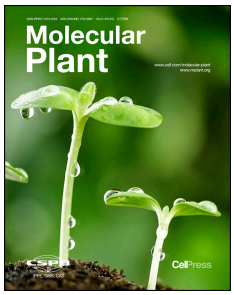
Accepted Manuscript



Genome and comparative transcriptomics of African wild rice Oryza longistaminata provide insights into molecular mechanism of rhizomatousness and self-incompatibility

Yesheng Zhang, Shilai Zhang, Hui Liu, Binying Fu, Lijuan Li, Min Xie, Yue Song, Xin Li, Jing Cai, Wenting Wan, Ling Kui, Hui Huang, Jun Lyu, Yang Dong, Wensheng Wang, Liyu Huang, Jing Zhang, Qinzhong Yang, Qinli Shan, Qiong Li, Wangqi Huang, Dayun Tao, Muhua Wang, Mingsheng Chen, Yeisoo Yu, Rod A. Wing, Wen Wang, Fengyi Hu

PII: S1674-2052(15)00352-4
DOI: 10.1016/j.molp.2015.08.006
Reference: MOLP 171

To appear in: *MOLECULAR PLANT*
Accepted Date: 12 August 2015

Please cite this article as: **Zhang Y., Zhang S., Liu H., Fu B., Li L., Xie M., Song Y., Li X., Cai J., Wan W., Kui L., Huang H., Lyu J., Dong Y., Wang W., Huang L., Zhang J., Yang Q., Shan Q., Li Q., Huang W., Tao D., Wang M., Chen M., Yu Y., Wing R.A., Wang W., and Hu F.** (2015). Genome and comparative transcriptomics of African wild rice Oryza longistaminata provide insights into molecular mechanism of rhizomatousness and self-incompatibility. Mol. Plant. doi: 10.1016/j.molp.2015.08.006.

Genome and comparative transcriptomics of African wild rice *Oryza longistaminata* provide insights into molecular mechanism of rhizomatousness and self-incompatibility

Short Summary

A reference genome of *O. longistaminata* is presented here. Comparative genomics and transcriptomics study identified genes and pathways that may be related to resistance, rhizomatousness and self-incompatibility.

Dear editor,

Oryza longistaminata (*O. longistaminata*) is an African wild rice species with AA genome type and possesses special traits that are highly valued for improving the cultivated rice, such as strong resistances to biotic and abiotic stresses (Song et al., 1995) for improving resistance of cultivars, rhizomatousness for perennial breeding (Glover et al., 2010) and self-incompatibility (SI) for new ways to produce hybrid seeds (Ghesquiere, 1986). Deciphering the genome of *O. longistaminata* will be the key to uncover the mechanism of above hallmark traits and further to improve the cultivated rice. However, it is a big challenge for us to decipher the genome of this species since its high heterozygosity (1.3% estimated by k-mer distribution and 2.7 polymorphic sites per kilobase in the final assembly, **Supplementary Figure 1, Supplementary Table 1, Supplementary Information S2.8**) as the result of its inherent self-incompatibility.

Here we *de novo* assembled a total length ~347Mb reference genome of *O. longistaminata* (contig N50 12.5 kb; scaffold N50 363 kb) based on a large quantity of combined sequencing data (~396× Illumina short reads and ~5.9× Roche GS FLX+ long reads) (**Supplementary Table 2-8; Supplementary Figure 2-6**). Further, we assembled scaffolds of *O. longistaminata* into 12 pseudo-chromosomes (**Figure 1A**)

based on its syntenic relationship with *Oryza glaberrima* (Wang et al., 2014) and the resulting assembled genome was referred to as Oryza_longistaminata_v1.0 (**Supplementary Information S8)**.

Annotations were conducted using the assembled genome. The transposable elements (TEs) annotation results show that 48.73% of the genome assembly were TEs (**Supplementary Table 9, 10; Supplementary Figure 7**), which is a markedly higher percentage than previously found in *O. glaberrima* (34.25%), *O. brachyantha* (29.2%) and *O. sativa* (34.8%). We annotated 32,502 protein-coding genes (**Supplementary Table 11, 12; Supplementary Figure 8, 9**) and the non-protein-coding RNA genes, including 3,954 putative miRNA, 720 tRNA, 34 rRNA, and 669 snRNA genes (**Supplementary Table 13**).

Divergent times analysis among 6 Gramineae species, including *Oryza sativa ssp. japonica*, *Oryza glaberrima*, *O. longistaminata*, *Oryza brachyantha*, *Phyllostachys heterocycla* and *Brachypodium distachyon* (**Figure 1B**) indicates that the genus *Oryza* shares a common ancestor ~15.2 million years ago (MYA), with *O. longistaminata* having potentially diverged from *O. glaberrima* ~1.9 MYA. 120 rapidly evolving genes between *O. longistaminata* and *O. glaberrima* were identified (**Supplementary Table 14**), and eight are transcription factor genes with zinc-finger motif, helix-loop-helix DNA-binding domain or GRAS transcription factor domain, while 3 are disease-resistance genes with NB-ARC domain.

Gene family clustering analysis based on pair-wise genes among 4 rice species (*O. longistaminata*, *O. sativa, O. glaberrima* and *O. brachyantha*) shows that 13,238 gene families were identified in these 4 species (**Figure 1C**). The expansion or contraction of gene families analysis using 7 species (4 rice species as described above, as well as *Brachypodium distachyon*, *P. heterocycla* and *Setaria italica*) showed 393 gene families contracted and 474 expanded in the genome of *O. longistaminata* (**Figure 1D**). Our assembly-collapsed genomic region (Supplementary Information S2.6, Supplementary Table 7) and genic region recovery analyses (Supplementary Information S 4.4) suggest that there are few false positive family expansion/contraction events in our results. In the contracted gene families, there is

one gene family annotated as putative receptor protein kinase ZmPK1, which is related to self-incompatabibilty in *Brassica* (Walker and Zhang, 1990), significantly contracted in the genome of *O. longistaminata* (the number of this gene family, *O. sativa*: 26, *O. brachyantha*: 22, *S. italic*: 12, *B. distachyon*: 12, *O. glaberrima*: 28, *P. heterocycla*: 9 and *O. longistaminata*: 4), indicating that this contracted gene families are probably related to the SI of *O. longistaminata*. In the expanded families, interestingly, 20 gene families related to resistance, such as disease-resistance, LRR receptor-like serine/threonine-protein kinase or other gene family IDs annotated as stimuli-response (lectin receptor kinase or glutathione S-transferases) appeared to have expanded in the *O. longistaminata* genome (**Supplementary Table 15**). Aside from the resistance (R) gene families, we also carefully identified all R genes in these 4 rice genomes, and again *O. longistaminata* showed a greater number of R genes than the other rice species (546 R genes in *O. longistaminata*, 480 in *O. glaberrima*, 466 in *O. sativa* and 439 in *O. brachyantha*) (**Supplementary Figure 11**). Although there exist copy number variations among individuals within a species, the conspicuous difference in R gene/family numbers between *O. longistaminata* and other rice species suggests that *O. longistaminata* as a species has more R genes/families than other rice species do.

To investigate the mechanism of rhizomatousness of *O. longistaminata*, we performed transcriptome sequencing for 2 groups of tissues (rhizome vs. stem and rhizome-tips vs. stem-tips) (**Supplementary Table 16**). In total, 672 and 151 differential expression (DE) genes with 4-fold expression differences were respectively identified in the two groups. KEGG enrichment analysis revealed pathways of photosynthesis, photosynthesis – antenna proteins, carbon fixation in photosynthetic organisms, carbon metabolism and metabolic pathways were enriched significantly in the DE genes between the rhizome and stem, and the pathway of taurine and hypotaurine metabolism, glyoxylate and dicarboxylate metabolism, carbon fixation in photosynthetic organisms and plant hormone signal transduction were enriched in the DE genes between rhizome-tips and stem-tips respectively (**Supplementary Table 17**). The results related to the plant hormones are consistent

with a previous study (He et al., 2014) which also found that hormone genes might play key roles in rhizome development. Via QTL mapping Hu et al. (Hu et al., 2003) identified two loci *Rhz2* and *Rhz3* that may be dominant-complementary for rhizome formation in *O. longistaminata*. Here, we carefully compared these two QTL genomic regions (*Rhz2*, OSR13-OSR16; *Rhz3*, RM119-RM273) between *O. longistaminata* and *O. sativa ssp. japonica* cv. Nipponbare. The correspondent *Rhz2* and *Rhz3* regions in the *O. longistaminata* are about 960 kb and 2.73 Mb, respectively and 122 and 331 respective annotated genes are in the *Rhz2* and *Rhz3* regions (**Supplementary Figure 12, 13;**expression patterns of these genes in the 2 tissue groups are listed in **Supplementary Text 1**). In the *Rhz2* region, there are no DE genes and in the *Rhz3* region, only 4 DE genes were identified (3 DE genes in the group of rhizomes vs. stems and 1 genes in the group of rhizome-tips vs. stem-tips, **Supplementary Table 19**). One of these 4 DE genes, *Olong01m10027813*, encodes a protein phosphatase 2C (PP2C) gene, and expresses > 4 folds greater in the rhizomes than stems. A previous study (Yu et al., 2003) indicated that PP2C involved in the CLAVATA pathways, controlling stem cell identity in both shoot and flower meristems of *Arabidopsis*. Aligning the putative protein sequence of *Olong01m10027813* to its homologs in *O. brachyantha*, *O. barthii*, *O. glaberrima*, *O. rufipogon*, *O. nivara* and *O. sativa*, all of which have no rhizome, revealed one non-synonymous substitution is consistently different between *O. longistaminata* and other non-rhizome species (**Supplementary Figure 14**). Based on these two pieces of evidence, it is plausible that *Olong01m10027813* is a likely candidate for the *Rhz3* gene, which offers a starting point for future functional explorations of the mechanism of rhizomatousness. Furthermore, we also identified DE genes in other 10 QTL regions that affect abundance of rhizomes identified by Hu et al. (2003). In total, 41and 12 DE genes identified in the respective rhizome vs. stem and rhizome-tips vs. stem-tips groups were mapped within these 10 QTLs regions (**Supplementary Table 19**). All the DE genes within the QTLs regions provide us the candidates to uncover the molecular mechanism of rhizomatousness of *O. longistaminata*.

The mechanism of SI in Gramineae is poorly understood and therefore exploring the SI in *O. longistaminata* may extend our understanding of the SI in Gramineae, with the potential promise of enhanced ability of producing hybrid seeds. To explore the molecular mechanisms of SI of *O. longistaminata*, we performed transcriptomic analysis for two tissues groups: stamen group (stamens of *O. longistaminata* vs. stamens of the self-compatibility hybrid line from cross between *O. longistaminata* and *O. sativa ssp. indica* RD23) and pistil group (pistils of *O. longistaminata* vs. pistils of the self-compatibility hybrid line) (**Supplementary Table 16**), of which 571 and 999 DE genes with 4 folds expression change were identified, respectively. KEGG analysis revealed that the pathways of pentose and glucuronate interconversions and carbon fixation in photosynthetic organisms were enriched in the DE genes of the stamen group, and no pathways were enriched in the DE genes of the pistil group significantly (**Supplementary Table 17**). In Gramineae, Yang et al. (2009) mapped the *S* and *Z* loci to the perennial ryegrass (*Lolium perenne*, a species with SI of Gramineae) linkage groups 1 and 2 respectively, which are syntenic to regions in the rice chromosome 5 and chromosome 4. Using the syntenic relationship between ryegrass and rice, we identified 15 DE genes within these 2 corresponding regions in *O. longistaminata* (**Supplementary Table 20**), of which one gene, *Olong01m10012815*, annotated as a EF-hand calcium-binding protein gene, expressed extremely higher in pistils of the hybrid line (FPKM value: 112.6) than *O. longistamianta* (FPKM value: 0.252249), indicating that calcium-dependent signaling, like the SI in Papaveraceae (Franklin‐Tong et al., 2002), may be involved in SI of *O. longistamianta*.

In summary, we present a reference genome of *O. longistaminata* here and comparative genomics and transcriptomics identified genes and pathways that may be related to resistance, rhizomatousness and SI. This work provides a basic evidentiary foundation for targeted studies into the genes underlying its valuable phenotypic traits, future gene mining or breeding efforts, and further study into the evolution of African rice and the *Oryza* genus.

Author contributions

F.H., W.Wang and R.W. conceived the project and its components, designed the studies and contributed to the original concept of the project. Y.Z., S.Z., L.L. and B.F. collected leaves of *O. longistaminata* for genomic sequencing and 8 tissues samples for RNA-Seq. W.Wan and Y.D. constructed short insert libraries and long insert libraries and sequenced the short insert libraries on Hiseq 2000 platform. H.L., X.L., J.C. and J.L. performed genome assembly analysis. M.X and Y.S. conducted genome annotation, divergent time analysis, gene families clustering analysis and expansion or contraction of gene families analysis. L.K. and H.H conducted flow cytometry experiment. Y.Z. analyzed heterozygosity, rapid evolving genes, resistance (R) genes, rhizomatousness and self-incompatibility. Wensheng Wang, L.H., J.Z., Q.Y., Q.S., Q.L., W.H. and D.T. took part in the samples preparation. M.W, M.C., Y.Y. and R.W. modified the manuscript. Y.Z., W.Wang and F.H. analyzed the data as a whole and wrote the manuscript.

Acknowledgments and founding information

This work was supported by the National Natural Science Foundation of China (U1302264) to F.H., the National Basic Research Program of China (2013CB835200, 2013CB835201) and the Department of Sciences and Technology of Yunnan Province (2013GA004) to W.W. and F.H..

We would like to thank Xueyan Li of the Kunming institude of Zoology, Chinese Academy of Science for helpful discussion. We would also like to thank Andrew Willden for English editing the manuscript. No conflict of interestdeclared.

Yesheng Zhang[1,3#], Shilai Zhang[2#], Hui Liu[1,3#], Binying Fu[4#], Lijuan Li[2#], Min Xie[5], Yue Song[5], Xin Li[1], Jing Cai[1], Wenting Wan[6], Ling Kui[1,3], Hui Huang[7], Jun Lyu[1], Yang Dong[1,6], Wensheng Wang[4], Liyu Huang[2], Jing Zhang[2], Qinzhong Yang[2], Qinli Shan[2], Qiong Li[2], Wangqi Huang[2], Dayun Tao[2], Muhua Wang[8], Mingsheng Chen[9], Yeisoo Yu[8], Rod A. Wing[8]*, Wen Wang[1*], Fengyi Hu[2*]

[1] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

[2] Food Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming 650205, China

[3] University of Chinese Academy of Sciences, Beijing 100039, China

[4] Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[5] BGI-Shenzhen, Shenzhen 518083, China

[6] Laboratory of Applied Genomics and Synthetic Biology, College of Life Science, Kunming University of Science and Technology, Kunming 650093, P. R. China

[7] Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwest China, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China

[8] Arizona Genomics of Institute, School of Plant Sciences, University of Arizona, Tucson, Arizona 85721, USA

[9] State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

 # These authors contributed equally to this work.

** To whom correspondence should be addressed.*

Correspondence should be addressed to Fengyi Hu (hfengyiyaas@aliyun.com) and *Wen Wang (wwang@mail.kiz.ac.cn) and Rod Wing (rwing@Ag.arizona.edu).*

The current address of some authors:

Lijuan Li: Laboratory of Molecular Breeding of Energy Plants, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, 88 Xuefu Road, Kunming 650223, Yunnan, China

Xin Li: Center for Epigenetics, Johns Hopkins University School of Medicine, Rangos 580, 855 N. Wolfe St., Baltimore, MD 21205

Jing Cai: Institute for Systematic Botany, University of Zurich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland

Fengyi Hu: School of Agriculture, Yunnan University, Kunming, Yunnan 650091, China

Reference

Franklin‐Tong, V. E.T. L. Holdaway‐ClarkeK. R. StraatmanJ. G. KunkelP. K. Hepler. (2002). Involvement of extracellular calcium influx in the self‐incompatibility response of Papaver rhoeas. The Plant Journal 29:333-345.

Ghesquiere, A. (1986). Evolution of Oryza longistaminata. Rice Genetics:15-27.

Glover, J.J. ReganoldL. BellJ. BorevitzE. BrummerE. BucklerC. CoxT. CoxT. CrewsS. Culman. (2010). Increased food and ecosystem security via perennial grains. Science(Washington) 328:1638-1639.

He, R. F.F. SalvatoJ. J. ParkM. J. KimW. NelsonT. S. BalbuenaM. WillerJ. A. CrowG. D. MayC. A. Soderlund, et al. (2014). A systems-wide comparison of red rice (Oryza longistaminata) tissues identifies rhizome specific genes and proteins that are targets for cultivated rice improvement. BMC plant biology 14.

Hu, F.D. TaoE. SacksB. FuP. XuJ. LiY. YangK. McNallyG. KhushA. Paterson. (2003). Convergent evolution of perenniality in rice and sorghum. Proceedings of the National Academy of Sciences 100:4050-4054.

Song, W. Y.G. L. WangL. L. ChenH. S. KimL. Y. PiT. HolstenJ. GardnerB. WangW. X. ZhaiL. H. Zhu, et al. (1995). A Receptor Kinase-Like Protein Encoded by the Rice Disease Resistance Gene, Xa21. Science 270:1804-1806.

Walker, J. C. and R. Zhang. (1990). Relationship of a Putative Receptor Protein-Kinase from Maize to the S-Locus Glycoproteins of Brassica. Nature 345:743-746.

Wang, M. H.Y. YuG. HabererP. R. MarriC. Z. FanJ. L. GoicoecheaA. ZuccoloX. SongD. KudrnaJ. S. S. Ammiraju, et al. (2014). The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. Nature Genetics 46:982-+.

Yang, B.D. ThorogoodI. P. ArmsteadF. C. FranklinS. Barth. (2009). Identification of genes expressed during the self-incompatibility response in perennial ryegrass (Lolium perenne L.). Plant Mol Biol 70:709-723.

Yu, L. P.A. K. MillerS. E. Clark. (2003). POLTERGEIST encodes a protein phosphatase 2C that regulates CLAVATA pathways controlling stem cell identity at Arabidopsis shoot and flower meristems. Current Biology 13:179-188.

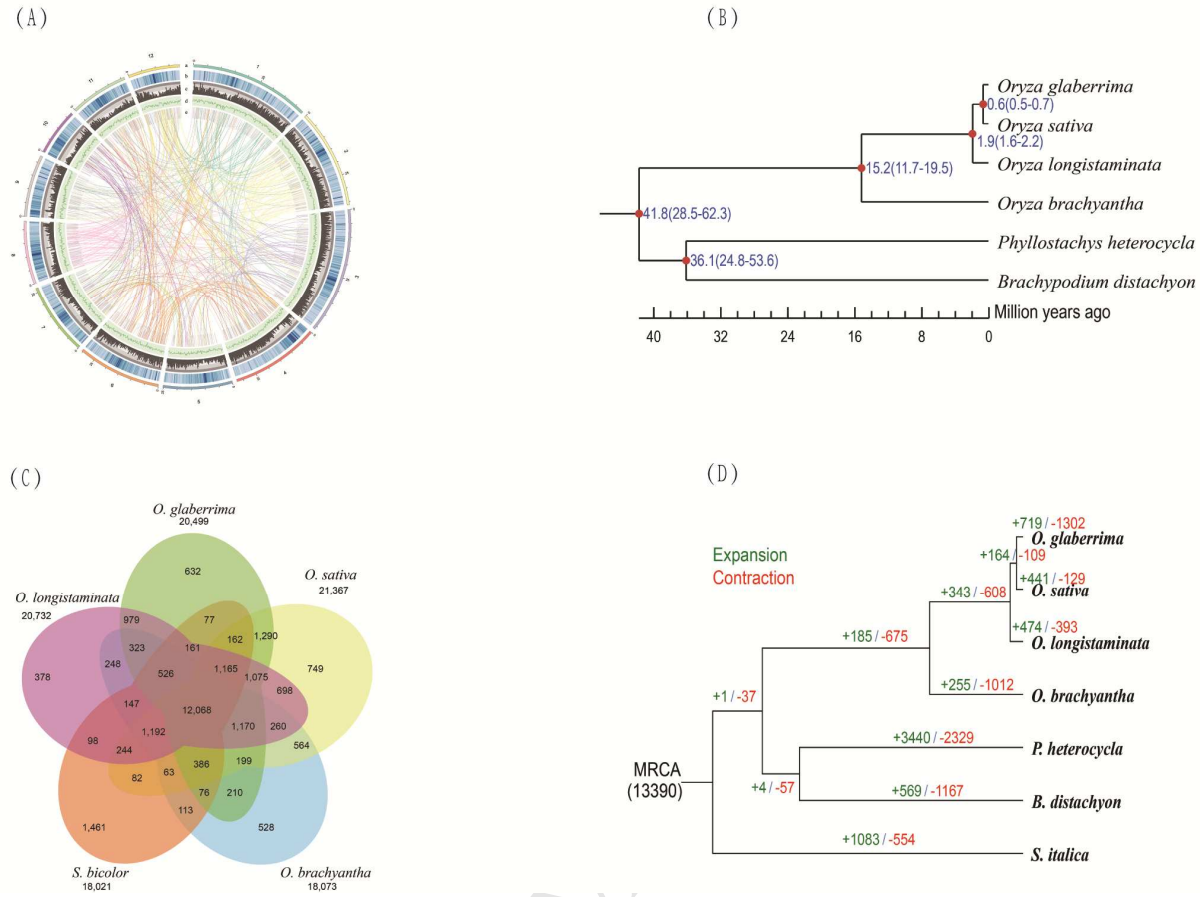


Figure 1 Genome of *O. longistaminata* and genomic analysis. (A) Assembly genome of *O. longistaminata* (Oryza_longistaminata_v1.0). Concentric circles show the structural, functional and evolutionary aspects of the genome: a, chromosome number; b, heat map view of repeats; c, density of genes; d, GC content and e, paralogous relationships between chromosomes of *O. longistaminata*. (**B**) Divergence time between *O. longistaminata* and other species. (**C**) Gene families clustering. (**D**) Expanded or contracted gene families.