

# A chromosome-scale genome assembly of mungbean (*Vigna radiata*)

Supaporn Khanbo<sup>1</sup>, Poompat Phadphon<sup>1</sup>, Chaiwat Naktang<sup>1</sup>, Duangjai Sangsarakru<sup>1</sup>, Pitchaporn Waiyamitra<sup>1</sup>, Nattapol Narong<sup>1</sup>, Chutintorn Yundaeng<sup>1</sup>, Sithichoke Tangphatsornruang<sup>1</sup>, Kularb Laosatit<sup>2</sup>, Prakit Somta<sup>2</sup> and Wirulda Pootakham<sup>1</sup>

<sup>1</sup>National Omics Center, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand

<sup>2</sup>Department of Agronomy, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University, Kamphaeng Saen Campus, Nakhon Pathom, Thailand

## ABSTRACT

**Background.** Mungbean (*Vigna radiata*) is one of the most socio-economically important leguminous food crops of Asia and a rich source of dietary protein and micronutrients. Understanding its genetic makeup is crucial for genetic improvement and cultivar development.

**Methods.** In this study, we combined single-tube long-fragment reads (stLFR) sequencing technology with high-throughput chromosome conformation capture (Hi-C) technique to obtain a chromosome-level assembly of *V. radiata* cultivar 'KUML4'.

**Results.** The final assembly of the *V. radiata* genome was 468.08 Mb in size, with a scaffold N50 of 40.75 Mb. This assembly comprised 11 pseudomolecules, covering 96.94% of the estimated genome size. The genome contained 253.85 Mb (54.76%) of repetitive sequences and 27,667 protein-coding genes. Our gene prediction recovered 98.3% of the highly conserved orthologs based on Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis. Comparative analyses using sequence data from single-copy orthologous genes indicated that *V. radiata* diverged from *V. mungo* approximately 4.17 million years ago. Moreover, gene family analysis revealed that major gene families associated with defense responses were significantly expanded in *V. radiata*.

**Conclusion.** Our chromosome-scale genome assembly of *V. radiata* cultivar KUML4 will provide a valuable genomic resource, supporting genetic improvement and molecular breeding. This data will also be valuable for future comparative genomics studies among legume species.

Submitted 17 July 2024  
Accepted 5 December 2024  
Published 23 December 2024

Corresponding authors  
Prakit Somta, agrpks@ku.ac.th  
Wirulda Pootakham,  
wirulda@alumni.stanford.edu

Academic editor  
Tong Geon Lee

Additional Information and  
Declarations can be found on  
page 15

DOI 10.7717/peerj.18771

© Copyright  
2024 Khanbo et al.

Distributed under  
Creative Commons CC-BY 4.0

OPEN ACCESS

**Subjects** Agricultural Science, Genetics, Genomics, Molecular Biology, Plant Science

**Keywords** Mungbean, *Vigna radiata*, Chromosome-scale, Genome assembly, Hi-C, Annotation

## INTRODUCTION

Mungbean (*Vigna radiata* (L.) R. Wilczek var. *radiata*) is a socio-economically important legume crop with high nutritional value. It is a self-pollinated and fast-growing diploid plant species with  $2n = 2x = 22$  chromosomes (Kang et al., 2014). This legume crop belongs to the papilionoid subfamily of the Fabaceae and is widely cultivated in South and Southeast Asian countries, including India, China, Indonesia, Thailand, Myanmar, and the

**How to cite this article** Khanbo S, Phadphon P, Naktang C, Sangsarakru D, Waiyamitra P, Narong N, Yundaeng C, Tangphatsornruang S, Laosatit K, Somta P, Pootakham W. 2024. A chromosome-scale genome assembly of mungbean (*Vigna radiata*). *PeerJ* 12:e18771 <http://doi.org/10.7717/peerj.18771>

Philippines ([Sokolkova et al., 2020](#)). Globally, mungbean cultivation covers approximately 7.3 million hectares, with an average yield of 721 kilograms per hectare. India and Myanmar each contribute 30% to the total production of 5.3 million tons ([Nair & Schreinemachers, 2020](#)). This crop is an inexpensive source of dietary protein, carbohydrates, vitamins, and contains higher levels of folate and iron than most other legumes ([Keatinge et al., 2011](#); [Liu et al., 2022](#)). The seeds of mungbean contain a high amount of protein and carbohydrates, comprising approximately 20–30% protein and 60–70% carbohydrates ([Somta et al., 2022](#)). Mungbean is popularly grown in crop rotation systems due to its short duration, relative drought tolerance, and ability to improve soil fertility through atmospheric nitrogen fixation in symbiosis with *Rhizobium* and *Bradyrhizobium* bacteria in the soil ([Somta et al., 2022](#); [Yimram, Somta & Srinivas, 2009](#)).

Reference genomes are key to understand genome organization, to describe evolutionary events and to precisely identify genomic regions/genes associated with agronomically important traits. With advances in DNA sequencing technologies, mungbean genomes have been sequenced and assembled. The first draft genome sequence of mungbean was published in 2014, with the initial assembly of the elite breeding line 'VC1973A' from Illumina short-read sequences ([Kang et al., 2014](#)). However, this draft assembly covered 80% (421 Mb) of the estimated genome size and consists of 2,748 scaffolds with an N50 length of 1.52 Mb. An improved mungbean genome of VC1973A using long sequence reads was then published ([Ha et al., 2021](#)). The genome assembly covered a total sequence of 476 Mb with an N50 length of 5.2 Mb. Recently, [Liu et al. \(2022\)](#) reported a high-quality reference genome assembly of the mungbean cultivar 'Jilv7', utilizing the PacBio Sequel II platform and Illumina paired-end (PE) read technology. The genome assembly was 479.35 Mb in size, with a contig N50 length of 10.34 Mb. The line VC1973A, developed by the World Vegetable Center (formerly the Asian Vegetable Research and Development Center) in 1982, has since been widely grown as elite cultivar in Korea, Thailand, Taiwan, Canada, and China ([Kang et al., 2014](#)). Different cultivars/genotypes within the same species are likely to have variations in genome content, structure, and gene numbers, making a single reference genome insufficient to represent the full spectrum of genomic diversity within the species ([Editorial, 2020](#); [Golicz, Bately & Edwards, 2016](#); [Saxena, Edwards & Varshney, 2014](#)). Having more than one reference genome for mungbean offers significant advantages for genetics improvement and breeding.

In this study, we employed single-tube long-fragment reads (stLFR) sequencing ([Wang et al., 2019](#)) combined with high-throughput chromosome conformation capture (Hi-C) technique ([Putnam et al., 2016](#)) to assemble and construct a chromosome-level genome assembly of the mungbean 'KUML4', a cultivar recently released in Thailand ([Somta et al., 2024](#)). We specifically chose this cultivar due to its agronomically desirable traits, including high yield, large seed size, early and uniform maturity, and resistance to leaf spot and powdery mildew diseases ([Somta et al., 2024](#)). The reference assembly provides valuable resource for genetic improvement in mungbean variety breeding programs. This genomic data will also be useful for future comparative genomic studies of legume species.

## MATERIALS & METHODS

### Sample collection and DNA extraction

We selected the mungbean cultivar KUML4 for genome sequencing and assembly. Fresh leaves were collected from a 4-week-old mungbean plant. The leaf samples were immediately frozen in liquid nitrogen and stored until use. High molecular weight (HMW) DNA was extracted from the frozen tissues using the Genomic-tip 100/G DNA preparation kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. After extraction, the concentration, integrity, and purity of the DNA were determined using the dsDNA HS assay on a Qubit fluorometer (Thermo Fisher Scientific, Waltham, USA) and pulsed-field gel electrophoresis (PFGE).

### stLFR library construction and sequencing

The stLFR sequencing library was constructed using a total of 10 ng of HMW DNA following the MGIEasy stLFR Library Prep Kit's instructions (MGI Tech, Shenzhen, China). In brief, the HMW DNA was first inserted with transposons. The transposon-inserted DNA was then incubated with clonal barcoded beads and subjected to random priming. Following that, the captured DNA molecules were disrupted to yield sub-fragments, each less than 1 kb in size. Adapters were ligated to the DNA fragments, and the library was developed by PCR. Finally, the library was sequenced on the DNBSEQ-G400 using the MGISEQ-2000RS Sequencing Flow Cell v3.0 (MGI Tech, Shenzhen, China).

### Hi-C library preparation and sequencing

A high-throughput chromosome conformation capture (Hi-C) technique ([Putnam et al., 2016](#)) was performed by Biomarker Technologies Corporation (Beijing, China) to construct the chromosome-level assembly of mungbean. Briefly, the chromatin was crosslinked with fresh formaldehyde, and the fixed chromatin was then digested with *HindIII* restriction endonuclease. Subsequently, the digested DNA and the overhanging 5' ends of the DNA fragments were repaired with biotinylated nucleotides. DNA fragments were then circularized by blunt-end ligation. After the cross-linking was reversed, the DNA was purified and sheared into smaller fragments of 300–700 bp in size, which were then pulled down with streptavidin beads. The purified fragments were used to generate a sequencing library that was then sequenced on the Illumina platform (PE150; Illumina, San Diego, CA, USA).

### Genome assembly and Hi-C scaffolding

The stLFR reads were used to construct the preliminary draft genome assembly following the single-tube long fragment read data analysis software stLFRdenovo version 1.0.5, with default settings (<https://github.com/BGI-biotools/stLFRdenovo/releases/tag/v1.0.5>, accessed on 12 January 2024). This preliminary assembly was further scaffolded into a chromosome-level assembly using the high-throughput chromosome conformation capture (Hi-C) technique. The preliminary draft genome assembly and Hi-C reads were used as the input data for HiRise version 2.1.9, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies ([Putnam et al., 2016](#)).

The Hi-C sequences were mapped to the draft input assembly using BWA version 0.7.17 (*Li & Durbin, 2009*) (<https://github.com/lh3/bwa>). HiC-Pro version 2.10.0 (*Servant et al., 2015*) was then applied for data filtering and quality assessment of the Hi-C reads. The separations of Hi-C read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, score prospective joins, and make joins above a threshold. Manual adjustment and curation were carried out using a Hi-C visualization tool to correct potential misplacements or orientation errors. The mungbean genome assembly was deposited in National Center for Biotechnology Information (NCBI) under the accession number [JAVCZA000000000](#). The raw reads stLFR data were deposited in the SRA database at the NCBI with the accession number [SRR29729411](#).

### Genome size estimation

To estimate the genome size, k-mer analysis was performed on the high-quality paired-end (PE) read distribution using Jellyfish software version 2.2.10, and the k-mer distribution was plotted with GenomeScope version 2.0 ( $k = 21$ ; <http://genomescope.org/genomescope2.0/>; latest accessed: September 18, 2024) (*Vurture et al., 2017*).

### Genome assembly evaluation

After assembly, we evaluated the quality of the final genome assembly by aligning stLFR reads using BWA version 0.7.17 (*Li & Durbin, 2009*). We also used Benchmarking Universal Single-Copy Orthologs (BUSCO) version 5.4.4 (*Manni et al., 2021*) to assess the completeness of the assembly. The BUSCO pipeline was used to test for the presence and completeness of orthologs using the plant-specific database from Embryophyta OrthoDB release 10 (*Kriventseva et al., 2015*). Additionally, the LTR Assembly Index (LAI) was computed and assessed using LTR\_retriever version 3.0.1 (*Ou & Jiang, 2018*).

### Repetitive sequence identification

The repetitive sequences in the assembled genome were identified with de novo and homology-based methods. First, we constructed a de novo repeat library of the mungbean using the software package RepeatModeler version 2.0.1 with default parameters (<https://github.com/Dfam-consortium/RepeatModeler>, accessed on 30 April 2024). This package includes distinct discovery algorithms, RECON version 1.08 and RepeatScout version 1.0.5 (*Bao & Eddy, 2002; Flynn et al., 2020; Price, Jones & Pevzner, 2005*). The programs were employed to identify the boundaries of repetitive elements and to build consensus models of interspersed repeats. After we obtained the repeat library, we aligned the repeat sequences to NCBI GenBank's non-redundant protein database using BLASTX with an e-value threshold of  $10^{-6}$  to verify that the library did not contain sequences belonging to large families of protein-coding sequences. Then, using the RepBase plant repeat database (*Jurka et al., 2005*) (<https://www.girinst.org/>), homology-based repetitive sequences of the mungbean genome were identified with RepeatMasker version 4.0.9\_p2 with default parameters (*Tempel, 2012*). Additionally, we employed Tandem Repeats Finder (TRF) version 4.09 (*Benson, 1999*) with default parameters to predict tandem repeats element. HelitronScanner version 1.0 (*Xiong et al., 2014*) and miteFinder version 3.0

(*Hu, Zheng & Shang, 2018*) were used to identify helitrons and miniature inverted-repeat transposable elements (MITEs), respectively.

### Gene annotation

Regarding gene annotation, we employed the BRAKER2 version 2.1.6 (*Brüna et al., 2021*) which performs gene prediction by using GeneMark-ES/ET/EP/ETP version 4.65 and AUGUSTUS version 3.4.0. BRAKER2 version 2.1.6 also enables the integration of transcript and protein homology information into the predictions. Six sets of RNA-Seq data from various tissues, [SRS652317](#) (roots, stems, leaves), [SRS1308070](#) (seed), [SRR27190622](#) (petal blooming flower), [SRS3960651](#) (V6-stage leaves), [SRS9581523](#) (Thai SUT1 4-days-old root) and [SRS702243](#) (5-day seedling hypocotyl) were retrieved from SRA database of NCBI. The RNA-Seq data were mapped onto the *V. radiata* assembly of this study with HiSAT2 version 2.2.1 (*Kim et al., 2019*) and used for BRAKER2 analysis. Protein data from Viridiplantae OrthoDB11 (*Kuznetsov et al., 2023*) and *V. radiata* cultivar VC1973A (GCF\_000741045.1) were included in the analysis. The predicted genes were filtered and the longest isoforms of overlapping transcripts were selected using TSEBRA version 1.0.3 (*Gabriel et al., 2021*) with the default evidence-weight setting. BUSCO version 5.4.4 (*Manni et al., 2021*) was used with eudicots\_odb10 to evaluate the quality of the annotation. Functional annotation of the predicted genes was performed using OmicsBox version 1.3.11 (<https://www.biobam.com/download-omicsbox/>, last accessed on 10 May 2024). We aligned the protein sequences against the UniProtKB/Swiss-Prot (*Bairoch & Apweiler, 2000*) and GenBank non-redundant (NR) databases *via* BLASTP (*Gish & States, 1993*) with an e-value cutoff of  $10^{-5}$ . Gene ontology (GO) terms were retrieved and assigned to *V. radiata* query sequences, while enzyme codes (EC) corresponding to *V. radiata* gene ontology were extracted and mapped to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations. We also predicted tRNAs using the tRNAscan-SE tool version 2.0 (*Chan Patricia et al., 2021*) (<http://lowelab.ucsc.edu/tRNAscan-SE/>). As rRNAs are highly conserved, we used rRNA sequences from closely related species as references and used BLAST to predict rRNA sequences. Other ncRNAs, including miRNAs and snRNAs, were predicted by searching against the Rfam version 14.1 database (*Griffiths-Jones et al., 2005*) (<http://infernald.janelia.org/>) using the Infernal software version 1.1.5 with the default parameters. The genome annotation of mungbean was deposited in the Figshare database ([https://figshare.com/articles/dataset/Vigna\\_radiata\\_annotation\\_files/27094231](https://figshare.com/articles/dataset/Vigna_radiata_annotation_files/27094231)).

### Comparative genomics and phylogenetic analyses

OrthoFinder version 2.4.0 (*Emms & Kelly, 2019*) was used to identify orthologous gene groups among the three mungbean cultivars (KUML4, Vrad\_JL7, and VC1973A v2) and 14 other plant species (Table S1). These included nine legumes [black gram (*Vigna mungo* (L.) Hepper), créole bean (*Vigna reflexo-pilosa* Hayata), cowpea (*Vigna unguiculata* (L.) Walp.), adzuki bean (*Vigna angularis* (Ohwi) Ohwi and Ohashi), common bean (*Phaseolus vulgaris* L.), barrel medic (*Medicago truncatula* Gaertn.), soybean (*Glycine max* (L.) Merr.), chickpea (*Cicer arietinum*) and peanut (*Arachis duranensis*)], three cucurbit species (cucumber (*Cucumis sativus* L.), melon (*Cucumis melo* L.) and watermelon (*Citrullus lanatus* L.)),

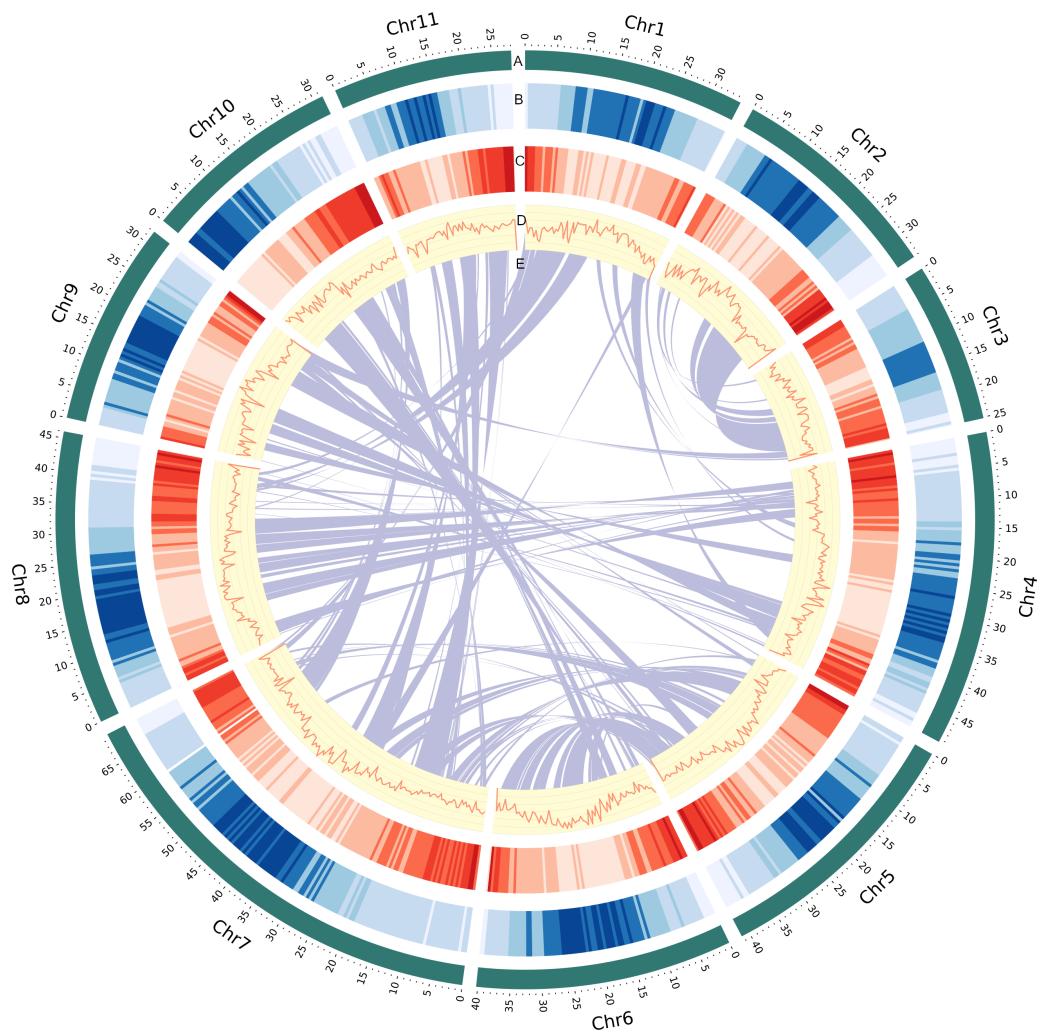
one rosid species (*Arabidopsis*), and one monocot species (rice (*Oryza sativa* L.)). The three mungbean cultivars were further clustered and subjected to Venn diagram analysis to explore the common and cultivar-specific genes across the three mungbean genomes. Single-copy orthologous protein sequences were aligned with MUSCLE version 3.8.1551 (Edgar, 2004), and alignment gaps were removed using trimAl version 1.4 rev15 (Capella-Gutiérrez, Silla-Martínez & Gabaldón, 2009). Alignment blocks were concatenated using catsequences program (<https://github.com/ChrisCreevey/catsequences>, last accessed on 22 May 2024), and the best substitution model for each block was estimated using ModelTest-NG program version 0.1.7 (Darriba et al., 2019). A maximum-likelihood phylogenetic tree was constructed using RAxML-NG program version 1.0.2 (Kozlov et al., 2019) with the best-fit substitution model, and 1,000 bootstrap replicates. *O. sativa* was designated as the outgroup of the phylogenetic tree. Divergence times of species within the phylogenetic tree were estimated with the MCMCTree program version 4.0 in the Phylogenetic Analysis by Maximum Likelihood software (PAML4 package) (Yang, 2007) using the relaxed-clock model with the known divergence time between cucumber and melon (8.4–11.8 million years ago; Sebastian et al., 2010). Finally, we employed CAFE software version 4.2 (Han et al., 2013) to predict the significant expansion and contraction of gene families across the phylogeny ( $p$ -value  $< 0.05$ ) with the gene birth-death ( $\lambda$ ) parameters estimated using the maximum-likelihood method. The expanded and contracted genes were then subjected to GO and KEGG functional annotation. Genome collinearity analysis was performed using the MCScan tool (Python version, <https://github.com/tanghaibaojcvi/>) from jcvi v1.3.885 (Tang et al., 2024; Wang et al., 2012). MCScan was employed to define syntenic blocks between the *V. radiata* KUML4, *V. radiata* JL7, and *V. mungo* genomes. Collinearity between the three genomes was visualized using collinearity plots.

## RESULTS

### Genome assembly and evaluation

To construct the chromosome-scale assembly of the *V. radiata* genome, we integrated datasets from stLFR reads and Hi-C sequences (Table S3). This initial genome assembly was generated from a total of 501,995,196 reads totaling 121.48 Gb of stLFR sequencing data (225.12  $\times$  coverage based on the estimated genome size obtained from k-mer analysis), resulting in a draft genome of 464.02 Mb with 6,685 scaffolds. The longest scaffold was 17.68 Mb, and the N50 length was 4.64 Mb. Additionally, we used 50.32 Gb of Hi-C data to generate chromosome-level scaffolds. The preliminary assembly was scaffolded using the long-range Hi-C technique, resulting in a final assembly containing 468.08 Mb in 5,834 scaffolds with an N50 length of 40.75 Mb (Table S2 and Fig. S1). We estimated the size of the *V. radiata* KUML4 genome based on the k-mer frequency histogram derived from short-read stLFR sequences (Fig. S2). The assembly size (468.08 Mb; Fig. 1) corresponded well with the analysis of k-mer distribution of PE reads (445.98 Mb). Our assembly size was about 2.35% smaller than the estimated size of the Vrad\_JL7 genome (479.35 Mb; Table 1) (Liu et al., 2022).

The final assembly was anchored into 11 pseudochromosomes larger than 25 Mb, corresponding to the haploid chromosome number in *V. radiata* ( $2n = 2x = 22$ ). These



**Figure 1** Genomic landscape of *V. radiata* chromosomes. (A) A physical map of 11 pseudomolecules (chromosomes) numbered according to [Liu et al. \(2022\)](#). (B) Repeat density represented by proportion of genomic regions covered by repetitive sequences in 500 kb windows. (C) Gene density represented by number of genes in 500 kb windows. (D) GC content represented by percentage of G + C bases in 500 kb windows. (E) Syntenic blocks in the genome are illustrated by connected lines.

Full-size DOI: [10.7717/peerj.18771/fig-1](https://doi.org/10.7717/peerj.18771/fig-1)

pseudomolecules covered 432.35 Mb, or 96.94% of the estimated genome size and were named in accordance with [Liu et al. \(2022\)](#). To evaluate the quality and accuracy of the *V. radiata* KUML4 assembly, we first mapped the stLFR reads back to the genome assembly and obtained a mapping rate of 98.09%. Second, we evaluated the assembly using 1,614 Benchmarking Universal Single Copy Orthologs (BUSCO) genes from Embryophyta to assess the completeness of the conserved orthologs in the *V. radiata* genome. Our assembly demonstrated high levels of completeness, with proportions of complete (C), complete and single-copy (S), complete and duplicated (D), fragmented (F), and missing (M) genes as follows: C: 98.3% [S: 95.0%, D: 3.1%], F: 0.8%, M: 1.0% ([Table S2](#)). Finally, the long terminal repeat (LTR) assembly index (LAI) was employed to assess the completeness of

**Table 1** Assembly statistical comparison of the KUML4, Vrad\_JL7, and VC1973A v2 genomes.

Genomic feature	KUML4	Vrad_JL7	VC1973A v2
Total assembly size, Mb	468.08	475.35	475.7
No. contigs	5,834	632	1,511
Largest contig, Mb	17.68	30.20	12.73
Contig N50, Mb	4.64	10.34	2.8
Scaffold N50, Mb	40.75	43.79	47.1
Percentage anchored to chromosomes	93.52	98.72	89.92
No. gaps	15,536	259	1,047
Length of gaps, Mb	14.45	0.11	1.81
GC content, %	33.27	33.45	33.27
Complete BUSCOs (genome), %	98.0	98.02	91.36
Complete BUSCOs (protein), %	94.1	96.9	80.01
LAI	11.57	15.67	14.65
Intact LTR-RTs	1,525	2,725	2,458
Repetitive sequences, %	54.76	53.45	52.79
Protein-coding genes	27,667	40,125	30,958

the assembled LTR as an indicator of the quality of the whole genome assembly. The LAI score of the assembly was 11.57 (Table 1), meeting the reference genome level according to developer classification standards (Ou, Chen & Jiang 2018).

### Gene annotation

A combination of several types of evidence, including ab initio prediction, RNA-seq supported evidence, and protein homology evidence, was integrated into the genome annotation pipeline to obtain reliable gene predictions. The genome annotation contained 28,593 predicted gene models, of which 27,667 (96.76%) were protein-coding genes. The BUSCO completeness for the predicted genes was 94.1%. The average gene length was 2,793 bp with 4.65 exons per gene, and the average exon and intron length were 236 bp and 464.93 bp, respectively (Table 2). We observed that average GC content of the *V. radiata* genome assembly was 33.27% (Table S2), which was close to the average GC content in introns (32.16%), whereas the average GC content in exons was higher at 44.50% (Table 2). A total of 18,334 predicted genes had been assigned gene ontology (GO) terms. The top three most prevalent GO terms associated with biological process were protein phosphorylation, regulation of DNA-templated transcription, and transmembrane transport, while the most prevalent terms associated with the cellular component were membrane followed by nucleus, and cytoplasm. The largest categories of genes assigned to molecular function were ATP binding followed by metal ion binding, and DNA binding (Fig. S3). Of 28,593 predicted gene models, 96.76%, 70.52%, and 64.12%, could be identified functionally using the NCBI nonredundant (NR) protein, SwissProt and GO databases, respectively (Table S4). For the non-coding RNAs, we also identified 28,017 microRNAs, 765 transfer RNAs, 483 ribosomal RNAs and 12,497 small nuclear RNAs in the *V. radiata* genome (Table S5).

**Table 2** Annotation statistics for *V. radiata* KUML4 genome.

Number of predicted gene models	28,593
Number of protein-coding genes	27,667
Total gene length (Mb)	79.88
Average gene size (bp)	2,793.58
Average number of exons/gene	4.65
Total exon length (Mb)	31.47
Average exon length (bp)	236.74
GC content of exons (%)	44.50
Average number of introns/gene	3.48
Total intron length (Mb)	46.27
Average intron length (bp)	464.93
GC content of introns (%)	32.16

### Identification of repetitive elements

The *V. radiata* KUML4 genome assembly contained a total of 253.85 Mb of repetitive elements, accounting 54.76% of the assembly (Table 3). These repeat elements comprised 25.72 Mb (10.12%) of DNA transposons, 8.04 Mb (3.17%) of simple sequence repeats, 119.09 Mb (46.92%) of retrotransposons, 35.49 Mb (13.98%) of tandem repeats, 18.53 Mb (7.30%) of miniature inverted-repeat transposable elements (MITEs), and 1.60 Mb (0.63%) of helitron elements. Retrotransposons constituted the majority of known repeats, comprising nearly half of the total repeat content in the *V. radiata* KUML4 genome. Within the retrotransposon class, long terminal repeat (LTR) retrotransposons were the most frequently detected, accounting for 46.23% of the repetitive sequences. The most abundant LTR superfamilies, *Gypsy* and *Copia*, accounted for 15.65% and 9.65% of the mungbean genome, respectively (Table 3), which differed from the 16.65% and 13.77% in the Vrad\_JL7 genome (Liu et al., 2022).

### Comparative genomic analysis

To explore the evolutionary relationships among the three mungbean cultivars (KUML4, Vrad\_JL7, and VC1973A v2) and other plant species, the gene sets from the five *Vigna* species (*V. radiata*, *V. mungo*, *V. reflexo-pilosa*, *V. unguiculata*, and *V. angularis*), five other legumes (*P. vulgaris*, *M. truncatula*, *G. max*, *C. arietinum*, and *A. duranensis*), four other eudicots (*A. thaliana*, *C. sativus*, *C. melo*, *C. lanatus*), as well as one monocot (*O. sativa*) were used for gene family clustering. Of the total 670,738 proteins in 17 genomes from 15 species, 641,506 proteins (95.6%) were clustered into 36,504 orthologous groups. Among the 27,625 protein-coding genes in KUML4 that had orthologs in the other species analyzed, only 256 (0.93%) were specific to KUML4, while 575 (2.08%) were shared among legume species (Fig. 2A). Comparison of gene families among the three mungbean genomes revealed that 238, 1,150, and 308 orthologous groups were unique to KUML4, Vrad\_JL7 and VC1973A v2, respectively, while 18,152 orthologous groups were shared among KUML4 and the two previously published mungbean genomes (Vrad\_JL7 and VC1973A v2; Fig. 2B). Sequence information from single-copy orthologous genes was

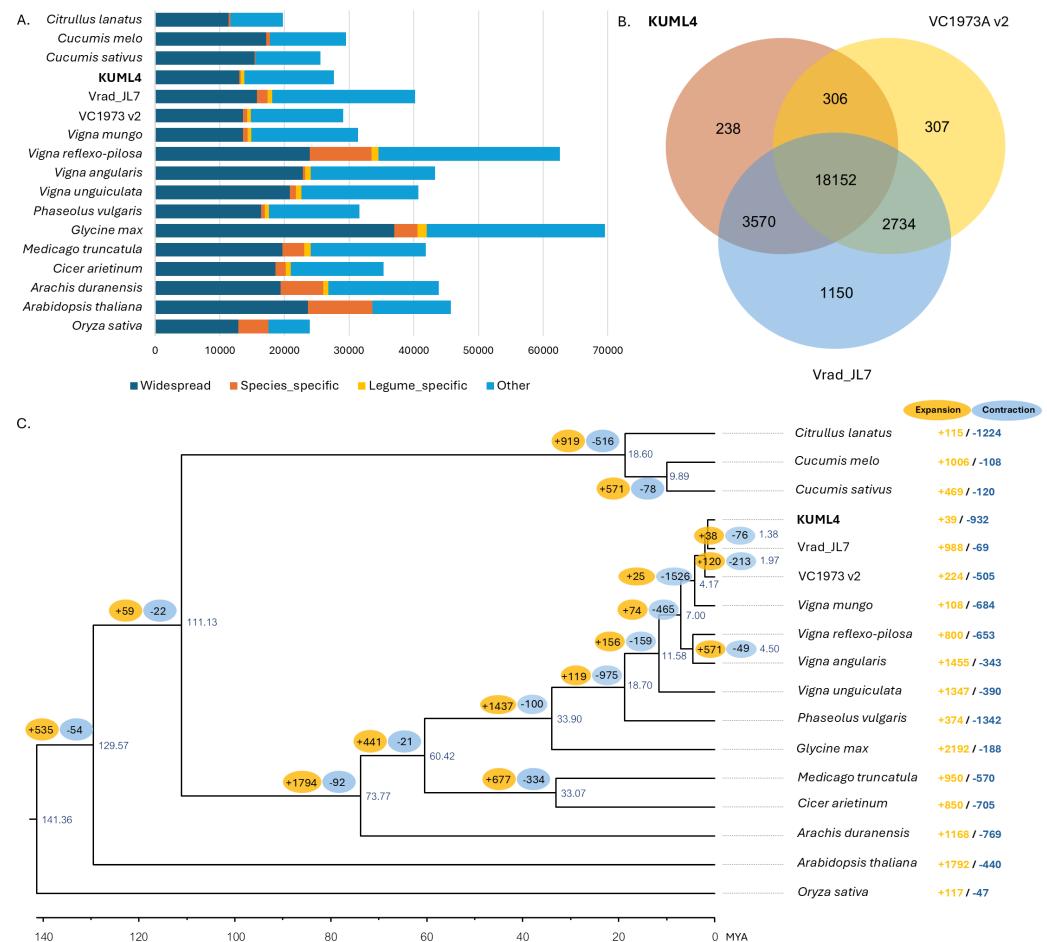
**Table 3** Repeat elements in the *V. radiata* KUML4 genome assembly.

Types of repeats	Bases (Mb)	% of the assembly	% of total repeats
<b>DNA transposons:</b>			
En-Spm	13.01	2.81	5.12
MULE-MuDR	4.06	0.88	1.60
DNA: Others	8.65	1.87	3.40
<b>Retrotransposons:</b>			
LINE	1.75	0.38	0.69
SINE	0.0015	0.00	0.00
LTR: <i>Copia</i>	44.63	9.65	17.58
LTR: <i>Gypsy</i>	72.34	15.65	28.50
LTR: Others	0.37	0.08	0.15
<b>Simple sequence repeats:</b>	8.04	1.74	3.17
<b>Tandem repeats:</b>	35.49	7.58	13.98
<b>MITES:</b>	18.53	3.96	7.30
<b>Helitrons:</b>	1.60	0.34	0.63
<b>Others:</b>	45.38	9.82	17.88
<b>Total</b>	253.85	57.46	

used to construct a maximum-likelihood tree with *O. sativa* as an outgroup, and the divergence time was estimated based on the topology and branch length. The phylogenetic analysis revealed that the three *V. radiata* cultivars (KUML4, Vrad\_JL7 and VC1973A v2) and *V. mungo* were closest and diverged approximately 4.17 million years ago (MYA). The divergence of KUML4 from Vrad\_JL7 occurred around 1.38 MYA. The ancestor of *V. radiata* and *V. mungo* formed a sister clade to the ancestor of *V. angularis* and *V. reflexo-pilosa*. The estimated divergence time between the two clades was about 7 MYA (Fig. 2C). Additionally, our analysis of gene family expansion and contraction across the three mungbean cultivars and other plant species identified 2,764 significant gene families. A total of 39 significantly expanded gene families, and 932 significantly contracted families were identified in KUML4 (Fig. 2C).

Functional categories indicated that a large number of expanded gene families were associated with secondary metabolite biosynthesis, including berberine bridge enzyme-like 13, isoflavone 2'-hydroxylase, UDP-glycosyltransferase 73D1. In addition, they were enriched in the defense responses, and signal transduction related to plant hormones like ethylene and cytokinin, including aspartic proteinase CDR1-like, GDSL esterase/lipase EXL3, ethylene-responsive transcription factor ERF098-like, zeatin O-glucosyltransferase-like, and cytokinin riboside 5'-monophosphate phosphoribohydrolase (Table S6). The contracted gene families were mainly involved in the cysteine-rich receptor-like protein kinase 10, proline-rich receptor-like protein kinase PERK4, shaggy-related protein kinase epsilon isoform X1, and MDIS1-interacting receptor-like kinase (Table S6).

We also explored genome synteny among *V. radiata* KUML4, *V. radiata* JL7, and *V. mungo*. Our genome assembly of KUML4 reveals a high level of synteny with the Vrad\_JL7 genome, as illustrated by the large syntenic blocks in Fig. 3. Although some



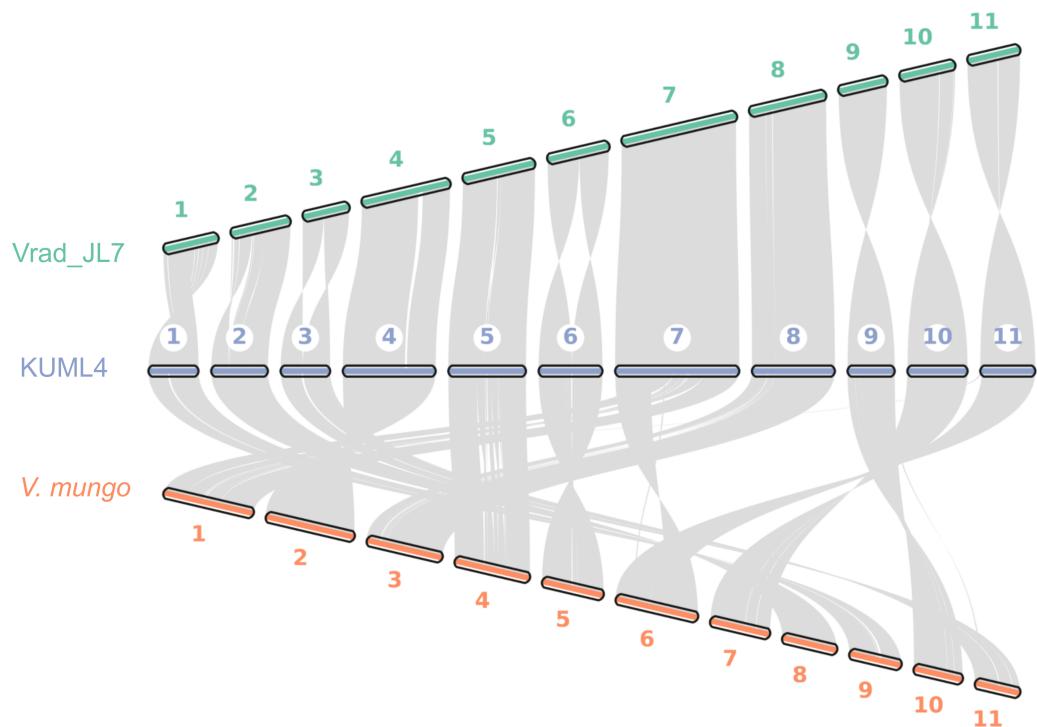
**Figure 2** Comparative genomics of mungbean and other plant species. (A) Bar charts display the number of proteins that were widespread, legume-specific and species-specific (having no detectable orthologues in other species). (B) Venn diagrams displaying clusters of shared and unique orthologous gene families in three mugbean genomes (KUM4, Vrad\_JL7, and VC1973A v2). (C) Inferred phylogenetic tree of mungbean and other plant species based on protein sequences of single-copy orthologous genes. Numbers at each node represent the estimated divergence time of each node in million years ago (MYA). The branch labels in yellow and blue represent the significantly expanded and contracted gene families, respectively, of each node. The right column shows significantly expanded and contracted gene families of individual species.

Full-size DOI: 10.7717/peerj.18771/fig2

structural variations were detected, including inverted regions, these were observed on chromosomes 1, 2, 3, 6, 9, 10, and 11 between the KUM4 and Vrad\_JL7 genomes.

## DISCUSSION

The mungbean is one of the most important edible legume crops, widely cultivated in many Asian countries (Hou *et al.*, 2019). It has significant health and economic benefits and also plays a significant role in climate resilience and increasing food security (Assefa, Dinku & Jemal, 2022). Genome assemblies of the mungbean are fundamental to understanding its genetic makeup, which is essential for genetic improvement and superior cultivar



**Figure 3** Chromosome collinearity between *V. radiata* cultivar KML4, *V. radiata* cultivar Vrad\_JL7, and *V. mungo*.

Full-size DOI: 10.7717/peerj.18771/fig-3

development. In the present study, we sequenced and assembled the reference genome of the mungbean cultivar KUML4, a new cultivar grown for dry seed in Thailand. By combining the stLFR technique with the high-throughput chromosome conformation capture (Hi-C) technique, we achieved a chromosome-scale assembly for the cultivar KUML4. The stLFR technology enables sequencing of data from long DNA molecules by adding the same barcode sequence to sub-fragments of the original long DNA molecule (Wang *et al.*, 2019). The Hi-C approach facilitates the analysis of chromosomal interactions through chromosome conformation capture. Hi-C data can provide long-range linkage information up to tens of megabases that could be used to generate chromosome-scale scaffolds (Burton *et al.*, 2013; Ghurye *et al.*, 2019; Marie-Nelly *et al.*, 2014). Our final assembly comprised 11 pseudochromosomes covering 432.35 Mb, accounting for 96.94% of the 445.98 Mb estimated genome size, with an N50 length of 40.75 Mb. Compared with three previously reported mungbean genome assemblies—VC1973A (421 Mb) (Kang *et al.*, 2014), the improved version of VC1973A (476 Mb) (Ha *et al.*, 2021), and Vrad\_JL7 (479 Mb) (Liu *et al.*, 2022)—our mungbean assembly (KUML4) is longer in size than the initial draft of the VC1973A genome and shorter than the improved version of VC1973A and the Vrad\_JL7 genomes. The completeness of the gene space measured by BUSCO in our assembly is higher than that of the two VC1973A assemblies (Ha *et al.*, 2021; Kang *et al.*, 2014) and comparable to the assembly of Vrad\_JL7 (98.3% for our assembly and 98.02% for Vrad\_JL7

([Liu et al., 2022](#)), suggesting that both the KUML4 and Vrad\_JL7 assemblies achieve a high level of completeness.

We also performed a synteny analysis between our assembled genome, Vrad\_JL7, and *V. mungo*. The genome comparison revealed high collinearity between the chromosomes of mungbean KUML4 and Vrad\_JL7 (Fig. 3). However, some genomic regions in KUML4, Vrad\_JL7, and *V. mungo* appeared inverted in the comparison, particularly on chromosomes 1, 2, 3, 6, 9, 10, and 11. [Liu et al. \(2022\)](#) suggested that these inversions observed in Vrad\_JL7, as well as in cowpea and adzuki bean, may have resulted from misassembly in Vrad\_JL7. The study also suggested that the density distribution of genes and transposable elements indicated that certain segments on chromosomes 1, 2, 3, 6, and 10 might need to be reversed in orientation.

The KUML4 assembly is not significantly improved compared to the Vrad\_JL7 assembly, with several evaluation scores, including contig number, scaffold N50 length, and gap count, showing both lower and higher values (Table 1). According to the LAI-based method for assessing genome assembly quality, our mungbean KUML4 genome assembly shows a slightly lower LAI score than the two previously published mungbean genomes (Vrad\_JL7 and VC1973A; Table 1). However, the KUML4 assembly still offers high-quality synteny and gene content, making it a valuable genome for comparative analysis. The LAI scale typically considers scores above 10 as indicative of reference-quality assemblies ([Ou, Chen & Jiang, 2018](#)). The differences in LAI scores may be attributed to several factors, such as variations in sequencing technology, assembly methods, or data quality. LAI is one of several metrics used to assess genome quality. However, based on contig N50 size, scaffold N50 size, BUSCO scores, and the completeness of the functional annotation (Table 1), our KUML4 genome assembly qualifies as a draft reference-quality genome. Its good contiguity and accuracy provide a valuable reference for future genetic research on other legume species, and the newly assembled KUML4 genome contributes significantly to the legume genome resource bank, offering a novel genomic resource for future mungbean improvement efforts. In addition, the number of predicted genes (28,593) in our assembly is higher than that predicted in the initial draft genome assembly of mungbean ([Kang et al., 2014](#)), but lower than that of the latest assembly ([Liu et al., 2022](#)). We annotated 27,667 protein-coding genes in our assembled genome, while the published Vrad\_JL7 genome annotated 40,125 genes. The discrepancy in gene numbers between our KUML4 assembly and Vrad\_JL7 can be attributed to several factors. Among the 40,125 protein-coding genes in the Vrad\_JL7 genome, approximately 32,000 were functionally annotated, while the remaining ~7,400 genes had minimal hits in the NCBI nr database during BLAST searches. According to orthogroup analysis of the three *V. radiata* genomes (KUML4, Vrad\_JL7, and VC1973A v2), 1,150 orthogroups consisting of 3,437 genes were unique to Vrad\_JL7 (Fig. 2B). Of these, ~1,900 genes were not functionally annotated, and ~2,700 of the 7,400 genes were not assigned to any orthogroup. Therefore, approximately 4,600 to 7,400 genes may be unique to Vrad\_JL7. Additionally, [Liu et al. \(2022\)](#) noted that some gene models in Vrad\_JL7 could be improved, highlighting that some gene models were short (less than 200 nt), while others were partial, fragmented, repetitive, or chimeric. The complexity of genetic variation among mungbean cultivars plays a significant role.

The different agronomic traits of KUML4 and Vrad\_JL7 cultivars likely contribute to differences in genome content, including gene number and structural variations, with each cultivar having cultivar-specific genes based on its breeding history. Furthermore, differences in sequencing technologies and assembly pipelines compared to previous studies may contribute to discrepancies in gene numbers and structural variations. The use of stLFR sequencing and Hi-C for genome assembly in KUML4 differs from the long-read sequencing employed in Vrad\_JL7. For instance, variations in read lengths, coverage depth, and assembly algorithms can influence the detection of repetitive sequences, gene fragmentation, and overall assembly completeness.

However, the BUSCO completeness for the predicted genes assessment revealed 94.1% in our assembly, which is close to the 96.9% reported in a previous study ([Liu et al., 2022](#)), indicating the high completeness of gene annotation. Compared with other sequenced legume genomes, the number of predicted genes in our mungbean genome is lower than that of black gram ([Pootakham et al., 2021](#)), adzuki bean ([Yang et al., 2015](#)), soybean ([Schmutz et al., 2010](#)) and pigeonpea ([Varshney et al., 2012](#)), and higher than that of common bean ([Schmutz et al., 2014](#)), but similar to chickpea ([Varshney et al., 2013](#)). The proportion of repetitive sequences identified in our genome assembly (54.76%) is slightly higher than the percentage reported for mungbean Vrad\_JL7 (53.45%) in [Liu et al. \(2022\)](#). LTR elements, including Gypsy and Copia elements were the predominant classes in our mungbean assembly. Compared to Copia elements (9.65%), Gypsy elements (15.65%) occupied relatively larger proportions of genomic sequence in mungbean, which is consistent with earlier reports about other *Vigna* species ([Guan et al., 2022](#); [Pootakham et al., 2021](#); [Pootakham et al., 2023](#)).

To explore the genetic distinctiveness of mungbean, we focused on expanded orthologous gene families and their gene functions. Changes in genetic makeup are potentially linked to unique phenotypic and functional changes. By examining expanded gene families in mungbean, we found an enrichment in genes associated with secondary metabolite biosynthesis, defense responses, and signal transduction related to plant hormones like ethylene and cytokinin. Notably, we observed that several genes associated with the auxin signaling pathway, including auxin-induced protein X10A-like, were expanded. This pathway plays a key role in regulating cell elongation and division, which directly influence seed size and yield. Additionally, genes involved in cytokinin response pathways, crucial for promoting cell proliferation and grain filling, were also enriched. These gene family expansions likely contribute to KUML4's high yield and large seed size by supporting growth-related processes. A phylogenetic analysis based on sequence information from single-copy orthologous genes revealed that mungbean and black gram were the most closely related, and the ancestor of mungbean and black gram formed a sister clade to the ancestor of créole bean and azuki bean, consistent with the results of previously published phylogenetic analyses ([Doi et al., 2002](#); [Pootakham et al., 2021](#)).

The KUML4 genome assembly provides valuable insights due to its distinct genetic background and unique agronomic traits. These traits set KUML4 apart from other *V. radiata* cultivars, making this genome a crucial addition to the understanding of the species' genetic diversity. Developing mungbean cultivars with desirable agronomic traits,

such as enhanced productivity and tolerance to biotic and abiotic stresses, has long been a key goal in breeding programs. We believe that having a chromosome-scale reference genome for KUML4 will significantly enhance our understanding of mungbean biology, thereby greatly aiding molecular breeding efforts to produce improved mungbean cultivars.

## CONCLUSIONS

Mungbean genome assemblies are key to understanding the genetic basis of its ability to adapt to various environments and produce high nutritional content. Utilizing the stLFR technique together with high-throughput chromosome conformation capture (Hi-C) technique, we achieved a chromosome-scale assembly of the mungbean cultivar KUML4. The assembly encompassed 11 pseudochromosomes with a total length of 468.08 Mb and a scaffold N50 of 40.75 Mb. The genomic resource developed here will provide valuable information for future mungbean genetic improvement. Furthermore, it serves as a valuable addition to ongoing comparative genomics research within legume species.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This project was funded by the National Research Council of Thailand (NRCT) (grant no. N42A650274) and Kasetsart University Research and Development Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:  
National Research Council of Thailand (NRCT): N42A650274.  
Kasetsart University Research and Development Institute.

### Competing Interests

The authors declare there are no competing interests.

### Author Contributions

- Supaporn Khanbo conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Poompat Phadphon analyzed the data, prepared figures and/or tables, and approved the final draft.
- Chaiwat Naktang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Duangjai Sangsakru performed the experiments, prepared figures and/or tables, and approved the final draft.
- Pitchaporn Waiyamitra performed the experiments, prepared figures and/or tables, and approved the final draft.

- Nattapol Narong performed the experiments, prepared figures and/or tables, and approved the final draft.
- Chutintorn Yundaeng analyzed the data, prepared figures and/or tables, and approved the final draft.
- Sithichoke Tangphatsornruang conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Kularb Laosatit conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Prakit Somta conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Wirulda Pootakham conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

## Data Availability

The following information was supplied regarding data availability:

The genome assembly data is available at NCBI: [JAVCZA000000000](https://www.ncbi.nlm.nih.gov/nuccore/JAVCZA000000000), [GCA\\_040583985](https://www.ncbi.nlm.nih.gov/nuccore/GCA_040583985).

The raw reads are available at NCBI SRA: [SRR29729411](https://www.ncbi.nlm.nih.gov/sra/SRR29729411).

The genome annotation of mungbean is available at Figshare: Sonthirod, Chutima (2024). *Vigna radiata* annotation files. figshare. Dataset. <https://doi.org/10.6084/m9.figshare.27094231.v1>

## Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.18771#supplemental-information>.

## REFERENCES

**Assefa ZB, Dinku AM, Jemal AM. 2022.** Value chain analysis of Mung Bean (*Vigna radiata* L. Wilczek thrive) in Kalu Woreda, Ethiopia. *Agriculture & Food Security* 11:54 DOI [10.1186/s40066-022-00393-2](https://doi.org/10.1186/s40066-022-00393-2).

**Bairoch A, Apweiler R. 2000.** The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28:45–48 DOI [10.1093/nar/28.1.45](https://doi.org/10.1093/nar/28.1.45).

**Bao Z, Eddy SR. 2002.** Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* 12:1269–1276 DOI [10.1101/gr.88502](https://doi.org/10.1101/gr.88502).

**Benson G. 1999.** Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27:573–580 DOI [10.1093/nar/27.2.573](https://doi.org/10.1093/nar/27.2.573).

**Brúna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021.** BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* 3:lqaa108 DOI [10.1093/nargab/lqaa108](https://doi.org/10.1093/nargab/lqaa108).

**Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013.** Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* 31:1119–1125 DOI [10.1038/nbt.2727](https://doi.org/10.1038/nbt.2727).

**Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T.** 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973 DOI [10.1093/bioinformatics/btp348](https://doi.org/10.1093/bioinformatics/btp348).

**Chan Patricia P, Lin Brian Y, Mak Allysia J, Lowe Todd M.** 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* 49:9077–9096 DOI [10.1093/nar/gkab688](https://doi.org/10.1093/nar/gkab688).

**Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T.** 2019. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution* 37:291–294 DOI [10.1093/molbev/msz189](https://doi.org/10.1093/molbev/msz189).

**Doi K, Kaga A, Tomooka N, Vaughan DA.** 2002. Molecular phylogeny of genus *vigna* subgenus *ceratotropis* based on rDNA ITS and *atpB-rbcL* intergenic spacer of cpDNA sequences. *Genetica* 114:129–145 DOI [10.1023/A:1015158408227](https://doi.org/10.1023/A:1015158408227).

**Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113 DOI [10.1186/1471-2105-5-113](https://doi.org/10.1186/1471-2105-5-113).

**Editorial N.** 2020. Crop genomes and beyond. *Nature Genetics* 52:865 DOI [10.1038/s41588-020-0691-5](https://doi.org/10.1038/s41588-020-0691-5).

**Emms DM, Kelly S.** 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* 20:238 DOI [10.1186/s13059-019-1832-y](https://doi.org/10.1186/s13059-019-1832-y).

**Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF.** 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* 117:9451–9457 DOI [10.1073/pnas.1921046117](https://doi.org/10.1073/pnas.1921046117).

**Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M.** 2021. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* 22:566 DOI [10.1186/s12859-021-04482-0](https://doi.org/10.1186/s12859-021-04482-0).

**Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S.** 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology* 15:e1007273 DOI [10.1371/journal.pcbi.1007273](https://doi.org/10.1371/journal.pcbi.1007273).

**Gish W, States DJ.** 1993. Identification of protein coding regions by database similarity search. *Nature Genetics* 3:266–272 DOI [10.1038/ng0393-266](https://doi.org/10.1038/ng0393-266).

**Golicz AA, Batley J, Edwards D.** 2016. Towards plant pangenomics. *Plant Biotechnology Journal* 14:1099–1105 DOI [10.1111/pbi.12499](https://doi.org/10.1111/pbi.12499).

**Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A.** 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* 33:D121–D124 DOI [10.1093/nar/gki081](https://doi.org/10.1093/nar/gki081).

**Guan J, Zhang J, Gong D, Zhang Z, Yu Y, Luo G, Somta P, Hu Z, Wang S, Yuan X, Zhang Y, Wang Y, Chen Y, Laosatit K, Chen X, Chen H, Sha A, Cheng X, Xie H, Wang L.** 2022. Genomic analyses of rice bean landraces reveal adaptation and yield related loci to accelerate breeding. *Nature Communications* 13:5707 DOI [10.1038/s41467-022-33515-2](https://doi.org/10.1038/s41467-022-33515-2).

**Ha J, Satyawan D, Jeong H, Lee E, Cho KH, Kim MY, Lee SH.** 2021. A near-complete genome sequence of mungbean (*Vigna radiata* L.) provides key insights into the modern breeding program. *Plant Genome* 14(e20121): DOI [10.1002/tpg2.20121](https://doi.org/10.1002/tpg2.20121).

**Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013.** Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution* **30**:1987–1997 DOI [10.1093/molbev/mst100](https://doi.org/10.1093/molbev/mst100).

**Hou D, Yousaf L, Xue Y, Hu J, Wu J, Hu X, Feng N, Shen Q. 2019.** Mung Bean (*Vigna radiata* L.): bioactive polyphenols, polysaccharides, peptides, and health benefits. *Nutrients* **11**:1238 DOI [10.3390/nu11061238](https://doi.org/10.3390/nu11061238).

**Hu J, Zheng Y, Shang X. 2018.** MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. *BMC Medical Genomics* **11**:101 DOI [10.1186/s12920-018-0418-y](https://doi.org/10.1186/s12920-018-0418-y).

**Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005.** Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**:462–467 DOI [10.1159/000084979](https://doi.org/10.1159/000084979).

**Kang YJ, Kim SK, Kim MY, Lestari P, Kim KH, Ha B-K, Jun TH, Hwang WJ, Lee T, Lee J, Shim S, Yoon MY, Jang YE, Han KS, Taeprayoon P, Yoon N, Somta P, Tanya P, Kim KS, Gwag J-G, Moon J-K, Lee Y-H, Park B-S, Bombarely A, Doyle JJ, Jackson SA, Schafleitner R, Srinivas P, Varshney RK, Lee S-H. 2014.** Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature Communications* **5**:5443 DOI [10.1038/ncomms6443](https://doi.org/10.1038/ncomms6443).

**Keatinge JDH, Easdown WJ, Yang RY, Chadha ML, Shanmugasundaram S. 2011.** Overcoming chronic malnutrition in a future warming world: the key importance of mungbean and vegetable soybean. *Euphytica* **180**:129–141 DOI [10.1007/s10681-011-0401-6](https://doi.org/10.1007/s10681-011-0401-6).

**Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019.** Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**:907–915 DOI [10.1038/s41587-019-0201-4](https://doi.org/10.1038/s41587-019-0201-4).

**Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019.** RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**:4453–4455 DOI [10.1093/bioinformatics/btz305](https://doi.org/10.1093/bioinformatics/btz305).

**Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM. 2015.** OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research* **43**:D250–256 DOI [10.1093/nar/gku1220](https://doi.org/10.1093/nar/gku1220).

**Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Evgenia VKriventseva, Zdobnov EM. 2023.** OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* **51**:D445–D451 DOI [10.1093/nar/gkac998](https://doi.org/10.1093/nar/gkac998).

**Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760 DOI [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324).

**Liu C, Wang Y, Peng J, Fan B, Xu D, Wu J, Cao Z, Gao Y, Wang X, Li S, Su Q, Zhang Z, Wang S, Wu X, Shang Q, Shi H, Shen Y, Wang B, Tian J. 2022.** High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Communications* **3**:100352 DOI [10.1016/j.xplc.2022.100352](https://doi.org/10.1016/j.xplc.2022.100352).

**Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021.** BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 38:4647–4654 DOI [10.1093/molbev/msab199](https://doi.org/10.1093/molbev/msab199).

**Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G, Parodi DP, Syan S, Guillén N, Margeot A, Zimmer C, Koszul R. 2014.** High-quality genome (re)assembly using chromosomal contact data. *Nature Communications* 5:5695 DOI [10.1038/ncomms6695](https://doi.org/10.1038/ncomms6695).

**Nair R, Schreinemachers P. 2020.** Global Status and Economic Importance of Mungbean. In: Nair RM, Schafleitner R, Lee S-H, eds. *The mungbean genome*. Cham: Springer International Publishing, 1–8.

**Ou S, Chen J, Jiang N. 2018.** Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research* 46:e126–e126 DOI [10.1093/nar/gky730](https://doi.org/10.1093/nar/gky730).

**Ou S, Jiang N. 2018.** LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* 176:1410–1422 DOI [10.1104/pp.17.01310](https://doi.org/10.1104/pp.17.01310).

**Pootakham W, Nawae W, Naktang C, Sonthirod C, Yoocha T, Kongkachana W, Sangsrakru D, Jomchai N, U-thoomporn S, Somta P, Laosatit K, Tangphatsornruang S. 2021.** A chromosome-scale assembly of the black gram (*Vigna mungo*) genome. *Molecular Ecology Resources* 21:238–250 DOI [10.1111/1755-0998.13243](https://doi.org/10.1111/1755-0998.13243).

**Pootakham W, Sonthirod C, Naktang C, Yundaeng C, Yoocha T, Kongkachana W, Sangsrakru D, Somta P, Tangphatsornruang S. 2023.** Genome assemblies of *Vigna reflexo-pilosa* (créole bean) and its progenitors, *Vigna hirtella* and *Vigna trinervia*, revealed homoeolog expression bias and expression-level dominance in the allotetraploid. *GigaScience* 12:giad050 DOI [10.1093/gigascience/giad050](https://doi.org/10.1093/gigascience/giad050).

**Price AL, Jones NC, Pevzner PA. 2005.** De novo identification of repeat families in large genomes. *Bioinformatics* 21:i351–i358 DOI [10.1093/bioinformatics/bti1018](https://doi.org/10.1093/bioinformatics/bti1018).

**Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE. 2016.** Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Research* 26:342–350 DOI [10.1101/gr.193474.115](https://doi.org/10.1101/gr.193474.115).

**Saxena RK, Edwards D, Varshney RK. 2014.** Structural variations in plant genomes. *Briefings in Functional Genomics* 13:296–307 DOI [10.1093/bfgp/elu016](https://doi.org/10.1093/bfgp/elu016).

**Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. 2010.** Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183 DOI [10.1038/nature08670](https://doi.org/10.1038/nature08670).

**Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D,**

Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MMS, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS, Jackson SA. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics* **46**:707–713 DOI [10.1038/ng.3008](https://doi.org/10.1038/ng.3008).

Sebastian P, Schaefer H, Telford IRH, Renner SS. 2010. Cucumber (*Cucumis sativus*) and melon (*C. melo*) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. *Proceedings of the National Academy of Sciences of the United States of America* **107**:14269–14273 DOI [10.1073/pnas.1005338107](https://doi.org/10.1073/pnas.1005338107).

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**:259 DOI [10.1186/s13059-015-0831-x](https://doi.org/10.1186/s13059-015-0831-x).

Sokolkova A, Burlyanova M, Valiannikova T, Vishnyakova M, Schafleitner R, Lee C-R, Ting C-T, Nair RM, Nuzhdin S, Samsonova M, Wettberg E. 2020. Genome-wide association study in accessions of the mini-core collection of mungbean (*Vigna radiata*) from the World Vegetable Gene Bank (Taiwan). *BMC Plant Biology* **20**:363 DOI [10.1186/s12870-020-02579-x](https://doi.org/10.1186/s12870-020-02579-x).

Somta P, Laosatit K, Yuan X, Chen X. 2022. Thirty years of mungbean genome research: where do we stand and what have we learned? *Frontiers in Plant Science* **13**:944721 DOI [10.3389/fpls.2022.944721](https://doi.org/10.3389/fpls.2022.944721).

Somta P, Sorajapinun W, Yimram T, Tanadul O-u-m, Laosatit K, Srinives P. 2024. Registration of 'KUML4' and 'KUML8' mungbean cultivars with high yield and large seeds. *Journal of Plant Registrations* **18**:33–41 DOI [10.1002/plr2.20308](https://doi.org/10.1002/plr2.20308).

Tang H, Krishnakumar V, Zeng X, Xu Z, Taranto A, Lomas JS, Zhang Y, Huang Y, Wang Y, Yim WC. 2024. JCVI: a versatile toolkit for comparative genomics analysis. *IMeta* e211.

Tempel S. 2012. Using and understanding RepeatMasker. *Methods in Molecular Biology* **859**:29–51 DOI [10.1007/978-1-61779-603-6\\_2](https://doi.org/10.1007/978-1-61779-603-6_2).

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM. 2012. Draft genome sequence of pigeonpea (*Cajanus cajan*). *Nature Biotechnology* **30**:83 DOI [10.1038/nbt.2022](https://doi.org/10.1038/nbt.2022).

Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B, Millan T, Zhang X, Ramsay LD, Iwata A, Wang Y, Nelson W, Farmer AD, Gaur PM, Soderlund C, Penmetsa RV, Xu C, Bharti AK, He W, Winter P, Zhao S, Hane JK, Carrasquilla-Garcia N, Condie JA, Upadhyaya HD, Luo M-C, Thudi M, Gowda CLL, Singh NP, Lichtenzveig J, Gali KK, Rubio J, Nadarajan N, Dolezel J, Bansal KC, Xu X, Edwards D, Zhang G, Kahl G, Gil J, Singh KB, Datta SK, Jackson SA, Wang J, Cook DR. 2013. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology* **31**:240–246 DOI [10.1038/nbt.2491](https://doi.org/10.1038/nbt.2491).

**Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC.** 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33:2202–2204 DOI [10.1093/bioinformatics/btx153](https://doi.org/10.1093/bioinformatics/btx153).

**Wang O, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, Sun Y, Anderson E, Lam HK, Chen D, Zhou Y, Wang L, Fan F, Zou Y, Xie Y, Zhang RY, Drmanac S, Nguyen D, Xu C, Villarosa C, Gablenz S, Barua N, Nguyen S, Tian W, Liu JS, Wang J, Liu X, Qi X, Chen A, Wang H, Dong Y, Zhang W, Alexeev A, Yang H, Wang J, Kristiansen K, Xu X, Drmanac R, Peters BA.** 2019. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Research* 29:798–808 DOI [10.1101/gr.245126.118](https://doi.org/10.1101/gr.245126.118).

**Wang Y, Tang H, De Barry JD, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H.** 2012. MCSpanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* 40:e49–e49 DOI [10.1093/nar/gkr1293](https://doi.org/10.1093/nar/gkr1293).

**Xiong W, He L, Lai J, Dooner HK, Du C.** 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences of the United States of America* 111:10263–10268 DOI [10.1073/pnas.1410068111](https://doi.org/10.1073/pnas.1410068111).

**Yang K, Tian Z, Chen C, Luo L, Zhao B, Wang Z, Yu L, Li Y, Sun Y, Li W, Chen Y, Li Y, Zhang Y, Ai D, Zhao J, Shang C, Ma Y, Wu B, Wang M, Gao L, Sun D, Zhang P, Guo F, Wang W, Li Y, Wang J, Varshney RK, Wang J, Ling H-Q, Wan P.** 2015. Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *Proceedings of the National Academy of Sciences of the United States of America* 112:13213–13218 DOI [10.1073/pnas.1420949112](https://doi.org/10.1073/pnas.1420949112).

**Yang Z.** 2007. PAML 4: phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24:1586–1591 DOI [10.1093/molbev/msm088](https://doi.org/10.1093/molbev/msm088).

**Yimram T, Somta P, Srinives P.** 2009. Genetic variation in cultivated mungbean germplasm and its implication in breeding for high yield. *Field Crops Research* 112:260–266 DOI [10.1016/j.fcr.2009.03.013](https://doi.org/10.1016/j.fcr.2009.03.013).