scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale genome assembly of *Trigonella corniculata* (L.)L. (Nagauri pan /Kasuri methi), an important spice

Ambika Baldev Gaikwad ✉, Sheel Yadav, Ratna Kumari, Wanchha Maurya, Parimalan Rangan, Rakesh Singh & Gyanendra Pratap Singh

Trigonella corniculata **(L) L. or Nagauri pan /Kasuri methi, is an important spice crop with high nutraceutical potential. We report the *de novo* chromosome-scale assembly of *T. corniculata* genome using high coverage PacBio, Illumina and Hi-C reads. The assembly spans 798 Mb (Megabases) in 282 scaffolds with a scaffold N50 of 99.6 Mb. More than 98% of the sequence length is captured in eight different pseudomolecules with an average length of 98 Mb. A BUSCO score of over 97% is suggestive of the high degree of completeness and contiguity of the genome. A total of 64,801 protein-coding genes are predicted. Genome-wide Simple Sequence Repeats (99,149) have been identified and wet lab validated at forty-eight loci. The chromosome-scale genome assembly of *T. corniculata* and the SSR markers identified in this study will provide a strong foundation for future structural and functional genomics studies in *T. corniculata* and other fenugreek species.**

Background & Summary

Trigonella, a genus belonging to the family Fabaceae, is reported to have more than 250 species, with the herb fenugreek as the best known member[1–5]. Fenugreek, popularly known as "methi" in Hindi is a multi-purpose crop, with different species being used for varied purposes[6]. A few noteworthy edible species are, *T. foenum-graecum* (common fenugreek), *T. corniculata* (Nagauri pan or Kasuri fenugreek), *T. caerulea* (blue fenugreek) and *T. suavissima* (sweet fenugreek)[6–11]. Most species are diploid with 2n = 2x = 16[12]. The term "fenugreek" is frequently used to denote *T. foenum-graecum*, while it extends to other fenugreek species as well. In India, the species *T. foenum-graecum* and *T. corniculata* are one of the oldest commercially cultivated spice crops. Unlike *T. foenum-graecum* where both leaves and seeds are consumed, *T. corniculata* is primarily used in the dried leaf form[12]. The two species differ remarkably in their morphology with most prominent differences for leaf and seed size, flower colour, etc. The leaves of *T. corniculata* are wedge shaped, 1–4 cm long and the flowers are 6–7 mm long, yellow in color which bloom in clusters of 8–20[12,13] (Fig. 1a). Significant differences are reported in the genome sizes of *T.corniculata* and *T. foenum-graecum*[14]. Much like *T. foenum-graecum*, *T. corniculata* also has significant nutraceutical potential, being abundant in phytochemicals like diosgenin, trigonelline, fenugreekine, galactomannan, 4-hydroxy isoleucine, isoorientin, orientin, vitexin, isovitexin, etc[15–27]. The medicinal value of fenugreek has long been known with references dating back to 1500 BC in the Ebers Papyrus, an ancient Egyptian medical text[16]. Despite their medicinal importance, these two species are largely unexplored for the genomic resources availability. There are only a few studies where through transcriptome sequencing, genes involved in key metabolic pathways have been identified[28–30]. The lack of genome sequence information severely impedes the crop improvement endeavours in fenugreek. The absence of sequence information, limits the availability of molecular markers for the species. Very few DNA-based molecular markers are available in fenugreek[31–34], more so in Kasuri fenugreek[3]. Hence, very little progress has been made in the direction of genetic improvement in fenugreek[35]. Here, we report upon a high quality, whole genome sequence (WGS) of 798 Mb (Megabases) for the genotype JKM-5 of *T. corniculata*. The genome has been assembled by using a hybrid approach which involves a combination of Illumina short reads, PacBio long reads, and high-throughput

ICAR-National Bureau of Plant Genetic Resources, New Delhi, 110 012, India. ✉e-mail: ambikabg@gmail.com; ambika.gaikwad@icar.gov.in

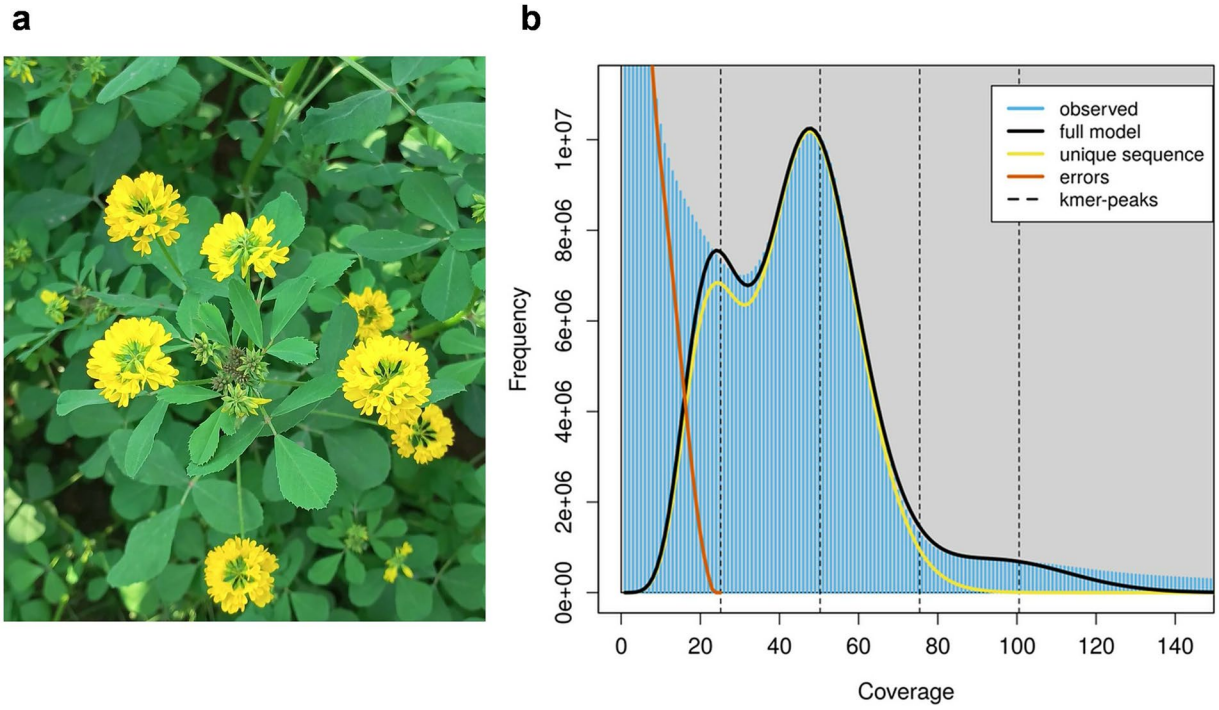


Fig. 1 (**a**) The plant morphology of *T. corniculata* with inflorescence shown. (**b**) The 23-mer frequency distribution of sequencing reads of the *T. corniculata* genome, used for genome size estimation.

chromosome conformation capture (Hi-C) data. The final version of the assembled genome consists of 282 scaffolds with a scaffold N50 of 99.6 Mb, and with more than 98% of the sequence length captured in eight pseudomolecules. The genome size of *T. corniculata* is estimated to be ~850 Mb[14]. This implies that >93% of the genome has been captured in this assembly. A BUSCO (Benchmarking Universal Single-Copy Orthologs) score of 97% is indicative of the completeness of the assembled genome. A total of 64,801 protein-coding genes were predicted. The WGS was further utilized to identify genome-wide SSRs (Simple Sequence Repeats or microsatellites) and a total of 99,149 SSRs were identified. These markers would serve as novel genomic resources for *T. corniculata* and can be utilized for various purposes like genetic mapping, marker-assisted selection (MAS) of traits.

Methods

DNA extraction. Genomic DNA was isolated from the leaves of the genotype JKM-5 (source SKNAU, Jobner, Rajasthan, India) using the CTAB (Cetyltrimethylammonium bromide) extraction method[36]. The quality and quantity of the isolated DNA was checked on a NanoDrop spectrophotometer (DS-11 spectrophotometer, DeNovix, Wilmington, Delaware) and through agarose gel electrophoresis.

Library preparation and sequencing. We deployed both long- and short-read sequencing chemistries for sequencing and assembly. Sequencing libraries were prepared according to the recommended protocols provided by the manufacturer as detailed below.

Illumina paired-end (PE) genomic libraries were prepared following the NEBNext® Ultra™ II DNA Library Prep Kit. In brief, genomic DNA was sheared using Covaris® M220 Focused-ultrasonicator TM (Covaris, Woburn, MA, USA). The sheared DNA was subjected to size selection with insert sizes of 350 and 550 bp, followed by adapter ligation and PCR enrichment of adapter ligated DNA. The quality of the libraries was checked on the Agilent 2100 Bioanalyzer system using the Agilent High Sensitivity DNA Kit and sequenced on Illumina HiSeq X Ten sequencer as 150 bp PE reads.

For PacBio library preparation, the SMRTbell® Express Template Prep Kit (Pacific Bioscience, catalog no.101-357-000) was used. Using the BluePippin Size selection system, the small fragments were removed. After annealing of the sequencing primer to the SMRTbell template, DNA polymerase was bound to the complex (Sequel II binding kit 2.0). The excess of primer and polymerase was removed before sequencing. The library was sequenced using 2 SMRT cells using the Sequel II Sequencing Kit 2.0.

For Hi-C libraries, the leaf tissue was treated with 1% final concentration fresh formaldehyde for cross-linking or fixing the chromatin (1 g leaf material per 100 ml) and quenching with 0.2 M final concentration glycine for 5 minutes. The fixed cells were lysed for nuclei isolation by treatment with lysis buffer (10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% NP-40, and complete protease inhibitors (Roche). The DNA in the nuclei was digested by adding 30 µl 10x New England Biolabs (NEB) buffer 2.1 (50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl, 100 µg/mL bovine serum albumin (BSA), pH 7.9) and 150 U of MboI restriction enzyme and incubated at 37 °C overnight. The following day, the enzyme MboI was inactivated at 65 °C for 20 minutes. The cohesive ends

Statistics	Illumina	PacBio	Hi-C
No. of reads	201,352,096	1,313,226	970,555,786
No. of bases (Gb)	60.4	82.3	145.5
Depth (X)	88.1	103	212.5
Mean length (bp)	150	15,451	150

Table 1. The details of sequencing data generated by using three different sequencing chemistries namely Illumina, PacBio and Hi-C.

Features	Minimum	Maximum
Heterozygosity	0.89%	0.89%
Genome haploid length (bp)	690,465,573	691,164,565
Genome repeat length (bp)	351,061,443	351,416,840
Genome unique length (bp)	339,404,130	339,747,725
Model fit	89.35%	96.82%

Table 2. Genome size estimation through K-mer (K = 23) analysis.

were filled with biotinylated nucleotides and universal adapter ligated. After fragmentation and size-selection, fragments containing biotin were captured to construct the paired-end library. The final library was sequenced on the Illumina NovaSeq 6000 platform (San Diego, CA, USA).

In total, we obtained 82.3 Gb of PacBio sequencing data (coverage: 103x) with an average read length of 15.45 Kb and 60.4 Gb of (coverage: 88.1x) short read data. A total of 145.5 Gb (coverage: 212x) of Hi-C data was generated (Table 1).

Genome size estimation. The Illumina paired-end (PE) reads were utilized for estimation of genome size by counting the K-mers. The abundance of 23-mer K-mer was calculated using Jellyfish 2.2.3[37]. The histogram was used as an input for estimating various genome characteristics such as repeat content, heterozygosity, genome size and duplications using Genomescope (v2.0)[38], to obtain K-mer frequency peaks (Fig. 1b). In the histogram, the middle and right peak which represent the heterozygous and homozygous fractions of the genome, respectively, showed depths at ~25 and 52, respectively. Based on this analysis, the haploid genome length for *T. corniculata* was estimated to be 691 Mb (691,164,565 bp) and the heterozygosity fraction was 0.89% (Table 2).

***De novo* genome assembly.** The reads obtained were checked for quality in terms of base quality, adapter and GC content, number of unambiguous bases. Only the reads which confirmed to the desired quality standards were used for assembly. The initial assembly was done using PacBio reads. The raw subreads generated by the PacBio Sequel II were converted to HiFi reads using CCS (Circular Consensus Sequencing, v6.2.0) and assembled using HiFiasm tool v0.16.1-r375 where excessive purging options were specified[39]. The duplicate contigs and contig regions arising due to haplotypes were purged from the primary contigs using purge_dups v1.2.5. To assess the contiguity and quality of genome assembly QUAST v5.2.0 was used to generate basic statistics[40]. The preliminary assembly consisted of 1738 contigs with a N50 of 1.2 Mb, and the largest contig length of 8 Mb (Table 3). This assembly was then polished using 88x Illumina reads via two iterations of Pilon v1.23[41]. Each iteration consisted of first the alignment of Illumina short reads to the assembly with BWA-MEM v0.7.17-r1188[42]. The resulting SAM file was converted to a BAM file and indexed with SAMtools v1.9[43].

The 970.5 million Hi-C reads were used for ordering and scaffolding the contigs of the polished assembly using the YaHS (https://github.com/c-zhou/yahs.git) tool. YaHS requires two input files: a FASTA format file with contig sequences which need to be indexed and a BAM file with the alignment results of Hi-C reads to the contigs. Juicer 1.6 tool was used for generating Hi-C maps from fastq raw data files and command line tools for feature annotation on the Hi-C maps[44]. It was observed that there were relatively independent Hi-C signals observed between the 8 scaffolds and the density of interactions was confirmed through heatmap analysis (Fig. 2). This indicated high degree of contiguity in the assembly. The final assembly consisted of 798 Mb captured on 282 scaffolds, with more than 98% of sequence scaffolded into eight pseudomolecules representing the eight chromosomes, which is the haploid chromosome number of *T. corniculata*[3,14] (Fig. 3a,b). The length of these eight scaffolds ranged from 77 Mb (scaffold 8) to 114 Mb (scaffold 1). The scaffold N50 was 99.6 Mb and L50 value was 4, which is indicative of the good quality and contiguity of the assembled genome. The average scaffold length was 2.8 Mb which implies a higher probability of generating complete gene models for the genome (Tables 3, 4).

Completeness of genome assembly. The completeness of the assembled genome was estimated by using Benchmarking Universal Single-Copy Orthologs[45] (BUSCOs; v 5.4.2, http://busco.ezlab.org/) based on single-copy orthologs selected from Viridiplantae OrthoDB v10. Its results are simplified into categories of 'Complete and single-copy', 'Complete and duplicated', 'Fragmented', or 'Missing' BUSCOs. A BUSCO score value of 97.1% was obtained, which depicts the high quality of the genome (Table 5).

Assembler	Contig assembly (PacBio) Preliminary assembly	YaHS assembly (Scaffolding) Final assembly
(>=0 bp)	1738	282
(>=1 Kb)	1738	282
(>=5 Kb)	1738	282
(>=10 Kb)	1737	281
(>=25 Kb)	1668	211
(>=50 Kb)	1373	8
Contigs/Scaffolds	1738	282
Largest (bp)	8083617	114574750
Total length (bp)	799108466	798664075
GC (%)	35.56	35.24
N50 (bp)	1293154	99641310
N70 (bp)	709802	94651279
L50	177	4
L70	344	6

Table 3. The descriptive statistics for the preliminary (contig) assembly and the final assembly.

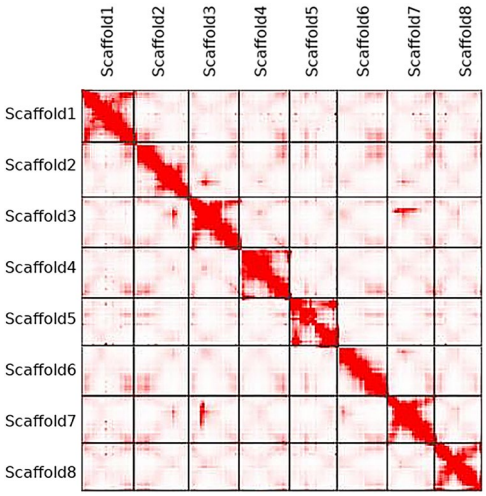
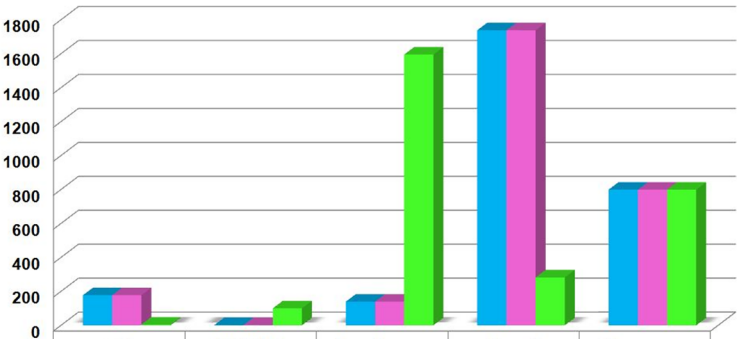


Fig. 2 The Hi-C interaction heatmap of the genome assembly (window size: 100Kb).

Genome annotation. This included annotation of the repeat sequences followed by gene structure and functional annotation. Tandem repeat finder was utilized to identify tandem repeats in the final assembled genome. RepeatModeler v2.0.1 (http://www.repeatmasker.org/RepeatModeler/) was used to create a repeat library *de novo* which was then used to predict transposable elements in the unannotated genome assembly[46]. We used three different *de novo* repeat-finding programs, namely RECON v1.08[47], RepeatScout v1.0.5[48] and LTR_retriever v2.9.0[49] in order to identify the boundaries of repetitive elements and to build consensus models of interspersed repeats. LAI (LTR Assembly Index), a metric for assessing assembly continuity, was calculated by LTR_retriever[49]. The LAI score of 9.93 for the assembled genome, which is very close to the threshold value of 10 for reference genomes[50]. For most of the sequenced genomes of the Fabaceae family, the LAI values are reportedly less than 10 except for the genomes of *Trifolium pratense* (GCA_020283565.1), *Medicago truncatula* (GCA_003473485.2) and *Vigna unguiculata* (GCA_004118075.2), suggestive of the good quality of the assembled genome in this study[51]. The interspersed repeats were annotated using RepeatMasker v 4.1.0 (https://www.repeatmasker.org/). The repeat content in the genome was estimated to be 42.1% with LTR (long terminal repeat) retrotransposons accounting for 38% of the repeats (Fig. 3b). The Gypsy (17.12%) and Ty1- Copia (16.26%) elements were most abundant LTR retrotransposons (Table 6).

This was followed by gene prediction and annotation. Seqping (https://sourceforge.net/projects/seqping/, version 0.1.48.1) was used for gene prediction of the masked sequence. It is an automated pipeline that performs gene prediction using self trained HMM (Hidden Markov model) models and transcriptomic data[52]. The program processes the genome sequences of a target species through GlimmerHMM[53], SNAP[54], and AUGUSTUS[55] training pipeline. MAKER2 was used to combine the predictions from the three models. Input data for the pipeline included genome sequence in FASTA format, transcriptome for generating training set (PRJNA544308, transcriptome assembly of *T. foenum-graecum*), reference protein file containing full length protein sequences of 20 species. Additionally, transcriptome data for four different tissues of Kasuri fenugreek (unpublished) was

a



	L50	N50 (Mb)	No. of gaps	No. of contigs/scaffolds	Sequence length (Mb)
Preliminary assembly	177	1.29	139	1738	799.1
Polished assembly	177	1.29	139	1738	798.64
Final assembly	4	99.64	1595	282	798.66

b

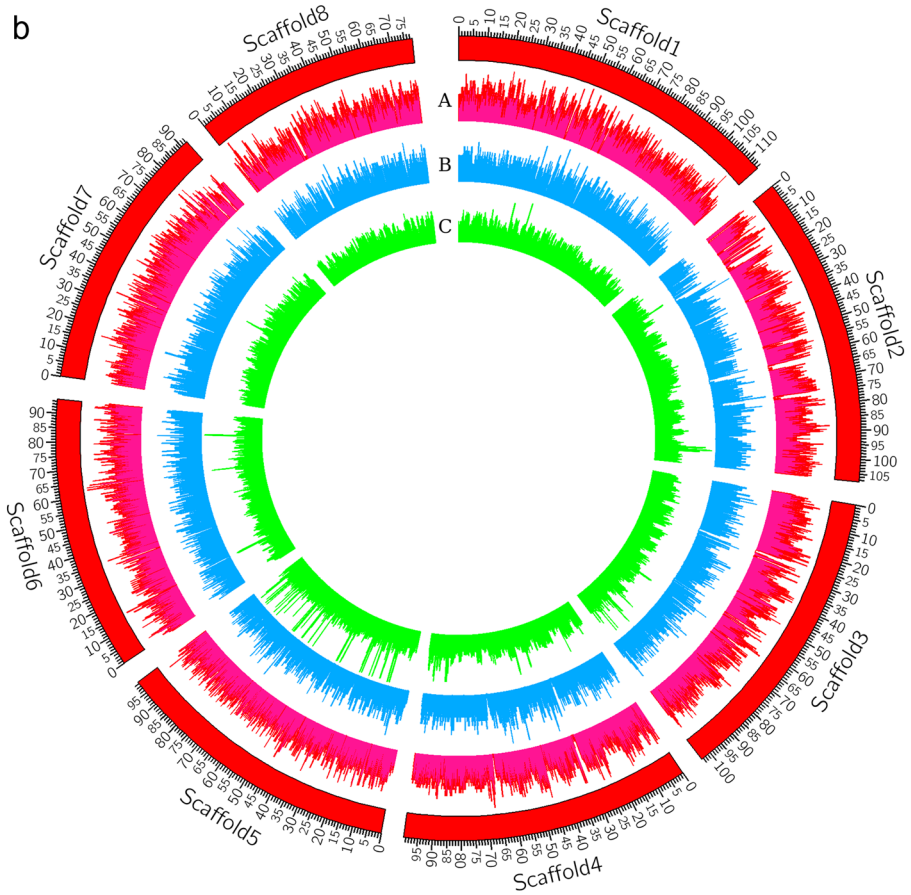


Fig. 3 The chromosome scale assembly of *T. corniculata*. (**a**) Descriptive statistics of the *de novo* assembled genome. (**b**) Circos plot depicting genomic features at window size of 100 Kb for the genome, with track A representing the gene density; track B transposon density and track C depicting the GC content.

also generated and utilized for gene prediction. The pipeline generated species-specific HMMs and is able to predict genes that are not biased to other model organisms. Genes were clustered using CDHIT v4.6 (http://cd-hit.org) software at a sequence similarity of 90%. Subsequently, functional annotation was performed and the predicted genes were mapped against known public databases Nr (updated December 2023) using NCBI blastx v 2.2.29 (e-value $\leq$ 1e−5), SwissProt, Pfam, GO[56] (Gene Ontology) and KEGG[57] (Kyoto Encyclopedia of Genes and Genomes). A total of 64,801 protein-coding genes were predicted, with an average length of 5.6 Kb. The gene annotation results depicted that out of the 64,801 genes predicted, 62079 genes (95.8%) were annotated

Statistics	Description
Total (bp)	798664075
Chromosome (bp)	790185390
Number of scaffolds	282
Largest scaffold (bp)	114574750
Anchoring ratio (%)	98.93
Gaps	1595
N50 (bp)	99641310
GC (%)	35.24
BUSCO (%)	97.00
LAI	9.93
Mapping rate (NGS) (%)	98.12

Table 4. The summary statistics for the *de novo* assembled genome of *T. corniculata*.

Category	Percentage (%)	Gene number
Complete BUSCOs (C)	97.0	412
Complete and single-copy BUSCOs (S)	87.1	370
Complete and duplicated BUSCOs (D)	9.9	42
Fragmented BUSCOs (F)	0.2	1
Missing BUSCOs (M)	2.8	12
Total BUSCO groups searched	—	425

Table 5. BUSCO assessment of genome assembly (Database used: Viridiplantae).

Category	Type	Number of elements	Length Occupied (bp)	Percentage (%) of sequence
Retrotransposons		320759	336936283	42.19
	Non- LTR Retrotransposon:			
	LINEs:	57018	32172762	4.03
	CRE/SLACS	2333	672636	0.08
	L2/CR1/Rex	113	66400	0.01
	RTE/Bov-B	7999	2083303	0.26
	L1/CIN4	46570	29349621	3.67
	LTR Retrotransposon:	263741	304763521	38.16
	BEL/Pao	1833	369718	0.05
	Ty1/Copia	96608	129874255	16.26
	Gypsy/DIRS1	87933	136718324	17.12
DNA transposons		41464	13930803	1.74
	hobo-Activator	14280	3025708	0.38
	Tc1-IS630-Pogo	2121	1085684	0.14
	Tourist/Harbinger	2678	851667	0.11

Table 6. The summary of the repetitive DNA sequences identified across the assembled genome.

Database	Number	Ratio (%)
Nr	62079	95.8
Pfam	33890	52.3
Swiss-Prot	23911	36.9
GO	33186	51.2
KEGG	4013	6.19
All	64801	100

Table 7. The statistical analysis of the predicted genes for the assembled genome depicting the number of genes annotated through various databases.

in the Nr database, followed by 33890 (52.3%) in Pfam and 23911 (36.9%) in SwissProt databases (Table 7). The GO term binding (GO:0005488) was most abundant (Fig. 4a). KEGG analysis depicted the highest number of

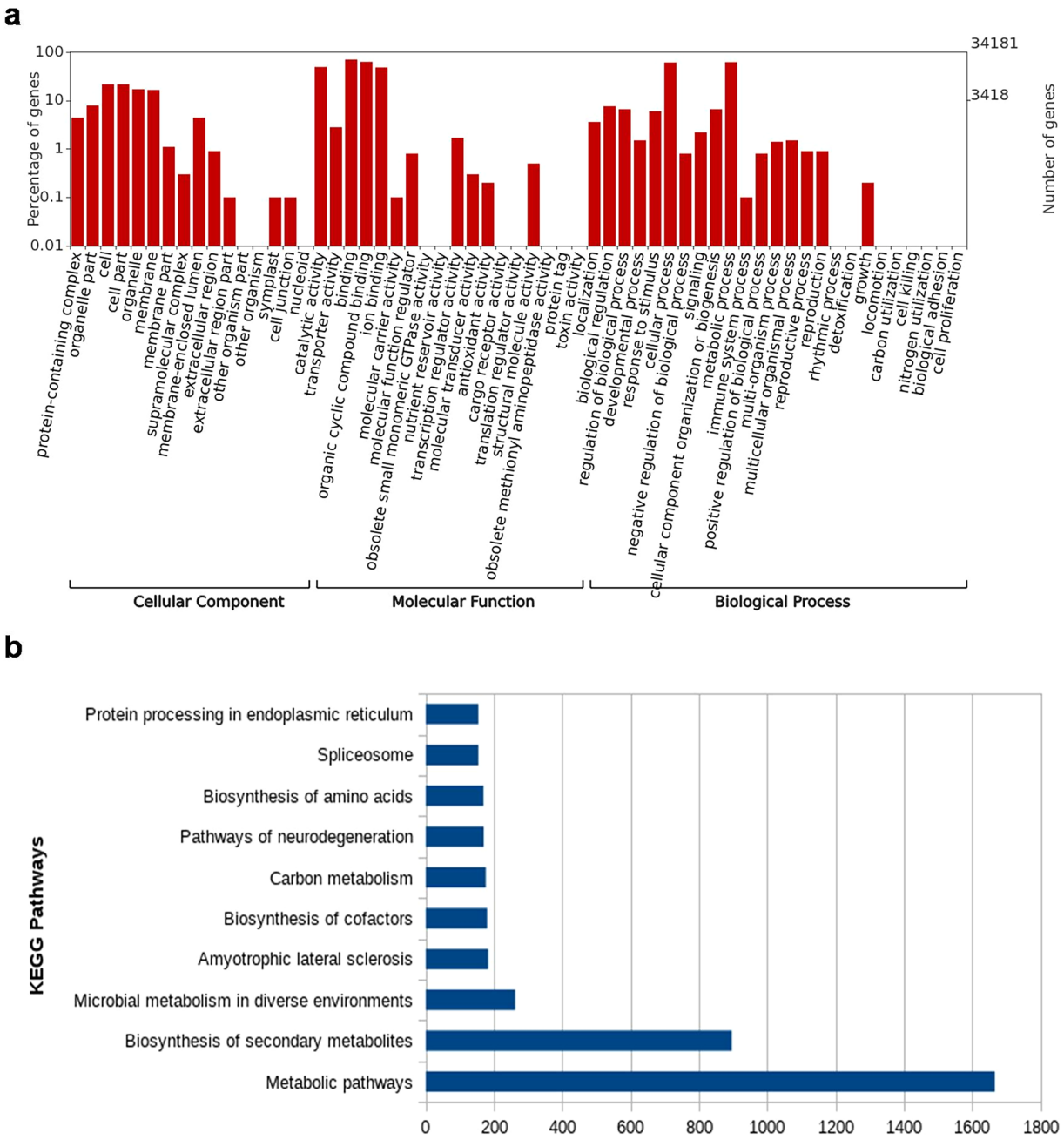


Fig. 4 The annotation of the assembled genome. (**a**) GO analysis depicting the distribution of the predicted genes across the GO terms of cellular component, molecular function and biological process. (**b**) The KEGG analysis of the predicted genes.

genes involved in metabolic pathways (1666) followed by biosynthesis of secondary metabolites (896) (Fig. 4b). A total of 1,716 tRNA genes were identified using tRNAscan-SE (v.1.3.1, http://lowelab.ucsc.edu/tRNAscan-SE/).

Microsatellite or simple sequence repeat (SSR) identification. The assembled genome sequence was searched for presence of simple sequence repeats (SSRs), with the core motif lengths of di to hexa nucleotides, following the default parameters of MISA (MIcroSAtellite identification tool;http://pgrc.ipk-gatersleben.de/misa/)[58]. The search criteria was set to include a minimum of six repeats for dinucleotides, minimum five repeats for trinucleotides, tetranucleotides, pentanucleotides and hexanucleotides. A total of 99,149 SSRs were identified across the genome. The most abundant repeat type was dinucleotide repeat (54,330) and least abundant was hexanucleotide (242) (Fig. 5a). The highest number of SSRs were identified on scaffold 1 while the least number of SSRs were identified on scaffold 8 (Fig. 5b).

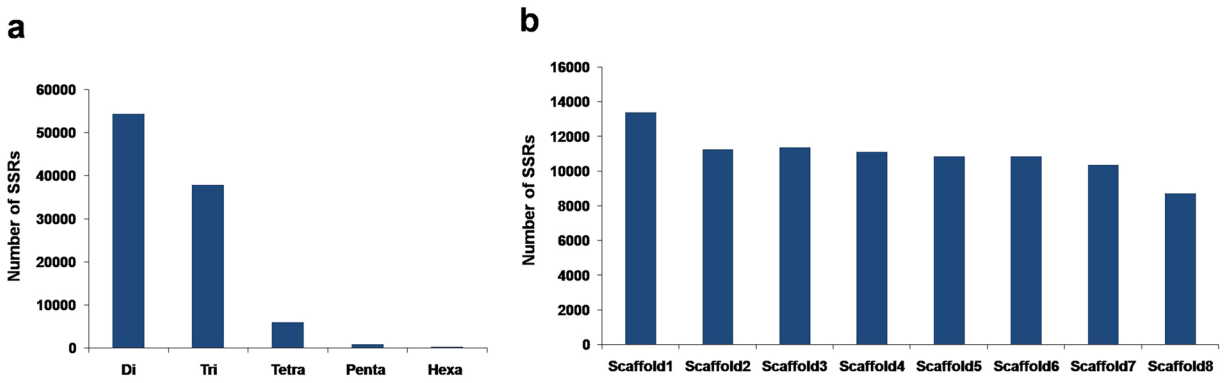


Fig. 5 *In silico* mining of SSRs and their distribution. (**a**) Frequency of SSR repeats identified. (**b**) Distribution of the SSR repeats across the eight different scaffolds.

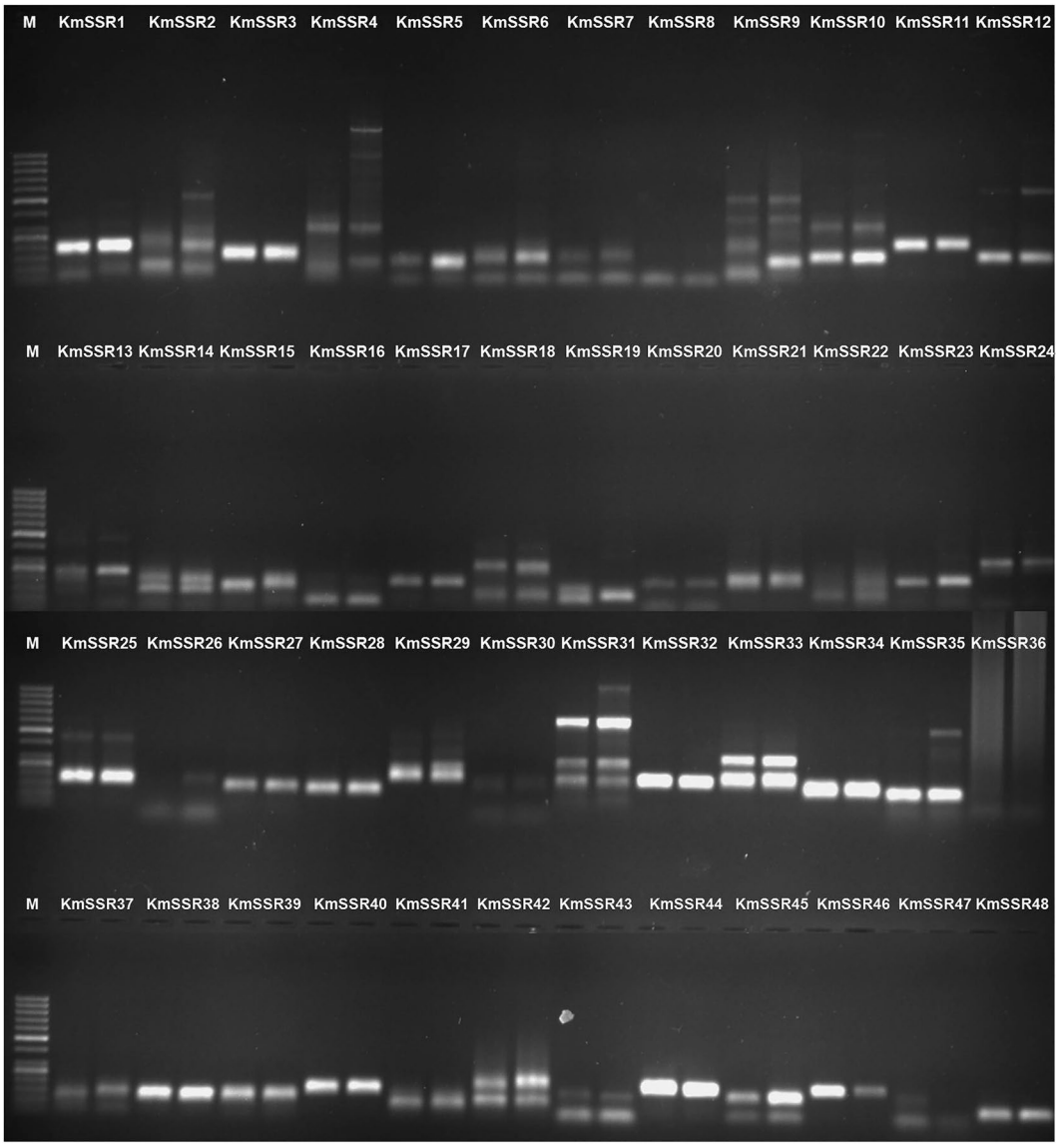


Fig. 6 SSR validation through PCR amplification. Gel image depicting PCR based amplification of the SSRs and resolution of the products on a 3% metaphor agarose gel. M: 100 bp DNA ladder; KmSSR1-KmSSR48 are the various Kasuri methi (Km) SSR markers used for amplification. Genomic DNA of the samples JKM-5 and JKM-6, were used for amplification with each of the 48 KmSSRs.

Wet lab validation of SSR markers. Following the identification of SSRs, the flanking sequences of the SSRs were used to design primers for PCR amplification using Primer 3[59,60]. Primer pairs were synthesized for a subset of 50 SSR loci for PCR amplification of genomic DNA of two samples of *T. corniculata*, namely JKM-5 and JKM-6 (Fig. 6). The reaction was carried out in a total volume of 20 µl, composed of 1X PCR buffer, 2.5 mM $MgCl_2$, 1 µM primer, 0.2 mM of each dNTPs, 1U Taq DNA polymerase (NEB) and 15 ng template DNA. The standard PCR amplification conditions were used[61], with standardization done for annealing temperatures for the PCR program.

Data Records

The raw sequence data (Illumina, PacBio and Hi-C) used for genome assembly have been deposited in the NCBI Sequence Read Archive (SRA) under the accession number SRP513070[62], with the BioProject accession number PRJNA1122319. The assembled genome has been deposited at the NCBI genbank with accession number JBLEBM000000000. The version described in this paper is version JBLEBM010000000[63]. The genome annotation file and FASTA sequences of the predicted genes, proteins, and transcripts have been deposited in the Figshare database[64].

Technical Validation

The quality of the genome assembly was analysed in the following aspects: (1) The BUSCO score of 97% depicted the degree of completeness of the assembled genome. (2) The LAI score of 9.93 for the genome, which is almost 10, depicts the high quality of the assembled genome, very close to the reference genome standards. (3) Most of the SSRs selected for wet laboratory validation through PCR produced sharp and reproducible amplification profiles with the desired product sizes, which is an indication of the high degree of genome integrity. (4) The high degree of mapping (98.1%) and anchoring rate (98.9%) of the reads, validates the contiguity of the assembled genome. (5) We used SeqKit v 2.8.2 to identify the presence of telomeric repeats across the eight pseudomolecules. For two pseudomolecules (5 and 8) telomere repeats were identified at both ends while for two pseudomolecules (2 and 6) telomeres were identified at only one end. This shows that the assembled genome is of a good quality and high integrity.

Code availability

All software employed for data processing was executed following the guidelines of the bioinformatic software cited above. If no detailed parameters are mentioned, the default parameters were used. Versions of the software have been described in Methods.

Received: 2 August 2024; Accepted: 19 March 2025;
Published online: 26 March 2025

References

1. Martin, E., Akan, H., Ekici, M. & Aytac, Z. Karyotype analyses of ten sections of *Trigonella* (Fabaceae). *Comp. Cytogenet.* **5**(2), 105–21 (2011).
2. Abozeid, A., Turki, Z., El-Shayeb, F. & Tang, Z. Embryo and seedling morphology of some *Trigonella* L. species (Fabaceae) and their taxonomic importance. *Flora* **230**, 57–65 (2017).
3. Chaudhary, S. & Chaudhary, P. Is Kasuri Methi Genetically Different from other Methi (Fenugreek) Varieties? *Journal of Genetics and Genetic Engineering* **3**(2), 15–18 (2019).
4. Hilles, A. R. & Mahmood, S. in Fenugreek. Historical background, origin, distribution, and economic importance of fenugreek pp. 3–11 (Singapore: Springer, 2021).
5. Yameen, B. *et al*. RbcLa marker based identification and phylogenetic analysis of Kasuri methi (*Trigonella foenum-graecum* L.): a native plant of Kasur district of Punjab (Pakistan). *Genet. Resour. Crop Evol.* https://doi.org/10.1007/s10722-024-02044-w (2024).
6. Acharya, S. N., Thomas, J. E. & Basu, S. K. Fenugreek: An "old world" crop for the "new world". *Biodiversity* **7**(3&4), 27–30, https://doi.org/10.1080/14888386.2006.9712808 (2006).
7. Thomas, J. E., Basu, S. K. & Acharya, S. N. Identification of *Trigonella* accessions which lack antimicrobial activity and are suitable for forage development. *Canadian Journal of Plant Science* **86**(3), 727–732, https://doi.org/10.4141/P05-155 (2006).
8. Ahmad, A., Alghamdi, S. S., Mahmood, K. & Afzal, M. Fenugreek a multipurpose crop: Potentialities and improvements. *Saudi J. Biol. Sci.* **23**, 300–310 (2016).
9. Basu, A. *et al*. Fenugreek (*Trigonella foenum-graecum* L.), a potential new crop for Latin America. *American Journal of Social Issues and Humanities* **4**(3), 148–162 (2014).
10. Aasim, M. *et al*. in Global Perspectives on Underutilized Crops (eds. Ozturk, M., Hakeem, K. R., Ashraf, M. & Ahmad, M. S. A.) Fenugreek (*Trigonella foenum-graecum* L.): An underutilized edible plant of modern world pp. 381–408 (Springer International Publishing; Cham, Switzerland, 2018).
11. Acharya, S. N., Thomas, J. E. & Basu, S. K. Fenugreek (*Trigonella foenum-graecum* L.) an alternative crop for semiarid regions of North America. *Crop Science* **48**, 841–853, https://doi.org/10.2135/cropsci2007.09.0519 (2008).
12. Kakani, R. K. & Anwer, M. M. in Fenugreek. Peter, K. V. (ed) Handbook of herbs and spices. Woodhead Publishing Limited, Sawston, pp. 286–295 (2012).
13. Chandan, T. K., Lakshminarayana, D., Seenivasan, N., Joshi, V. & Kumar, S. P. Growth and yield of Kasuri methi (*Trigonella corniculata* L.) var. Pusa Kasuri as influenced by different organic manures and biofertilizers under Telangana conditions. *Int J Chem Stud* **9**, 280–284 (2021).
14. Yadav, S. *et al*. Variability in genome size of *Trigonella foenum-graecum*, *Trigonella corniculata* and *Trigonella caerulea* as estimated by flow cytometry indicates complex evolutionary history of fenugreek. *Mol. Biol. Rep.* **51**, 489 (2024).
15. Altuntaş, E., Özgöz, E. & Taşer, Ö. F. Some physical properties of fenugreek (*Trigonella foenum-graceum* L.) seeds. *J. Food Eng.* **71**, 37–43, https://doi.org/10.1016/j.jfoodeng.2004.10.015 (2005).
16. Acharya, S., Srichamroen, A., Basu, S., Ooraikul, B. & Basu, T. Improvement in the nutraceutical properties of fenugreek (*Trigonella foenum-graecum* L.). *Songklanakarin Journal of Science and Technology* **28**(1), 1–9 (2006).
17. Mehrafarin, A. *et al*. Bioengineering of important secondary metabolites and metabolic pathways in fenugreek (*Trigonella foenum graecum* L.). *Journal of Medicinal Plants* **9**(35), 1–18 (2010).

18. Chaudhary, S. *et al*. Elicitation Diosgenin Prod *Trigonella foenum-graecum* (Fenugreek) Seedlings Methyl Jasmonate. *Int J Mol Sci.* **16**(12), 29889–99 (2015).
19. Wani, S. A. & Kumar, P. Fenugreek: A review on its nutraceutical properties and utilization in various food products. *J. Saudi Soc. Agric. Sci.* **17**, 97–106, https://doi.org/10.1016/j.jssas.2016.01.007 (2016).
20. Zandi, P. *et al*. Fenugreek (*Trigonella foenum-graecum* L.): An Important Medicinal and Aromatic Crop. Active ingredients from aromatic and medicinal plants. InTech. https://doi.org/10.5772/66506 (2017).
21. Basu, S. K., Zandi, P. & Cetzal-Ix, W. Fenugreek (*Trigonella foenum*-graecum L.): distribution, genetic diversity, and potential to serve as an industrial crop for the global pharmaceutical, nutraceutical, and functional food industries. The role of functional food security in global health pp. 471–497 (Academic Press, 2019).
22. Singh, P. *et al*. Determination of Bioactive Compounds of Fenugreek (*Trigonella foenum*-graecum) Seeds Using LC-MS Techniques. *Methods Mol. Biol.* **2107**, 377–393 (2020).
23. Mohamadi, M., Ebrahimi, A. & Amerian, M. The expression enhancement of some genes involved in diosgenin biosynthesis pathway in fenugreek treated with different levels of melatonin under salinity stress. *Iran J. Field Crop Sci.* **52**(4), 235–47 (2021).
24. Maroufi, A., Lotfi, M., Esmaeli, A. & Dastan, D. Relative expression of the key genes of diosegnin biosynthesis in fenugreek (*Trigonella foenum-graecum*) in response to salicylic acid and methyl jasmonate. *Cellular and Molecular Research (Iranian J Biology)* **34**(3), 440–54 (2021).
25. Sun, W., Shahrajabian, M. H. & Cheng, Q. Fenugreek Cultivation with Emphasis on Historical Aspects and its uses in Traditional Medicine and Modern Pharmaceutical Science. *Mini Rev. Med. Chem.* **21**, 724–730 (2021).
26. Ayvazyan, A., Stegemann, T., Galarza Pérez, M., Pramsohler, M. & Çiçek, S. S. Phytochemical Profile of *Trigonella caerulea* (Blue Fenugreek) Herb and Quantification of Aroma-Determining Constituents. *Plants* **12**, 1154 (2023).
27. Pasricha, V. & Gupta, K. R. Nutraceutical potential of Methi (*Trigonella foenum-graecum* L.) and Kasuri methi (*Trigonella corniculata* L.). *J. Pharmacogn. Phytochem.* **3**(4), 47–57 (2014).
28. Ciura, J., Szeliga, M., Grzesik, M. & Tyrka, M. Changes in fenugreek transcriptome induced by methyl jasmonate and steroid precursors revealed by RNA-Seq. *Genomics* **S0888-7543**(17), 30132–5 (2017).
29. Zhou, C., Li, X., Zhou, Z., Li, C. & Zhang, Y. Comparative Transcriptome Analysis Identifies Genes Involved in Diosgenin Biosynthesis in *Trigonella foenum-graecum* L. *Molecules* **24**(1), 140 (2019).
30. Naika, M. B. N. *et al*. Exploring the medicinally important secondary metabolites landscape through the lens of transcriptome data in fenugreek (*Trigonella foenum-graecum* L.). *Sci. Rep.* **12**, 13534 (2022).
31. Dangi, R. S., Lagu, M. D., Choudhary, L. B., Ranjekar, P. K. & Gupta, V. S. Assessment of genetic diversity in *Trigonella foenum-graecum* and *Trigonella caerulea* using ISSR and RAPD markers. *BMC Plant Biol.* **4**, 13 (2004).
32. Amiriyan, M., Shojaeiyan, A., Yadollahi, A., Maleki, M. & Bahari, Z. Genetic diversity analysis and population structure of some Iranian Fenugreek (*Trigonella foenum-graecum* L.) landraces using SRAP Markers. *Mol. Biol. Res. Commun.* **8**(4), 181–190 (2019).
33. Maloo, S. R., Sharma, R., Jain, D., Chaudhary, S. & Soan, H. Assessment of genetic diversity in fenugreek (*Trigonella foenum-graecum*) genotypes using morphological and molecular markers. *The. Indian Journal of Agricultural Sciences* **90**(1), 25–30 (2020).
34. Abd El-Wahab, M. M. H. *et al*. High-Density SNP-Based Association Mapping of Seed Traits in Fenugreek Reveals Homology with Clover. *Genes (Basel)* **11**, 893 (2020).
35. Zandi, P. *et al*. Fenugreek (*Trigonella foenum-graecum* L.) seed: A review of physiological and biochemical properties and their genetic improvement. *Acta Physiol. Plant* **37**, 1714 (2015).
36. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. *Phytochem. Bull.* **19**, 11–15 (1987).
37. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
38. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
39. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
40. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8), 1072–5, https://doi.org/10.1093/bioinformatics/btt086 (2013).
41. Walker, B. J. *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* **9**(11), e112963 (2014).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows– wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009).
43. Li, H. *et al*. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16), 2078–9 (2009).
44. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**(1), 95–8 (2016).
45. Simão, F. A. *et al*. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
46. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**(17), 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
47. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**(8), 1269–76 (2002).
48. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, 351–358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
49. Shujun, O. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology* **176**(2), 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
50. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research* **46**, e126–e126 (2018).
51. Mokhtar, M. M., Abd-Elhalim, H. M. & El Allali, A. A large-scale assessment of the quality of plant genome assemblies using the LTR assembly index. *AoB Plants* **15**(3), plad015 (2023).
52. Chan, K. L. *et al*. Seqping: gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data. *BMC Bioinformatics* **18** (Suppl 1), 1–7 (2017).
53. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
54. Johnson, A. D. *et al*. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
55. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–439 (2006).
56. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
57. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
58. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**(16), 2583–2585 (2017).

59. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–86 (2000).
60. Untergasser, A. *et al*. Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**(15), e115 (2012).
61. Gaikwad, A. B. *et al*. Small cardamom genome: development and utilization of microsatellite markers from a draft genome sequence of *Elettaria cardamomum* Maton. *Front. Plant Sci.* **14**, 1161499 (2023).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP513070 (2025).
63. Gaikwad, A. B. *Trigonella corniculata isolate JKM-5, whole genome shotgun sequencing project GenBank* https://identifiers.org/ncbi/insdc:JBLEBM000000000 (2025).
64. Gaikwad, A. B. *et al*. Chromosome-scale genome assembly of *Trigonella corniculata* (L.)L. (Nagauri pan/Kasuri methi), an important spice. *figshare* https://doi.org/10.6084/m9.figshare.26355829.v2 (2025).

Acknowledgements

The authors acknowledge the funding provided by ICAR-Consortium Research Platform on Genomics (Project number 1007341) and the facilities provided at ICAR-National Bureau of Plant Genetic Resources.

Author contributions

A.B.G. conceptualized and designed the study, procured the grants, drafted and edited the manuscript. S.Y. performed the experiments, analysed the data, and drafted the manuscript. R.K. performed the experiments and analysed the data. W.M. and P.R. analysed the data. R.S. and G.P.S. edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.B.G.

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.