (2018).
- Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258. <https://doi.org/10.1186/s13059-022-02823-7> (2022).
- Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275. <https://doi.org/10.1186/s13059-019-1905-y> (2019).
- Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126. <https://doi.org/10.1093/nar/gky730> (2018).
- Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422. <https://doi.org/10.1104/pp.17.01310> (2018).
- Gabriel, L. et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res* **34**, 769–777. <https://doi.org/10.1101/gr.278090.123> (2024).
- Wang, Y. et al. shinyCircos-V2.0: Leveraging the creation of Circos plot with enhanced usability and advanced features. *iMeta* **2**, e109. <https://doi.org/10.1002/imt2.109> (2023).
- Sun, J. et al. OrthoVenn3: An integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Res.* **51**, W397–W403. <https://doi.org/10.1093/nar/gkad313> (2023).
- Kourelis, J., Sakai, T., Adachi, H. & Kamoun, S. RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. *PLoS Biol.* **19**, e3001124. <https://doi.org/10.1371/journal.pbio.3001124> (2021).
- Simon, E. V. et al. Available cloned genes and markers for genetic improvement of biotic stress resistance in rice. *Front Plant Sci.* **14**, 1247014. <https://doi.org/10.3389/fpls.2023.1247014> (2023).
- Iyengar, G. A. S. & Sen, S. K. Nuclear DNA content of several wild and cultivated *Oryza* species. *Environ. Exp. Bot.* **18**, 219–224. [https://doi.org/10.1016/0098-8472\(78\)90047-3](https://doi.org/10.1016/0098-8472(78)90047-3) (1978).
- Hu, H. et al. Genomic signatures of barley breeding for environmental adaptation to the new continents. *Plant Biotechnol. J.* **21**, 1719–1721. <https://doi.org/10.1111/pbi.14077> (2023).
- Zhang, W. et al. Genome architecture and diverged selection shaping pattern of genomic differentiation in wild barley. *Plant Biotechnol. J.* **21**, 46–62. <https://doi.org/10.1111/pbi.13917> (2023).
- Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e113. <https://doi.org/10.1016/j.cell.2020.05.023> (2020).
- Hu, H. et al. Unravelling inversions: Technological advances, challenges, and potential impact on crop breeding. *Plant Biotechnol. J.* **22**, 544–554. <https://doi.org/10.1111/pbi.14224> (2024).
- Li, H. et al. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat. Commun.* **13**, 682. <https://doi.org/10.1038/s41467-022-28362-0> (2022).
- Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation?. *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42. <https://doi.org/10.1146/annurev.ecolsys.39.110707.173532> (2008).
- Huang, K. & Rieseberg, L. H. Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Front Plant Sci.* **11**, 296. <https://doi.org/10.3389/fpls.2020.00296> (2020).

38. Wilkinson, M. J. et al. Centromeres are hotspots for chromosomal inversions and breeding traits in mango. *New Phytol.* **245**, 899–913. <https://doi.org/10.1111/nph.20252> (2025).
39. Abdullah, M., Furtado, A., Masouleh, A. K., Okemo, P. & Henry, R. The genomes of the most diverse AA genome rice species provide a resource for rice improvement and studies of rice evolution and domestication. *BMC Genom.* **26**, 54. <https://doi.org/10.1186/s12864-025-11246-0> (2025).
40. Gill, N. et al. Dynamic Oryza genomes: Repetitive DNA sequences as genome modeling agents. *Rice* **3**, 251–269. <https://doi.org/10.1007/s12284-010-9054-7> (2010).
41. Uozu, S. et al. Repetitive sequences: Cause for variation in genome size and chromosome morphology in the genus Oryza. *Plant Mol. Biol.* **35**, 791–799. <https://doi.org/10.1023/a:1005823124989> (1997).
42. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M. A. LTR-retrotransposons in plants: Engines of evolution. *Gene* **626**, 14–25. <https://doi.org/10.1016/j.gene.2017.04.051> (2017).
43. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta Gene Regul. Mech.* **1860**, 157–165. <https://doi.org/10.1016/j.bbaggm.2016.05.010> (2017).
44. Ramakrishnan, M. et al. Transposable elements in plants: Recent advancements, tools and prospects. *Plant Mol. Biol. Report.* **40**, 628–645. <https://doi.org/10.1007/s11105-022-01342-w> (2022).
45. Xu, T., Niu, J. & Jiang, Z. Sensing mechanisms: Calcium signaling mediated abiotic stress in plants. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2022.925863> (2022).
46. Jiang, Y. & Ding, P. Calcium signaling in plant immunity: A spatiotemporally controlled symphony. *Trends Plant Sci.* **28**, 74–89. <https://doi.org/10.1016/j.tplants.2022.11.001> (2023).
47. Cackett, L. et al. Salt-specific gene expression reveals elevated auxin levels in arabidopsis thaliana plants grown under saline conditions. *Front Plant Sci.* **13**, 804716. <https://doi.org/10.3389/fpls.2022.804716> (2022).
48. Ning, X., Yunyu, W. & Aihong, L. Strategy for use of rice blast resistance genes in rice molecular breeding. *Rice Sci.* **27**, 263–277. <https://doi.org/10.1016/j.rsci.2020.05.003> (2020).
49. Tan, Q. et al. Integrated genetic analysis of leaf blast resistance in upland rice: QTL mapping, bulked segregant analysis and transcriptome sequencing. *AoB PLANTS* <https://doi.org/10.1093/aobpla/plac047> (2022).
50. Muduli, L. et al. Understanding brown planthopper resistance in rice: Genetics, biochemical and molecular breeding approaches. *Rice Sci.* **28**, 532–546. <https://doi.org/10.1016/j.rsci.2021.05.013> (2021).
51. Jamaluddin, M. et al. in *Rice Improvement: Physiological, Molecular Breeding and Genetic Perspectives* (eds Jauhar Ali & Shabir Hussain Wani) 315–378 (Springer, 2021).
52. Akita, M. & Valkonen, J. P. T. A novel gene family in moss (*Physcomitrella patens*) shows sequence homology and a phylogenetic relationship with the TIR-NBS class of plant disease resistance genes. *J. Mol. Evol.* **55**, 595–605. <https://doi.org/10.1007/s00239-002-2355-8> (2002).
53. Bai, J. et al. Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res.* **12**, 1871–1884. <https://doi.org/10.1101/gr.454902> (2002).
54. Cannon, S. B. et al. Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. *J. Mol. Evol.* **54**, 548–562. <https://doi.org/10.1007/s0023901-0057-2> (2002).
55. Li, P. et al. RGAugury: A pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genom.* **17**, 852. <https://doi.org/10.1186/s12864-016-3197-x> (2016).
56. Meyers, B. C. et al. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.* **20**, 317–332. <https://doi.org/10.1046/j.1365-313x.1999.t01-1-00606.x> (1999).
57. Gottin, C. et al. A new comprehensive annotation of leucine-rich repeat-containing receptors in rice. *Plant J.* **108**, 492–508. <https://doi.org/10.1111/tpj.15456> (2021).
58. Zhou, T. et al. Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol. Genet. Genom.* **271**, 402–415. <https://doi.org/10.1007/s00438-004-0990-z> (2004).
59. Li, J. et al. Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. *Mol. Genet. Genom.* **283**, 427–438. <https://doi.org/10.1007/s00438-010-0527-6> (2010).
60. Imam, J., Mandal, N. P., Variar, M. & Shukla, P. Allele mining and selective patterns of Pi9 gene in a set of rice landraces from India. *Front Plant Sci.* **7**, 1846. <https://doi.org/10.3389/fpls.2016.01846> (2016).
61. Liu, J. et al. Genetic variation and evolution of the Pi9 blast resistance locus in the AA genome Oryza species. *J. Plant Biol.* **54**, 294–302. <https://doi.org/10.1007/s12374-011-9166-7> (2011).
62. Zhou, Y. et al. Identification of novel alleles of the rice blast-resistance gene Pi9 through sequence-based allele mining. *Rice (N. Y.)* **13**, 80. <https://doi.org/10.1186/s12284-020-00442-z> (2020).
63. Beşer, N. et al. Marker-assisted introgression of a broad-spectrum resistance gene, Pi40 improved blast resistance of two elite rice (*Oryza sativa* L.) cultivars of Turkey. *Mol. Plant Breed.* **7**, 1–15 (2016).
64. Ishii, T., Brar, D. S., Multani, D. S. & Khush, G. S. Molecular tagging of genes for brown planthopper resistance and earliness introgressed from *Oryza australiensis* into cultivated rice, *O. sativa*. *Genome* **37**, 217–221. <https://doi.org/10.1139/g94-030> (1994).
65. Furtado, A. DNA extraction from vegetative tissue for next-generation sequencing. *Methods Mol. Biol.* **1099**, 1–5. https://doi.org/10.1007/978-1-62703-715-0_1 (2014).
66. Galbraith, D. W. et al. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220**, 1049–1051. <https://doi.org/10.1126/science.220.4601.1049> (1983).
67. Arumuganathan, K. & Earle, E. D. Estimation of nuclear DNA content of plants by flow cytometry. *Plant Mol. Biol. Report.* **9**, 229–241. <https://doi.org/10.1007/BF02672073> (1991).
68. Doležel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2**, 2233–2244. <https://doi.org/10.1038/nprot.2007.310> (2007).
69. 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. <https://doi.org/10.1093/bioinformatics/btv351> (2015).
70. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312. <https://doi.org/10.1093/nar/gkh379> (2004).
71. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. <https://doi.org/10.1093/bioinformatics/btt086> (2013).
72. Zhang, H. et al. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat. Commun.* **12**, 6566. <https://doi.org/10.1038/s41467-021-26865-w> (2021).
73. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962. <https://doi.org/10.1371/journal.pone.0163962> (2016).
74. Lin, Y. et al. quarTeT: A telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* <https://doi.org/10.1093/hr/uhad127> (2023).
75. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770. <https://doi.org/10.1093/bioinformatics/btr011> (2011).
76. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432. <https://doi.org/10.1038/s41467-020-14998-3> (2020).

77. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457. <https://doi.org/10.1073/pnas.1921046117> (2020).
78. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*, <<http://www.repeatmasker.org>> (2013–2015).
79. Krueger, F. et al. FelixKrueger/TrimGalore: v0.6.10. (2023). <https://doi.org/https://zenodo.org/records/7598955>
80. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915. <https://doi.org/10.1038/s41587-019-0201-4> (2019).
81. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108. <https://doi.org/10.1093/nar/gkt214> (2013).
82. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930. <https://doi.org/10.1093/bioinformatics/btt656> (2014).
83. Shang, L. et al. A complete assembly of the rice Nipponbare reference genome. *Mol. Plant* **16**, 1232–1236. <https://doi.org/10.1016/j.molp.2023.08.003> (2023).
84. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277. <https://doi.org/10.1186/s13059-019-1911-0> (2019).
85. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191> (2018).
86. Quevillon, E. et al. InterProScan: Protein domains identifier. *Nucleic Acids Res.* **33**, W116–120. <https://doi.org/10.1093/nar/gki442> (2005).
87. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genom.* **2008**, 619832. <https://doi.org/10.1155/2008/619832> (2008).
88. Wang, L. et al. CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74–e74. <https://doi.org/10.1093/nar/gkt006> (2013).
89. Al-Shahrour, F., Díaz-Uriarte, R. & Dopazo, J. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578–580. <https://doi.org/10.1093/bioinformatics/btg455> (2004).
90. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).
91. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238. <https://doi.org/10.1186/s13059-019-1832-y> (2019).
92. Ohyanagi, H. et al. The rice annotation project database (RAP-DB): Hub for *Oryza sativa* ssp. *japonica* genome information. *Nucleic Acids Res.* **34**, D741–D744. <https://doi.org/10.1093/nar/gkj094> (2006).
93. Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N. Y.)* **6**, 4. <https://doi.org/10.1186/1939-8433-6-4> (2013).
94. Hamilton, J. P., Li, C. & Buell, C. R. The rice genome annotation project: An updated database for mining the rice genome. *Nucleic Acids Res.* **53**, D1614–D1622. <https://doi.org/10.1093/nar/gkae1061> (2024).
95. Steuernagel, B. et al. The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire1 [OPEN]. *Plant Physiol.* **183**, 468–482. <https://doi.org/10.1104/pp.19.01273> (2020).
96. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214. <https://doi.org/10.1089/10665270050081478> (2000).
97. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164. <https://doi.org/10.1126/science.252.5009.1162> (1991).
98. Finn, R. D. et al. The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–222. <https://doi.org/10.1093/nar/gkp985> (2010).
99. Käll, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036. <https://doi.org/10.1016/j.jmb.2004.03.016> (2004).
100. Wang, Y. et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49. <https://doi.org/10.1093/nar/gkr1293> (2012).
101. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368. <https://doi.org/10.1038/s41592-021-01101-x> (2021).

Acknowledgements

The authors acknowledge The University of Queensland Research Computing Centre (UQ-RCC) staff for the computing resources and support provided during this study. The authors would also like to acknowledge the contributions of Upuli Nakandala and Upendra Kumari Wijesundara for their support in this project.

Author contributions

RH, AKM, RW, and OP supervised, managed the project, advised and supported data analysis, and data interpretation. AKM supported the genome assembly work. SM performed DNA extractions, data analysis and data interpretation. LC contributed flow cytometry analysis. The manuscript was organized and written by SM. LC contributed to the manuscript with data interpretation of flow cytometry analysis. All authors approved the submitted version.

Funding

The sample collection, sequencing, and analyses associated with this work were funded by The Australian Research Council Centre of Excellence for Plant Success in Nature and Agriculture (CE200100015).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-13310-x>.

Correspondence and requests for materials should be addressed to R.J.H.

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

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025