




DATA NOTE

REVISED First *de novo* draft genome sequence of *Oryza coarctata*, the only halophytic species in the genus *Oryza* [version 2; referees: 3 approved]

Tapan Kumar Mondal , Hukam Chand Rawal, Kishor Gaikwad, Tilak Raj Sharma, Nagendra Kumar Singh

National Research Centre on Plant Biotechnology (ICAR), PUSA, New Delhi, 110012, India

v2 First published: 25 Sep 2017, 6:1750 (doi: [10.12688/f1000research.12414.1](https://doi.org/10.12688/f1000research.12414.1))
 Latest published: 15 Dec 2017, 6:1750 (doi: [10.12688/f1000research.12414.2](https://doi.org/10.12688/f1000research.12414.2))

Abstract

Oryza coarctata plant, collected from Sundarban delta of West Bengal, India, has been used in the present study to generate draft genome sequences, employing the hybrid genome assembly with Illumina reads and third generation Oxford Nanopore sequencing technology. We report for the first time the draft genome with the coverage of 85.71 % and deposited the raw data in NCBI SRA, with BioProject ID [PRJNA396417](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA396417).





This article is included in the [Global Open Data for Agriculture and Nutrition gateway](#).

Open Peer Review

Referee Status: 

	Invited Referees		
	1	2	3
REVISED			
version 2		report	
published			
15 Dec 2017			
version 1			
published	report	report	report
25 Sep 2017			

- 1 **Sandip Das** , University of Delhi, India
- 2 **Stephen P. Moose** , University of Illinois at Urbana-Champaign, USA
- 3 **Kashmir Singh**, Panjab University, India

Discuss this article

Comments (0)

Corresponding author: Tapan Kumar Mondal (mondalk@rediffmail.com)

Author roles: **Mondal TK:** Conceptualization, Data Curation, Investigation, Methodology, Resources, Writing – Review & Editing; **Rawal HC:** Data Curation, Formal Analysis; **Gaikwad K:** Data Curation, Formal Analysis, Supervision, Validation; **Sharma TR:** Conceptualization; **Singh NK:** Conceptualization, Data Curation, Formal Analysis, Project Administration

Competing interests: No competing interests were disclosed.

How to cite this article: Mondal TK, Rawal HC, Gaikwad K *et al.* **First *de novo* draft genome sequence of *Oryza coarctata*, the only halophytic species in the genus *Oryza* [version 2; referees: 3 approved]** *F1000Research* 2017, **6**:1750 (doi: [10.12688/f1000research.12414.2](https://doi.org/10.12688/f1000research.12414.2))

Copyright: © 2017 Mondal TK *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: The author(s) declared that no grants were involved in supporting this work.

First published: 25 Sep 2017, **6**:1750 (doi: [10.12688/f1000research.12414.1](https://doi.org/10.12688/f1000research.12414.1))

REVISED Amendments from Version 1

In this revision, we have addressed all the issues raised by the referee. In additions, we have made few grammatical corrections. We revised the manuscript as a result of reanalysis of the data. We have also improved clarity in the Methods regarding assembly of the genome. Further, as per the suggestions of the referee as well as requirement, we incorporated 6 more references.

See referee reports

Introduction

Soil salinity is a major abiotic stress of rice cultivation globally (Molla *et al.*, 2015), and rice cultivation areas under soil salinity stress are increasing gradually. Genetic potential for salt tolerance of rice that exists among the natural population has been largely exploited, and alternative useful alleles may further enhance salinity tolerance. Wild species are a potential source of many useful genes and QTLs that may not be present in the primary gene pool of the domesticated species.

Oryza coarctata, known as Asian wild rice, grows naturally in the coastal region of South-East Asian countries. It flowers and set seeds under as high as 40 E.Ce dS m⁻¹ saline soil (Bal & Dutt, 1986). It is the only species in the genus *Oryza* that is halophyte in nature. However, with the exception of one transcriptomic (Garg *et al.*, 2014) and one miRNA (Mondal *et al.*, 2015) experiment, no large scale generation of any other genomic resource is available for this important species, although several pinitol biosynthesis pathway genes have been cloned to study the functional genomics (Sengupta & Majumder, 2009).

Methods

The plant was collected from its native place, Sundarban delta of West Bengal, India (21°36'N and 88°15' E) and established at our institute Net house through clonal propagation. To determine the genome size, 20 mg of young leaf tissue from Net house grown plants was chopped into small pieces and stained with RNase containing propidium iodide (50 µg/ml) (BD Science, India) as per the protocol of Dolezel *et al.* (2007). The samples were filtered through a 40-µm mesh sieve (Corning, USA), before analysis in (CFM) BD FACS Calibur (BD Biosciences, San Jose, CA, USA). *Pisium sativum* leaf was used as standard for calculating the genome size. Further, high-quality genomic DNA from 100 mg young leaf of a single plant was extracted using CTAB method (Ganie *et al.*, 2016) for the preparation of various genomic DNA libraries. We used standard Illumina HiSeq 4000 platform (San Diego, CA, USA) to construct 151-bp paired-end libraries and four mate-pair libraries of four different sizes (average of 2, 4, 6 and 8 kb size). In addition, we also used third generation sequencing (Oxford Nanopore) technology for better assembly. Sequencing was performed on MinION Mk1b (Oxford Nanopore Technologies, Oxford, UK) using SpotON flow cell (R9.4) in a 48h sequencing protocol on MinKNOW 1.4.32. Base calling was performed using Albacore. Base called reads were processed using poretools version 0.24 (Watson *et al.*, 2015) and poretools version 0.6.0 (Loman & Quinlan, 2014). Assembly of the high quality reads was performed using

PLATANUS v1.2.4 (Kajitani *et al.*, 2014) and SSPACE v3.0 (Boetzer *et al.*, 2011) with default parameter. The simple sequence repeats (SSRs) of each scaffold were identified by MISA perl script (Thiel *et al.*, 2003). Gene model prediction was done by ab initio gene predictor AUGUSTUS 3.1 (Stanke & Waak, 2003) and sequence evidence based annotation pipeline, MAKER v2.31.8 (Campbell *et al.*, 2014) with *O. sativa ssp. japonica* as reference gene model. The protein-coding genes were annotated by using BLAST based approach against a database containing functional plant genes downloaded from NCBI with Blast2GO (version 4.01) (Conesa & Gotz, 2008). Genes with significant hits were assigned with GO (Gene Ontology) terms and EC (Enzyme Commission) numbers. InterProScan search and pathway analyses with KEGG database were also performed by using Blast2GO. Non-coding RNAs, such as miRNA, tRNA, rRNA, snoRNA, snRNA, were identified by adopting Infernal v1.1.2 (Nawrocki & Eddy, 2013) using Rfam database (release 9.1) (Nawrocki *et al.*, 2015) and snoscan distribution. Transfer RNA was predicted using tRNAscan-SE v 1.23 (Lowe & Eddy, 1997)

Discussion

The *O. coarctata* genome (2n=4X=48; KKLL; Sanchez *et al.*, 2013) is self-pollinated, (Sarkar *et al.*, 1993) tetraploid plant with a genome size estimated by flow cytometry is found to be approximately 665Mb. The Illumina 4000 GA Iix sequencer pair-end generated 123.78 Gb data. Further four mate-pair libraries together generated 36.54 Gb and Nanopore generated 6.35 Gb sequence data. Hence, we achieved 250.66 X depth of the genome of *O. coarctata*. The final assembly generated 58362 numbers of scaffolds with a minimum length of 200 bp to maximum length of 7,855,609 bp and 1,858,627 bp N50 value, making a total scaffold length of 569994164 (around 570 Mb) assembled genome, resulting in 85.71% genome coverage. It has been calculated that data contain very small amount of non-ATGC character. Further, we also found that the 19.89% of the assembled genome is repetitive in nature. We also identified approximately 5512 different non-coding RNAs and around 230,968 SSRs. Gene ontology analysis identified several salt responsive genes.

Data availability

Raw sequence data are available at NCBI SRA under the BioProject ID: PRJNA396417.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

TKM is grateful to Mr Sukdev Nath, who provided the planting material. TRS is thankful to the DST, Govt. of India for JC Bose National Fellowship. The authors are thankful to M/S Genotypic Technology Private Limited, Bengaluru, India for sequencing work and M/S BD Biosciences, India for Flow Cytometer work.

References

- Bal AR, Dutt SK: **Mechanism of salt tolerance in wild rice (*Oryza coarctata* Roxb).** *Plant Soil*. 1986; **92**(3): 399–404.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Boetzer M, Henkel CV, Jansen HJ, *et al.*: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics*. 2011; **27**(4): 578–579.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Campbell MS, Law M, Holt C, *et al.*: **MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations.** *Plant Physiol*. 2014; **164**(2): 513–524.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Conesa A, Götz S: **Blast2GO: A comprehensive suite for functional analysis in plant genomics.** *Inter J Plant Genomics*. 2008; 619832.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dolezel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants using flow cytometry.** *Nat Protoc*. 2007; **2**(9): 2233–2244.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ganie SA, Borgohain MJ, Kritika K, *et al.*: **Assessment of genetic diversity of *Saltol* QTL among the rice (*Oryza sativa* L.) genotypes.** *Physiol Mol Biol Plants*. 2016; **22**(1): 107–114.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Garg R, Verma M, Agrawal S, *et al.*: **Deep transcriptome sequencing of wild halophyte rice, *Porteresia coarctata*, provides novel insights into the salinity and submergence tolerance factors.** *DNA Res*. 2014; **21**(1): 69–84.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kajitani R, Toshimoto K, Noguchi H, *et al.*: **Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short read.** *Genome Res*. 2014; **24**(8): 1384–95.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Loman NJ, Quinlan AR: **Poretools: a toolkit for analyzing nanopore sequence data.** *Bioinformatics*. 2014; **30**(23): 3399–3401.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res*. 1997; **25**(5): 955–964.
[PubMed Abstract](#) | [Free Full Text](#)
- Molla KA, Debnath AB, Ganie SA, *et al.*: **Identification and analysis of novel salt responsive candidate gene based SSRs (cgSSRs) from rice (*Oryza sativa* L.).** *BMC Plant Biol*. 2015; **15**: 122.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mondal TK, Ganie SA, Debnath AB: **Identification of novel and conserved miRNAs from extreme halophyte, *Oryza coarctata*, a wild relative of rice.** *PLoS One*. 2015; **10**(10): e0140675.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics*. 2013; **29**(22): 2933–2935.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nawrocki EP, Burge SW, Bateman A, *et al.*: **Rfam 12.0: updates to the RNA families database.** *Nucleic Acids Res*. 2015; **43**(Database issue): D130–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sanchez PL, Wing RA, Brar DS: **The wild relative of rice: genomes and genomics.** In: Q. Zhang and RA. Wing (eds.), *Genetics and genomics of rice, plant genetics and genomics: crops and models*. Springer Science Business Media New York. 2013; 9–25.
[Publisher Full Text](#)
- Sarkar RH, Samad MA, Seraj ZI, *et al.*: **Pollen tube growth in crosses between *Porteresia coarctata* and *Oryza sativa*.** *Euphytica*. 1993; **69**: 129–134.
[Publisher Full Text](#)
- Sengupta S, Majumder AL: **Insight into the salt tolerance factors of a wild halophytic rice, *Porteresia coarctata*: a physiological and proteomic approach.** *Planta*. 2009; **229**(4): 911–929.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics*. 2003; **19**(Suppl 2): ii215–225.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Thiel T, Michalek W, Varshney RK, *et al.*: **Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.).** *Theor Appl Genet*. 2003; **106**(3): 411–422.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Watson M, Thomson M, Risse J, *et al.*: **poRe: an R package for the visualization and analysis of nanopore sequencing data.** *Bioinformatics*. 2015; **31**(1): 114–115.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Referee Status:   

Version 2

Referee Report 27 December 2017

doi:10.5256/f1000research.14545.r29080



Stephen P. Moose 

Department of Crop Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA

The authors have included the requested details about the source of plant materials, estimate of genome size, and genome assembly methods. Although still not sure how they arrived at the estimate of 19.86% repeat sequences, this is a minor point.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 20 October 2017

doi:10.5256/f1000research.13443.r26359



Kashmir Singh

Department of Biotechnology, Panjab University, Chandigarh, India

The work describe the whole genome sequence of wild species of *Oryza coarctata* species that exclusively grow under saline water and thus will be an important source of salinity tolerance genes. These genes can later be used to introduce salinity tolerance in commercial cultivars of rice. The authors used Illumina and Oxford nanopore sequencing platforms to generate 372.48X data.

The genome sequencing methods seems good enough but authors have discussed very little about the annotation of the genome data. I can understand that there is word limit under Data Note in F1000Research, but still by looking at the discussion, I think analysis portion is weak point in this paper. Authors should provide a comparative note on the genome of *Oryza sativa* and *Oryza coarctata*. How this species is tolerating such a high saline conditions, which kind of genes/osmoregulators are involved in this adaptation should be discussed along with comparison to *O. sativa*. How many different genes were predicted should be mentioned. Authors found approximately 1605 non-coding RNAs? I am not sure, what are trying to tell here, this number should be high as per my opinion.

There are some minor mistakes like; in the affiliation the word “Delhi” is not required. The word, “Primary” should be inserted in the first paragraph last line of Introduction. So the correct sentence will be "...in the primary gene pool...".

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 13 October 2017

doi:[10.5256/f1000research.13443.r26486](https://doi.org/10.5256/f1000research.13443.r26486)



Stephen P. Moose 

Department of Crop Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA

The authors report a whole genome sequence dataset for a halophytic wild rice species. These data will be useful for discovery of novel alleles for rice improvement, and for comparative/evolutionary genomics within the *Oryza* genus.

1. The report would benefit from more details on the plant accession used as source of DNA for sequencing. It is stated *O. coarctata* is tetraploid. Was that determined by the authors, or is there a citation to include? Is it known whether *O. coarctata* is typically self or cross-pollinated, or other information about expected degree of heterozygosity? When grown in greenhouse to generate the plant tissue used for DNA extraction, were the plant(s) established from seeds, or via clonal propagation? Was the genomic DNA used to prepare sequencing libraries from a single plant, or a pool from multiple plants? This information is important to assess expected frequencies of variant types such as alleles or homeologs due to tetraploidy, which are likely collapsed to varying degrees in the subsequent assembly.
2. There is mention of an assembly and its quality, but not about the method(s) used to produce it or key parameters that guided the assembly. Can the authors provide that information, so that others have a benchmark upon which to compare future assemblies using the datasets?
3. The sentence “Further, we also found that the repeat contain 19.89% of the genome.” Is not completely clear. I believe what the authors intend to say is that approximately 20% of the genome assembly is comprised of repeats. How was this sequence fraction defined as repeats, via tool for matching to known repeat sequences, or a de novo approach? By inference, it is also likely that the

approximately 100-kb of the estimated genome size not covered by the assembly is comprised of high-copy repeats, leading to an estimate of about 30% total repeat content.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 11 October 2017

doi:10.5256/f1000research.13443.r26360



Sandip Das 

Department of Botany, University of Delhi, Delhi, India

The authors report a draft genome sequence of a halophyte *Oryza* species collected from Sunderbans, and provides a glimpse into the adaptive strategies employed by *Oryza* against salinity stress. Undoubtedly, it will be an useful resource for future functional characterization, comparative genomic studies, and developing salinity tolerance in rice. I understand that the present format is only for reporting, and look forward to reading the full manuscript with all the analysis. There are small language edits that that authors need to incorporate.

Comments:

Please add a reference or sufficient information for general readers as to how genome types of rice (for instance, KKLL in case of *O. coarctata*) was assigned.

Language corrections:

1. Change “have been used” to “has been used”
2. Change “We report for the first time that more than 85.71 % of the genome coverage and the data have been deposited in NCBI SRA, with BioProject ID PRJNA396417” to “deposited in NCBI SRA, with BioProject ID PRJNA396417”
3. Change “and established to our institute NET” to “and established at our institute NET”
4. Change “resulting 85.71 % genome coverage” to “resulting in 85.71 % genome coverage”
5. Change “we also found that the repeat contain 19.89% of the genome.” to “we also found that the 19.89% of the genome is repetitive in nature”.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Comparative genomics, brassica, polyploidy, regulatory evolution

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research