scientific data

OPEN

DATA DESCRIPTOR

Chromosome genome assembly and annotation of Adzuki Bean (*Vigna angularis*)

Wan Li[1,2,6], Fanglei He[3,4,6], Xueyang Wang[2,6], Qi Liu[2], Xiaoqing Zhang[2,4], Zhiquan Yang[3] ✉, Chao Fang[3] ✉ & Hongtao Xiang[1,5] ✉

Adzuki bean (*Vigna angularis*) is a significant dietary legume crop that is prevalent in East Asia. It also holds traditional medicinal importance in China. In this study, we report a high-quality, chromosome-level genome assembly of adzuki bean obtained by employing Illumina short-read sequencing, PacBio long-read sequencing, and Hi-C technology. The assembly spans 447.8 Mb, encompassing 96.32% of the estimated genome, with contig and scaffold N50 values of 16.5 and 41.0 Mb, respectively. More than 98.2% of the 1,614 BUSCO genes were fully identified, and 25,939 genes were annotated, with 98.23% of them being functionally identifiable. *Vigna angularis* was estimated to diverge successively from *Vigna unguiculata* and *Vigna radiata* about 15.3 and 8.7 million years ago (Ma), respectively. This chromosome-level reference genome of *Vigna angularis* provides a robust foundation for exploring the functional genomics and genome evolution of adzuki bean, thereby facilitating advancements in molecular breeding of adzuki bean.

Background & Summary

Adzuki bean [*Vigna angularis* (Willd.) Ohwi & Ohashi] is an annual cultivated crop belonging to the genus Vigna and subgenus Ceratotropis[1]. The grains can exhibit a range of colors including red, white, black, gray, and others[2,3]. The ideal temperature range for its growth is between 20–24 °C. Temperatures that are excessively high will result in elongated seedlings, while lower temperatures will impede developmental progress[4]. As a warm-season pulse crop, it is extensively cultivated in East Asia, particularly in China, Japan, and Korea. Presently, adzuki beans are grown in over 20 countries, with China and Japan being the leading producers[5]. Adzuki beans can be sown in spring, summer, or autumn, depending on climatic conditions. However, they are predominantly planted in spring in the northeastern regions of China, which represent the primary production areas[4]. In Japan, the adzuki bean ranks as the second most significant legume, following the soybean[6]. The annual cultivation areas of adzuki beans are estimated to be 670,000 hectares in China, 120,000 hectares in Japan, and 30,000 hectares on the Korean Peninsula[7].

The exact origin of the adzuki bean remains unclear; however, wild species such as *V. angularis var. nipponensis*, *V. nakashimae*, and *V. nepalensis* are broadly distributed throughout East Asia and the Himalayan countries[6]. The likely wild progenitor of the cultivated adzuki bean is *V. angularisvar* var. *nipponensis*, found in Japan, Korea, China, Nepal, Bhutan, and the Himalayan region, which exhibit substantial genetic diversity[8]. Additionally, archaeological evidence indicates that northeastern Asia was the primary site of adzuki bean domestication[6].

Adzuki bean is a diploid legume crop with 22 chromosomes (2n = 2x = 22)[7]. Several adzuki bean genomes have been published[7,9–11]. These assemblies range in size from 291 Mb to 522 Mb, with the contig N50 sizes varying from 13 kb to approximately 16 Mb (Table 1). These genome sequences have facilitated biological and genetic research on adzuki bean and other legume crops. However, it is important to note that the relatively short N50 contig sizes in these published genomes indicate limitations in the quality and continuity of the assembled sequences (Table 1). Moreover, one or a few reference genomes cannot represent the genomic diversity of an

[1]Institute of Crop Cultivation and Tillage, Heilongjiang Academy of Agricultural Sciences, Harbin, 150086, China. [2]Heilongjiang Academy of Agricultural Sciences, Harbin, 150086, China. [3]Innovative Center of Molecular Genetics and Evolution, School of Life Sciences Guangzhou University, Guangzhou, 510405, China. [4]College of Agriculture and Biotechnology, Yunnan Agricultural University, Kunming, 650201, China. [5]Suihua Branch, Heilongjiang Academy of Agricultural Machinery Sciences, Suihua, 152054, China. [6]These authors contributed equally: Wan Li, Fanglei He, Xueyang Wang. ✉e-mail: yang_zq@foxmail.com; fangchao@gzhu.edu.cn; zpszls@aliyun.com

Accession	Total length (Mb)	Contigs N50 (bp)	Scaffolds N50 (bp)	Number of coding genes	Assembly level
Gyeongwon[7]	444.4	26,637	8,174,047	26,857	Chromosome
Shumari[9]	522.8	1,575,115	38,860,970	31,310	Chromosome
Jingnong6[10]	467.3	38,390	1,292,063	34,183	Chromosome
Jingnong6[11]	489.7	16,063,027	41,615,786	32,748	Chromosome

Table 1. The published *Vigna angularis* genome.

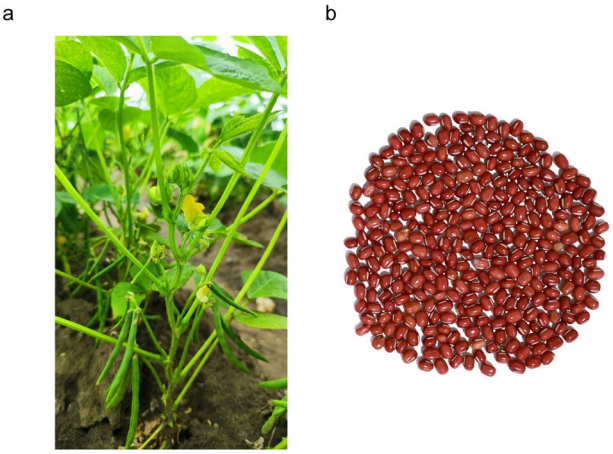

Fig. 1 The appearance of *Vigna angularis* (cultivar Longxiaodou 4). (**a**) Plant morphology of *Vigna angularis*. (**b**) Seeds of V*igna angularis*.

entire species[12]. Assembling one more high-quality reference genome for adzuki bean will promote further evolutionary and functional genomics studies and understanding of adzuki bean.

To obtain high-quality chromosome sequences of *V. angularis*, we applied Illumina, PacBio, and Hi-C technologies to sequence the genome. We report a 447.80 Mb high-quality genome of adzuki bean, with the contig N50 size of 16.53 Mb and 99.9% of the assembled bases associated with the 11 chromosomes. The BUSCO completeness (92.22%) and LAI index (15.23) both demonstrate the completeness and accuracy of this assembled genome. Based on our genome assembly, 54.50% (243.62 Mb) of the sequences were repeat sequences, with a dominance of long terminal repeats (LTRs), accounting for 40.54% of the repeat sequences. This genome will facilitate the functional gene mapping of adzuki bean's economically important traits, enhance the comprehension of their underlying biological mechanisms, and thereby expedite genomics applications in its breeding.

Methods

Sample collection and sequencing. Longxiaodou 4, an adzuki bean cultivar, is extensively cultivated in Heilongjiang Province, China. Its plant is compact, stands upright, and resistant to lodging. Longxiaodou 4 has oval-shaped grains and red seed coats, and it weighs about 20 grams per 100 seeds. In this study, Longxiaodou 4 (Fig. 1) was utilized for genome sequencing and assembly. Plants were cultivated under long-day conditions (16 hours light, 8 hours dark) at 24 °C in a controlled growth cabinet (Xunneng Instrument, Beijing, China). Genomic DNA was extracted from leaf tissue using the Qiagen DNA Purification Kit (Qiagen, Valencia, CA, USA).

To obtain sufficient read data for genome assembly, both the PacBio SEQUEL II (Pacific Biosciences, California, USA) and the Illumina HiSeq 4000 platforms (Illumina, California, USA) were employed. Long reads from the PacBio platform were used for genome assembly, while the short, precise reads from the Illumina platform were used for genome survey and base-level correction after the assembly. For the PacBio platform, 20-kb genomic sequencing libraries were constructed according to the PacBio-suggested protocol, yielding 65.71 Gb of long sequencing reads. After adaptor removal, 65.58 Gb of subreads (coverage of 149.34x) were obtained, with subread N50 and average lengths of 22.16 kb and 13.90 kb, respectively (Table 2). Besides. the DNA was utilized to construct sequencing library using the Illumina TruSeq DNA Sample Prep Kit (Illumina, San Diego, CA, USA). Paired-end sequencing with a read length of 150 base pairs (bp) was conducted on an Illumina HiSeq 4000 system (Illumina, San Diego, CA, USA). This process generated a total of 98.10 Gb of short sequencing reads (coverage of 222.95x). Reads containing adaptors and those with quality scores below 20 were excluded. Consequently, 96.03 Gb of high-quality reads were obtained for *de novo* genome assembly (Table 2).

The same individual used for genomic sequencing was also employed for transcriptome sequencing to provide essential gene expression data for genome annotation. Gene expression showed distinct tissue specificity, and therefore, RNA was extracted from root, stem, and leaf tissues. RNA extraction was performed using the RNAiso Pure RNA Isolation Kit (Takara, Japan), followed by DNase I treatment to remove DNA contamination. RNA quality was assessed with a NanoVue Plus spectrophotometer (GE Healthcare, NJ, USA). RNA-seq libraries

Library type	Insert size (bp)	Raw data (Gb)	Clean data (Gb)	Read length (bp)	Sequence coverage (X)
Illumina reads	250	98.10	96.03	150	222.95
Pacbio reads	20,000	65.71	65.58	13,902	149.34
Hi-C	—	69.83	68.24	150	158.70
RNA reads	250	7.89	7.72	150	17.93
Total	—	241.53	237.57	—	548.93

Table 2. Sequencing data used for the *Vigna angularis* (cultivar Longxiaodou 4) genome.

Kmer	Genome size (Mb)	Heterozygous ratio (%)	Repeat (%)
21	469.94	0.54	43.878

Table 3. Statistics of the 21-mer analysis of the *Vigna angularis* (cultivar Longxiaodou 4) genome. All 21-mers used for the evaluation of the *Vigna angularis* (cultivar Longxiaodou 4) genome were extracted from the reads that passed strict quality control (i.e., Q20 greater than 98%).

Chromosome	Front end of the chromosome	Back end of the chromosome
Chromosome 1	0	586
Chromosome 2	0	173
Chromosome 3	32	0
Chromosome 4	0	378
Chromosome 5	29	0
Chromosome 6	0	72
Chromosome 7	0	0
Chromosome 8	0	0
Chromosome 9	104	0
Chromosome 10	151	2349
Chromosome 11	0	31

Table 4. Statistical results of telomere identification.

were prepared and sequenced on the Illumina HiSeq 4000 platform, yielding a total of 7.89 Gb of transcriptome data. A summary of the genome and transcriptome sequencing data is presented in Table 2.

***De novo* assembly of the adzuki bean genome.** Short next-generation sequencing (NGS) reads were employed for estimating the genome size, heterozygosity, and repeat content of *V. angularis* before the *de novo* genome assembly. Jellyfish (v2.1.3)[13] was adopted to count the number of 21-mers, which was used to calculate the basic information of the genome (Table 3). The genome size of *V. angularis* was estimated at 464.9 Mb, with heterozygosity of 0.54% and a repeat content percentage of 43.878%.

The long reads from the PacBio SEQUAL sequencing platform were utilized for the contig assembly using Canu (v2.0)[14], with the parameter of the corrected ErrorRate set to 0.045 and corOutCoverage set to 40, respectively. Approximately 150-fold coverage of the estimated genome size was generated after self-correction. The primary assembled genome size was 495 Mb, with a contig N50 of 16.14 Mb. To revise the random error introduced by the PacBio sequencing reads, this assembled genome sequence was polished with the long reads obtained with Racon (v1.3.3)[15] and then further polished with the short reads obtained with Pilon (v1.23)[16]. Purge_haplotigs (v1.0.4)[17] was used to purge the heterozygous and redundancy regions of the polished sequences. Ultimately, a high-quality genome of *Vigna angularis* was obtained, featuring a total size of 447.80 Mb, with a contig N50 of 16.53 Mb and a total of 47 contigs.

The completeness and accuracy of the assembled genome were then evaluated with multiple methods. BUSCO (Benchmarking Universal Single-Copy Orthologs, v3.0.0)[18] was employed to assess the completeness of the single-copy genes from the orthologs database, with 95.42% complete and 0.68% partial of a total of 1,614 genes in the embryophyta_odb10 database identified, respectively. LTR_FINDER (v1.0.7)[19] and LTR_retriever (v2.7)[20] were used to search the LTR elements and calculate the LAI score of the genome, which was 15.23. The short NGS reads and long PacBio reads were aligned into the genome. Of the short reads, 97.95% were mapped to the genome, and the coverage was 99.98%, with 96.24 and 99.97% for the long reads for these two values. The BUSCO, LAI index, and read mapping ratio results proved the completeness and accuracy of this assembled genome.

Telomere sequence identification was performed based on the characteristic base repeat sequences in the telomere regions (signature sequences: CCCTAAA/TTTAGGG). The details are presented in Table 4.

Chromosome construction using the interaction information from Hi-C data. The Hi-C technique has proven its efficacy in chromosome assembly and has been successfully employed in numerous genomic

	Length (bp) (Contigs)	Number (Contigs)	Length (bp) (Scaffolds)	Number (Scaffolds)
Max	32,757,015	—	65,407,200	
N10	31,838,529	2	65,407,200	1
N20	28,356,209	3	52,846,339	2
N30	24,060,476	5	42,681,439	3
N40	17,512,071	8	42,641,100	4
N50	16,516,761	10	41,033,850	5
N60	14,791,398	13	38,121,061	6
N70	11,394,282	16	37,163,307	7
N80	9,817,593	21	32,215,300	9
N90	6,667,416	26	31,493,818	10
Total	447,803,293	—	447,806,493	—
Number >= 100 bp	—	66	—	34
Number >= 2000 bp	—	56	—	25
GC_rate	0.336	—	0.336	—
Total N bases	3,200 (0.0007%)	—	—	—

Table 5. Assembly statistics of the *Vigna angularis* (cultivar Longxiaodou 4) genome. Note that the contigs here refer to the continuous sequences after the Hi-C data–based chromosome construction.

	Length (bp) (RepBase TEs)	% in genome (RepBase TEs)	Length (bp) (TE Proteins)	% in genome (TE Proteins)	Length (bp) (*De novo* TEs)	% in genome (*De novo* TEs)	Length (bp) (Combined TEs)	% in genome (Combined TEs)
DNA	13,884,521	3.10	4,101,497	0.92	43,782,983	9.78	51,548,926	11.51
LINE	1,585,845	0.35	70,942	0.02	1,823,273	0.41	3,321,407	0.74
SINE	27,721	0.01	0	0.00	65,063	0.01	82,278	0.02
LTR	46,931,771	10.48	34,285,845	7.66	173,532,636	38.75	181,528,806	40.54
Satellite	473,622	0.11	0	0.00	8,875	0.00	482,369	0.11
Simple_repeat	0	0.00	0	0.00	11,922	0.00	11,922	0.00
Other	4,076	0.00	0	0.00	0	0.00	4,076	0.00
Unknown	51,238	0.01	10,830	0.00	26,552,182	5.93	26,613,717	5.94
Total	61,391,689	13.71	38,466,903	8.59	229,681,274	51.29	244,042,507	54.50

Table 6. Summary statistics of the repeats' annotation of the *Vigna angularis* (cultivar Longxiaodou 4) genome. Note that RepBase TEs and TE Proteins represent the results of RepeatMasker and RepeatProteinMask based on Repbase; *de novo* TEs are the result of RepeatMasker based on RepeatModeler, RepeatScout, and LTR_FINDER; combined TEs refer to the combined results of *de novo* + Repbase and TE proteins.

projects[21]. In this work, we used leaves from the same individual as in the genome sequencing for the Hi-C library construction and sequencing. Approximately 69.8 GB of the raw reads were generated from the Illumina platform, filtered, and subsequently utilized for further analyses. The sequencing reads were mapped to the polished adzuki bean genome with BWA 0.7.17[22]. Pair-end short reads were mapped to the genome, and only the uniquely mapped read pairs were selected. Juicer 1.5.6[23] was applied to process the Hi-C reads, and the interaction frequency was quantified and normalized. Then, 3D-DNA[24] was applied to identify and correct the errors in the initial assembly and orient and cluster the contigs according to the Hi-C contact matrix. Consequently, 11 groups were successfully clustered, which were further ordered and oriented into chromosomes. Finally, 447.5 Mb contigs were reliably anchored on chromosomes, accounting for 99.9% of the total genome. The contig and scaffold N50 reached 16.5 and 40.0 Mb (Table 5), respectively, providing a high-quality chromosomal genome assembly for adzuki bean.

Repetitive element annotation. Repetitive sequences of the adzuki bean genome were identified through a combination of ab initio and homology-based prediction approaches. For the ab initio repeat annotation, LTR_FINDER[15], RepeatScout[25], and RepeatModeler (http://repeatmasker.org/RepeatModeler/) were used to construct a *de novo* repetitive element database, and RepeatMasker[26] (http://repeatmasker.org/RMDownload.html) was used to annotate the repeat elements with the database. RepeatMasker and RepeatProteinMask were used to identify repeats at the DNA and protein level by mapping to the Repbase database[27]. Tandem repeats were also ab initio annotated with Tandem Repeat Finder[28]. A total of 243.62 Mb repeat sequences were identified, accounting for 54.50% of the genome (Table 6).

Protein-coding gene prediction and functional annotation. A combined approach involving *de novo* prediction, homology-based prediction, and transcriptome-based prediction was used for the protein-coding gene prediction. The RNA-seq reads from multiple tissues such as root, stem, and leaf were cleaned and mapped to the *Vigna angularis* genome using HISAT2[29]. Subsequently, StringTie[30,31] was employed to identify the potential

Methods/Tools	Average transcript length (bp)	Average CDS length (bp)	Average Exons length (bp)	Average Introns length (bp)	Exon number per gene
Ab initio (Augustus)	3,898	1,154	208	602	5.56
Ab initio (Genscan)	9,866	1,199	206	1,797	5.82
Homolog (*Abrus precatorius*)	5,183	953	273	1,698	3.49
Homolog (*Vigna radiata var radiata*)	2,891	828	284	1,079	2.91
Homolog (*Vigna angularis*)	3,722	980	288	1,144	3.4
Homolog (*Vigna unguiculata*)	1,029	589	449	1,420	1.31
Homolog (*Arachis hypogaea*)	1,184	749	546	1,170	1.37
RNA-seq	5,574	1,484	303	577	6.99
BUSCO	5,801	1,586	170	505	9.35
MAKER	5,597	1,337	254	726	6.45
Final set	4,564	1,265	281	614	5.79

Table 7. Statistics of the gene models of the protein-coding genes annotated in the *Vigna angularis* (cultivar Longxiaodou 4) genome.

Type	Proteins	Percentage (%)
Complete BUSCOs (C)	1,585	98.2
Complete and single-copy BUSCOs (S)	1,540	95.42
Complete and duplicated BUSCOs (D)	45	2.79
Fragmented BUSCOs (F)	11	0.68
Missing BUSCOs (M)	18	1.12
Total BUSCO groups searched	1,614	100.00

Table 8. BUSCO results for the gene model with the embryophyta_odb10 database.

exon regions, and TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki) was utilized to predict the Open Reading Frames (ORFs) based on the transcript sequences. The homologous protein sequences of *Abrus precatorius*, *Vigna radiata var. radiate*, *Vigna angularis* (old version), *Vigna unguiculata*, and *Arachis hypogaea* were downloaded from NCBI and mapped to the adzuki bean genome using TBLASTN[32]. The blast results were conjoined, and accurate coding sequences of the corresponding genomic region on each blast hit were predicted using Exonerate (https://github.com/nathanweeks/exonerate). The *de novo* gene structures were predicted using AUGUSTUS[33] and Genscan[34] based on the repeat-masked genome sequence. As a result, a gene set of 25,939 high-quality protein-coding genes was obtained after integrating all the gene structure results from the ab initio, homology, and transcriptome results by MAKER[35] (Table 7). Gene annotation completeness was assessed using embryophyta BUSCOs, finding 98.2% completeness[18] (Table 8). The distribution of gene element length was compared to the homology species above (Fig. 2).

Using the publicly available databases TrEMBL, Swiss-Prot[36], InterPro[37], NCBI non-redundant protein, euKaryotic Orthologous Groups[38], Gene Ontology[39], and Kyoto Encyclopedia of Genes and Genomes[40], 25,479 predicted genes (approximately 98.23% of all) were functionally annotated with at least one of these databases.

Data Records

The sequencing datasets and genome assembly were deposited in NCBI under the accession PRJNA629451. This whole genome shotgun project has been deposited in GenBank under the accession JABFOF000000000[41]. The version described in this paper is version JABFOF010000000[41]. The Illumina genomic sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) under the project number SRR11787767[42]. The PacBio genome sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) under the project number SRR11787766[43]. The transcriptome Illumina sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) under the project number SRR11787768[44]. The genomic Hi-C sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) under the project number SRR11787765[45].

Technical Validation

The quality of the DNA and RNA molecules and libraries for genomic sequencing and transcriptome sequencing were validated before the sequencing. The extracted DNA spectrophotometer ratios were $260/280 \geq 1.6$, both for the Illumina and PacBio sequencing. DNA > 2 and $20\,\mu g$ was used for the Illumina and PacBio sequencing. The concentration and quality of the total RNA were evaluated using the NanoVue Plus spectrophotometer (GE Healthcare, NJ, USA). RNAs samples with a total RNA amount $\geq 10\,\mu g$, RNA integrity number ≥ 8, and rRNA ratio ≥ 1.5 were finally used to construct the sequencing library.The genome assembly demonstrated a BUSCO completeness of 98.2%, with 95.42% single-copy BUSCOs, 2.79% duplicated BUSCOs, 0.68% fragmented BUSCOs, and 1.12% missing BUSCOs.

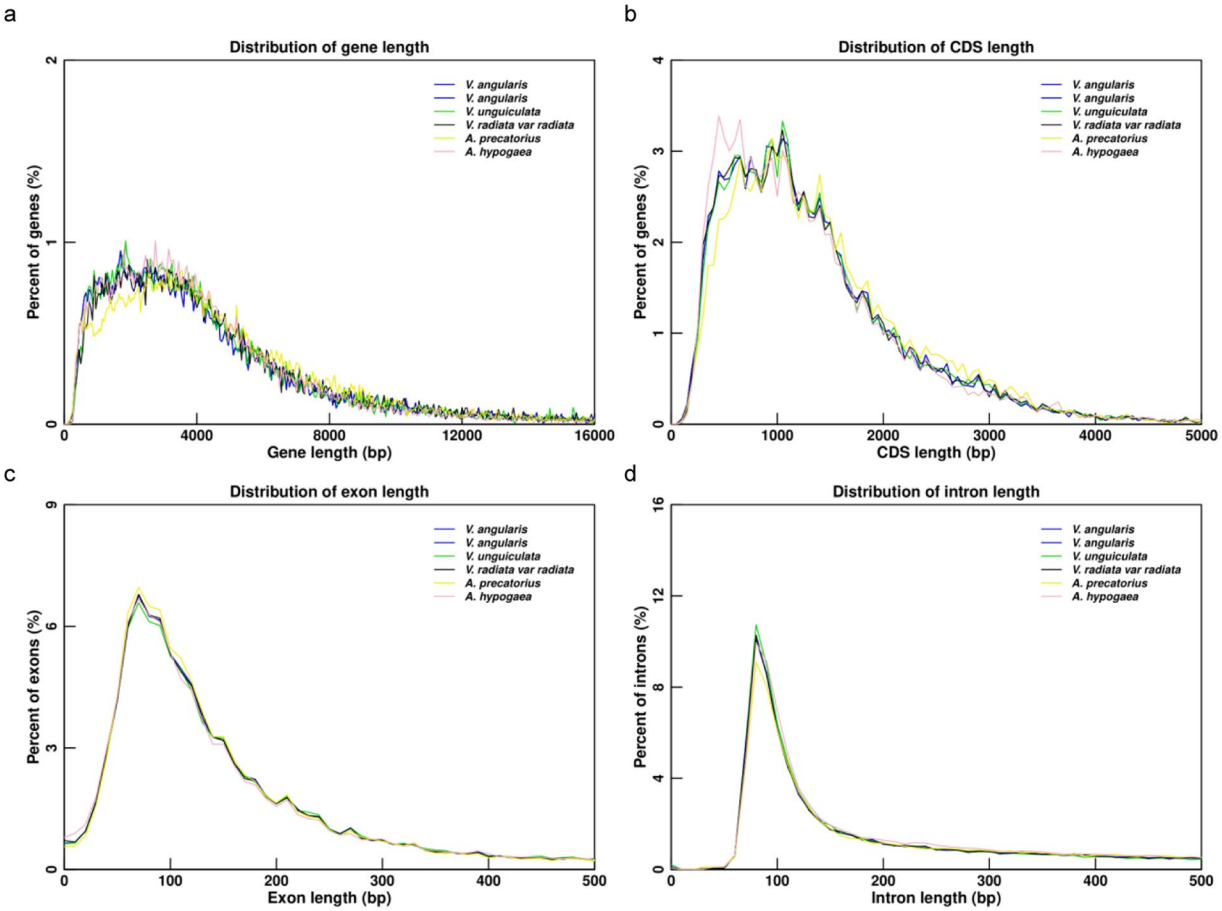

Fig. 2 Comparisons of the prediction gene models in the *Vigna angularis* (cultivar Longxiaodou 4) genome to other species. (**a**) Comparison of gene length between *Vigna angularis* and other species. (**b**) Comparison of CDS length between *Vigna angularis* and other species. (**c**) Comparison of exon length between *Vigna angularis* and other species. (**d**) Comparison of intron length between *Vigna angularis* and other species.

Code availability

No specific code or script were used in this work. All commands used in the data processing were executed according to the manual of the instrument of the corresponding bioinformatics software.

Received: 19 June 2023; Accepted: 23 September 2024;
Published online: 02 October 2024

References

1. Xie, Y., Xu, J. H., Lu, W. Y. & Lin, G. Q. Adzuki bean: a new resource of biocatalyst for asymmetric reduction of aromatic ketones with high stereoselectivity and substrate tolerance. *Bioresour Technol.* **100**, 2463–8 (2009).
2. Yook, J. S. *et al.* Black Adzuki bean (*Vigna angularis*) attenuates high-fat diet-induced colon inflammation in mice. *J Med Food.* **20**, 367–375 (2017).
3. Chu, L. *et al.* Genetic analysis of seed coat colour in adzuki bean (*Vigna angularis* L.). *Plant Genet Resour.* **19**, 67–73 (2021).
4. Xiang, H. *et al.* Uniconazole foliar spray treatment alleviates cold stress in adzuki bean (*Vigna angularis)* seedlings. *Intl J Agric Biol.* **23**, 235–240 (2020).
5. Kramer, C. *et al.* Control of volunteer adzuki bean in soybean. *Agri Sci.* **3**, 501–509 (2012).
6. Jameel, M., Al-Khayri, ShriMohan Jain, Dennis V. Johnson. Advances in plant breeding strategies: Legumes. Springer Nature Switzerland AG. Chapter 1 (2019)
7. Kang, Y. J. *et al.* Draft genome squence of adzuki bean, *Vigna angularis*. *Sci Rep.* **5**, 8069 (2015).
8. Yamaguchi, H. Wild and weed azuki beans in Japan. *Econ Bot.* **46**, 384–394 (1992).
9. Sakai, H. *et al.* The power of single molecule real-time sequencing technology in the *de novo* assembly of a eukaryotic genome. *Sci. Rep.* **5**, 1–13 (2015).
10. Yang, K. *et al.* Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *Proc. Natl. Acad. Sci. USA* **112**, 13213–13218 (2015).
11. Chu, L. *et al.* Chromosome-level reference genome and resequencing of 322 accessions reveal evolution, genomic imprint and key agronomic traits in adzuki bean. *Plant Biotechnol. J.* https://doi.org/10.1111/pbi.14337 (2024).
12. Liu, Y. *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162–176 (2020).
13. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764 (2011).
14. Sergey, K. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
15. Robert, V. *et al.* Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

16. Bruce, W. *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS One*. **9**, 112 (2014).
17. Roach, M. J. *et al*. Purge Haplotigs: Synteny Reduction for Third-gen Diploid Genome Assemblies. *BMC Bioinformatics*. **19**, 460 (2018).
18. Simao, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. **31**, 3210–3212 (2015).
19. Zhao, X. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. **35**, 265–268 (2007).
20. Ou, S. J. & Jian, N. LTR_retriever: a highly accurate and sensitive program for identification of 2 long terminal-repeat retrotransposons. *Plant Physiol*. **176**, 1410–1422 (2017).
21. Nicolas, S. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. **16**, 259 (2015).
22. Jung, Y. & Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics*. **38**, 2404–2413 (2022).
23. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98 (2016).
24. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*. **356**, 92–95 (2017).
25. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics*. **21**, i351–i358 (2005).
26. Tempel, S. Using and Understanding RepeatMasker. *Methods Mol Biol*. **859**, 29–51 (2012).
27. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. **6**, 11 (2015).
28. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. **27**, 573–580 (1999).
29. Kim, D., Langmead, B. & Salzberg, S. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. **12**, 357–360 (2015).
30. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. **33**, 290–295 (2015).
31. Pertea, M. *et al*. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. **11**, 1650–1667 (2016).
32. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol*. **4**, 41 (2006).
33. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*. **33**, W465–W467 (2005).
34. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. **268**, 78–94 (1997).
35. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 22, **12**, 491 (2011).
36. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. **45**, D158–D169 (2016).
37. Finn, R. D. *et al*. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res*. **45**, D190–D199 (2016).
38. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. **11**, 41 (2003).
39. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. **47**, D330–D338 (2019).
40. Kanehisa, M. *et al*. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids Res*. **42**, D199–205 (2014).
41. Xiang, H. whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JABFOF000000000 (2020).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11787767 (2020).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11787766 (2020).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11787768 (2020).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR11787765 (2020).

Acknowledgements

This study was supported by Heilongjiang Key R&D Program project (GA21B009-14), China Agriculture Research System (CARS-08-G8), Heilongjiang Provincial Natural Science Foundation of China (LH2021C078).

Author contributions

X.H. and F.C. conceived and devised the experimental design. L.W. and H.F. conducted the experiments. H.F., W.X., L.Q., Z.X. and Y.Z. performed the data analysis. F.C. and L.W. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.Y., C.F. or H.X.

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

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