De Novo Assemblies of Three *Oryza glaberrima* Accessions Provide First Insights about Pan-Genome of African Rices

Cécile Monat[1], Bérengère Pera[1,2], Marie-Noelle Ndjiondjop[3], Mounirou Sow[4], Christine Tranchant-Dubreuil[1], Leila Bastianelli[5], Alain Ghesquière[1], and Francois Sabot*[,1]

[1]RICE Team, DIADE UMR 232 IRD/UM, IRD France Sud, Montpellier, France

[2]CEA/Genoscope, Evry, France

[3]AfricaRice Center, Cotonou Station, Benin

[4]AfricaRice Center, Ibadan Station, Nigeria

[5]Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, Montpellier, France

*Corresponding author: E-mail: francois.sabot@ird.fr.

Accepted: October 12, 2016

Data deposition: Raw sequence data are available through GenBank BioProject AC SAMN02404669–SAMN02404669, SAMN05462099–SAMN05462101. A Genome browser is available at http://irigin.org/; last accessed November 2, 2016

Abstract

Oryza glaberrima is one of the two cultivated species of rice, and harbors various interesting agronomic traits, especially in biotic and abiotic resistance, compared with its Asian cousin *O. sativa*. A previous reference genome was published but newer studies highlighted some missing parts. Moreover, global species diversity is known nowadays to be represented by more than one single individual. For that purpose, we sequenced, assembled and annotated *de novo* three different cultivars from *O. glaberrima*. After validating our assemblies, we were able to better solve complex regions than the previous assembly and to provide a first insight in pan-genomic divergence between individuals. The three assemblies shown large common regions, but almost 25% of the genome present collinearity breakpoints or are even individual specific.

Key words: assembly, African rice, pan-genome.

Introduction

Rice is the first world human food resource, with 80% of the human population relying on rice for 20% for its daily nutrient intake. From the 24 species belonging to the *Oryza* genus, only two are cultivated: *Oryza sativa* and *Oryza glaberrima*, the Asian and African rice, respectively. The Asian is cultivated worldwide, whereas the African is endemic and restricted to West Africa (Vaughan et al. 2005, 2008).

African rice domestication probably occurred near Dia, in the Niger delta in Mali around 2,500 and 3,500 years ago from the wild rice *O. barthii* (Portères 1962). This wild species originated probably from the same genome AA ancestor than Asian rice more than a million years ago, in Asia, with a divergence between the proto-*O. rufipogon* and the proto-*O. barthii* through a transfer of *O. barthii* ancestor to Africa. Following, the African ancestor evolved in *O. barthii* with the Sahara appearance (Vaughan et al. 2005). The African Rice is thus the result of double evolutionary bottleneck throttling: the first associated to the divergence of ancestor of AA genome species in direction of Africa, and the second, associated with the African domestication. Such evolution explains why the African rice has a much lower genetic diversity than *O. sativa* (Vaughan et al. 2008).

The great diversity of the *Oryza* genus could serve as a pool of genes for the improvement of cultivated varieties (Ge et al. 1999), and in this regard the African rice would be a good tool for varietal improvement of *O. sativa*. However, despite strong cytological similarities, Asian and African rices are reproductively isolated due to an extremely high sterility barrier, managed mainly by the S_1 locus (Sano 1990; Garavito et al. 2010). Introgressions were nevertheless achieved by AfricaRice in the early 2000s, through the cross of some *O. sativa* with some varieties of *O. glaberrima*, through recurrent back-crosses on the Asian rice parent, called NEw RICe for Africa (NERICA;

Gridley et al. 2002). These lines present a better resistance to drought than Asian rice and a better yield than African rice. The original parents were chosen first based on their use in breeding schemes from AfricaRice for *O. sativa*, and second on their supposed ecology for *O. glaberrima*. Thus, no specific traits were targeted in those crosses, neither any idea of genetic diversity. Then, in order to improve such crosses, we need to know trait-related genes and alleles, and genomic sequences for each of these two species are required.

O. sativa was sequenced in 2005 (International Rice Genome Sequencing Project 2005) using the Nipponbare cultivar from the subspecies *japonica*, and is the best annotated plant genome in its current version (Kawahara et al. 2013). Four other reference-like sequences are nowadays available so far for this species (Sakai et al. 2014; Pan et al. 2013; Schatz et al. 2014). The comparison of those accessions with the others have shown that a not negligible part of the whole genome is not shared between two given individual of the same species. Indeed, between Nipponbare (*japonica* subgroup) and Kasalath (*aus* subgroup), more than 10% of sequences are either present or absent of the genome (Sakai et al. 2014). In the same way, around 6% of the genome is individual-specific when comparing two *indica* varietes (Zhang *et al.* 2016). Such data highlights the need to have more than one reference per species in order to understand the full variability of a given species.

For *O. glaberrima*, few data are available so far, as only the CG14 cultivar sequence (parent of the first generation of NERICA) was released in 2014 by the OMAP consortium (Wang et al. 2014) using a reference guided approach based on *O. sativa ssp japonica* cv NipponBare. This BAC-based assembly, of good overall quality, nevertheless presents abnormalities in terms of specific African sequences (such as the *SWEET14* gene duplication, Hutin et al. 2015). Moreover, various analyses shown that the CG14 OMAP sequence have important gaps (between 20% and 30% of the whole sequence is missing; Nabholz et al., 2014; Orjuela et al. 2014; Hutin et al. 2015; Ta et al. 2016). In order to extend the potentiality of allele and gene mining in *O. glaberrima*, we decided to sequence, assemble and annotate three accessions of *O. glaberrima*: TOG5681 and CG14 (both parents of two NERICA generations and representing each one extremity of the variability of the species), as well as G22 because of its centric position in this variability (Orjuela et al. 2014). We re-assembled CG14 cultivar in order to be able to compare the three accessions without introducing biases such as sequencing technologies or assembly tools. The three assemblies were found to be highly similar, but each of them harbors a dedicated set of unique scaffolds.

Materials and Methods

Plant Material

Three accessions of *Oryza glaberrima* (CG14, TOG5681 and G22) were used in this study. Information such as identifier and synonymous or original location are available in Orjuela et al. (2014). The plants were grown at IRD greenhouses at Montpellier (France) under normal conditions, and DNA extracted from fresh leaves using *QIAGEN* Genomic Tip/20 kit (Germany), as recommended by suppliers.

DNA Sequencing

Different *Illumina* sequencing were performed for CG14, TOG5681, and G22. For CG14 and TOG5681, Montpellier GenomiX (Montpellier, France) performed the library construction (1 library for TOG5681 and 2 for CG14) using the *Truseq* DNA sample prep kit (*Illumina*) and the sequencing on a Hiseq2000 using *Truseq* v3 clustering and SBS kit (Illumina). A single lane was used for 100 bases paired-end reads. An additional 75 paired-end sequencing for TOG5681 was performed by Integragen (Evry, France) on a GaIIx sequencer and a *Truseq* DNA sample prep. kit. For G22, the sequencing was performed by CEA/Genoscope (Evry, France) on a HiSeq2500 sequencer using *TruSeq* DNA sample prep. kit, 100 bases paired-end reads (table 1).

Genome Assembly

The *ABySS* software (Simpson et al. 2009) was used for generating the three assemblies. After tests, a *k-mer* size of 48 and the number minimal of pairs needed to consider link between two reads of 8 were fixed.

Only sequences larger or equal to 200 bp were conserved for further analyses. N50 and N90 calculation were performed as the statistical value so that 50% and 90% of all contigs/scaffolds have a size higher or equal to this value, respectively.

Genome Annotation

The *MAKER*-P software (Campbell et al. 2014) was used for the structural annotation on the scaffolds of the three assemblies. ESTs and full-length cDNA from *O. sativa* (*indica* and *japonica* data) were used as support for annotation, as well as proteins of *O. sativa* from *SwissProt* and local databases. Repeated elements were masked using the *Oryza* model of *RepeatMasker* (Smit et al. 1996–2010) and using *RepBase* (Bao et al. 2015) *Oryza* data. For gene prediction, *O. sativa SNAP hmm* file was used. The other parameters of *MAKER*-P were conserved as default. Functional annotations were added upon the *MAKER*-P structural annotations, using *BLAST2GO* software. In that purpose, we performed a *BLASTx* analysis upon the Plant Protein database v137 from PlantGDB (http://www.plantgdb.org/) using standard values except for an *e*-value of e10^{-3}. The *InterProScan* tool v5 (Jones et al. 2014) was used to obtain the motifs in all protein model using standard conditions.

Table 1

Assemblies Statistics

	CG14	TOG5681	G22
Number of reads	91,173,341	141,028,786	64,710,458
Number of sequences	64,988 (88,310)	51,262 (65,834)	49,662 (74,592)
N50	10,233 (7,735)	13,404 (10,725)	14,556 (9,686)
N90	2,025 (1,351)	2,733 (1,940)	5,884 (3,898)
k-mer divergence	0.70%	1.20%	0.40%
Longest Sequence	90,835 (90,602)	105,329 (92,240)	112,369 (87,329)
Total assembly size	299,704,894	292,222,046	305,237,265

Note.—In bracket are shown the values for contigs, otherwise for scaffolds. Size values are in bp.

K-mer Analysis

Ten sampling of 400 Mb of raw reads were performed using *RandomFQ* from *ea-utils* suite (Aronesty 2013) and their 31 *k-mer* content estimated using *Jellyfish* (Marçais and Kingsford 2011). The mean value for each *k-mer* was then recovered and compared with its value in the assembly (also calculated using *Jellyfish*). A *g*-test with a single freedom of liberty was performed to estimate the significant differences at $P < 0.05$.

Whole Genome Comparisons

MUMmer (Kurtz et al. 2004) was used for whole genome comparisons as follows: first *NUCmer* alignment package calculated the delta file, then the *show-coords* utility parsed this delta alignment output. Thirdly, *mummerplot* utility converted the output from *NUCmer* to a format suitable for plotting with *gnuplot*. Only alignments which represent the best one-to-one mapping of reference and query subsequences were displayed. Finally *show*-snps utility reported polymorphism contained in delta output format from *NUCmer*, without SNP with ambiguous mapping.

BLAST Inter-References Comparisons

To evaluate the presence of divergent regions among the three accessions, *BLASTn* analyses between the three assemblies against each other were performed. *BLAST* results were parsed to class the scaffolds into three different categories:

- similar (to the reference tested), scaffold similarity is min 80% and the cumulative hit length higher than 80% of scaffold length;
- not similar, at least one of the previous defined criteria is not filled;
- not referenced, no hits on the reference used.

Results and Discussion

Generalities about the Three Assemblies

Basic Statistics

The three assemblies provide as expected a large number of scaffolds, ranging from 49,622 for G22 to 64,988 for CG14 (table 1). Their N50 and N90 value are high, allowing large-scale analyses or molecular targeting (PCR primer design, sequence capture, etc), ranging from 10 to 15 kb for N50 and 2 to 5.8 kb for N90. In order to validate the structure of the assembly, in terms of representativity of the initial data set, we performed a *k-mer* comparison with original datasets (see "Materials and Methods" for more details). The divergence is lower than 1% for CG14 and G22, and less than 1.2% for TOG5681 (table 1), showing that our assemblies reflect correctly the initial datasets. The variation for TOG5681 may be explained because of the use of a 75 bp sequence set. Each of the assembly covered ~300 Mb (table 1), a little less than the CG14 OMAP assembly (Wang et al. 2014). However, short reads data are known to provoke lower assembly size due to incomplete assembly of repeated regions (Paszkiewicz and Studholme 2010). Nevertheless, the assemblies were enough complete to identify specific events such as the duplication of the *SWEET14* gene (Hutin et al. 2015) in African rices. At this position, our assemblies are split in four scaffolds, two per side (fig. 1A), indicating that a duplicated sequence in two different locations. In the same way, we obtained a better resolution on the *RYMV1* locus (Albar et al. 2003) than on the OMAP reference (fig. 1B).

Annotation

The *MAKER*-P structural annotation provided between 49,000 and 51,262 gene locations per assembly (table 2), with more than 160,000 models (mean 3.2 mRNA possibilities per gene location) per assembly, which is quite similar to what was found for *Oryza sativa ssp. japonica* cv. Nipponbare. The *Blast2GO* functional annotation was able to annotate almost half of those models. The majority of these assemblies is genome-related (93%), the organite-related part being between 2,071 models related to mitochondria and 4,814 to the chloroplast (CG14 values, similar for the two other assemblies; table 2). At the opposite of Wang et al. (2014), we were able to identify the *Hd1* gene (also shown previously to be normally absent from *O. glaberrima*, Sanyal et al. 2010), in all three assemblies.

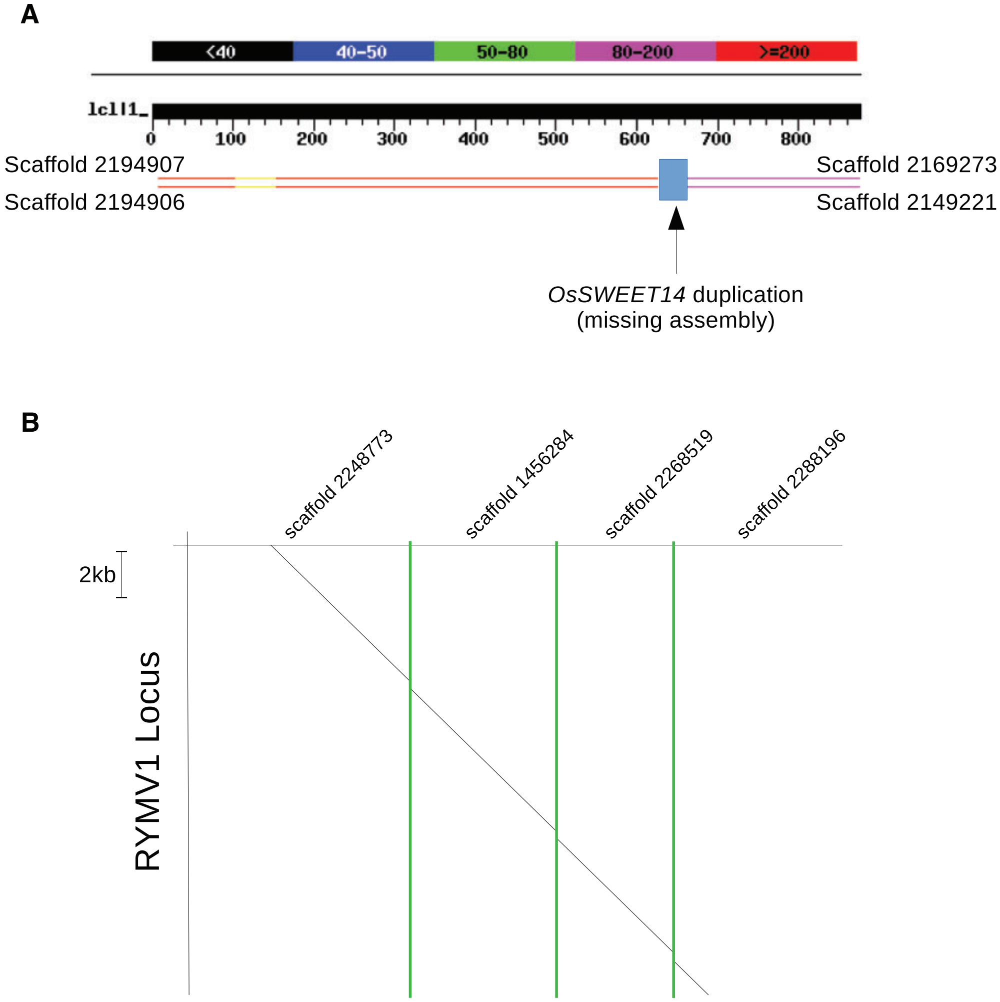

Fig. 1.—Structural validation of assemblies. (*A*) *SWEET14* duplication, *BLAST* output representation on TOG5681 sequence. (*B*) Dot plot of CG14 scaffolds matching 20 kb around the *RYMV1* locus from MSU7 *O. sativa* ssp *japonica* NipponBare genome.

Table 2
Annotation Statistics

	CG14	TOG5681	G22
Number of gene location	50,000	51,262	49,662
Number of GO	82,248	93,396	87,605
Number of organite-related models	6,885	1,532	2,385

Comparison between Our Three Assemblies

Global Analyses

We decided to compare our assemblies in terms of what is non-collinear or either absent. *MUMmer* analyses shown that the three assemblies seemed coherent when compared at a genome wide scale (supplementary fig. S1, Supplementary Material online). *MUMmer* analyses shown that global collinearity is almost conserved between our three independent assemblies. Moreover, the mean SNP density for *O. glaberrima* is of 2.4 SNP per kb (similar as the ~2.3/kb for Wang et al. 2014), which is quite lower than in its Asian cousin *O. sativa* (around 7/kb, Zhang et al. 2016). However, *MUMmer* analyses are only at macro-scale, and we decided to be more precise and to use *BLASTn* to identify non-collinear and absent regions when compared one to another.

Micro-Collinearity Analyses

Each assembly was compared with other ones using *BLASTn* (see “Materials and Methods” for detailed procedures). Basic statistics about the CG14 vs. TOG5681 comparison are provided in table 3.

As expected for a poorly diverse species, most of the scaffolds are almost collinear (55–65%), and 45–35% are not. Collinearity breaks may be due to transposable elements activities or large genomic modification (duplication, deletion, inversion). The GO type of genes between the similar and not similar scaffolds here are not significantly different (data not shown).

Table 3
Micro-Collinearity Statistics for CG14 vs. TOG5681

	Valid Scaffolds	Not Valid Scaffolds	Not Referenced Scaffolds
Number of sequences	48223	16672	93
Minimal size	200	201	202
Maximal size	86103	90835	3041
Mean size	4110	6087	447
Median size	1942	2592	320
Number of functionally annotated gene model	10685	2147	2
Number of GO	23634	4817	4

NOTE.—Sizes are given in bp.

Interestingly, around 100 scaffolds per assembly (<0.2%) are stand-alone in this analysis, i.e., they are not referenced (not detected) in the other one. Very few genes are located on those small sequences (from 202 to 3041 bases long; table 3), but with very few information about their potential function or involvement in metabolic pathway (only 2 GO are identified). More analyses (transcriptomic, phylogenetic) are requested to identify a potential role for those genes.

Discussion

In this study, we propose three new assemblies for the African cultivated rice species, *O. glaberrima*. When compared with current available reference genome CG14 OMAP v1.1 (Wang et al. 2014), our sequences are smaller and more fragmented. However, as we use a completely *de novo* approach, it resulted in a closer assembly to reality, as we did not bias our assembly toward Nipponbare genome, but with smaller genomic size and more scaffolds (as expected). Indeed, we obtained better sequences on some loci such as *RYMV1* (Albar et al. 2003), and we can identify the duplication leading to *SWEET14*-controlled *Xanthomonas* resistance (Hutin et al. 2015). However, we cannot found all genes and sequences found in the OMAP assembly. Interestingly, we found the *Hd1* gene, supposed to be absent in the African rice in a *sativa*-anchored BAC-based analysis (Sanyal et al. 2010). However, complementary analyses have to be conducted to determine if this gene is truly active in the species. These assemblies are thus imperfect but can provide good complement for genomic analyses in the cultivated African rice.

Those three assemblies are quite similar at the whole genome-scale (supplementary fig. S1, Supplementary Material online) and present a very low SNP density, as expected. We then aimed to find large divergence within those sequences, and decided to check the (micro-)collinearity status of our assemblies, and found that around 25% of the genome present enough variation to lead to collinearity breakdown. As shown on Asian rice (Sakai et al. 2014; Schatz et al. 2014; Zhang et al. 2016), we identified thus large scale variations and thus a larger pan-genomic space than what could have been expected based only on SNP data (Nabholz et al. 2014; Orjuela et al. 2014). Interestingly, we found around 0.6% of our scaffolds to be absent in one assembly regarding another one. Very few genes exist on those scaffolds, and no specific GO enrichment or gene structure was found for them. They may be either old sequences surviving only in one of the three samples analyzed here, or at the opposite new sequences arising in the African rice genome. More individuals from this species have to be studied to conclude on such sequences.

The three individuals were chosen because of their location in the *O. glaberrima* diversity tree (Orjuela et al. 2014), i.e., on each side for CG14 and TOG5681, and in the middle for G22. Their divergence, even small, can drive to reproductive isolation, and thus to sympatric speciation (Dieckmann and Doebeli 1999). A better knowledge of the physical differences between subgroup will help us to better understand such a mechanism.

In addition, these differences highlight the need to better known the pan-genomic structure of the cultivated plant, in order to optimize breeding. In this regard, the example of the *Sub1A* locus (Xu et al. 2006) is highly relevant: this gene allows a higher tolerance for submergence but is present only in some subgroups of the *O. sativa* ssp *indica*. The current sequencing cost and facilities allow any research group to sequence massively various samples. Thus, when we will have access to the whole pan-genome of *O. glaberrima* through massive resequencing, we would be able to empowered new NERICA type crosses to transfer more efficiently the *O. glaberrima* characters of interest to *O. sativa*.

Supplementary Material

Supplementary figure S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Authors want to thank Sophie Cheron-Perez, Harold Chrestin and Julie Orjuela-Bouniol for their help in plant management and DNA extraction. C Monat was supported by grants from French ANR (AfriCrop project #ANR-13-BSV7-0017) and NUMEV Labex (LandPanToggle #2015-1-030-LARMANDE) grants. The whole project was funded through a AfricaRice/IRD LoA (GLASS project).

Literature Cited

Albar L, et al. 2003. Fine genetic mapping of a gene required for Rice yellow mottle virus cell-to-cell movement. Theor Appl Genet. 107:371–378.

Aronesty E. 2013. Comparison of sequencing utility programs. Open Bioinforma J. 7:1–8.

Bao W, et al. 2015. *Repbase Update*, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.

Campbell MS, et al. 2014. *MAKER*-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164:513–524.

Dieckmann U, Doebeli M. 1999. On the origin of species by sympatric speciation. Nature 400:354–357.

Garavito A, et al. 2010. A genetic model for the female sterility barrier between Asian and African cultivated rice species. Genetics 185: 1425–1440.

Ge S, et al. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc Natl Acad Sci U S A. 96:14400–14405.

Gridley HE, Jones MP, Wopereis-Pura M. 2002. Development of new rice for Africa (NERICA) and participatory varietal selection. In Breeding rainfed rice for drought-prone environments: integrating conventional and participatory plant breeding in South and Southeast Asia: proceedings of a DFID Plant Sciences Research Programme/IRRI Conference. p. 12–15.

Hutin M, et al. 2015. A knowledge-based molecular screen uncovers a broad spectrum OsSWEET14 resistance allele to bacterial blight from wild rice. Plant J. 84:694–703.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature 436:793–800.

Jones P, et al. 2014. *InterProScan* 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.

Kawahara Y, et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data - Springer. Rice 6:4.

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k-mers*. Bioinformatics 27:764–770.

Nabholz B, et al. 2014. Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*O. glaberrima*). Mol Ecol. 23:2210–2227.

Orjuela J, et al. 2014. An extensive analysis of the African rice genetic diversity through a global genotyping—Springer. Theor Appl Genet. 127:2211–2223.

Pan Y, et al. 2013. Comparative BAC-based physical mapping of *Oryza sativa* ssp. *indica* var. 93-11 and evaluation of the two rice reference sequence assemblies. Plant J. 77:195–805.

Paszkiewicz K, Studholme DJ. 2010. De novo assembly of short sequence reads. Brief Bioinform. 11:457–472.

Portères R. 1962. Primary cradles of agriculture in the African continent. J Afr Hist. 3:195–210.

Sakai H, et al. 2014. Construction of pseudomolecule sequences of the *aus* rice cultivar Kasalath for comparative genomics of Asian cultivated rice. DNA Res. 21:397–405.

Sano Y. 1990. The genic nature of gamete eliminator in rice. Genetics 125:183–191.

Sanyal A, et al. 2010. Orthologous comparisons of the *Hd1* region across genera reveal *Hd1* gene lability within diploid *Oryza* species and disruptions to microsynteny in Sorghum. Mol Biol Evol. 27:2487–2506.

Schatz MC, et al. 2014. New whole genome *de novo* assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*. Genome Biol. 15:506.

Simpson JT, et al. 2009. *ABySS*: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.

Smit AFA, Hubley R, Green P. (1996–2010) RepeatMasker Open-3.0. Available at: http://www.repeatmasker.org.

Ta KN, et al. 2016. miR2118-triggered phased siRNAs are differentially expressed during the panicle development of wild and domesticated African rice species. Rice (N Y) 9:10.

Vaughan DA, et al. 2005. On the phylogeny and biogeography of the genus *Oryza*. Breed Sci. 55:113–122.

Vaughan DA, et al. 2008. The evolving story of rice evolution. Plant Sci. 174:394–408.

Wang M, et al. 2014. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat Genet. 46:982–988.

Xu K, et al. 2006. *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442:705–708.

Zhang J, et al. 2016. Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. Proc Natl Acad Sci U S A. 113:E5163–E5171.

Associate editor: Howard Ochman