



OPEN

Draft genome sequence of the pulse crop blackgram [*Vigna mungo* (L.) Hepper] reveals potential R-genes

Souframanien Jegadeesan^{1,2}, Avi Raizada^{1,2}, Punniyamoorthy Dhanasekar¹ & Penna Suprasanna^{1,2}

Blackgram [*Vigna mungo* (L.) Hepper] ($2n = 2x = 22$), an important Asiatic legume crop, is a major source of dietary protein for the predominantly vegetarian population. Here we construct a draft genome sequence of blackgram, for the first time, by employing hybrid genome assembly with Illumina reads and third generation Oxford Nanopore sequencing technology. The final de novo whole genome of blackgram is ~ 475 Mb (82% of the genome) and has maximum scaffold length of 6.3 Mb with scaffold N50 of 1.42 Mb. Genome analysis identified 42,115 genes with mean coding sequence length of 1131 bp. Around 80.6% of predicted genes were annotated. Nearly half of the assembled sequence is composed of repetitive elements with retrotransposons as major (47.3% of genome) transposable elements, whereas, DNA transposons made up only 2.29% of the genome. A total of 166,014 SSRs, including 65,180 compound SSRs, were identified and primer pairs for 34,816 SSRs were designed. Out of the 33,959 proteins, 1659 proteins showed presence of R-gene related domains. KIN class was found in majority of the proteins (905) followed by RLK (239) and RLP (188). The genome sequence of blackgram will facilitate identification of agronomically important genes and accelerate the genetic improvement of blackgram.

Blackgram [*Vigna mungo* (L.) Hepper] is an annual leguminous crop belonging to the family *Fabaceae* and sub-family *Papilionaceae*. This crop is a major constituent of the genus *Vigna* Savi (subgenus *Ceratotropis*) grouped under the key tribe Phaseoleae that is known to accommodate other economically significant grain legumes like soybean (*Glycine max* (L.) Merr.), common bean (*Phaseolus vulgaris* L.), pigeonpea (*Cajanus cajan* (L.) Millsp.), mungbean (*Vigna radiata* (L.) R. Wilczek), cowpea (*Vigna unguiculata* (L.) Walp) and adzuki bean (*Vigna angularis* (Willd.) Ohwi & Ohashi). Blackgram is a self pollinated diploid ($2n = 2x = 22$) with genome size estimated to be 0.59 pg/1C (574 Mbp)¹. It is popularly known as 'urd bean', 'urd' or 'mash' and is an excellent source of easily digestible good quality protein (25–26%), carbohydrates (60%), fat (1.5%), minerals, amino-acids and vitamins. In addition to being an important source of human food and animal feed, it also plays a significant role in sustaining soil fertility by improving soil physical properties and fixing atmospheric nitrogen. As a hardy legume tolerant to drought, blackgram is suitable for dry land farming and is predominantly grown as an intercrop or as a sole crop under residual moisture conditions post rice harvest. Blackgram is extensively grown in south and south-east Asia from ancient times. It originated in India and has been domesticated from its wild ancestral form *V. mungo* var. *silvestris*². India is the largest producer of blackgram, where about 5.0 million hectares are cultivated with an annual production of 3.8 million tonnes³.

In spite of its economic importance and surging demand for improved blackgram varieties, susceptibility to multiple diseases, including mungbean yellow mosaic, powdery mildew, *Cercospora* leaf spot and leaf crinkle hinders cultivation and reduces produce yield and quality. In this regard, it is important to study plant disease resistance mechanisms and identify genes to develop varieties with durable resistance. Plant disease resistance genes (R-genes) play a key role in recognizing proteins expressed by specific avirulence (Avr) genes of pathogens⁴. The proteins encoded by the resistance genes share common domains such as coiled-coils (CC), nucleotide binding regions (NB), toll-interleukin regions (TIR), leucine rich regions (LRR) and kinases (K). Hundreds of NBS-LRR, RLK and RLP genes have been reported in plants^{5–8}. Pyramiding of plant resistance genes in new cultivars is the most effective and environment friendly approach for plant disease control and reduction of

¹Nuclear Agriculture and Biotechnology Division, BARC, Trombay, Mumbai 400085, India. ²Homi Bhabha National Institute, Training School Complex, Anushaktinagar, Mumbai 400094, India.  email: souf@barc.gov.in

Scaffolds generated	1085
Maximum Scaffold length (bp)	63,43,804
Minimum Scaffold length (bp)	510
Average Scaffold length (bp)	438,629
Median Scaffold length (bp)	67,909
Total Scaffolds length (bp)	47,59,13,455
Scaffolds \geq 100 bp	1085
Scaffolds \geq 200 bp	1085
Scaffolds \geq 500 bp	1085
Scaffolds \geq 1 Kbp	1048
Scaffolds \geq 10 Kbp	920
Scaffolds \geq 1 Mbp	168
N50 value	14,26,686
Number of genes	42,115
Average gene length	1131 bp
Maximum gene length	23.17 kb
Minimum gene length	120 bp
Number of genes annotated	33,959

Table 1. De novo assembly and annotation statistics of the blackgram genome.

yield losses. Such useful information is lacking in blackgram. This could be attributed to the lack of genomic resources coupled with limited understanding of the molecular basis of gene expression and phenotypic variation. Whole genome assemblies support genome wide association studies (GWAS) to identify trait-specific loci and for genomic-based selective breeding⁹. Whole-genome sequencing has been conducted on several commercial *Vigna* species such as mungbean, adzuki bean, cowpea, beach pea^{5,10–13}. Elucidation of the genome sequence of *V. mungo* var. *mungo* could reveal the general genome structure, repetitive sequences and R-gene composition of this legume species in comparison to closely related genomes and greatly assist comparative genomics with other well-studied legume genomes.

Next-generation sequencing (NGS) reads are too short to resolve abundant repeats particularly in plant genomes, leading to incomplete or ambiguous assemblies¹⁴. Construction of highly contiguous genomes has been possible in recent years owing to expeditious advances in sequencing technologies and substantial refinements in assembly algorithms. The advent of third generation sequencing technologies capable of delivering long reads over several kilobases for haplotype phasing have significantly enhanced the possibility of de novo assemblies^{15–17}. In view of the importance of this pulse crop in the Asiatic region and the need for molecular detailing of trait based selection, we assembled a draft genome of *Vigna mungo* var. *mungo* using next-generation platform Illumina paired end and mate pair reads combined with third generation Oxford Nanopore sequencing.

Results

Illumina and nanoporesequencing of blackgram. We prepared three libraries for sequencing by Illumina HiSeqX Ten sequencer including 150 bp paired-end library and 5–7 kb and 7–10 kb mate-pair libraries. Whole genome sequencing using Illumina paired-end (PE) long insert generated 154,940,012 reads representing \sim 98x genome coverage. Sequencing of 2 mate-pairs of 5–7, and 7–10 kb yielded, 33,617,232 and 10,247,813 reads, respectively, with an approximate coverage of 21.2x and 6.5x, respectively, and a grand total of 43 million mate-pair reads representing \sim 28x coverage (Table S1). In addition, longread sequencing by Oxford Nanopore sequencing technology (ONT) was used to generate 1,633,898 long reads, having 10,425,220,236 bp and coverage of \sim 22x. A total of 11.5 Gb data was generated from whole genome library with an average read length of 6.4 kb and a maximum read length of 128.7 kb using Nanopore sequencer (Table S2). The complete genome was sequenced to a depth of \sim 148x, using both Illumina and ONT platforms.

De novo assembly of blackgram genome and gene annotation. The raw reads generated from Illumina paired end, mate-pair and nanopore sequencing were processed and good quality reads were retained. Hybrid assembly was performed using Illumina and Nanopore reads by MaSuRCA v3.3.4 hybrid assembler. Scaffolds were further processed for super-scaffolds using PyScaf producing 1085 scaffolds with a N50 of 1.42 Mb (Table 1). Overall, the maximum scaffold assembled length was 6343.0 kb with median scaffold length of 67.9 kb. The total length of the produced scaffolds was 475 Mb (82% of genome) for *Vigna mungo* cultivar Pant U-31. Read utilization was also performed to ascertain the correctness of the assembly. Illumina reads were mapped against assembly of 280,233,560 total processed reads with 279,112,626 mapped reads (99.60%). Similarly, Nanopore reads were mapped against assembly of 1,633,786 total processed reads with 1,626,270 mapped reads (99.53%).

The gene prediction and annotation of the assembled genome was carried out using repeat masked assembly genome and reference transcriptome data of *Vigna mungo* using BRAKER tool. In total 42,115 genes were identified with average coding sequence length of 1131 bp. The maximum and minimum sequence lengths

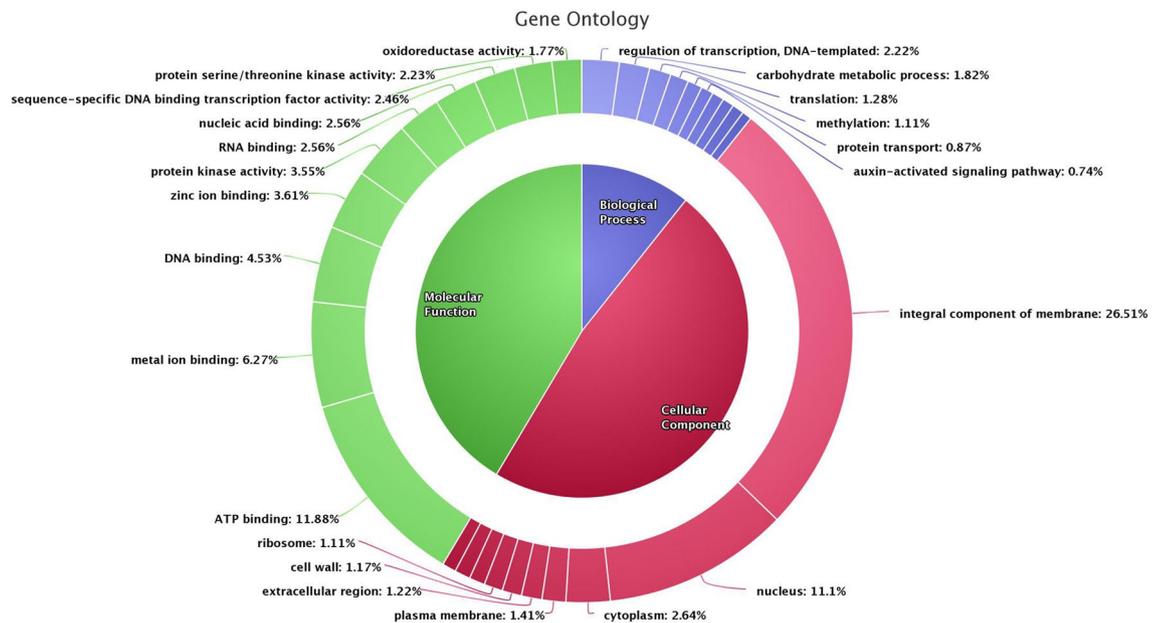


Figure 1. Gene ontology chart of *Vigna mungo*.

were 23.17 kb and 120 bp, respectively (Table 1). A total of 33,959 genes of the predicted genes (80.6%) could be functionally annotated with gene ontology and pathway information (Table S3). Gene ontology provides a system to categorize description of gene products according to three ontologies: molecular function, biological process and cellular component. Of the 33,959 annotated genes, majority (45.2%) were assigned with cellular components, followed by molecular (41.4%) and biological functions (8.0%). Among the assignment made to the cellular component function, the majority represented integral component of membranes (26.5%) followed by nucleus (11.1%) and cytoplasm (2.6%). Among those with molecular function, a large proportion of the sequences represented ATP binding (11.8%) followed by metal ion binding (6.3%) and DNA binding (4.5%). Under the biological process category, more sequences were assigned to regulation of transcription (2.2%) followed by carbohydrate metabolic process (1.8%) and translation process (1.3%) (Fig. 1). To assess the completeness of *Vigna mungo* genome assembly and gene annotation, we performed the BUSCO analysis with summarized benchmarking: C: 96.8% (S: 94.6%, D: 2.2%, m: 2.5%, n: 5366) and ~97% of genes were observed to be complete which also validates the completeness of draft assembly genome. Pathway assignments were carried out according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. A total number of 16,404 unique KEGG pathways were identified (Table S4), of which the majority of sequences were grouped into protein families (8715) followed by carbohydrate metabolism (1158) and transcription (954). Orthologous gene comparison studies using genes from *Vigna mungo* (Pant U-31), *Vigna radiata*, *Vigna unguiculata* and *Vigna angularis* were carried out using Ortho Venn13. A total of 19,095 gene clusters were shared by all four species, while 1970 gene clusters were specific to *Vigna mungo* (Fig. 2).

Prediction of transposons. The presence of transposons in the assembled genome was predicted using TREP (TRansposable Elements Platform). Repetitive sequences occupy 49.6% of the *V. mungo* genome as revealed by homology- and structure-based surveys. Majority of the transposable elements were retrotransposons (47.3% of genome), whereas DNA transposons made up only 2.29% of the genome (Table 2). Long terminal repeat (LTR) retrotransposons forming the predominant class of transposable elements in *V. mungo* genome showed homologies with that of *Metrosideros polymorpha*, *Blumeria graminis tritici*, *Sorghum bicolor*, *Triticum aestivum*, *Hordeum vulgare*, *Brachypodium distachyon*, *Arabidopsis thaliana* and *Oryza sativa* genomes. Overall, 47.3% of the repetitive DNA was long terminal repeat retrotransposons of which 13.4% were Gypsy type and 34.5% were Copia type elements. In contrast, class II DNA transposons, including Mutator, PIF-Harbinger, hAT, Helitron, and Tc1-Mariner, accounted for 2.3% of the blackgram genome. The rolling-circle Helitron (DHH) superfamily is relatively abundant at 1.3% of the genome (Table S5). Only 3.1% of the TE sequences were unclassified.

Simple sequence repeats (SSR) prediction. SSRs were detected using Microsatellite Identification Tool (MISA v1.0). A total of 166,014 SSRs were identified from 989 scaffolds (Table 3). More than one SSR were present in 953 scaffolds and 65,180 SSRs were of the compound type. SSR loci with di- and tri-nucleotides constituted 103,955 (62.6%) of the identified loci. The proportions of di-, tri-, tetra-, penta-, and hexa-nucleotide repeats were 38.1%, 24.5%, 36.4%, 0.69%, and 0.24%, respectively (Table S6). The number of repeats varied from 6–61 for di-nucleotides, 5–361 for tri-nucleotides, 3–7 for tetra-nucleotides, 5–19 for penta-nucleotides and 5–14 for hexa-nucleotides. The most prevalent di-, tri-, tetra-, penta-, and hexa-nucleotide repeats were AT (22.6%), AAT (3.9%), TTTA (5.1%), AAAAT (4.6%) and ATGTTG (1.9%), respectively (Table S7). Of the

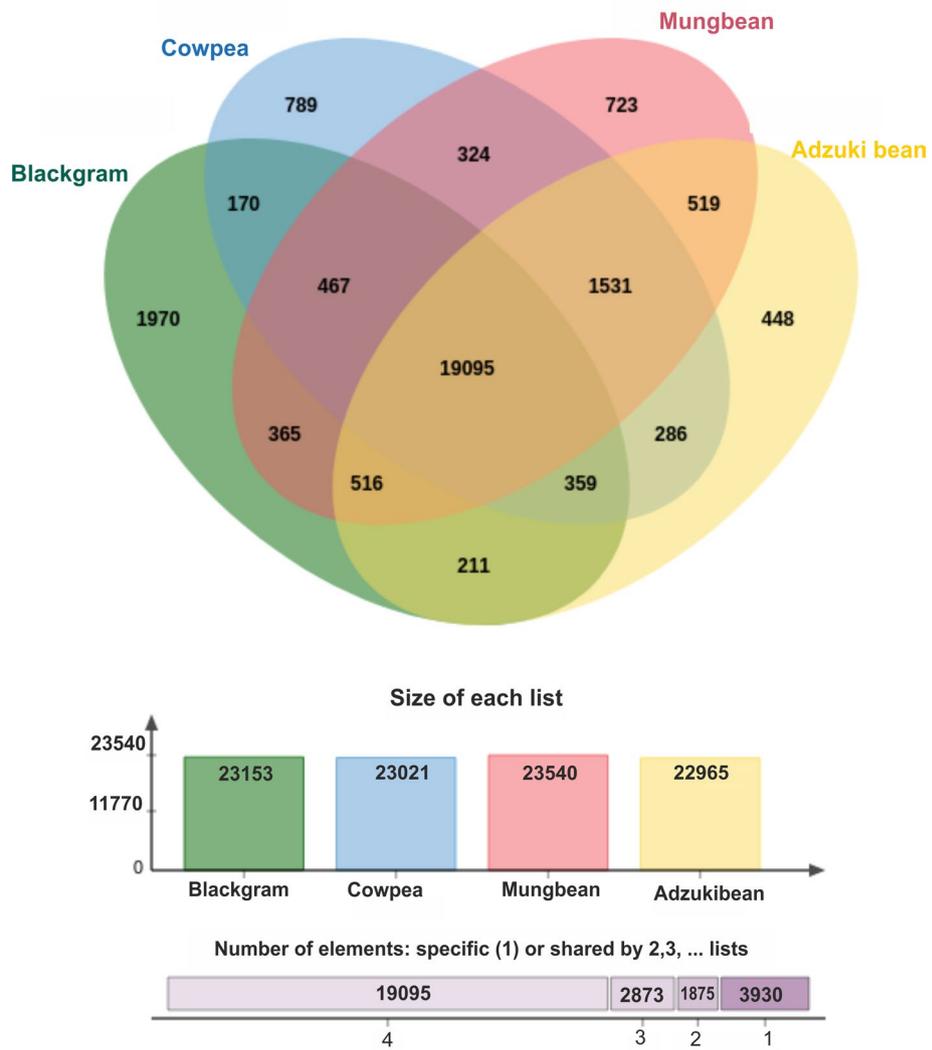


Figure 2. Venn diagram showing shared orthologous gene clusters among *V. mungo*, *V. radiata*, *V. unguiculata* and *V. angularis*.

	<i>V. mungo</i> (% genome)
Genome assembly size (Mbp)	475.91
Transposable elements	49.6
Class I: LTR Retrotransposon (RLX)	47.3
<i>Gypsy</i> (RLG)	13.38
<i>Copia</i> (RLC)	31.46
unclassified LTR (RLX)	3.15
Class II: TIR DNA transposon (DXX)	2.29
Helitron (DHH)	1.3
PIF-Harbinger (DTH)	0.40
Mariner (DTT)	0.33
Mutator (DTM)	0.11
hAT (DTA)	0.1
Class I/class II ratio	15.6
<i>Gypsy/Copia</i> ratio	0.6

Table 2. Annotated repeat abundances in blackgram. The major represented classes, super-families, and subgroups of transposable elements as determined by automated annotation and classified according to the scheme of Wicker et al.²⁷, as well as other major repeat types are presented.

Description	<i>V. mungo</i> genome
Total number of sequences examined	1085
Total size of examined sequences (bp)	475,913,455
Total number of identified SSRs	166,014
Number of SSR containing sequences	989
Number of sequences containing more than 1 SSR	953
Number of compound SSRs (i.e. c)	65,180
p2	63,220
p3	40,735
p4	60,512
p5	1146
p6	402

Table 3. Number and distribution of SSRs identified in the blackgram (*Vigna mungo*) cv. Pant U-31 genome.

166,014 SSR motifs identified, PCR primer pairs were successfully designed for 34,816 SSR loci. Details about primer sequences and expected product sizes for 34,816 SSR loci are provided in supplementary table (Table S7).

Identification of disease resistance genes. A total of 33,959 protein sequences were analysed for resistance (R) genes related domains and motifs with the help of DRAGO 2 (Disease Resistance Analysis and Gene Orthology) pipeline of plant resistance gene database (PRGDB). Out of 33,959 proteins, 1659 proteins showed presence of R-gene related domains. Majority of the proteins (688) contained TM-kinase domains and kinase formed the major class (905 proteins) (Table S8). One hundred and forty-two proteins (8.6%) were found to have Nucleotide Binding Sites (NBS) (Table 4). A total of 294 proteins showed a single type domain (219 kinase, 40 LRR, 16 NBS, 10 TIR and 9 TM) (Table 4), while remaining proteins harboured more than one domain types such as NBS-TM, TIR-NBS etc. The LRR-TM-Kinase-CC, NBS-LRR-TM, NBS-CC-TM-TIR-LRR, NBS-LRR-TM-TIR and NBS-CC-TM-LRR domain combinations were found in 3, 23, 2, 7 and 17 proteins, respectively. Among the different classic R-gene classes majority were found to be of kinases (KIN) (54.5%) followed by transmembrane receptors (RLP or RLK) (25.7%) and twenty-seven proteins were found to represent the class of cytoplasmic proteins (CNL and TNL). The classic R-gene classes RLP (Ser/Thr-LRR) and RLK (Kin-LLR) were found in 188 and 239 proteins, respectively. R-domain occurrence in the full dataset showed that the NBS and LRR domains were found in 8 and 9 classes, respectively, followed by the KIN domain in 5, and TIR domains in 6 classes. Likewise, proteins showing other classes such as TN, TRAN, NL, CNL, C, CTNL and CLK were found in 1, 2, 31, 18, 1, 2 and 3 proteins, respectively (Table 4). Seventy-one R-genes were identified based on their homologies with mungbean, cowpea and adzuki bean sequences (Table S9).

Discussion

A better understanding of blackgram genetics is crucial for more efficient breeding in light of an anticipated increase in biotic and abiotic stresses that may accompany climate change. Whole-genome sequences are an important resource for evolutionary geneticists studying plant domestication, as well as breeders aiming to improve crop varieties. We sequenced *V. mungo* using Illumina PE and Nanopore with a coverage of 148x and assembled genome using MaSuRCA hybrid assembler. The final assembly comprised of 1085 scaffolds (N50 = 1.43 Mb). Hybrid assembly through combinational sequencing is a useful approach in obtaining accurate sequence data. Moreover, the production of long-reads while using third generation sequencing (Nanopore) overcomes the weakness of assembling short-reads by minimizing the generation of gaps or covering the repetitive sequences that appear in the plant genomes. In addition, while only considering the accuracy, short-reads can be used for error-correction by aligning them to long-reads, which enable the increased accuracy of the genome assembly¹⁸. We constructed 475 Mb (82%) of the total estimated *V. mungo* var. *mungo* genome and identified 42,115 protein-coding genes and 1970 *Vigna mungo* specific gene clusters. The assembly generated will also advance comparative genomics in *Vigna* species, as whole genome sequences of prominent *Vigna* species including mung bean, adzuki bean and cowpea are already available^{5,11,12}. Of the 42,115 predicted genes, 33,959 could be functionally annotated. In *V. radiata* genome, 22,427 genes were annotated with high confidence⁵. Most of the gene annotations were comparable to the annotation of immature seed transcriptome sequence of blackgram¹⁹. Orthologous gene comparison studies using genes from *Vigna mungo* (Pant U-31), *Vigna radiata*, *Vigna unguiculata* and *Vigna angularis* revealed that a total of 19,095 gene clusters were shared by all four species. High degree of conservation and collinearity between blackgram and adzuki bean was revealed through comparative mapping²⁰. Gene order conservation between closely related legume species (*V. angularis* var. *angularis*, *V. radiata* var. *radiata*, and *P. vulgaris*) has been exploited in synteny based scaffolding approach in genome assembly¹¹. Similarly, Cowpea chromosomes Vu02, Vu03 and Vu08 also have one-to one relationship with the other two *Vigna* species (mungbean and common bean) suggesting that these chromosome rearrangements are characteristic of the divergence of *Vigna* from *Phaseolus*¹².

Domain/motif types	Number of proteins	Class	Number of proteins
TM-Kinase	688	KIN	905
TM-Kinase-LRR	235	RLK	239
Kinase	219	RLP	188
LRR-TM	188	CK	102
CC-TM-Kinase	82	N	66
NBS-TM	47	L	40
LRR	40	NL	31
NBS-LRR-TM	23	CN	20
CC-Kinase	20	CNL	18
NBS-CC-TM	17	CL	14
NBS-CC-TM-LRR	17	T	12
NBS	16	TNL	9
TIR	10	CLK	3
TM	9	TL	3
CC-LRR-TM	8	CTNL	2
NBS-LRR-TM-TIR	7	NK	2
CC-LRR	6	TRAN	2
NBS-LRR	6	C	1
NBS-CC	3	CT	1
LRR-TM-Kinase-CC	3	TN	1
LRR-Kinase	2	Total proteins	1659
TM-TIR	2		
NBS-TM-TIR	2		
LRR-TM-TIR	2		
NBS-CC-TM-TIR-LRR	2		
LRR-TIR	1		
NBS-CC-LRR	1		
NBS-LRR-TIR	1		
CC-TM	1		
CC-TM-TIR	1		
Total proteins	1659		

Table 4. Prediction of resistance genes domains/motifs present in proteins identified from whole genome sequencing of blackgram cultivar Pant U-31 with the help of DRAGO pipeline of Plant resistance gene database.

Transposable elements (TEs). In plants, transposable elements are a major driver of genome expansion. Retrotransposons are the predominant TEs in large plant genomes and are further divided into class I, those flanked by long terminal repeats (LTRs) and those devoid of them. The class II elements, on the other hand, transpose via DNA intermediate and possess terminal inverted repeats (TIRs), which serves as sites of excision and re-integration by element-encoded transposase²¹. Homology and structure based analysis revealed that LTRs are the predominant class of transposable elements in the *Vigna mungo* genome, consistent with other legume species^{5,22–26}. Of the long terminal repeat (LTR) retrotransposons, elements of the *Copia* superfamily²⁷ (code RLC) are 0.6 times more abundant than *Gypsy* (code RLC) elements in blackgram. However, *Gypsy* element was found to be more abundant in the related *Vigna* species such as mung bean, adzuki bean and cowpea^{5,11,12}. The DNA, or class II, transposons comprise 2.3% of the genome, with Mutator, PIF-Harbinger, hAT, Helitron, and Tc1-Mariner being the major groups of classical ‘cut-and-paste’ transposons in blackgram. The rolling-circle Helitron (DHH) superfamily relatively abundant in blackgram is consistent with cowpea¹².

TEs are potential reservoirs of phenotypic variation and phenotypic plasticity²⁸. Moreover, TEs can directly assist the crop improvement programs through molecular marker approach. The presence of TEs, often close to or within the stress responsive quantitative trait loci (QTLs), especially plant defense genes, along with the traditional attributes of a molecular marker, make them the markers of choice for diversity studies and trait mapping^{29,30}. While more studies would be necessary to understand the functional effects of these insertions, long-read sequences have greatly improved the assembly and identification of repeat types.

Simple sequence repeats. The development of genomic resources is critical for crop improvement programmes. NGS has allowed the discovery of a large number of DNA polymorphisms, such as SNP and InDels markers, in a relatively short time at low cost³¹. Among 166,014 SSRs (excluding mono nucleotide repeats) identified, the proportions of dinucleotide repeats were higher (38.1%) compared to other repeats in *V. mungo*.

Similarly, dinucleotide repeats were found to be higher (71.3%) compared to other repeats in *V. radiata*⁵. Proportion of tri-, tetra-, penta-, and hexa-nucleotide SSRs were more or less same in comparison to *V. radiata* (24.6%, 2.5%, 1.2%, 0.2%) and lower than *V. marina* (49%, 3%, 7%, 5%) except for tetra-nucleotide repeats. Tetra-nucleotide repeats in *V. mungo* were found to be higher (36.4%) in comparison to *V. radiata* (2.5%) and *V. marina* (3.0%). Likewise, the number of compound SSRs was higher (39.2%) than that in *V. radiata* (35.9%) and *V. marina* (10.08%)^{5,13}. To date, few efforts have been made to develop sufficient genomic resources in *Vigna*. This pioneer genome sequencing effort in *V. mungo* has generated SSRs and functional annotations for a huge set of genes. This information holds great promise for use in trait mapping, genomic selections, and diversity assessment.

Disease resistance genes. Whole genome sequencing has enabled genome-level investigation of the R-gene family in crop plants such as mungbean, chickpea, rice, tomato^{5–8}. In blackgram, 3.9% of the total genes were found to contain R-genes which is higher (1.2%) than that reported for *Medicago*³² and lower (5.27%) than that reported for *Arabidopsis*³³. Plants possess a sophisticated immune system based on their ability to recognize phytopathogens. The activation of this system is based on the presence of specific receptors encoded by R-genes. Resistance genes are grouped as either nucleotide binding site leucine rich repeat (NBS-LRR) or transmembrane leucine rich repeat (TM-LRR)³⁴. NBS-LRR proteins encoded by resistance (R) genes play an important role in pathogen recognition process and the activation of signal transduction in the response to pathogen attack. NBS-LRR can be further classified as toll/interleukin receptor (TIR)-NBS-LRR (TNL) or non-TNL/coiled coil-NBS-LRR (CNL)³⁴. Both TNL and CNL specifically target pathogenic effector proteins inside the host cell, and thus mediate effector triggered immunity (ETI) response³⁵. In *Vigna mungo* 8.6% of total identified R-gene related sequence showed NBS domain. In *Vigna mungo* transmembrane leucine rich repeat (TM-LRR) class such as receptor like kinase (RLK) and receptor like protein (RLP) accounted for 25.7% of the R-genes identified. RLPs and RLKs are pattern recognition receptors (PRRs) that mediate pathogen/microbe associated molecular pattern (PAMP/MAMP) triggered immunity (PTI/MTI) to allow recognition of a broad range of pathogens³⁵. Development of diagnostic molecular markers associated with key disease resistance gene would aid in molecular resistance breeding.

In this study, the black gram genome was assembled using hybrid approach with the size of 475 Mb. This has potential for developing gold standard reference assembly in future. A total of 42,115 genes were predicted from the assembled genome. Further, the predicted genes were annotated with gene ontology and pathway information. The presence of transposons and SSRs in the assembled genome was also predicted. Blackgram is grown mostly in developing countries and lack of genome sequence has delayed the implementation of molecular breeding in this *Vigna* species. The whole-genome sequence and SSR discovery will thus boost genomics-assisted selection for blackgram genetic improvement.

Methods

DNA extraction. Blackgram (*V. mungo* var. *mungo*) cultivar Pant U-31 developed by GB Pant University of Agriculture and Technology, is a popular yellow moosaic virus resistant cultivar in public domain. The pure lines of this cultivar are maintained at Nuclear Agriculture and Biotechnology Division, Bhabha Atomic Research Centre, Trombay, Mumbai, India. Pant U-31 was used for whole genome sequencing. DNA was extracted from 50 to 100 mg young leaves using Qiagen DNeasy Plant Mini kit following manufacturer's instructions. Extracted genomic DNA was quantified and assessed for quality using Nanodrop2000 (Thermo Scientific, USA), Qubit (Thermo Scientific, USA) and agarose gel electrophoresis.

Illumina library preparation and sequencing. Whole genome sequencing (WGS) libraries were prepared using Illumina-compatible NEXTflex Rapid DNA sequencing Bundle (BIOO Scientific, Inc. U.S.A.) at Genotypic Technology Pvt. Ltd., Bangalore, India. Briefly, 300 ng of Qubit quantified DNA was sheared using Covaris S220 sonicator (Covaris, Inc. USA) to generate specific fragments in the size range of 300–400 bp. The fragment size distribution was verified on Agilent 2200 TapeStation and subsequently purified using High prep magnetic beads (Magbio Genomics). Purified fragments were end-repaired, adenylated and ligated to Illumina multiplex barcode adaptors as per NEXTflex Rapid DNA sequencing bundle kit protocol³⁶.

Matepair illumina library preparation. Mate pair sequencing library was prepared using Illumina-compatible Nextera Mate Pair Sample Preparation Kit (Illumina Inc., Austin, TX, U.S.A.). About 4 µg of genomic DNA was simultaneously fragmented and tagged with mate pair adapters in a transposon based tagmentation step. Tagmented DNA was then purified using AMPure XP magnetic beads (Beckman Coulter, USA) followed by strand displacement to fill gaps in the tagmented DNA. Strand displaced DNA was further purified with AMPure XP beads before size-selecting the fragments on low melting agarose gel. Size selected fragments were circularized in an overnight blunt-end intra-molecular ligation step that resulted in circular DNA with the insert flanked mate pair adapter junction. Circularized DNA was sheared using Covaris S220 sonicator (Covaris, Woburn, Massachusetts, USA) to generate fragment size distribution from 300 to 1000 bp. Sheared DNA was purified to collect the Mate pair junction positive fragments using Dynabeads M-280 streptavidin magnetic beads (Thermo Fisher Scientific, Waltham, MA, USA). Purified fragments were end-repaired, adenylated and ligated to Illumina multiplex barcode adaptors as per Nextera Mate Pair Sample Preparation Kit protocol. Sequencing library, thus constructed, was quantified using Qubit fluorometer (Thermo Fisher Scientific, MA, USA) and its fragment size distribution was analyzed on Agilent 2200 TapeStation. The libraries were sequenced on Illumina HiSeq X Ten sequencer (Illumina, San Diego, USA) using 150 bp paired-end chemistry following manufacturer's instructions.

Nanopore library preparation and sequencing. A total of 1.5 µg of gDNA was end-repaired (NEBnext ultra II end repair kit, New England Biolabs, MA, USA) and purified using 1 × AmPure beads (Beckmann Coulter, USA). Adapter ligation (AMX) was performed at RT (20 °C) for 20 min using NEB Quick T4 DNA Ligase (New England Biolabs, MA, USA). The reaction mixture was purified using 0.6 × AmPure beads (Beckmann Coulter, USA) and sequencing library was eluted in 15 µl of elution buffer provided in the ligation sequencing kit (SQK-LSK109) from Oxford Nanopore Technology (ONT). Sequencing was performed on GridION X5 (Oxford Nanopore Technologies, Oxford, UK) using SpotON flow cell R9.4 (FLO-MIN106) in 48 h sequencing protocol on MinKNOW (version 1.1.20, ONT) with Albacore (v1.1.2)³⁷ live base calling enabled with default parameters.

Primary data analysis. The data obtained from the Illumina sequencing run was demultiplexed using Bcl2fastq software v2.20 (https://sapac.support.illumina.com/sequencing/sequencing_software/bcl2fastqconversion-software.html) and FastQ files were generated based on the unique dual barcode sequences. The sequencing quality was assessed using FastQC v0.11.8 software³⁸. The adapter sequences were trimmed using Trimalore v0.4.0³⁹ and bases above Q30 were considered and low quality bases were filtered off during read pre-processing and used for downstream analysis. Similarly, the Nanopore reads were processed with default settings using Porechop tool (<https://github.com/rrwick/Porechop>). The pre-processing of Nanopore data retained 99.9% of data.

De novo genome assembly and gene annotation. Hybrid assembly was performed using Illumina and Nanopore processed reads by MaSuRCA v3.3.4 hybrid Assembler⁴⁰ with standard parameters. The assembled contigs were utilized to generate larger scaffolds using pyScaf(v1) software (<https://github.com/lpryszcz/pyScaf>). The generated assembled genome of ~475 MB size was used for further analysis. The correctness of the assembly was ascertained by mapping short and long reads to the assembly. For gene prediction and annotation of the assembled genome, we used combination of ab initio prediction and transcriptome data of *Vigna mungo* using BRAKER⁴¹ version 3.0.2. It helped in the identification of protein-coding genes and their exonic-intronic structure in the genome in order to improve the accuracy and completeness of the annotation. BRAKER predicted proteins were annotated against all Fabaceae protein sequences from Uniprot database⁴² using DIAMOND blastp⁴³ program with an e-value of 1e-5 for gene ontology and annotation. To assess the completeness of our *Vigna mungo* genome assembly and annotation, we employed the BUSCO software⁴⁴ to check the gene content using a plant specific database. Pathway analysis was performed using KAAS server⁴⁵. KAAS (KEGG Automatic Annotation Server) provides functional annotations of genes in a genome by amino acid sequence comparisons against a manually curated set of ortholog groups in KEGG genes. Comparative analysis of the organization of orthologous gene clusters were carried out using genes of *Vigna mungo*, *Vigna radiata*, *Vigna unguiculata* and *Vigna angularis* through OrthoVenn⁴⁶ with E-value of 0.01 and inflation value of 1.5.

Identification of transposable elements (TEs) and simple sequence repeats (SSRs). Transposable elements analysis was performed against TREP (TRansposable Elements Platform)⁴⁷ which is a curated database of TEs (<http://botserv2.uzh.ch/kelldata/trep-db/index.html>). Each consensus representing a structural variant of a TE family was classified according to its structural and functional features. TEs classifications were based on its ability to replicate in a host genome using various transposition mechanisms and are divided into two classes based on their replication mechanism. Retrotransposons (class I) use an RNA intermediate for transposition, while DNA transposons (class II) use a DNA intermediate for transposition²⁷. The genome sequence was checked for homology with TREP database using BLASTn⁴⁸ and the genomic positions having homology with known TEs were identified.

SSRs were identified from the genome sequence using MicroSatellite identification tool (MISA)⁴⁹ (<http://pgrc.ipk-gatersleben.de/misa/>). This predicted polymorphic loci of 1–6 bp length in nucleotide sequences. Repeats were identified in each scaffold sequences using MISA Perl script. In this study, the SSRs were considered to contain motifs with two to six nucleotides in size and a minimum of 6, 6, 3, 5, 5 contiguous repeat units for di-, tri-, tetra-, penta- and hexa-nucleotides, respectively. Mononucleotide repeats were not included in the SSR search criteria. Based on MISA results, primers were designed for SSR motifs using either WebSat (<http://purl.oclc.org/NET/websat/>) online software⁵⁰ or batch primer3 ver1.0⁵¹. For designing PCR primers, parameter for optimum primer length was 22 mer (range: 18–27 mer), optimum annealing temperature was 60 °C (range: 57–68 °C), GC content was 40–80%, and other parameter values as default.

Identification of disease resistance genes. Disease Resistance Analysis and Gene Orthology (DRAGO v.2) pipeline was used to predict and annotate the disease resistance genes from the Plant Resistance Genes database (PRGdb 3.0; <http://prgdb.org>) with curated reference R-genes^{52,53}. DRAGO was executed with peptide sequence file from *V. mungo* var. *mungo* as an input to define the normalization value and the minimum score thresholds. Specifically, the previously created 60 HMM (hidden Markov model) modules were used by DRAGO 2 to detect LRR, Kinase, NBS and TIR domains and compute the alignment score of the different hits based on a BLOSUM62 matrix. The normalization value was the absolute smallest similarity score found among the input sequences considering all domains. The minimum score thresholds were calculated from the smallest similarity score reported in a specific domain among the input sequences. DRAGO 2 generated files with numeric matrix that represented the similarity score of every single protein input to each HMM profile, the domain name, start position, end position, resistance class and identification for every putative plant resistance protein.

Data availability

The de novo genome assembly has been deposited at GenBank under submission ID, Bioproject PRJNA644765, biosample SAMN15473271 (SRA accessions: SRX9175307 and SRX9175306).

Received: 21 June 2020; Accepted: 17 May 2021

Published online: 27 May 2021

References

- Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–215 (1991).
- Lukoki, L., Marechal, R. & Otoul, E. Les ancetres sauvages des haricots cultives: *Vigna radiata* (L.) Wilczek et *V. mungo* (L.) Hepper. *Bull. Jard. Bot. Nat. Belgique* **50**, 385–391 (1980).
- Anonymous, Area, production, productivity of blackgram in India. Directorate of Economics and Statistics. Ministry of Agri. & FW, Govt. of India, 2017–18 (2018).
- Ellis, J., Dodds, P. & Pryor, T. The generation of plant disease resistance gene specificities. *Trends Plant Sci.* **5**, 373–379 (2000).
- Kang, Y. J. *et al.* Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* **5**, 5443. <https://doi.org/10.1038/ncomms6443> (2014).
- Li, Y. *et al.* Genome analysis identified novel candidate genes for *Ascochyta* blight resistance in chickpea using whole genome re-sequencing data. *Front Plant Sci.* **8**, 359. <https://doi.org/10.3389/fpls.2017.00359> (2017).
- Arafa, R. A. *et al.* Rapid identification of candidate genes for resistance to tomato late blight disease using next-generation sequencing technologies. *PLoS ONE* **12**, e0189951 (2017).
- Read, A. C. *et al.* Genome assembly and characterization of a complex zfBED-NLR gene-containing disease resistance locus in Carolina Gold Select rice with Nanopore sequencing. *PLoS Genet.* **16**, e1008571 (2020).
- Varshney, R. K., Terauchi, R. & McCouch, S. R. Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLOS Biol.* **12**, e1001883 (2014).
- Yang, K. *et al.* Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *PNAS* **112**, 13213–13218 (2015).
- Kang, Y. *et al.* Draft genome sequence of adzuki bean, *Vigna angularis*. *Sci. Rep.* **5**, 8069 (2015).
- Lonardi, S. *et al.* The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *Plant J.* **98**, 767–782 (2019).
- Singh, A. K. *et al.* Draft genome sequence of a less-known wild *Vigna*: Beach pea (*V. marina* cv. ANBp-14–03). *Crop J.* **7**, 660–666 (2019).
- Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2020).
- Jiao, W. B. & Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* **36**, 64–70 (2017).
- Li, C., Lin, F., An, D., Wang, W. & Huang, R. Genome sequencing and assembly by long reads in plants. *Genes* **9**, 6. <https://doi.org/10.3390/genes9010006> (2018).
- Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
- Jayakumar, V. & Sakakibara, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief. Bioinform.* **20**, 866–876 (2017).
- Souframanien, J. & Reddy, K. S. D. novo assembly, characterization of immature seed transcriptome and development of genic-SSR markers in blackgram [*Vigna mungo* (L.) Hepper]. *PLoS ONE* **10**, e0128748 (2015).
- Gupta, S. K., Souframanien, J. & Gopalakrishna, T. Construction of a genetic linkage map of black gram, [*Vigna mungo* (L.) Hepper] based on molecular markers and comparative studies. *Genome* **51**, 628–637 (2008).
- Finnegan, D. Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5**, 103–107. [https://doi.org/10.1016/0168-9525\(89\)90039-5](https://doi.org/10.1016/0168-9525(89)90039-5) (1989).
- Sato, S. *et al.* Genome structure of the legume *Lotus japonicus*. *DNA Res.* **15**, 227–239 (2008).
- Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
- Young, N. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520–524 (2011).
- Varshney, R. *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89. <https://doi.org/10.1038/nbt.2022> (2012).
- Varshney, R. *et al.* Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246 (2013).
- Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
- Paszowski, J. Controlled activation of retrotrans position for plant breeding. *Curr. Opin. Biotechnol.* **32**, 200–206. <https://doi.org/10.1016/j.copbio.2015.01.003> (2015).
- Kalendar, R. *et al.* Analysis of plant diversity with retrotransposon-based molecular markers. *Heredity* **106**, 520–530 (2011).
- Alzohairy, A. *et al.* Retrotransposon-based molecular markers for assessment of genomic diversity. *Funct. Plant Biol.* **41**, 781–789 (2014).
- Varshney, R. K., Nayak, S. N., May, G. D. & Jackson, S. A. Next generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **27**, 522–530 (2009).
- Yu, J. *et al.* Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana*. *BMC Genom.* **15**, 3–21 (2014).
- Meyers, B. C. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15**, 809–834 (2003).
- Hammond-Kosack, K. E. & Jones, J. D. Plant disease resistance genes. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 575–607 (1997).
- Chisholm, S. T., Coaker, G., Day, B. & Staskawicz, B. J. Host-microbe interactions: Shaping the evolution of the plant immune response. *Cell* **124**, 803–814 (2006).
- Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protoc.* **6**, 5448 (2010).
- Sahoo, N. *Sequence Base-calling through Albacore software: A part of the Oxford Nanopore Technology* (Doctoral dissertation) (2017).
- Andrews, S. FastQC: A quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
- Krueger, F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2015).
- Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
- Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-seq based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
- The UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* **36**, D190–D195 (2008).
- Buchfink, B., Xie, C. & Huson, D. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

44. Simao, F. A., Waterhouse, R. M., Lonnidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
45. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
46. Wang, Y., Coleman-Derr, D., Chen, G. & Gu, Y. Q. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **43**(W1), W78–W84 (2015).
47. Thomas, W., Matthews, D. E. & Keller, B. TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.* **7**, P561–562 (2002).
48. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
49. Sebastian, B., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
50. Martins, W. S., Lucas, D. C., Neves, K. F. & Bertioli, D. J. WebSat: A web software for microsatellite marker development. *Bioinformatics* **3**, 282–283 (2009).
51. Frank, M. Y. *et al.* Batch Primer3: A high throughput web application for PCR and sequencing primer designing. *BMC Bioinform.* **9**, 253. <https://doi.org/10.1186/1471-2105-9-253> (2008).
52. Sanseverino, W. *et al.* PRGdb: A bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res.* **38**, D814–D821. <https://doi.org/10.1093/nar/gkp978> (2010).
53. Osuna-Cruz, C. M. *et al.* PRGdb 3.0: A comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* **46**, D1197–D1201 (2018).

Acknowledgements

Authors thank Dr. P. Venugopalan, Associate Director, Biosciences Group, Bhabha Atomic Research Centre, Trombay, Mumbai, for his kind support and encouragement for execution of the project.

Author contributions

J.S. conceived the idea, coordinated the sequencing and wrote the manuscript. A.R. contributed to R-gene analysis. P.D. contributed to SSR analysis and primer designing. P.S. supervised the study. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-90683-9>.

Correspondence and requests for materials should be addressed to S.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021