Open access • Posted Content • DOI:10.1101/2020.05.16.098434

Chromosome scale reference genome of Cluster bean (Cyamopsis tetragonoloba (L.) Taub.) — Source link

Kishor Gaikwad, Guda Ramakrishna, Harsha Srivastava, Swati Saxena ...+13 more authors

Institutions: Indian Council of Agricultural Research, Central Arid Zone Research Institute, Indian Agricultural Research Institute, Indian Agricultural Statistics Research Institute

Published on: 18 May 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Guar, Reference genome, Guar gum and Genome

Chromosome scale reference genome of Cluster bean (*Cyamopsis tetragonoloba* (L.) Taub.)

Kishor Gaikwad*[1], G. Ramakrishna[1], Harsha Srivastava[1], Swati Saxena[1],Tanvi Kaila[1], Anshika Tyagi[1], Priya Sharma[1], Sandhya Sharma[1], R Sharma[2], HR Mahla[2], Amitha Mithra SV[1], Amol Solanke[1], Pritam Kalia[3], AR Rao[4], Anil Rai[4],TR Sharma[5], NK Singh[1].

[1]ICAR-National Institute for Plant Biotechnology, New Delhi, India

[2]ICAR-Central Arid Zone Research Institute, Jodhpur, India

[3]Div. of Vegetable Sciences, ICAR-Indian Agricultural Research Institute, New Delhi, India

[4]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

[5]Indian Council of Agricultural Research, New Delhi, India

*Correspondence:kish2012@nrcpb.org

Abstract: Clusterbean (*Cyamopsis tetragonoloba* (L.) Taub.), also known as Guar is a widely cultivated dryland legume of Western India and parts of Africa. Apart from being a vegetable crop, it is also an abundant source of a natural hetero-polysaccharide called guar gum or galactomannan which is widely used in cosmetics, pharmaceuticals, food processing, shale gas drilling etc. Here, for the first time we are reporting a chromosome-scale reference genome assembly of clusterbean, from a high galactomannan containing popular guar cultivar, RGC-936, by combining sequenced reads from Illumina, 10x Chromium and Oxford Nanopore technologies. The initial assembly of 1580 scaffolds with an N50 value of 7.12 Mbp was generated. Then, the final genome assembly was obtained by anchoring these scaffolds to a high density SNP map. Finally, a genome assembly of 550.31 Mbp was obtained in 7 pseudomolecules corresponding to 7 chromosomes with a very high N50 of 78.27 Mbp. We finally predicted 34,680 protein-coding genes in the guar genome. The high-quality chromosome-scale cluster bean genome assembly will facilitate understanding of the molecular basis of galactomannan biosynthesis and aid in genomics-assisted breeding of superior cultivars.

Introduction:

Clusterbean (*Cyamopsis tetragonoloba* (L.) Taub.), also known as guar[1] is a member of Leguminosae family. The common name clusterbean is attributed to its pods which appear in

clusters. Previous reports suggest that guar originated in Africa and later spread to the entire South Asian region. In India and Pakistan, clusterbean is cultivated since ancient times for its tender pods which are used as fresh vegetable and the remaining plant serves as fodder[2]. Clusterbean is a climate-resilient annual legume and a high potential alternative crop in the marginal lands of arid and semi-arid regions[3]. The genus *Cyamopsis* includes four species i.e., one cultivated *C. tetragonoloba* (L.) Taub., two wild relatives *C. serrata Schinz,* and *C. senegalensis Guill&Perr, and C. dentate Tarre*, and an interspecies hybrid of *C. serrata and C. senegalensis*[4.] A mature clusterbean seed is composed of three parts: germ (43-47%), endosperm (35-42%), and seed coat (14-17%). About 80-90% of the endosperm is composed of highly viscous water-soluble hetero-polysaccharide called gaur gum (or) galactomannan, having a 1:2 ratio of galactose to mannose[5]. Guar gum is extensively utilized as natural thickener, emulsifier and stabilizers in the food, textile, paper, petroleum and pharmaceutical industry with increasing global demand[6–9]. With the annual production of ~1-1.25 million tons of clusterbean seeds, India accounts for 80% of the global production, with several other countries, like Pakistan, United States, China, Australia and Africa contributing the rest. About 45% of total world demand is due to industrial application of guar gum[10]. Apart from being a rich source of commercial product like gum, clusterbean is also a highly nutritious legume crop, predominantly composed of protein (18%) and dietary fiber (32%)

Earlier cytogenetic studies in clusterbean revealed that ~580.9 Mb of the genome is organized in 2n=14 number of chromosomes[11,12]. Despite the considerable industrial importance, only a few studies have been carried out at genome level which includes genome size estimation (cultivated vs. wild type)[12], plastid genome sequencing (Chloroplast)[13], transcriptome analysis[14,15], small RNA sequencing[16] to identify novel miRNA associated with galactomannan biosynthesis as well as genetic diversity analysis based on SSRs[17] ,mostly from our group. Therefore, it was necessary to sequence a chromosome-scale high-quality reference genome to understand the molecular basis of galactomannan biosynthesis, synteny with other legumes and discovery of genes for other important traits. This will also enhance cluster bean genetic improvement via genomics assisted breeding.

Experimental Methods and Results:

Plant sample collection, genomic library preparation and sequencing

Seeds of the pure homozygous cluster bean variety, 'RGC-936', obtained from ICAR-CAZRI were sown in pots at ICAR-NIPB, New Delhi, India. Leaf samples were collected and snap frozen in liquid nitrogen and stored at -80^{o}C till further use. The genomic DNA was extracted using CTAB method[18] and the integrity and quantity of DNA were tested by separating the DNA on a 0.8% agarose gel and DeNovix DS-11 spectrophotometer, respectively. The high-quality DNA was used for genome sequencing by Illumina and HMW DNA was used for 10X Genomics and Oxford Nanopore sequencing. Similarly leaf samples of a F2 population (RGC 936 x CAZRI-15-3-8) were processed for GBS sequencing (Illumina).

Genome sequencing:

In the present study, we selected both short and long-read sequencing methods such as Illumina, 10X Genomics and Oxford Nanopore sequencing technology (ONT) for clusterbean genome sequencing.

For Illumina short-read sequencing, high-quality genomic DNA was randomly fragmented by the M220 Focused-ultra sonicator system (Covaris Inc, USA). Two genomic DNA libraries of 500-1000bp insert size were prepared using the TruSeq DNA PCR-Free Sample Preparation Kit, as per the manufacturer's guidelines. To realize sequence variation and high genome coverage (length), two separate Mate-Pair (MP) libraries of 3Kb and 7Kb insert size were prepared using Nextera Mate-Pair Sample Preparation Kit (Illumina, San Diego, CA). Both the Illumina PE and MP libraries were sequenced on Illumina HiSeqX-ten platform which produced 16.09Gbp and 42.3Gbp of (2X150bp) sequencing data respectively.

For 10X Genomics sequencing, a total of 8 Gemcode libraries were prepared from high-quality DNA fragments longer than 50Kb, using the Chromium instrument. Sequencing of these libraries was performed on Illumina, HiSeqX-ten platform, generating 2x150bp reads; resulting in a total of 92.767 Gbp of 10X Genomics linked read raw sequencing data.

For long-read Nanopore sequencing, genomic DNA was size-selected using BluePippin BLF7510 cassette (Sage Science) and high-pass mode (> 20 kb) and library was prepared using Oxford Nanopore Technologies (ONT) standard ligation sequencing kit SQK-LSK109 following the SQK-MAP005 PromethION protocol. Two libraries were prepared and sequenced using PromethION. A total of 50.64 Gbp of Nanopore sequencing data was generated.

As a result, we generated a total of 201.8 Gbp raw sequencing data corresponding to 366.73X genomic coverage (depth), of cluster bean genome (543.22 Mbp estimated genome size using *k*-mer frequency distribution analysis). Details of sequencing data are represented in **Table 1**.

Genome size estimation:

In the present study, the cluster bean genome size was estimated using the *k*-mer frequency distribution analysis of short reads sequencing data with a *k*-mer size of 31. The 31-mer abundance was calculated using Jellyfish[19] v2 and the cluster bean genome size was then estimated using Genomescope software[20]. Total 1,187,123,081 *k*-mers (31-mer) were counted and their frequency distributions were analyzed. In the 31-mer frequency distribution histogram, the main peak was observed at depth of 65 corresponding to homozygous haploid sequences. A small peak was also observed at half of the main peak depth representing heterozygous fraction of the genome (**Figure 1a**). The *k*-mer frequency distribution histogram produced by Jellyfish was then subjected to Genomescope (http://qb.cshl.edu/genomescope/genomescope2.0/) for estimating size and heterozygosity of the genome. Thus, cluster bean genome size was estimated to be 543.2 Mb (543,218,592 bp), and the fraction of heterozygosity in cluster bean genome estimated to be in the range of 0.70 - 0.71%.

***De novo* genome assembly:**

We followed a hybrid assembly approach for developing high-quality chromosome-level genome of cluster bean. Initially, we generated two separate *de novo* assemblies by using 10X genomics linked reads and Oxford Nanopore long reads. We then merged both the assemblies to get minimal sequencing errors and a high degree of sequence continuity for scaffolds.

The raw 10X genomics sequencing data was assembled using Supernova (v2.1.1)[21], which generates raw assembly as well as scaffold assembly. Removal of vector and mitochondrial contamination was done using seqclean tool through univec vector database (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/). We generated a scaffold assembly of 617.76 Mbp comprising 1616 contigs with an N50 size of 4.27Mbp, which was 13.7% longer than the estimated cluster bean genome size.

The long Nanopore sequencing reads were utilized to generate a *de novo* assembly using Canu (v1.6)[22], which generated 3904 contigs spanning 419.66Mbp with N50 of 775kbp. This primary

assembly was further polished with Illumina shotgun and Mate-Pair data to generate an improved assembly of length 441.85Mbp, which comprised of 1548 contigs with N50 of 544.474kbp.

The primary assemblies from 10X supernova and Oxford Nanopore (Canu) were merged with npScarf[23] and generated a highly contiguous assembly of 550.16Mb comprising 1580 scaffolds with an N50 value of 7.12Mb and longest scaffold of 35.03Mb.

Further, GBS reads of 142 members of a F_2 population developed from RGC-936/ CAZRI-15-3-8 were aligned against the genome assembly using bwa mapping software and SNP calling was done using Unified Genotyper from the Genome Analysis Toolkit GATK (v3.6). According to UGbs-Flex pipeline, SNPs with allele frequencies <0.1 and >0.9 and adjacent SNPs were discarded. Further markers showing segregation distortion from the expected 1:2:1 Mendelian ratio were discarded on the basis of $\chi 2$ test ($p < 0.05$). We obtained 6113 markers that were imported in JoinMap (v4)[24], program in Kyazma software package for creating the linkage map containing seven linkage groups (https://www.kyazma.nl/index.php/). Also, the linkage groups were determined at logarithm of odds (LOD) score of 6.0. A total of 1529 scaffolds having 6113 markers in seven linkage groups were merged into specific pseudomolecules using in-house Perl script. Finally, seven genetically anchored pseudomolecules/ chromosomes along with 51 unanchored contigs resulting in the final assembly of 550.30Mb for the clusterbean genome. The chromosome length of the clusterbean genome ranged from 61.32 Mbp (Chr7) to 93.95 Mbp (Chr1) with a scaffold N50 of 78.27Mb **(Table 2)**. The organization of clusterbean genome assembly, gene density, DNA repeat elements, SSRs and duplicate genes were shown in **Figure 1b.** Further analysis is being carried out to improve the assembly to final completion.

Identification and annotation of repetitive DNA sequences:

For identification and annotation of repetitive DNA sequences in the cluster bean genome assembly, we used a *de novo* repeat library and Dfam[25] (v3.1) database. First, RepeatModeler (v1.0.10, http://www.repeatmasker.org/RepeatModeler/) was employed to make a *de novo* repeat library of cluster bean genome assembly. Next, we annotated the cluster bean *de novo* repeat library by using Repeatmasker (v4.0.7). Then BLASTn search was performed to annotate unclassified elements from Repeatmasker with the repetitive elements in the Dfam database (https://www.dfam.org/releases/Dfam_3.1/). We identified 582339 DNA repeat sequences

covering 42.14% of the cluster bean genome. The most abundant repetitive element type was the retrotransposons making up 29.73% of the genome, including 8.08% of LINES, 21.62% of long terminal repeats (LTRs), 23.3% of DNA transposons and 5.56% of unclassified repetitive elements **(Table 3)**.

Simple sequence repeats (SSRs) or microsatellites in the cluster bean genome were identified by MISA[26] program with the following parameters: monomer ($n \geq 10$), dimer ($n \geq 6$), trimer ($n \geq 5$), tetramer ($n \geq 5$), pentamer ($n \geq 5$), and hexamer ($n \geq 5$). A total of 238176 SSRs, covering 0.46% (2.52 Mb) of the cluster bean genome were detected. Among the SSRs, monomers were the most abundant type (71.57%), followed by dimer (16.91%), trimer (9.84%), tetramer (1.31%), pentamer (0.20%), and hexamer (0.16%), respectively.

Gene model prediction and functional annotation:

From the cluster bean genome assembly, the protein-coding genes were predicted using the Seqping (version 0.1.45)[27] pipeline. Seqping provides species-specific, unbiased gene predictions, thus most suitable for gene prediction of non-model plant genomes like cluster bean. The RNA sequencing data generated earlier by us[15] was used for transcript assembly with Trinity (version 2.1.1)[28] and utilized in the Seqping pipeline. Initially, Seqping predicted 37509 protein-coding genes and after clustering with CD-HIT (version 4.6)[29], 34680 non-redundant gene predictions were obtained.

The efficiency of the gene prediction was evaluated with Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis, using the Plant embryophyta_odb10 lineage with Arabidopsis species (BUSCO v3.1.0)[30]. The BUSCO reported 96.90% of gene predictions as complete including 64% complete single-copy and 32.9% duplicated genes, 2.6% of missing gene models and 0.5% of fragmented gene models **(Figure 1c)**, indicating a high efficiency of gene prediction.

The functional annotation and Gene ontology (GO) terms for each predicted gene model were allocated via InterProScan 5 (version 5.25-64.0)[31] and Blast2GO (version 4.1)[32] respectively. About 28955 (78.93%) genes were annotated successfully.

Identification and annotation of Non-coding RNA genes:

Non-coding RNAs (ncRNAs) including tRNAs, rRNAs, and snRNAs in cluster bean genome were identified and annotated using various software packages and databases. First, tRNAscan-

SE (version 1.3.1) with default parameters was used to identify and annotate tRNAs and their secondary structures. Total 474 tRNA genes corresponding to a total length of 33.4 kbp were identified. Further, to annotate the ribosomal RNA (rRNA) and small nuclear RNA (snRNA), BLASTn search against the Rfam database (version 14.1) was performed. We found 922 rRNA genes with a total length of 654.07 kbp and 347 snRNA, with a total length of 34.8 kbp.

Comparative Genome analysis:

We used OrthoMCL[33] (https://orthomcl.org/) to identify ortholog genes among cluster bean and other important crop and model plants including, *Glycine max, Cajanus cajan, Cicer arietinum, Arabidopsis thaliana, and* a monocot *Oryza sativa*. The details of orthologous gene families are mentioned in **Table 4**.

Data Records:

The Guar *(Cyamopsis tetragonoloba)* genome assembly data is deposited in NCBI under BioProject and will be made available soon.

Code availability:

The bioinformatics tools/packages used in this research are described below along with their versions, settings and parameters.

1) Supernova: version 2.1.1, default parameters; **(2) Canu:** version 1.6, default parameters; **(3) Seqclean:** with default input sequence in fasta format -v /usr/db/adapters, /usr/db/UniVec, /usr/db/linkers\,-s/usr/db/ecoli_genome,/usr/db/mito_ribo_seqs; **(4) npScarf(jsa.np.npscarf):** default parameters ; **(5) UGbs-Flex pipeline:** with available python and perl script; **(6) GATK**: version 3.6; **(7) BUSCO:** version 3.1.0, -l eukaryota_odb10, -e 1e-05, --augustus_species Arabidopsis; **(8) RepeatModeler:** version 1.0.10; **(9) RepeatMasker:** version 4.0.7, search engine= rmblast; **(10) Jellyfish:** version 2.0 ; **(11) GenomeScope**: version 2.0; **(12) MISA perl script**: ; **(13) Rfam**: version14.1, (January 2019, 3016 families); **(14) Seqping**: version 1.45, -e 1e-10; **(15) GlimmerHMM**: version 3.0.4; **(16) Maker:** version 2.31.9; **(17) tRNAscan-SE**: version 1.3.1; **(18) CD-HIT:** version 4.6, default parameters; **(19) Trinity:** version 2.1.1, --seqType fq --full_cleanup --min_contig_length 250; **(20) InterProScan 5:** version 5.25-64.0; **(21) Dfam:** version 3.1 (June 2019, 6915 entries); **(22) JoinMap:** version 4, default parameters; **(23) OrthoMCL:** version 1.2

Acknowledgements:

Clusterbean Sequencing Initiative was funded completely by ICAR-CRP on Genomics. We are grateful to Dr J. K. Jena (Coordinator & DDG (FS), ICAR, New Delhi, India and Dr Vindhya Mohindra (Co-Coordinator, CRPG, NBFGR, Lucknow, India) for providing constant support, encouragement and efficient coordination of the entire program objectives. The authors acknowledge Dr T Mohapatra (Secretary,DARE & DG, ICAR, New Delhi) for constantly encouraging and giving critical suggestions for improvement of the program activities.

References:

1. Purohit, J., Arun Kumar, Hynniewta, M. & Satyawada, R. R. Karyomorphological Studies in Guar (Cyamopsis tetragonoloba (Linn.) Taub.) —An Important Gum Yielding Plant of Rajasthan, India. *Cytologia (Tokyo).* **76**, 163–169 (2011).
2. Gillett, J. B. Indigofera (Microcharis) in tropical Africa with the related genera Cyamopsis and Rhynchotropis. *H.M.S.O* **Kew Bull.**, 1–166 (1958).
3. Hymowitz, T. & Whistler, R. L. Guar: agronomy, production, industrial use, and nutrition. *Purdue Univ. Press. West Lafayette* (1979).
4. Bhatt, R. K., Jukanti, A. K. & Roy, M. M. Cluster bean [Cyamopsis tetragonoloba (L.) Taub.], an important industrial arid legume: A review. *Legum. Res.* **40**, 207–214 (2017).
5. B. Das & Arora, S. K. 'Guar seed–its chemistry and industrial utilization of gum.' Guar–Its improvement and management. *Forage Res.* **4**, 79–101 (1978).
6. Mulimani, H. V. & Sirigeri, J. P. Investigating Plant Galactomannans. *Biochem. Mol. Biol. Educ.* 101–103 (2002).
7. Sandra, E. K., Morris, J. B. & Yookung, Ki. Total and Soluble Dietary Fiber Variation in Cyamopsis tetragonoloba (L.) Taub. (Guar) genotypes. *J. Food Qual.* **29**, 383–391 (2006).
8. Prosky, L. & Lee, S. C. Classification of complex carbohydrates. *Adv. Exp. Med. Biol.* **427**, 55–61 (1997).
9. Mudgil, D., Barak, S. & Khatkar, B. S. Guar gum: Processing, properties and food applications - A Review. *J. Food Sci. Technol.* **51**, 409–418 (2014).
10. Tripathy, S. & Das, M. K. Guar Gum: Present Status and Applications. *J. Pharm. Sci. Innov.* **4**, 24–28 (2013).
11. Ayyangar-Rangaswami, G. N. & Krishnswami, N. A note on the chromosome numbers in cluster beans, Cyamopsis psoralioides DC. *Indian J. Agric. Sci.* **3**, 934–935 (1933).

12. Tyagi, A. *et al.* The genome size of clusterbean (Cyamopsis tetragonoloba) is significantly smaller compared to its wild relatives as estimated by flow cytometry. *Gene* **707**, 205–211 (2019).
13. Kaila, T. *et al.* Chloroplast genome sequence of clusterbean (Cyamopsis tetragonoloba L.): Genome structure and comparative analysis. *Genes (Basel).* **8**, (2017).
14. Chaudhury, A., Kaila, T. & Gaikwad, K. Elucidation of Galactomannan Biosynthesis Pathway Genes through Transcriptome Sequencing of Seeds Collected at Different Developmental Stages of Commercially Important Indian Varieties of Cluster Bean (Cyamopsis tetragonoloba L.). *Sci. Rep.* **9**, 1–17 (2019).
15. Rawal, H. C. *et al.* High quality unigenes and microsatellite markers from tissue specific transcriptome and development of a database in clusterbean (Cyamopsis tetragonoloba (L.) Taub.). *Genes (Basel).* **8**, (2017).
16. Tyagi, A. *et al.* Genome-wide discovery of tissue-specific miRNAs in clusterbean (Cyamopsis tetragonoloba) indicates their association with galactomannan biosynthesis. *Plant Biotechnol. J.* **16**, 1241–1257 (2018).
17. Tribhuvan, K. U. *et al.* Identification of genomic SSRs in cluster bean (Cyamopsis tetragonoloba) and demonstration of their utility in genetic diversity analysis. *Ind. Crops Prod.* **133**, 221–231 (2019).
18. Mace, E. S., Buhariwalla, H. K. & Crouch, J. H. A high-throughput DNA extraction protocol for tropical molecular breeding programs. *Plant Mol. Biol. Report.* **21**, (2003).
19. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
20. Vurture, G. W. *et al.* GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
21. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Corrigendum: Direct determination of diploid genome sequences. *Genome Res.* **28**, 757–767 (2018).
22. Sergey, K. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Black Hat Briefings* **25**, 1–11 (2014).
23. Cao, M. D. *et al.* Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* **8**, (2017).
24. Stam, P. Construction of integrated genetic linkage maps by means of a new computer

282 package: JOINMAP. *Plant J.* **3**, 739–744 (1993).
283 25. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**,
284 D81–D89 (2016).
285 26. Thiel, T., Michalek, W., Varshney, R. & Graner, A. Exploiting EST databases for the
286 development and characterization of gene-derived SSR-markers in barley (Hordeum
287 vulgare L.). *Theor. Appl. Genet.* **106**, 411–422 (2003).
288 27. Chan, K. L. *et al.* Seqping: Gene prediction pipeline for plant genomes using self-training
289 gene models and transcriptomic data. *BMC Bioinformatics* **18**, 1–7 (2017).
290 28. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a
291 reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
292 29. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of
293 protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
294 30. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
295 BUSCO: Assessing genome assembly and annotation completeness with single-copy
296 orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
297 31. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification.
298 *Bioinformatics* **30**, 1236–1240 (2014).
299 32. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in
300 functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
301 33. Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for
302 Eukaryotic Genomes. *Genome Res.* **13**, 2178–2189 (2003).
303
304
305
306
307
308
309

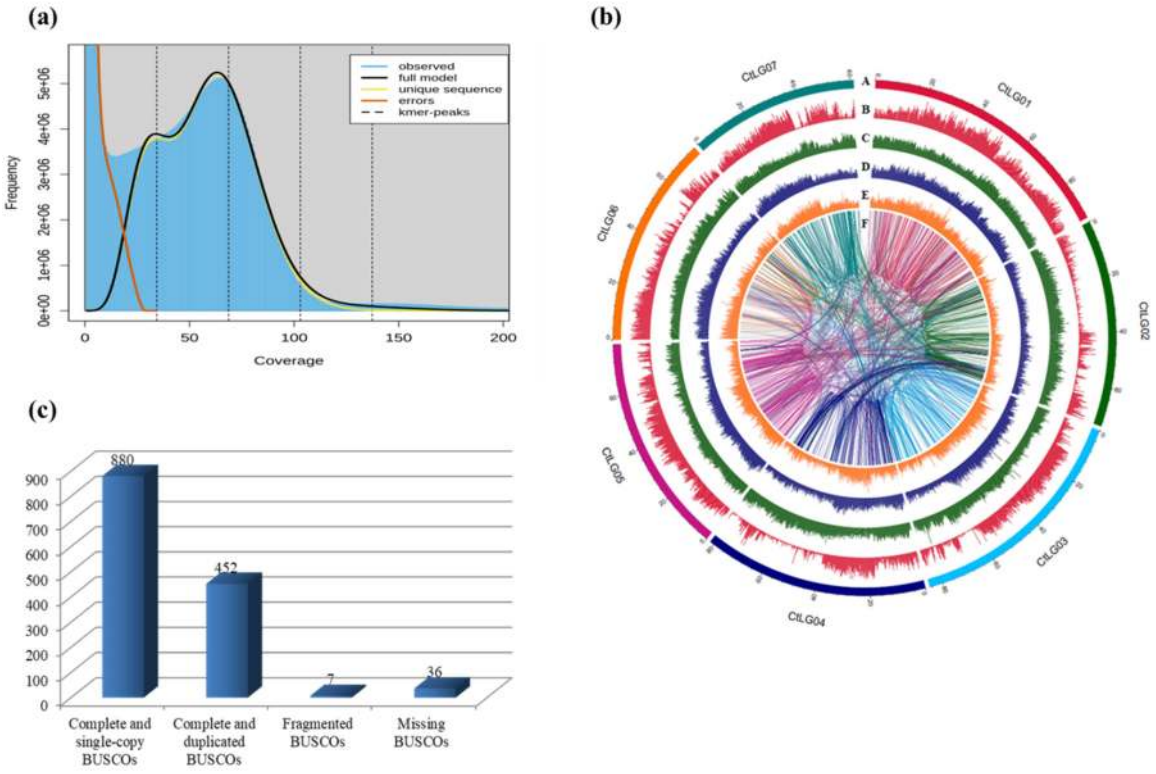

Figure 1: (a) The 31-mer frequency distributions of the sequencing reads. The sharp peak on left with low depths represents the random sequencing errors. The middle and right peaks indicate the heterozygous and homozygous peaks, the depths of which are 32 and 65, respectively. **(b)** The cluster bean genome features. Track A represents 7 pseudochromosomes. Track B to E represent the distribution of protein-coding genes, Retrotransposons, DNA elements and Simple sequence repeats, respectively. Track F represents gene duplications in the cluster bean genome. **(c)** Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of cluster bean annotated genes.

Table 1: Raw sequencing data generated for cluster bean genome assembly

Library type	**Reads length (bp)**	**Insert size (bp)**	**No. of libraries**	**No. of Reads**	**Sequence output (bp)**	**Genome coverage**
Illumina pair-ended/mate-pair sequencing						
PCR-free	2x150	470	1	54,304,615	10,427,940,295	18.95 X
PCR-free	2x150	800	1	29,507,608	5,666,395,556	10.30 X
Mate-Pair	2x160	3000	1	84,083,542	24,261,975,442	44.09 X
Mate-Pair	2x160	7000	1	63,517,764	18,046,975,201	32.79 X
10 x Genomics Linked Reads						
Chromium	2x150		8	309,223,705	92,767,111,500	168.57 X
Oxford Nanopore						
PromithION	2			6,712,699	50,647,765,074	92.03 X
Total					**201,818,163,068**	**366.73 X**

Table 2: Genome assembly and gene annotation statistics of the cluster bean genome

Assembly features	
Number of scaffolds	58
Total length of the assembly (bp)	550,317,005
GC content (%)	32.64
Scaffold (pseudomolecules) N50	78,271,499
Longest scaffold (pseudomolecule)	93,956,920
Number of contigs	1580
Max contig length	35,034,496
Min contig length	970
Contig N50	7,127,297
Gene models	
Number of gene models	34680
Mean transcript length	3823.77
Mean coding sequence length	289.115
Mean number of exons per gene	5.35043
Mean exon length	249.843
Mean intron length	681.044
Number of genes annotated	28,955
Number of genes unannotated	5725

Table 3: Organization of repetitive elements in the cluster bean genome

		Number	**Length occupied (bp)**	**Proportion of repeats (%)**	**Proportion of genome (%)**
Retrotransposons (total)		378584	163602743	70.55	29.73
	LINE	155598	44439166	19.16	8.08
	SINE	1100	176409	0.08	0.03
	LTRs	221886	118987168	51.31	21.62
	Gypsy	105335	78656908	33.92	14.29
	Copia	27099	15596297	6.72	2.83
DNA transposons		185613	54034406	23.30	9.82
Unclassified elements		13555	12889753	5.56	2.34
Satellites		1611	412566	0.18	0.07
Total repetitive elements		582339	231879649	100.00	42.14

Table 4: Gene family analysis of the predicted cluster bean genes in comparison to other plant genomes

Species	**No. of predicted genes**	**No. of Genes in Orthologous groups**	**No. of Genes not in orthologous groups**	**Total no. of orthologous groups**	**No. of Inparalogs genes**	**No. of Single copy Orthologs**	**Average genes group**
C. tetragonoloba	**34680**	**13584**	**21096**	**10639**	**3356**	**8555**	**1.28**
G. max	56044	43158	12886	24376	34900	5555	1.77
C. cajan	31841	23355	8486	20812	9329	12521	1.12
C. arietinum	26164	15505	10659	11639	4835	9939	1.33
O. sativa	55986	25591	30395	14108	19797	5522	1.81
A. thaliana	28775	21830	6945	14447	14471	6908	1.51