High-quality genome assembly, annotation and evolutionary analysis of the mungbean (*Vigna radiata*) genome

Qiang Yan[1], Qiong Wang[1], Cheng Xuzhen[2], Lixia Wang[2], Prakit Somta[3], Chenchen Xue[1], Jingbin Chen[1], Ranran Wu[1], Yun Lin[1], Xingxing Yuan[1], and Xin Chen[1]

[1]Jiangsu Academy of Agricultural Sciences
[2]Chinese Academy of Agricultural Sciences
[3]Kasetsart University Kamphaeng Saen Campus

November 20, 2020

Abstract

Mungbean (*Vigna radiata* [L.]) is an important economic crop grown in South, and East Asia. The low contiguity of the current assembly of *V. radiata* genome has limited its application. Here, we report a high-quality chromosome-scale assembled genome of *V. radiata* to facilitate the investigation of its genome characteristics and evolution. By combination of Nanopore long reads, Illumina short reads and Hi-C data, we generated a high-quality genome assembly of *V. radiata*, with 473.67 megabases assembled into 11 chromosomes with contig N50 and scaffold N50 of 11.3 and 42.4 megabases, respectively. A total of 52.8% of the genome was annotated as repetitive sequences, among which LTRs (long terminal repeats) were predominant (33.9%). The genome of *V. radiata* was predicted to contain 33,924 genes, 32,470 (95.7%) of which could be functionally annotated. Evolutionary analysis revealed an estimated divergence time of *V. radiata* from its close relative V. angularis of ˜11.66 million years ago. In addition, 277 *V. radiata* specific gene families, 18 positively selected genes were detected and functionally annotated. This high-quality mungbean genome will provide valuable resources for further genetic analysis and crop improvement of mungbean and other legume species.

High-quality genome assembly, annotation and evolutionary analysis of the mungbean (*Vigna radiata)* genome

Qiang Yan[1#], Qiong Wang[1#], Xuzhen Cheng[3], Lixia Wang[3], Prakit Somta[2], Chenchen Xue[1], Jingbin Chen[1], Ranran Wu[1], Yun Lin[1], Xingxing Yuan[1*], Xin Chen[1,4*]

[1]Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China

[2]Department of Agronomy, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University, Nakhon Pathom, Thailand

[3]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[4]School of Food and Biological Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang, Jiangsu 212013, China

[#]These authors contributed equally to this work

*Corresponding authors

Xingxing Yuan (yxx@jaas.ac.cn)

Xin Chen (cx@jaas.ac.cn)

Abstract

Mungbean (*Vigna radiata* [L.]) is an important economic crop grown in South and East Asia. The low contiguity of the current assembly of *V. radiata* genome has limited its application. Here, we report a high-quality chromosome-scale assembled genome of *V. radiata* to facilitate the investigation of its genome characteristics and evolution. By combination of Nanopore long reads, Illumina short reads and Hi-C data, we generated a high-quality genome assembly of *V. radiata* , with 473.67 megabases assembled into 11 chromosomes with contig N50 and scaffold N50 of 11.3 and 42.4 megabases, respectively. A total of 52.8% of the genome was annotated as repetitive sequences, among which LTRs (long terminal repeats) were predominant (33.9%). The genome of *V. radiata* was predicted to contain 33,924 genes, 32,470 (95.7%) of which could be functionally annotated. Evolutionary analysis revealed an estimated divergence time of *V. radiata* from its close relative *V. angularis* of ˜11.66 million years ago. In addition, 277 *V. radiata* specific gene families, 18 positively selected genes were detected and functionally annotated. This high-quality mungbean genome will provide valuable resources for further genetic analysis and crop improvement of mungbean and other legume species.

Running title: High-quality mungbean genome assembly

Keywords: Mungbean (*Vigna radiata*), Hi-C, genome assembly, genome annotation

1 | INTRODUCTION

Mungbean (*Vigna radiata* [L.]) is an important legume crop which widely cultivated in South and East Asia countries including India, Myanmar, China, Thailand, Bangladesh, Pakistan and Indonesia (Breria et al. 2020a; Keatinge et al. 2011). It is also widely planted in Tanzania and Kenya in recent years, but the average yield is still low (Nair & Schreinemachers 2020). Mungbean is also act as an important rotation crop due to short-duration (maturing in 60 to 75 days), drought tolerance, and ability to fix nitrogen as other legume crops. Mungbean seeds contain relatively high proportion of easily digestible proteins (24%), act as important sources of hunman dietary proteins and carbohydrates, while its sprouts are popular and inexpensive vegetable rich with vitamin C and folate (Keatinge et al. 2011).

The draft genome sequence of mungbean was assembly of a widely grown cultivar VC1973A (Kang et al. 2014). This initial reference enabled rapid progress in genetic and genomic researches with the aim to understand leaf development (Jiao et al. 2016), Powdery Mildew Resistance (Yundaeng et al. 2020), bruchid resistance (Chotechung et al. 2016; Kaewwongwal et al. 2017), salinity tolerance (Breria et al. 2020b), genomic diversity and Genome-Wide Association Studies (GWAS) investigated seed coat luster (Breria et al. 2020a). The assembly based on Illumina short-reads technology and consists of 2,748 scaffolds with a N50 of 1.52 Mb, there still about 130 Mb of unmapped scaffolds (Kang & Ha 2020). The low contiguity of the current assembly has limited its application for further fine mapping and candidate genes clone. Thus, a high-quality re-sequencing genome assembly of mungbean is needed.

In this study, we constructed a highly accurate, contiguous, chromosome-scale de novo assembly of the mungbean genome obtained by integrating short-read sequencing, Oxford Nanopore sequencing based gap closure, scaffolding, and orientation based on 3D proximity information derived from chromosome conformation capture (Hi-C) data. The contiguity of the newly assembled genome was 27.86-fold greater than that of the published draft genomes of *Vigna radiata* (scaffold N50 = 42.35 Mb versus 1.52 Mb). The total size of connective N sequences in the oriented genome assembly was dramatically reduced to 7.2 Kb. The genomic data will provide valuable resources for genetic study of mungbean.

2 | MATERIALS AND METHODS

2.1 | Plant materials

The pure line of an elite cultivar, Sulv 1 was chosen for genome sequencing. The seedings were grown in a greenhouse under 25 °C and 16 h photoperiod condition. The fresh leaves were collected from one-month old plants, the samples were rapidly frozen in liquid nitrogen and stored at -70°C until use.

2.2 | Genome size estimation

The Sulv 1 genome size was calculated with the following formula: genome size = total k-mer number/average k-mer depth, and total k-mer number is the total number of k-mers from all reads. 350 bp insert size clean reads were used to perform the k-mer (k = 19) analysis. A total of 54,027,628,001 k-mers were counted from these clean reads. A k-mer depth distribution was obtained from paired end reads, and the peak depth was clearly observed from the distribution data. Based on this distribution, the size of the Sulv 1 genome was estimated to be ˜ 539.8 Mb and the heterozygosity was estimated to be 0.03% (Figure S1).

2.3 | Genome sequencing and assembly

Total DNA was isolated by using the CTAB method to construct Nanopore and Illumina libraries. Libraries were generated and sequenced on the PromethION sequencer platform (Oxford Nanopore Technologies, UK) at the Biomarker Technologies Corporation (Beijing, China).

We first corrected the errors in the Nanopore sequencing reads with the help of Canu (Koren et al. 2017) software. Based on this corrected data, Sulv 1 genome was assembled using wtdbg2 (https://github.com/ruanjue/wtdbg2) software platform.Wtdbg2 chops reads into 1024 bp segments, merges similar segments into a vertex and connects vertices based on the segment adjacency on reads. The resulting graph is called fuzzy Bruijn graph (FBG) which is similar to De Bruijn graph but permits mismatches/gaps and keeps read paths when collapsing k-mers. The draft genome was first calibrated using Racon (Vaser et al. 2017) with Nanopore reads through 3 rounds of calibration, and Pilon (v1.21, Bruce et al. 2014) was then used to calibrate the draft genome again with the help of short Illumina HiSeq reads in a 3 rounds of calibration process too.

2.4 | Hi-C sequencing and chromosomes anchoring

We constructed Hi-C fragmented libraries (300-700 bp insert size) as illustrated in Rao et al (Rao et al. 2014) and libraries were sequenced through Illumina sequencing platform. Briefly, adapter sequences of raw Hi-C reads were trimmed and low quality paired end reads were removed in order to obtain clean data. The clean Hi-C reads were first truncated at the putative Hi-C junctions and then the resulting trimmed reads were aligned to the draft assembly with the help of bwa aligner (Li et al. 2013). Only uniquely aligned read pairs whose mapping quality was greater than 20 were retained for further analysis. Invalid read pairs, including Dangling-End and Self-cycle, Re-ligation and Dumped products, were filtered by HiC-Pro (v2.8.1, Servant et al. 2015).

The uniquely mapped read pairs were valid interaction pairs and were used for the correction of scaffolds and to order and orientate scaffolds onto chromosomes by LACHESIS (Burton et al. 2013).

Before chromosomes assembly, we first performed a pre-assembly for the error correction of scaffolds which required the splitting of scaffolds into segments of 50 kb on average. The Hi-C data were mapped to these segments using BWA (version 0.7.10-r789, Li et al. 2009) software. The uniquely mapped data were retained to perform assembly by using LACHESIS software. Any two segments which showed inconsistent connection with information from the raw scaffolds were checked manually. These corrected scaffolds were then assembled with LACHESIS. Parameters for running LACHESIS software included: CLUSTER_MIN_RE_SITES=186, CLUSTER_MAX_LINK_DENSITY=2, CLUSTER_NONINFORMATIVE_RATIO=1.3, ORDER_MIN_N_RES_IN_TRUN=98, ORDER_MIN_N_RES_IN_SHREDS=100. After this step, placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted. Finally, 470.45 Mb of the sequences (representing 99.3% total length) were anchored to 11 chromosomes.

2.5 | Gene and repetitive sequence annotation

Repeats were masked on the assembled Sulv 1 genome using de novo-based and homology-based strategies. We used RepeatMasker (Tarailo-Graovac et al. 2009) for de novo repeat prediction based on a custom library produced by RepeatModeler. Repbase (Jurka et al. 2005) was downloaded from

http://www.girinst.org/repbase/ and was used for homology-based repeat detection. Repbase and the de novo repeat library were merged together to annotate the repetitive elements in the assembled Sulv 1 genome by using RepeatMasker. Parameters for running RepeatMasker were: "-nolow -no_is -norna -engine wublast". And default parameters were used for LTR_FINDER (Xu et al. 2007), RepeatScout (Price et al. 2005) and PASTEClassifier (Hoede et al. 2014) softwares.

Protein-coding genes prediction of the assembled Sulv 1 genome was conducted using three different strategies: *ab initio* prediction, predictions based on homologous species, and based on unigenes. EVM (v1.1.1, Haas et al. 2008) software was used to integrate these three prediction results using default parameters. Softwares used for *ab initio* prediction were Genscan (Burge et al. 1997), Augustus (v2.4, Stanke et al. 2003), GlimmerHMM (v3.0.4, Majoros et al. 2004), GeneID (v1.4, Blanco et al. 2007), SNAP (version 2006-07-28, Korf et al. 2004), and default parameters were used. GeMoMa (v1.3.1, Keilwagen et al. 2016, Keilwagen et al. 2018, with default parameters) was used for homology based prediction with protein sequences from homologous species including *Arabidopsis thaliana* , *Vigna radiata* , *Vigna unguiculata* and *Glycine max* . For unigene based prediction, PASA (v2.0.2, Campbell et al. 2006) software was used on the basis of assembled RNA-seq unigenes with the following parameters: "-align_tools gmap -maxIntronLen 20000". Specifically, Hisat (v2.0.4, Kim et al. 2015) and Stringtie (v1.2.3, Pertea et al. 2015) were used for the assembly of transcripts; TransDecoder (v2.0, available online: https://transdecoder.github.io/) and GeneMarkS-T (v5.1, Tang et al. 2015) were used for gene prediction. Parameters used for Hisat software were: "–max-intronlen 20000 –min-intronlen 20" and default parameters were used for Stringtie, TransDecoder and GeneMarkS-T.

For the annotation of noncoding RNAs, Blastn was used for genome-wide comparison to identify microRNAs and rRNAs based on the Rfam (Griffiths-Jones et al. 2005) database, and tRNAs were identified using tRNAscan-SE (Lowe et al. 1997) sofaware.

Using the predicted protein sequences, BLAT (Kent et al. 2002) alignment was conducted to find homologous gene sequences (possible genes) in the assembled Sulv 1 genome, and then GeneWise (Birney et al. 2004) was used to detect immature stop codons and frameshift mutations in the gene sequences to identify pseudogenes with default parameters. E-value cutoff for GenBlastA (She et al. 2009) was set to 1e-5.

The sequences of the predicted protein-coding genes were searched against commonly used Nr (Marchler-Bauer et al. 2011), KOG (Koonin et al. 2004), GO (Dimmer et al. 2012), KEGG (Kanehisa et al. 2000) and TrEMBL (Boeckmann et al. 2003) databases for gene function annotation with BLAST software (v2.2.31, Altschul et al. 1990, e-value cutoff 1e-5). Motif annotation was performed through comparison against PROSITE (Bairoch et al. 1991), HAMAP (Lima et al. 2009), Pfam (Finn et al. 2006), PRINTS (Attwood et al. 1994), ProDom (Bru et al. 2005), SMART (Letunic et al. 2004), TIGRFAMs (Haft et al. 2003), PIRSF (Wu et al. 2004), SUPERFAMILY (Gough et al. 2002), CATH-Gene3D (Lees et al. 2012) and PANTHER (Thomas et al. 2003) databases using InterProScan (Zdobnov et al. 2001) software.

2.6 | Gene family analysis

The protein sequences of Sulv 1 and 10 related species including*Vigna radiata* , cowpea, common bean, soybean, *Vigna angularis* , *Lablab purpureus* , *Medicago truncatula* ,*Lotus japonicus* , *Vigna subterranea* and *Arabidopsis thaliana* (downloaded from NCBI database) were used for gene family clustering through OrthoMCL (Li et al. 2003) software with default parameter settings.

2.7 | Phylogenetic tree reconstruction and divergence time prediction

Single-copy orthologues identified from the analyzed genomes were used for subsequent phylogenetic tree reconstruction and divergence time evaluation. Multiple sequence alignment was performed using MUSCLE (Edgar et al. 2004), and then a phylogenetic tree was constructed using PHYML (Guindon et al. 2010) software based on the alignment. MCMCTREE implemented in the PAML package (v4.7b, Yang et al. 1997) was used to estimate the speciation time.

2.8 | Expansion and contraction of gene family

We used CAFE (v 2.0, De et al. 2006) to infer gene family sizes of the most recent common ancestor (MRCA) and to analyze expansion and contraction of gene family based on the phylogenetic tree.

2.9 | Positively selected gene analysis

Codeml (Schabauer et al. 2012) software implemented in the PAML program package was used to identify positively selected genes in the assembled Sulv 1 genome with a branch model (model=2, NSsites=0). The positively selected genes were annotated by GO and KEGG analyses.

2.10 | Estimation of LTR insertion time

We used LTR_FINDER software accompanied by PS SCAN (Prestridge et al. 1991) software to identify LTR sequences whose score was greater than or equal to 6 in the assembled Sulv 1 genome, and duplicate results were filtered. Then the flanking sequences on both sides of the LTR were extracted. After aligned with MUSCLE, DistMat was used to calculate the distance based on Kimura model with the molecular clock selected as 7.3*10-9.

3 | RESULTS AND DISCUSSION

3.1 | Genome sequencing

A single plant of *Vigna radiata* var. Sulv 1 was used for genome sequencing. To achieve a high-quality *Vigna radiata* var. Sulv 1 genome assembly, we used a combination of sequencing methods including Oxford Nanopore sequencing Technology (ONT), Illumina sequencing and Hi-C mapping. A total of ˜ 122.9 Gb sequencing data (equivalent to 259.5 X genomic coverage) was generated.

3.2 | Genome assembly

For the Nanopore sequencing data, we first corrected the errors in the sequencing reads with the help of Canu software, then genome assembly was conducted using wtdbg2 software platform based on the corrected data. This draft genome assembly was first calibrated through Racon with the help of Nanopore reads by 3 rounds of calibration, and we then used Pilon (v1.21) software to calibrate the draft genome again with the help of Illumina HiSeq short reads in a 3 rounds of calibration process. The resulting Nanopore genome assembly was 473.67 Mb in length, composed of 359 contigs, and the contig N50 was 11.32 Mb. The Nanopore assembly results were summarized in table S1.

Raw Hi-C sequencing data were first filtered to remove adapter sequences and low quality reads to obtain high quality clean data. BWA aln was used to map the Hi-C clean data reads against the draft genome with default parameters. The reads that can be aligned to the assembled genome are mapped reads. There are 111.97 million unique mapped read pairs, accounting for 60.45% of the total reads. These unique mapped read pairs were used to identify the valid interaction pairs and the invalid interaction pairs mapped to the draft genome by HiC-Pro. The preliminary assembled draft genome sequence was then further assembled using valid Hi-C data through LACHESIS, including the grouping, sorting and orientation of the draft genome sequence, and finally the genome sequence of Sulv 1 at the chromosome level is obtained. After Hi-C assembly and manual adjustment, the final assembly of mungbean genome consists of 470.45 Mb assigned into 11 individual chromosomes that accounted for 99.32% of the genome (Figure 1), and is highly contiguous with scaffold N50 at 42.35 Mb and contig N50 at 11.32 Mb (Table 1 and table S2).

3.3 | Evaluation of assembly

The Nanopore assembly result was evaluated from the following three aspects. First, to assess the assembly integrity and genome coverage, we used bwa software to map the short sequence reads obtained from the Illumina HiSeq sequencing platform to the reference genome. The percent of reads mapped to the reference genome was 99.07%. Then CEGMA (v2.5, Parra et al. 2007) software was used to assess the integrity of the genome assembly. 449 (98.03%) of the 458 conserved core genes for eukaryotes were present in the assembled genome. Furthermore, the completeness of our assembled genome was assessed through BUSCO (Felipe et al. 2015) analysis using generic model. Approximately 92.43% of the plant orthologs were included in

the assembled Sulv 1 mungbean genome sequences (table S3). These results indicated a high accuracy and integrity of the mungbean genome assembly.

For the Hi-C data assembled to the chromosome, the genome sequences were cut into 100 Kb bins with equal length, and then the number of Hi-C read pairs between any two bins is used as the intensity of the interaction between the two bins. Within each chromosome group, the intensity of interaction at the diagonal position is higher than the off-diagonal position, indicating efficient chromosome assembly of Hi-C data (figure S2).

3.4 | Protein-coding gene prediction

Protein-coding region identification and gene prediction were conducted by a combination of *ab initio* , homology-based, and unigene-based prediction methods, and aided by the software EVM for the integration of the prediction results. We used Genscan, Augustus (v2.4), GlimmerHMM (v3.0.4), GeneID (v1.4) and SNAP (version 2006-07-28) for *ab initio* gene prediction. The homology-based prediction was conducted with GeMoMa (v1.3.1) software. For the unigene-based prediction, TransDecoder (v2.0) and GeneMarkS-T (v5.1) were used to predict coding genes after reference genome based mapping using Hisat (v2.0.4) and Stringtie (v1.2.3), and PASA (v2.0.2) was used for the prediction of coding genes after de novo transcriptomes assembly. Finally, we used EVM (v1.1.1) to integrate the prediction results obtained by all the above three methods, and after modified with PASA (v2.0.2), a total of 33,924 protein-coding regions were constructed (table S4 and table S5). Among these predicted coding genes, 20,446 were constructed by all three methods. 6,222 genes can be predicted by both *ab initio* and homology-based methods but can't be constructed by unigene-based method. In addition, 1,291, 5,248 and 7 coding genes were found to be specific to *ab initio* , homology-based, and unigene-based prediction methods respectively (figure S3).

3.5 | Gene function annotation

We assigned the functions of predicted protein-coding genes through BLAST (v2.2.31) against NR, KOG, GO, KEGG and TrEMBL database, performed KEGG pathway gene annotation analysis, KOG functional annotation analysis and GO gene function annotation analysis (figure S4 and figure S5). A total of 32,470 genes can be annotated, accounting for 95.71% of all predicted genes (table S6). Among them, about 56.6% predicted genes have GO annotations. GO enrichment analysis of gene sets was performed in Blast2GO against Sulv 1 mungbean genome as reference. Statistical significance was tested by Fisher's exact test corrected in multiple tests using Bonferroni method under false discovery rate (FDR) threshold of 0.05.

Motifs are short conservation sequences that homologous to regions in other sequences and perform a similar function. We annotated motifs of the *Vigna radiata* genome using InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) with default parameters. InterProScan software combines several protein motifs/domains search tools together. It allows users to scan protein sequences at one time against several signature databases including Prosite, PRINTS, PFAM, ProDom, Smart, TIGRFAMs, SignlP, Trans memberane etc., and also gives GO annotation. Analysis of protein domains by InterProScan software and motif searching identified 2,765 motifs and 35,154 domains in the *Vigna radiata* var. Sulv 1 genome.

3.6 | Non-coding RNA annotation

Non-coding RNA includes RNA with a variety of known functions such as microRNA, rRNA, and tRNA. Different strategies were used to predict non-coding RNAs according to the structural characteristics of different kinds of non-coding RNAs. To identify microRNA and rRNA, we used blastn to perform a genome-wide comparison based on the Rfam database, and tRNA was identified using tRNAscan-SE software. We finally identified 86 miRNA, 352 rRNA and 653 tRNA belonging to 23, 4 and 22 families respectively (table S7).

3.7 | Pseudogenes prediction

Pseudogenes have sequences similar to functional genes, but they have lost their original functions due to

mutations such as insertions and deletions. We searched for possible homologous gene sequences in the genome with the help of the predicted protein sequences through BLAT alignment, and then GeneWise was used to search for immature stop codons and frameshift mutations in the gene sequences to obtain pseudogenes. A total of 4,320 pseudogenes were identified, with an average length of 2,237 bp.

3.8 | Annotation of repetitive elements

Due to the relatively low conservation of repetitive sequences between species, it is necessary to construct a specific repetitive sequence database when predicting repetitive sequences for *Vigna radiata* . We used Repbase and a constructed de novo repeat library to annotate repeat DNA sequences in the *Vigna radiata* genome. A de novo repeat library from the assembled *Vigna radiata* genome was constructed using LTR_FINDER and RepeatModeler (version open-1.0.8, http://repeatmasker.org/RepeatModeler/) and Repbase was downloaded from http://www.girinst.org/repbase/. The database was classified through PASTEClassifier, and then merged with the Repbase database as the final repeated sequence database. The repetitive elements in the*Vigna radiata* de novo repeat library and Repbase database were annotated by RepeatMasker. About 52.83% of the *Vigna radiata*genome was annotated as repetitive sequences based on RepeatMasker output (table S8). The length of the repetitive element type ranged from 46.4 Kb to 215.1 Mb. The most abundant repetitive element repeat type is long terminal repeat (LTR), making up 33.92% of the genome, including 56.6% Gypsy LTRs, 39.77% Copia LTRs and 3.63% other types of LTRs.

Simple sequence repeats (SSRs) are another type of important tandem repetitive sequences. We used MISA software to detect SSRs in the mungbean genome. A total of 224,409 SSRs (136,045 mono-, 56,033 di-, 28,959 tri-, 1,977 tetra-, 1,098 penta-, and 297 hexa-nucleotide repeats) were detected (table S9). The total length of the SSR sequences was 3,252,656 bp, accounting for ˜0.69% of the assembled Sulv 1 mungbean genome.

3.9 | Phylogenetic analysis and estimation of divergence time

The assembled and annotated mungbean genome allowed us to investigate its evolutionary history. Single-copy orthologs among taxa were used to achieve robust phylogenetic reconstruction with high confidence and concordance. We identified a set of single-copy orthologs from mungbean and 10 closely related species including *Vigna radiata* , cowpea, common bean, soybean, *Vigna angularis* , *Lablab purpureus* ,*Medicago truncatula* , *Lotus japonicus* , *Vigna subterranea* and *Arabidopsis thaliana* using OrthoMCL software (table S10). Based on this ortholog set, a phylogenetic tree of the eleven plant species was constructed as follows: for each single-copy gene, a coding sequence alignment was created using MUSCLE and then all coding sequence alignments were concatenated in MEGA. The concatenated alignment was then used to construct a maximum likelihood phylogenetic tree using PHYML. Species divergence time was then estimated by using the maximum likelihood tree as a starting tree through Mcmctree (Figure 2). We used a fossil calibration with a strict clock rate for the divergence time estimation.

3.10 | Whole genome duplication in mungbean genome

To investigate the evolution of mungbean, we compared its genome with four other eudicots: *Vigna radiata* , *Arabidopsis thaliana*(Arabidopsis), *Vigna unguiculata* and *Phaseolus vulgaris* . The orthologs between mungbean and these species were identified using analysis described above. We searched for genome wide duplications in assembled mungbean genome to study mungbean genome evolution. 4DTv (4-fold degenerate synonymous sites of the third codons) values were calculated based on the homologous gene pairs between two species or within the species itself. The analysis revealed whole-genome replication events in mungbean genome. A divergence peak was observed for *Vigna radiata* vs *Arabidopsis thaliana* , and another lower peak was found for *Vigna radiata* vs common bean (Figure 3), which suggested that the divergence of mungbean and*Arabidopsis thaliana* occurred earlier than the divergence of mungbean and common bean.

3.11 | Estimation of LTR insertion time

LTRs were identified in the Sulv 1 genome using LTR_FINDER software. Mutation rates were used to estimate LTR insertion times. The results indicated that LTR insertions are not very active in Sulv 1 (Figure 4).

3.12 | Genes underwent positively seletion

To detect positively selected genes in Sulv 1 genome, we evaluated the Ka/Ks ratios of single copy genes by using branch model. In total, we detected 18 genes that probably have experienced positive selection (table S11). GO enrichment revealed that a majority of these genes were involved in membrane-enclosed lumen and cell junction.

4 | CONCLUSIONS

In this study, we combined sequencing technologies including Oxford Nanopore (142.4X), Illumina sequencing and Hi-C mapping to upgrade the *Vigna radiata* genome assembly. We present a high-quality genome assembly and gene annotation of *Vigna radiata* var. Sulv 1 with 33,924 protein-coding genes. The final genome assembly of 473.67Mb covers 87.8 % of the estimated genome size, and 99.32% of sequences have been assigned into 11 individual chromosomes. This work provides valuable chromosome-level genomic data for mungbean. The identified genomic features of Sulv 1, including gene families, WGD events, and genome-specific genes, provide rich data for comparative genomic studies in legume plants.

COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENTS

This study was supported by the National Key R&D Program of China (grant number: 2019YFD1001301/ 2019YFD1001300) and China Agriculture Research System (grant number: CARS08-G15)

DATA AVAILABILITY

All the data (Illumina, Nanopore, Hi-C and RNAseq) support the findings of this study are openly available in the NCBI SRA (Sequence Read Archive) database under the Bioproject ID: PRJNA660308.

AUTHOR CONTRIBUTIONS

C.X., Y.X.X., Y.Q. and C.X.Z. designed the study;

X.C.C. and C.J.B. collected the samples;

W.Q., S.P. and W.R.R. analyzed the results;

W.Q. and Y.Q. wrote the manuscript;

W.L.X., S.P. and L.Y. revised the manuscript;

all authors read and approved the final version of the manuscript.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* , 215, 403-410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Attwood, T., & Beck, M. (1994). Prints–a protein motif fingerprint database. *Protein Engineering* , 7, 841-848. http://dx.doi.org/10.1093/protein/7.7.841

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Research* , 19. 2241. http://dx.doi.org/10.1093/nar/19.suppl.2241

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. *Genome Research* , 14, 988-995. http://dx.doi.org/10.1101/gr.1865504

Blanco, E., Parra, G., & Guigo., R. (2007). Using geneid to identify genes. *Current Protocols in Bioinformatics* , 4.3.1-4.3.28. http://dx.doi.org/10.1002/0471250953.bi0403s18

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M-C., Estreicher, A., Gasteiger, E., ... Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.*Nucleic Acids Research* , 31, 365-370. http://dx.doi.org/10.1093/nar/gkg095

Breria, C. M., Hsieh, C. H., Yen, J. Y., Nair, R., Lin, C. Y., Huang, S. M., ... Schafleitner, R. (2020a). Population structure of the world vegetable center mungbean mini core collection and genome-wide association mapping of loci associated with variation of seed coat luster. *Tropical Plant Biology* , 13(1), 1-12. doi:10.1007/s12042-019-09236-0

Breria, C. M., Hsieh, C. H., Yen, T. B., Yen, J. Y., Noble, T. J., & Schafleitner, R. (2020b). A SNP-based genome-wide association study to mine genetic loci associated to salinity tolerance in mungbean (*Vigna radiata* L.). *Genes* (Basel), 11(7). doi:10.3390/genes11070759

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005) The ProDom database of protein domain families: more emphasis on 3D.*Nucleic Acids Research* , 33, 212-215. doi:10.1093/nar/gki034

Bruce, J., Walker, T. A., Terrance, S., Margaret, P., Amr, A., Sharadha, S., ... Ashlee, M. E. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* , 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* , 268, 78-94. https://doi.org/10.1006/jmbi.1997.0951

Joshua, N. B., Andrew, A., Rupali, P. P., Ruolan, Q., Jacob, O. K., & Jay, S. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Naturenature Biotechnology* , 31(12), 1119-1125. https://doi.org/10.1038/nbt.2727

Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M., & Buell, C. R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with*Arabidopsis* . *BMC Genomics* , 7, 327. http://dx.doi.org/10.1186/1471-2164-7-327

Chotechung, S., Somta, P., Chen, J., Yimram, T., Chen, X., & Srinives, P. (2016). A gene encoding a polygalacturonase-inhibiting protein (PGIP) is a candidate gene for bruchid (Coleoptera: bruchidae) resistance in mungbean (*Vigna radiata*). *Theor Appl Genet* , 129(9), 1673-1683. doi:10.1007/s00122-016-2731-1

De, B. T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution.*Bioinformatics* , 22, 1269-1271. doi: 10.1093/bioinformatics/btl097

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., ... Eberhardt, R. (2012). The UniProt-GO annotation database in 2011. *Nucleic Acids Research* , 40, 565-570. http://dx.doi.org/10.1093/nar/gkr1048

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput.*Nucleic Acids Research* , 32, 1792-1797. http://dx.doi.org/10.1093/nar/gkh340

Felipe, A. S., Robert, M. W., Panagiotis, I., Evgenia, V. K., & Evgeny, M. Z. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* , 31(19), 3210-3212. DOI: 10.1007/978-1-4939-9173-0_14

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., ... Durbin, R. (2006). Pfam: clans, web tools and services. *Nucleic Acids Research,* 34, 247-251. http://dx.doi.org/10.1093/nar/gkj149

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.*Bioinformatics* , 23, 1061-1067. http://dx.doi.org/10.1093/bioinformatics/btm071

Gough, J., & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Research* , 30, 268-272. http://dx.doi.org/10.1093/nar/30.1.268

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., & Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* , 33, 121-124. http://dx.doi.org/10.1093/nar/gki081

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* , 59, 307-321. http://dx.doi.org/10.1093/sysbio/syq010

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., ... Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* , 9, R7. http://dx.doi.org/10.1186/gb-2008-9-1-r7

Haft, D. H., Selengut, J. D., & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research* , 31, 371-373. http://dx.doi.org/10.1093/nar/gkg128

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., & Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. *PLoS One* , 9(5), e91929. https://doi.org/10.1371/journal.pone.0091929

Jiao, K. Y., Li, X., Guo, W. X., Yuan, X. X., Cui, X. Y., & Chen, X. (2016). Genome re-sequencing of two accessions and fine mapping the locus of lobed leaflet margins in mungbean. *Molecular Breeding* , 36(9). doi:ARTN 128 10.1007/s11032-016-0552-1

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* , 110, 462-467. http://dx.doi.org/10.1159/000084979

Kaewwongwal, A., Chen, J., Somta, P., Kongjaimun, A., Yimram, T., Chen, X., & Srinives, P. (2017). Novel alleles of two tightly linked genes encoding polygalacturonase-inhibiting proteins (VrPGIP1 and VrPGIP2) associated with the br locus that confer bruchid (*Callosobruchus* spp.) resistance to mungbean (*Vigna radiata*) accession V2709.*Frontiers in Plant Science* , 8, 1692. doi:10.3389/fpls.2017.01692

Keatinge, J. D. H., Easdown, W. J., Yang, R. Y., Chadha, M. L., & Shanmugasundaram, S. (2011). Overcoming chronic malnutrition in a future warming world: the key importance of mungbean and vegetable soybean. *Euphytica* , 180(1), 129-141. doi:10.1007/s10681-011-0401-6

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics* , 19(1), 189. http://dx.doi.org/10.1186/s12859-018-2203-5

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* , 28, 27-30. http://dx.doi.org/10.1093/nar/28.1.27

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B. K., ... Lee, S. H. (2014). Genome sequence of mungbean and insights into evolution within *Vigna*species. *Nature Communications* , 5, 5443. http://dx.doi.org/10.1038/ncomms6443

Kang. Y., J., & Ha, J. (2020) Mungbean genome and synteny with other genomes. In: Nair R, Schafleitner R, Lee SH, editors. *The mungbean genome. Compendium of plant genomes* . Berlin: Springer; 2020.125-127. https://doi.org/10.1007/978-3-030-20008-4

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Jan, G., & Frank, H. (2016). Using intron position conservation for homology-based gene prediction.*Nucleic Acids Research* , 44, e89. http://dx.doi.org/10.1093/nar/gkw092

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Research* , 12(4), 656-664. http://dx.doi.org/10.1101/gr.229202

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* , 12, 357-360. http://dx.doi.org/10.1038/nmeth.3317

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., ... Natale, D. A. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* , 5, R7. https://doi.org/10.1186/gb-2004-5-2-r7

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* , 27(5), 722-736. http://dx.doi.org/10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* , 5, 59. DOI: 10.1186/1471-2105-5-59

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., & Orengo, C. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Research* , 40, 465-471. http://dx.doi.org/10.1093/nar/gkr1181

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., ... Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Research* , 32, 142-144. http://dx.doi.org/10.1093/nar/gkh088

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv e-prints.

Li, H., & Richard, D. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform.*Bioinformatics* , 25(14), 1754-1760. http://dx.doi.org/10.1093/bioinformatics/btp324

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* , 13, 2178-2189. http://dx.doi.org/10.1101/gr.1224503

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., ... Baratin, D. (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot.*Nucleic Acids Research* , 37, 471-478. http://dx.doi.org/10.1093/nar/gkn661

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* , 25, 0955-0964. http://dx.doi.org/10.1093/nar/25.5.955

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.*Bioinformatics* , 20, 2878-2879. doi: 10.1093/bioinformatics/bth315

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., ... Gonzales, N. R. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins.*Nucleic Acids Research* , 39, 225-229. http://dx.doi.org/10.1093/nar/gkq1189

Nair, R., & Schreinemachers, P. (2020). Global status and economic importance of mungbean. In: Nair R, Schafleitner R, Lee SH, editors. *The mungbean genome. Compendium of plant genomes* . Berlin: Springer; 2020.1-6. https://doi.org/10.1007/978-3-030-20008-4

Pertea, M., Pertea, G. M., Antonescu, Corina, M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.*Naturenature Biotechnology* , 33, 290-295. http://dx.doi.org/10.1038/nbt.3122

Prestridge, D. S. (1991). Signal scan: a computer program that scans DNA sequences for eukaryotic transcriptional elements. *Computer Applications in the Bioences Cabios* , 7(2), 203. http://dx.doi.org/10.1093/bioinformatics/7.2.203

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes.*Bioinformatics* , 21, 351-358. http://dx.doi.org/10.1093/bioinformatics/bti1018

Rao, S., Huntley, M., Durand, N., Stamenova, E., Bochkov, I., Robinson, J., . . . Aiden, E. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* , 159(7), 1665-1680. http://dx.doi.org/10.1016/j.cell.2014.11.021

Schabauer, H., Valle, M., Pacher, C., Stockinger, H., Stamatakis, A., Robinson-Rechavi, M., ... Salamin, N. (2012). SlimCodeML: an optimized version of CodeML for the Branch-Site Model. *Paper presented at the 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum* , 706-714

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., ... Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* , 16, 259. http://dx.doi.org/10.1186/s13059-015-0831-x

She, R., Chu, J. S., Wang, K., Pei, J., & Chen, N. (2009). GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Research* , 19(1), 143-149. http://genome.cshlp.org/cgi/doi/10.1101/gr.082081.108

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* , 19, 215-225. https://doi.org/10.1093/bioinformatics/btg1080

Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* , 43(12), e78. https://doi.org/10.1093/nar/gkv227

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* , 25(1), 4.10.11-14.10.14. https://doi.org/10.1002/0471250953.bi0410s25

Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., ... Doremieux, O. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Research* , 31(7), 2024-2024. https://doi.org/10.1093/nar/gkg115

Vaser, R., Sovic, I., Nagarajan, N., & Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* , 27(5), 737-746. doi:10.1101/gr.214270.116

Wu, C. H., Nikolskaya, A., Huang, H. Z., Yeh, L. S. L., Natale, D. A., Vinayaka, C. R., ... Barker, W. C. (2004). PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research* , 32, D112-D114. DOI:10.1093/nar/gkh097

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* , 35, W265-W268. https://doi.org/10.1093/nar/gkm286

Yang, Z. H. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* , 13(5), 555-556. https://doi.org/10.1093/bioinformatics/13.5.555

Yundaeng, C., Somta, P., Chen, J., Yuan, X., Chankaew, S., Srinives, P., & Chen, X. (2020). Candidate gene mapping reveals *VrMLO12* (*MLO* Clade II) is associated with powdery mildew resistance in mungbean (*Vigna radiata*[L.] Wilczek). *Plant Science* , 298, 110594. doi:10.1016/j.plantsci.2020.110594

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* , 17(9), 847-848. doi: 10.1093/bioinformatics/17.9.847

Table 1. Assembly and annotation statistics of Sulv 1 genome

Total number of contigs	362
Assembly size	473.7 Mb
N50	11.3 Mb
N90	2.33 Mb

Total number of contigs	362
Largest contig	22.6 Mb
Total number of scaffolds	290
Assembly size	473.7 Mb
N50	42.4 Mb
N90	30.1 Mb
Largest scaffold	72.8 Mb
GC content	33.3%
Repeat density	52.8%
Number of protein-coding genes	33,924
Average length of protein-coding genes	3,623
Supported by RNA-seq or homologs	32,633

FIGURE LEGENDS

Figure 1. Characteristics of the 11 chromosomes of Sulv 1 genome. Characteristics of the 11 chromosomes of Sulv 1. Tracks a to e represent the distribution of FPKM, gene density, density of Copia retrotransposable elements, density of Gypsy retrotransposable elements and GC density, respectively, with densities calculated in 200-kb windows. Track f shows syntenic blocks.

Figure 2. Phylogenetic analysis of Sulv 1 and other representative plant genomes. A Phylogenetic tree of Sulv 1 and 10 other species based on a concatenated alignment of single-copy orthologues. B Estimates of gene family expansions and contractions based on CAFÉ. The red and blue numbers indicate expanded and contracted gene families, respectively.

Figure 3. 4DTv distribution in Sulv 1 and other representative plant species. 4DTv distributions for Sulv 1 with other representative plant species are represented with colored lines as indicated.

Figure 4. LTR insertion events in Sulv 1 genome. The LTR insertion times of Sulv 1 and 10 other related plant species was calculated.

SUPPLEMENTARY FILES

Supplemental Figure 1. The distribution of Sulv 1 19-mers.

Supplemental Figure 2. Hi-C heatmap of the Sulv 1 genome.

Supplemental Figure 3. Venn diagram of annotated genes in Sulv 1 genome.

Supplemental Figure 4. Enriched GO terms of Sulv 1 genes.

Supplemental Figure 5. KOG functional classification of Sulv 1 genes.

Supplemental Table 1. Statistics of the Nanopore assembly of Sulv 1 genome.

Supplemental Table 2. Statistics of the genome assembly of Sulv 1.

Supplemental Table 3. Quality assessment of the assembled genome of Sulv 1.

Supplemental Table 4. Summary statistics of annotated genes in Sulv 1 genome.

Supplemental Table 5. Summary statistics of the functional genes of Sulv 1.

Supplemental Table 6. Annotation of the protein-coding genes of Sulv 1.

Supplemental Table 7. Identification of non-coding genes in Sulv 1 genome.

Supplemental Table 8. Summary statistics of the annotated repetitive sequences in the Sulv 1 genome.

Supplemental Table 9. Specific statistics of the annotated SSRs in the Sulv 1 genome.

Supplemental Table 10. Summary statistics of the gene families of Sulv 1 and other 10 angiosperm species.

Supplemental Table 11. Genes underwent positively seletion.