Draft genome sequence of *Oryza sativa* elite indica cultivar RP Bio-226

Mettu M. Reddy and Kandasamy Ulaganathan*

Center for Plant Molecular Biology, Osmania University, Hyderabad, India

INTRODUCTION

Rice cultivars show diversity in various factors influencing yield like nitrogen use efficiency, root development, and stress tolerance (Suryapriya et al., 2009; Singh et al., 2014; Begum et al., 2015; Mickelbart et al., 2015). For better understanding of the factors influencing yield of specific rice cultivars it is necessary to apply the high throughput genomic methods like whole genome sequencing, RNA-seq, smallRNA-seq, and ChIP-Seq. Improved Samba Mahsuri (RP Bio-226) is a bacterial leaf blight resistant indica rice cultivar developed through marker assisted selection from BPT5204 and SS1113. It is widely cultivated in Southern India because of its high yield, premium grain quality, and excellent cooking qualities. For understanding this cultivar in terms of nitrogen use efficiency recently we have sequenced the urea nutrition responsive transcriptome of this cultivar (Reddy and Ulaganathan, 2015). Further, with the aim of studying the genomic basis of yield of RP Bio-226 cultivar, we have sequenced the genome and analyzing it. Here we report the genomic dataset available publicly.

MATERIALS AND METHODS

Plant Material

RP Bio-226 rice seeds (*O. sativa* sp. Indica) were surface- sterilized with 0.1% HgCl2 for 5 min and rinsed thoroughly with distilled water. These seeds were then transferred to Murashige and Skoog basal agar medium (Murashige and Skoog, 1962) for germination. Plants were grown in a culture room at 25°C, 60–80% relative humidity and 16 h (light)/8 h (dark) photo-period for 10 days. After this time, seedlings were harvested and used for genomic DNA isolation. The genomic DNA was isolated from the young leaves using cetyl trimethyl ammonium bromide (CTAB) method (Murray and Thompson, 1980). DNA concentration and purity were estimated using Nanodrop Spectrophotometer and Qubit Fluorometer.

Library Preparation and Sequencing

For library preparation ∼3 µg of genomic DNA was sonicated using Covaris to obtain 200–500 bp fragment size. The size distribution was checked by running an aliquot of the sample on Agilent HS DNA Chip. The resulting fragmented DNA was cleaned up using High Prep PCR clean up beads. Fragmented DNA was subjected to a series of enzymatic reactions that repair ends, phosphorylate the fragments, and add a single ∼350–600 bp fragments was size selected on 2% low melting agarose gel and cleaned using MinElute column (QIAGEN). PCR (10 cycles) amplification of adapter ligated fragments was done and cleaned up using High Prep PCR beads. The prepared libraries were quantified using

Qubit fluorometer and validated for quality by running an aliquot on High Sensitivity Bioanalyzer Chip (Agilent). Whole genome sequencing was carried out with Illumina_ Nextseq500 system (Illumina, San Diego, CA, USA). The raw read files in Fastq format were used for the genome assembly.

Preprocessing and Genome Assembly

The preprocessing of raw reads was done with FastQC and the adapters were removed with Cut adapt (Andrews, 2010; Martin, 2011). After preprocessing the reads were aligned to the reference genome by using Bowtie2 (version 2.2.4) (Langmead and Salzberg, 2012). The indica rice 93-11 genome downloaded from BGI was used as the reference genome. Reference based assembly of the reads against the reference genome involved, indexing of the reference genome and alignment of reads to the reference and creation of SAM file. Samtools (version 0.1.18) and SnpEff (version 4.1) were used for further analysis (Li et al., 2009, 2012). SAM file was converted into binary BAM file, sorted and indexed by using the 'view,' 'sort,' and 'index' functions of Samtools. The quality of the assembly was checked by viewing the BAM file with 'Bamview' tool and used for variant calling. Samtools created a variation report from the Bowtie assembly with a mapping quality of >30 and read depth of >20 as cutoffs by using the mpileup function. It created a 'bcf' file which was converted into 'vcf' file with Bcftools. The duplicate variants were removed by varFilter in vcfutils. The consensus sequence was created with Samtools. Further annotation was done with snpEff (version 4.1) using the database built from the latest version of representative genes of RAP-DB (Sasaki et al., 2013). Gene prediction was carried out using Fgenesh program (Solovyev et al., 2006).

TABLE 1 | *Oryza sativa* indica cultivar RP Bio-226 genome characteristics and resources.

Name	RP Bio-226 genome characteristic/resource
NCBI bioproject ID	PRJNA285384
NCBI biosample ID	SAMN03751782
NCBI SRA accession No.	SRP062659
NCBI genome accession numbers	CP012609-CP012620
Sequence type	Illumina_ Next seq500
Total number of reads	69,377,450
Read length	151
Overall coverage	20×
Mapped reads	90%
Estimated genome size	347MB
Predicted protein coding genes	35700
Genome download link	RP Bio-226 genome
Browse the genome	RP Bio-226 Gbrowse

RESULTS

Whole Genome Sequencing of RP Bio-226

The sequencing produced a total of 32,49,1164 paired-end reads of 151 bp length. These reads were checked for low quality and adapter contamination and the low quality reads and adapter sequences were removed using FastQC and Cutadapt tools. The remaining reads were assembled onto the reference genome, 93-11 indica rice genome, using Bowtie2 (Langmead and Salzberg, 2012). Over 90% reads were aligned to the reference genome and the overall coverage was estimated to be more than $20\times$. The Sam file produced by this reference based assembly was used for generating the variation (Single nucleotide polymorphism and indels) report using Samtools (Li et al., 2009). The variants reported by the Samtools were annotated with SnpEff tool (Li et al., 2012). The assembled genome size was estimated to be 347 MB and a total of 35,700 protein coding genes were predicted using Soft berry (Fgenesh) gene finding tool (Solovyev et al., 2006) (**Table 1**).

Direct Link to Deposited Data and Information to Users

The dataset submitted to NCBI include the assembled chromosomal sequences of *O. sativa* indica cultivar RP Bio-226 in Fasta format and the raw reads. The assembled chromosomal sequences in Fasta format and the raw reads can be accessed at NCBI with the following links CP012609–CP012620 and SRP062659, respectively. Users can download and use the data freely for research purpose only with acknowledgment to us and quoting this paper as reference to the data.

REFERENCES

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Begum, H., Spindel, J. E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., et al. (2015). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical Rice (*Oryza sativa*). *PLoS ONE* 10:e0119873. doi: 10.1371/journal.pone.0119873

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H., Hansaker, B., Cingolani, P., Platts, A., Wang, L., Coon, M., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Martin, M. (2011). Cutadapt removes sequences from high-throughput sequencing reads. *EMBnet* 17, 10–12. doi: 10.14806/ej.17.1.200

Mickelbart, M. V., Hasegawa, P. M., and Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nat. Rev. Genet.* 16, 237–251. doi: 10.1038/nrg3901

Murashige, T., and Skoog, F. (1962). A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant.* 15, 473–497. doi: 10.1111/j.1399-3054.1962.tb08052.x

Murray, M. G., and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–4325. doi: 10.1093/nar/8.19.4321

Reddy, M. M., and Ulaganathan, K. (2015). RNA-Seq analysis of urea nutrition responsive transcriptome of *Oryza sativa* elite indica cultivar RP Bio 226. *Genom. Data* 6, 112–113. doi: 10.1016/j.gdata.2015.08.025

Sasaki, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013). Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54:e6. doi: 10.1093/pcp/pcs183

Singh, H., Verma, A., Ansari, M. W., and Shukla, A. (2014). Physiological response of rice (*Oryza sativa* L.) genotypes to elevated nitrogen applied under field conditions. *Plant Signal. Behav.* 9:e29015. doi: 10.4161/psb.29015

Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* 7, 10.1–10.12. doi: 10.1186/gb-2006-7-s1-s10

Suryapriya, P., Snehalatha, A., Kayalvili, U., Krishna, R., Singh, S., and Ulaganathan, K. (2009). Genome-wide analyses of rice root development QTLs and development of an online resource, Rootbrowse. *Bioinformation* 3, 279–281. doi: 10.6026/97320630003279