DATA NOTE

First de novo draft genome sequence of Oryza coarctata, the only halophytic species in the genus Oryza [version 1; referees: 1 approved, 1 approved with reservations]

Tapan Kumar Mondal, Hukam Chand Rawal, Kishor Gaikwad, Tilak Raj Sharma, Nagendra Kumar Singh
National Research Centre on Plant Biotechnology (ICAR), PUSA, New Delhi, Delhi, 110012, India

Abstract

Oryza coarctata plants, collected from Sundarban delta of West Bengal, India, have 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 that more than 85.71% of the genome coverage and the data have been deposited in NCBI SRA, with BioProject ID PRJNA396417.

This article is included in the Global Open Data for Agriculture and Nutrition gateway.

Corresponding author: Tapan Kumar Mondal (mondaltk@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 1; referees: 1 approved, 1 approved with reservations] F1000Research 2017, 6:1750 (doi: 10.12688/f1000research.12414.1)

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.

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 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., 2014) 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 plants were collected from its native place, Sundarban delta of West Bengal, India (21°.36’N and 88°.15’E) and established to our institute NET house. 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., 2014) 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).

The final assembly generated 58362 number of contigs 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 contig length of 56994164 (around 570 Mb) assembled genome, resulting 85.71 % genome coverage. It has been calculated that data contain very small amount of non-ATGC character. Further, we also found that the repeat contain 19.89% of the genome. We also identified approximately 1605 different non-coding RNAs and also found that the repeat contain 19.89% of the genome. The genome (KKLL) of O. coarctata is tetraploid (2n=4X=48) with a genome size estimated by flow cytometer is found to be approximately 665Mb. The Illumina 4000 GA IIx sequencer pair-end generated 137 Gb data. Further four mate-pair libraries together generated 104.35 Gb and Nanopore generated 6.35 Gb sequence data. Hence, we achieved 372.48 X depth of the genome of O. coarctata. The final assembly generated 58362 number of contigs 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 contig length of 56994164 (around 570 Mb) assembled genome, resulting 85.71 % genome coverage. It has been calculated that data contain very small amount of non-ATGC character. Further, we also found that the repeat contain 19.89% of the genome. We also identified approximately 1605 different non-coding RNAs and around 105673 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 also 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
P ublisher Full Text
P ubMed Abstract | Publisher Full Text | Free Full Text
Dolezel J, Greilhuber J, Suda J: Estimation of nuclear DNA content in plants
Pubmed Abstract | Publisher Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
Pubmed Abstract | Publisher Full Text | Free Full Text
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.