Improved hybrid *de novo* genome assembly and annotation of African wild rice, *Oryza longistaminata*, from Illumina and PacBio sequencing reads

Wei Li¹* | Kui Li¹* | Qun-jie Zhang¹,²* | Ting Zhu²,³* | Yun Zhang²* | Cong Shi¹ | Yun-long Liu² | En-hua Xia² | Jian-jun Jiang² | Chao Shi²,⁴ | Li-ping Zhang² | Hui Huang² | Yan Tong² | Yuan Liu² | Dan Zhang¹ | Yuan Zhao² | Wen-kai Jiang² | You-jie Zhao⁵ | Shu-yan Mao² | Jun-ying Jiao² | Ping-zhen Xu² | Li-li Yang² | Guo-ying Yin¹ | Li-zhi Gao¹,²

¹Inst. of Genomics and Bioinformatics, South China Agricultural Univ., Guangzhou, 510642, China
²Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwestern China, Kunming Inst. of Botany, Chinese Acad. of Sciences, Kunming, 650204, China
³College of Life Science, Liaoning Normal Univ., Dalian, 116081, China
⁴Univ. of the Chinese Acad. of Sciences, Beijing, 100039, China
⁵Yunnan Agricultural University, Kunming, 650201, China

Correspondence
Li-zhi Gao, Inst. of Genomics and Bioinformatics, South China Agricultural Univ., Guangzhou, 510642, China.
Email: Lgaogenomics@163.com

Funding information
Natural Science Foundation of China, Grant/Award Numbers: 31501025, 31601045; Key Project of the Natural Science Foundation of Yunnan Province; Yunnan Innovation Team Project, Grant/Award Number: 2015FA030

*Equal contribution.

Abstract
African wild rice *Oryza longistaminata*, one of the eight AA-genome species in the genus *Oryza*, possesses highly valued traits, such as the rhizomatousness for perennial rice breeding, strong tolerance to biotic and abiotic stresses, and high biomass production on poor soils. To obtain the high-quality reference genome for *O. longistaminata* we employed a hybrid assembly approach through incorporating Illumina and PacBio sequencing datasets. The final genome assembly comprised only 107 scaffolds and was approximately ~363.5 Mb, representing ~92.7% of the estimated African wild rice genome (~392 Mb). The N50 lengths of the assembled contigs and scaffolds were ~46.49 Kb and ~6.83 Mb, indicating ~3.72-fold and ~18.8-fold improvement in length compared to the earlier released assembly (~12.5 Kb and 364 Kb, respectively). Aided with Hi-C data and syntenic relationship with *O. sativa*, these assembled scaffolds were anchored into 12 pseudo-chromosomes. Genome annotation and comparative genomic analysis reveal that lineage-specific expansion of gene families that respond to biotic- and abiotic stresses are of great potential for mining novel alleles to overcome major diseases and abiotic adaptation in rice breeding programs.

Abbreviations: BUSCO, benchmarking universal single-copy orthologs; BWA, burrows wheeler aligner; EVM, EVidenceModeler; GO, gene ontology; LTR, long terminal repeat; ncRNA, non-coding RNA; ORF, open reading frame; PacBio, Pacific Biosciences; PASA, Program to Assemble Spliced Alignments; SMRT, Single Molecule Real Time; SSR, simple sequence repeat; TE, transposable element.
This reference genome of African wild rice will greatly enlarge the existing database of rice genome resources and unquestionably form a solid base to understand genomic basis underlying highly valued phenotypic traits and search for novel gene sources in *O. longistaminata* for the future rice breeding programs.

## 1 | INTRODUCTION

*Oryza longistaminata* A. Chev. & Roehr. is among the eight AA- genome species in the genus *Oryza*. This wild rice species, which is native to Africa, owns highly prized traits for genetic improvement of cultivated rice, including well-made rhizomes for perennial rice breeding (Glover et al., 2010), strong tolerance to biotic and abiotic stresses (Brar & Khush, 1997, Tanksley & McCouch, 1997, Xiao et al., 1998), and high biomass production on poor soils. Decoding the *O. longistaminata* genome is undoubtedly a key to uncover molecular mechanisms that determine these remarkable agronomic traits and enhance rice genetic improvement. We recently sequenced the five AA- genome rice species, and obtained the first insights into the rice genome adaptation (Zhang et al., 2014). As a result of the nature of inherent self-incompatibility in *O. longistaminata*, however, it remains a great challenge to decipher this genome because of its relatively high heterozygosity of 1.05% estimated by 17 k-mer distribution (Figure S1).

A high-quality *O. longistaminata* genome assembly is fundamental to mine rich source of novel alleles that are of great importance for breeding new rice cultivars in the future. Based on a large quantity of combined sequencing data (~396 × Illumina short reads and ~5.9 × Roche GS FLX reads), Zhang et al. (Zhang et al., 2015) reported a draft genome assembly of *O. longistaminata* with a short (12.5 Kb) contig N50 length. The disadvantage of low-quality genome assembly makes it difficult to gain worthwhile information of genomic variation, functional genomic studies as well as efficient germplasm utilization. Here we report a high-quality genome assembly as well as large transcriptome datasets for *O. longistaminata* using Illumina and PacBio sequencing platforms.

## 2 | MATERIALS AND METHODS

### 2.1 | Plant materials

An individual plant of *O. longistaminata* (IRGC 81967) originally from Botswana was introduced and grown in the greenhouse of Kunming Institute of Botany, Chinese Academy of Sciences. The collected leaf samples were immediately frozen in liquid nitrogen and stored at −80°C in the laboratory.

### 2.2 | DNA extraction, library construction and sequencing

The nuclear DNA was isolated from young leaves using a modified CTAB method. The extracted DNA was examined using a NanoDrop D-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and electrophoresis on a 0.8% agarose gel, respectively. According to the Illumina’s protocol, three short-insert (300 bp, 360 bp and 500 bp) paired-end and five long-insert (2 Kb, 4 Kb, 5 Kb, 6 Kb and 8 Kb) mate-pair genomic DNA libraries were constructed and sequenced using Hiseq2000 platform. 20 μg of high-quality DNA was used to construct 20-Kb SMRT Bell library. The library was then sequenced on the PacBio RSII platform with P6-C4 chemistry combination.

### 2.3 | Estimation of the genome size

Two orthogonal methods were used to estimate the genome size of *O. longistaminata*, including k-mer frequency distribution and flow cytometric analyses. We generated the 17-mer occurrence distribution of sequencing reads from short libraries (≤500 bp) using Jellyfish (version 2.1.3) (Marcais & Kingsford, 2011). The genome size and heterozygosity were then estimated using GCE version 1.0 (Liu, Shi, Yuan, Hu, & Wei, 2013) based on the k-mer counting.

The genome size of the *O. longistaminata* was further assessed by flow cytometry analysis. Approximately 50 mg of fresh leaves was used for sample preparation according to Huang et al. (Huang, Tong, Zhang, & Gao, 2013) using Otto’s lysis buffer (Loureiro, Rodriguez, Dolezel, & Santos, 2006). A BD FACSCalibur (USA) flow cytometer was used to estimate the nuclear DNA amount. The Cellquest software (version 5.1) was employed to analyze the mean ratio of $G_{0/1}$ (Nuclei) peak and the internal standard. We selected the rice cultivar *Nipponbare* as inner standard with an estimated genome size of 389 Mb (IRGSP., 2005), and estimated the genome size of *O. longistaminata* based on the linear relationship of 2C peaks.
2.4 Hybrid de novo genome assembly and quality assessment

Low-quality raw reads were trimmed using Trimomatic (version 0.33) (Bolger, Lohse, & Usadel, 2014). ALLPATHS-LG (Gnerre et al., 2011) was then used to assemble the *O. longistaminata* genome. Gaps between scaffolds were then filled with GapCloser (version 1.12) (Luo et al., 2012) based on all pair-end reads. Potential bacterial-derived scaffolds were removed by Blast searches against bacterial genomes with an identity of \( \geq 95\% \) and E-value \( \leq 1 \times 10^{-5} \). Based on the above-generated de novo genome assembly using Illumina sequencing reads, we used \( \sim 7.28 \text{ Gb} \) (18.57 \( \times \)) PacBio RS reads longer than 500 bp in the hybrid assembly using the PBJelly 2 (English et al., 2012) with default parameters. Finally, Lachesis (Burton et al., 2013) was used to link the scaffolds into pseudo-chromosomes with default parameters.

Three approaches were used to evaluate the assembly quality of the *O. longistaminata* genome. First, we mapped all high-quality reads (\( \sim 21.6 \text{ Gb} \)) to the assembled genome using BWA (version 0.7.15) (Li & Durbin, 2009). Second, the completeness of the assembly was assessed with BUSCO (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) collected from Embryophyta lineage. Finally, RNA sequencing reads generated in this study were assembled into transcripts using Trinity (version 2.0.6) (Grabherr et al., 2011), which were then aligned back to our genome assembly using GMAP (Wu & Watanabe, 2005) at 90% coverage and 90% identity thresholds.

2.5 RNA extraction and transcriptome sequencing

Total RNA was isolated from the four tissues of *O. longistaminata*, including 30-d-roots, 30-d-shoots, panicles at booting stage and flag leaves at booting stage, using a Water Saturated Phenol method (Zhang et al., 2014). Following the Illumina’s recommendations, RNA sequencing libraries were constructed and paired-end sequenced on the Illumina Hiseq2000 platform, generating \( \sim 26.73 \text{ Gb} \) of RNA sequencing (RNA-Seq) data.

2.6 Genome annotation

Repetitive sequences of the *O. longistaminata* genome assembly were masked prior to gene prediction. A combined strategy was adopted to predict the protein-coding genes of *O. longistaminata*. Augustus (version 3.0.3) (Stanke, Steinkamp, Waack, & Morgenstern, 2004), GlimmerHMM (version 3.0.3) (Majoros, Pertea, & Salzberg, 2004) and GeneMarkHMM (version 3.47) (Lukashin & Borodovsky, 1998) were used to detect the hypothetical gene-coding regions in the *O. longistaminata* genome. The protein sequences from *Oryza*, *Zea mays*, *Sorghum bicolor*, and *Brachypodium distachyon* were aligned to the *O. longistaminata* genome assembly using GenBlastA (version 1.0.1) (She, Chu, Wang, Pei, & Chen, 2009) and further refined using GeneWise (version 2.2.0) (Birney, Clamp, & Durbin, 2004). To further aid the gene annotation Illumina RNA-seq reads (\( \sim 26.73 \text{ Gb} \)) were first assembled into transcripts using Trinity (version 2.0.6) (Grabherr et al., 2011), and then aligned to the genome assembly using PASA (Program to Assemble Spliced Alignments) (Haas et al., 2003) to determine the potential gene structures. EVidenceModeler (EVM) (Haas et al., 2008) was used to combine all predicted results from ab initio, protein and EST evidences into consensus gene predictions. We filtered out gene models with their peptide lengths \( \leq 50 \text{aa} \) and/or harboring stop codons to obtain the final predicted genes. The functional motifs and domains were determined using InterProScan (version 5.3) (Jones et al., 2014). PFAM domains and GO terms were directly retrieved from the corresponding InterPro entries.

TEs were annotated integrating RepeatMasker (www.repeatmasker.org), LTR_STRUC (McCarthly & McDonald, 2003), RECON (Bao & Eddy, 2002), and LTR_Finder (Xu & Wang, 2007). All intact LTR retrotransposons generated by LTR_STRUC were classified into Ty1-copia, Ty3-gypsy, and unclassified superfamilies according to the order of ORFs using PFAM (Punta et al., 2012). The 5′- and 3′-LTR were aligned using ClustalW (Larkin et al., 2007) and the K-value (the average number of substitutions per aligned site) was calculated by PAML (Yang, 2007). The insertion times (\( T \)) were calculated using the formula: \( T = K/2r \), where \( r \) represents the substitution rate, which is \( 1.3 \times 10^{-8} \) per site per year (Baucom, Estill, Leebens-Mack, & Bennetzen, 2009, Vitte, Panaud, & Quesneville, 2007). SSRs in the
**O. longistaminata** genome were identified using microsatellite identification tool (MISA) (Thiel, Michalek, Varshney, & Graner, 2003).

### 2.7 Gene family evolution analysis

To avoid potential contamination introduced by TE-related proteins, the predicted genes were blasted against the MSU *Oryza* Repeat Database. OrthoMCL (Li, Stoeckert, & Roos, 2003) was used to obtain paralogous and orthologous gene families with the finished genomes: *O. sativa*, *O. longistaminata*, *B. distachyon*, and *Arabidopsis thaliana*. Then, we used RAxML package (Stamatakis, 2006) to construct a phylogenetic tree of single-copy gene families. Subsequently, we used CAFÉ (version 2.2) (De Bie, Crisianini, Demuth, & Hahn, 2006) to detect the expansion and contraction of gene families. To further examine the evolutionary relationships in the genus *Oryza*, MCScan (Haibao et al., 2008) was used to construct genome collinearity blocks. The predicted genes for the two *Oryza* species were used to all-vs-all comparisons.

### 2.8 Data availability

The raw sequence data and genome assembly reported in this paper have been deposited in BIG Data Center (Nucleic Acids Res 2018) and NCBI (National Center for Biotechnology Information) Data Center under accession numbers PRICA000900 and PRJNA545798, respectively.

### RESULTS AND DISCUSSION

#### 3.1 Genome sequencing

The nuclear genome of *O. longistaminata* was sequenced using whole-genome shotgun sequencing (WGS). Based on the 17-mer frequency distribution of sequencing reads, the genome size of *O. longistaminata* was estimated to be ~392 Mb, and flow cytometry analysis further showed that the *O. longistaminata* genome was approximately estimated to be ~363 Mb. We obtained a total of 166.25 Gb of clean Illumina data, representing ~424-fold coverage (Supplemental Table S1). Then we performed single-molecule long read sequencing on the PacBio platform. This process produced ~7.28 Gb data, consisting of 873,994 reads with a length of 8,324 bp on average (Supplemental Figure S2; Supplemental Table S1).

#### 3.2 Hybrid de novo genome assembly and quality assessment

Heterozygosity is a challenging problem that complicates to assemble genomes, resulting in highly fragmented genome assemblies (Michael & VanBuren, 2015). A hybrid assembly method was employed to confront high heterozygosity in the *O. longistaminata* genome (Supplemental Figure S3). Firstly, the genome of *O. longistaminata* was assembled using ALLPATHS-LG (Gnerre et al., 2011), yielding an initial *O. longistaminata* genome assembly of ~331 Mb. The N50 lengths of the assembled contigs and scaffolds were ~17.06
Kb and ~1.11 Mb, respectively (Supplemental Table S2). Secondly, the SMRT reads were further added, yielding a genome assembly of ~371.3 Mb, representing ~94.7% of the estimated genome size (~392 Mb) (Supplemental Table S3). The N50 lengths of the assembled contigs and scaffolds increased to ~45.65 Kb and ~1.24 Mb, respectively (Supplemental Table S2). A total of 433 scaffolds larger than 100 Kb in length occupied 97.50% of the genome assembly (Figure S4a; Figure S4b). Aided with ~116 Gb (296× genome coverage) Hi-C data we further linked these scaffolds using Lachesis (Burton et al., 2013) with default parameters. This produced only 107 scaffolds with a total length of the final genome assembly of ~363.5 Mb, representing ~92.7% of the estimated genome size for the African wild rice (Table 1; Supplemental Table S4). The contig N50 and scaffold N50 sizes represent ~3.72-fold and ~18.8-fold improvement compared with previously reported ~12.5 Kb and 364 Kb (Zhang et al., 2015), respectively. It is apparent that the genome assembly were largely improved after adding long SMRT reads, making it feasible to accurately predict protein-coding genes and sufficiently annotate repeat sequences. We finally assembled scaffolds of O. longistaminata into 12 pseudo-chromosomes (Figure 1a; Supplemental Table S5) based on Hi-C data and syntenic relationships with O. sativa ssp. japonica cv. Nipponbare (MSU 7.0). The assembled genome was referred to as Oryza_longistaminata_v2.0.

For the validation of genome assembly, we first mapped all high-quality reads (~21.6 Gb) to the assembled genome sequence, and obtained an average mapping rate of 86.48%
3.3 | Transcriptome sequencing and de novo assembly

We performed transcriptome sequencing to improve the quality of the *O. longistaminata* genome annotation. A total of ~26.73 Gb of RNA sequencing (RNA-Seq) data was generated, representing major tissue types and developmental stages, including 30-d-shoots, 30-d-roots, panicles at booting stage and flag leaves at booting stage (Supplemental Table S7). The clean data was assembled using Trinity (version r20140717) (Grabherr et al., 2011) and only the longest contigs per component were counted, yielding a total of 111,105 transcripts with a N50 value of 1,064 bp (Supplemental Table S8).

3.4 | Genome annotation

A total of 35,440 protein-coding genes were predicted in the *O. longistaminata* genome assembly (Supplemental Table S9; Figure 1b). We aligned the annotated protein-coding genes of *O. sativa* and RNA-seq data of *O. longistaminata* generated in this study to assess the quality of gene prediction. Our results showed that 73.7% of the predicted gene models was supported by transcript or protein evidences (Figure 1c), demonstrating a high accuracy of the gene prediction of the *O. longistaminata* genome.
Non-coding RNA (ncRNA) genes were predicted using \textit{ab initio} and/or homology search methods (Zhang et al., 2014). We identified 755 transfer RNA (tRNA) genes, 4 ribosomal RNA (rRNA) genes, 274 small nucleolar RNA (snoRNA) genes, 143 small nuclear RNA (snRNA) genes, and 299 microRNA (miRNA) genes (Table 1; Supplemental Table S10).

The annotation of repeat sequences showed that approximately 33.09\% of the \textit{O. longistaminata} genome consisted of transposable elements (TEs) (Supplemental Table S11; Figure 2a), which is relatively lower than in the \textit{Nipponbare} genome (~39.40\%) (Figure 2b). Long terminal repeat (LTR) retrotransposons occupied ~20.79\% of the \textit{O. longistaminata} genome (Supplemental Table S11). Ty3-gypsy superfamilies appeared more abundant (6.01\%) than Ty1-copia superfamilies (2.03\%) (Supplemental Table S11; Figure 2c). Molecular dating suggested that nearly 60\% of LTR retrotransposons came to the \textit{O. longistaminata} genome within the last 2 MYA (Figure 2d), which is consistent with earlier findings in the rice genome (Gao, McCarthy, Ganko, & McDonald, 2004; Tian et al., 2009).

A total of 197,053 simple sequence repeats (SSRs) were identified in the \textit{O. longistaminata} genome, and the frequency of occurrence was estimated to be one SSR every 1.88 kb (Supplemental Table S12). Mono-, di-, tri-, tetra-, penta-, and hexa-mer accounted for 13.93\%, 13.94\%, 34.31\%, 25.81\%,
7.71%, and 4.30% of the identified SSRs, respectively (Supplementary Table S12). These identified SSRs can be used as valuable genetic markers for the germplasm characterization and utilization of *O. longistaminata*.

### 3.5 Rice gene family evolution

The predicted protein-coding genes of *O. longistaminata* were compared with those from rice (IRGSP., 2005), *A. thaliana* (The Arabidopsis Genome Initiative, 2000), and *B. distachyon* (The International Brachypodium Initiative, 2010). A total of 134,448 genes were clustered into 22,189 gene families (Figure 3a). Of these, 7,995 gene families, which contained 50,443 genes, were common to all four plant species, representing ancestral gene families (Figure 3b). There were 380 gene families containing 861 genes unique to *O. longistaminata*, of which some were potentially related to important biological processes (Figure 3b). Further functional annotation based on PFAM terms revealed that a total of 380 lineage-specific gene families were enriched with genes associated with wall-associated receptor kinase galacturonan-binding (PF13947), cathepsin propeptide inhibitor domain (PF08246) and pectinesterase (PF01095) (Supplemental Figure S5). Our data analysis revealed that a total of 1,556 and 4,300 gene families underwent significant expansion and contraction, respectively (Figure 3c). PFAM annotation showed that the defense genes were highly enriched in the expanded gene families, such as NB-ARC domain (PF00391) and leucine rich repeat (PF00560) (Figure 3d). This is consistent with earlier findings that *O. longistaminata* exhibits a high resistance to insect pests and diseases (Song et al., 1995), suggesting a great potential for mining novel alleles to overcome major diseases and abiotic adaptation in rice breeding program. A total of 1,200 syntenic blocks covered 56.73% of gene repertoire of *O. longistaminata* were identified (Figure 3e and Figure S6), indicating a highly conserved collinearity between the two *Oryza* species.

### 4 CONCLUSIONS

We report and annotate a high-quality reference genome of *O. longistaminata* to enlarge the existing database of rice genome resources. A large collection of genomic sequences of Asian cultivated rice and its wild relatives has indisputably formed a solid base for hunting novel gene sources from the gene pool of wild rice germplasms. Thus, it is our hope that the generation of high-quality reference genome of this African wild rice will enhance studies on comparative genomics of the genus *Oryza*, genomic basis underlying highly valued phenotypic traits, such as resistance, rhizomatousness and self-incompatibility, and germplasm exploitation of useful genes in *O. longistaminata* for the future rice breeding programs.

### ACKNOWLEDGMENTS

We thank International Rice Research Institute (Manila, Philippines) for kindly providing rice germplasms. This work was supported by the Key Project of the Natural Science Foundation of Yunnan Province (2015FA030), the start-up grant from South China Agricultural University and Yunnan Innovation Team Project (to L. G.), Natural Science Foundation of China (31501025) (to Y. L.), and Natural Science Foundation of China (31601045) (to Q. Z.).

### CONFLICT OF INTEREST DISCLOSURE

The authors declare that there are no conflicts of interest related to this manuscript.

### ORCID

Li-zhi Gao [https://orcid.org/0000-0003-0772-2770](https://orcid.org/0000-0003-0772-2770)

### REFERENCES

Bao, Z., & Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research, 12*, 1269–1276. [https://doi.org/10.1101/gr.88502](https://doi.org/10.1101/gr.88502).

Baucom, R. S., Estill, J. C., Leebens-Mack, J., & Bennezen, J. L. (2009). Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Research, 19*, 243–254. [https://doi.org/10.1101/gr.083360.108](https://doi.org/10.1101/gr.083360.108).

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Research, 14*, 988–995. [https://doi.org/10.1101/gr.1865504](https://doi.org/10.1101/gr.1865504).

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimer for Illumina sequence data. *Bioinformatics, 30*, 2114–2120. [https://doi.org/10.1093/bioinformatics/btu170](https://doi.org/10.1093/bioinformatics/btu170).

Brar, D. S., & Khush, G. S. (1997). Alien introgression in rice. *Plant Molecular Biology, 35*, 35–47.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology, 31*, 1119–1125. [https://doi.org/10.1038/nbt.2727](https://doi.org/10.1038/nbt.2727).

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics, 22*, 1269–1271. [https://doi.org/10.1093/bioinformatics/bti097](https://doi.org/10.1093/bioinformatics/bti097).

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., … Gibbs, R. A. (2012). Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *Plos One, 7*, e47768. [https://doi.org/10.1371/journal.pone.0047768](https://doi.org/10.1371/journal.pone.0047768).

Gao, L., McCarthy, E. M., Ganko, E. W., & McDonald, J. F. (2004). Evolutionary history of *Oryza sativa* LTR retrotransposons: A preliminary survey of the rice genome sequences. *Bmc Genomics [Electronic Resource], 5*, 18. [https://doi.org/10.1186/1471-2164-5-18](https://doi.org/10.1186/1471-2164-5-18).
Glover, J. D., Reganold, J. P., Bell, L. W., Borevitz, J., Brummer, E. C., Buckler, E. S., … Xu, Y. (2010). Agriculture. Increased food and ecosystem security via perennial grains. *Science*, 328, 1638–1639. https://doi.org/10.1126/science.1188761.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., … Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS*, 108, 1513–1518. https://doi.org/10.1073/pnas.1017351108.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644–652. https://doi.org/10.1038/nbt.1883.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., … White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31, 5654–5666.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., … Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, 9, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

Haibo, T., Bowers, J. E., Xiyin, W., Ray, M., Maqsudul, A., & Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, 320, 486–488.

Huang, H., Tong, Y., Zhang, Q. J., & Gao, L. Z. (2013). Genome size variation among and within *Camellia* species by using flow cytometric analysis. *Plos One*, 8, e64981. https://doi.org/10.1371/journal.pone.0064981.

IRGSP. (2005). The map-based sequence of the rice genome. *Nature*, 436, 793–800. https://doi.org/10.1038/nature03895.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McNaula, C., … Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGregor, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, L., Stoeckert, C. J., Jr., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13, 2178–2189. https://doi.org/10.1101/gr.1224503.

Liu, B., Shi, Y., Yuan, J., Hu, X., & Wei, F. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology*, 35, 62–67.

Loureiro, J., Rodriguez, E., Dolezel, J., & Santos, C. (2006). Comparison of four nuclear isolation buffers for plant DNA flow cytometry. *Annals of Botany*, 98, 679–689. https://doi.org/10.1093/aob/mcl141.

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research*, 26, 1107–1115.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1, 18. https://doi.org/10.1186/2047-217X-1-18.

Majors, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, 20, 2878–2879. https://doi.org/10.1093/bioinformatics/bth315.

Marcais, G., & Kingsford, C. (2011). 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.

McCarthy, E. M., & McDonald, J. F. (2003). LTR_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics*, 19, 362–367.

Michael, T. P., & VanBuren, R. (2015). Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology*, 24, 71–81. https://doi.org/10.1016/j.pbi.2015.02.002.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, 40, D290–301. https://doi.org/10.1093/nar/gkr1065.

She, R., Chu, J. S., Wang, K., Pei, J., & Chen, N. (2009). GenBlastA: Enabling BLAST to identify homologous gene sequences. *Genome Research*, 19, 143–149. https://doi.org/10.1101/gr.082081.108.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Song, W. Y., Wang, G. L., Chen, L. L., Kim, H. S., Pi, L. Y., Holsten, T., … Ronald, P. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science*, 270, 1804–1806.

Stamatakis, A. (2006). RAXML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22, 2688–2690. https://doi.org/10.1093/bioinformatics/btl446.

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32, W309–312. https://doi.org/10.1093/nar/gkh379.

Tanksley, S. D., & McCouch, S. R. (1997). Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science*, 277, 1063–1066.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408, 796–815. https://doi.org/10.1038/35048692.

The International Brachypodium Initiative. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, 463, 763–768. https://doi.org/10.1038/nature08747.

Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*, 106, 411–422. https://doi.org/10.1007/s00122-002-1031-0.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J. L., Jackson, S. A., … Ma, J. (2009). Do genomic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Research*, 19, 2221–2230. https://doi.org/10.1101/gr.083899.108.

Vitte, C., Panaud, O., & Quesneville, H. (2007). LTR retrotransposons in rice (*Oryza sativa*, L.): Recent burst amplifications followed by rapid DNA loss. *BMC Genomics [Electronic Resource]*, 8, 218. https://doi.org/10.1186/1471-2164-8-218.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310.
Xiao, J., Li, J., Grandillo, S., Ahn, S. N., Yuan, L., Tanksley, S. D. et al. (1998). Identification of trait-improving quantitative trait loci alleles from a wild rice relative, *Oryza rufipogon*. *Genetics, 150*, 899–909.

Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research, 35*, W265–268. https://doi.org/10.1093/nar/gkm286.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution, 24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.

Zhang, Q. J., Zhu, T., Xia, E. H., Shi, C., Liu, Y. L., Zhang, Y., … Gao, L. Z. (2014). Rapid diversification of five *Oryza* AA genomes associated with rice adaptation. *PNAS, 111*, E4954–4962. https://doi.org/10.1073/pnas.1418307111.

Zhang, Y., Zhang, S., Liu, H., Fu, B., Li, L., Xie, M., … Hu, F. (2015). Genome and comparative transcriptomics of African wild rice *Oryza longistaminata* provide insights into molecular mechanism of rhizomatousness and self-incompatibility. *Molecular Plant, 8*, 1683–1686. https://doi.org/10.1016/j.molp.2015.08.006.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

---

**How to cite this article:** Li W, Li K, Zhang Q-j, et al. Improved hybrid de novo genome assembly and annotation of African wild rice, *Oryza longistaminata*, from Illumina and PacBio sequencing reads. *Plant Genome*. 2020;13:e20001. https://doi.org/10.1002/tpg2.20001