



Genetic Insights into Cumin (*Cuminum cyminum* L.): First Draft Genome Sequencing and Mining of SSR Markers

R.S. Tomar^{1,4}, R.K. Solanki³, M.V. Parakhia¹, S.B. Bhatt¹, S.M. Padhiyar¹, M.K. Mahatma², S. Kumari², V. Sharma², S. Choudhary², S. John², M.R. Sharma², M. Sharma², M.K. Vishal², R.S. Meena², R.D. Meena², K. Singh², Y.K. Sharma², V. Pandey⁵, S. Pandey⁵, S.K. Singh⁵ and V. Bhardwaj^{2*}

Reviewed by:

Dr. Vipulkumar B. Patel
College of Forestry Navsari
Agricultural University, Gujarat
Email: vipulbiotech@nau.in

*Correspondence:

V. Bhardwaj
Email: vinaycpri@gmail.com

Received: 24 September 2026

Revision: 29 September 2026

Accepted: 30 September 2026

Citation:

Tomar, R.S., Solanki, R.K., Parakhia, M.V., Bhatt, S.B., Padhiyar, S.M., Mahatma, M.K., Kumari, S., Sharma, V., Choudhary, S., John, S., Sharma, M.R., Sharma, M., Vishal, M.K., Meena, R.S., Meena, R.D., Singh, K., Sharma, Y.K., Pandey, V., Pandey, S., Singh S.K., & Bhardwaj V., (2026). Genetic Insights into Cumin (*Cuminum cyminum* L.): First Draft Genome Sequencing and Mining of SSR Markers, *Int J Seed Spice*. 16 (1): 39-53

DOI:

<https://doi.org/10.56093/IJSS.v16i1.5>

Affiliation:

¹Department of Biochemistry and Biotechnology, Junagadh Agricultural University, Junagadh-362 001, Gujarat, India

Abstract

Cumin (*Cuminum cyminum* L.) is a globally significant spice crop, highly valued for its distinctive aromatic volatile oil content and its natural drought tolerance. To accelerate molecular breeding for these traits, this paper presents the first de novo draft genome assembly of the widely cultivated Indian variety Gujarat Cumin 4 (GC-4), a cultivar known for its high yield. The genome was sequenced using a hybrid strategy, combining high-accuracy Ion S5 short reads with long-read PacBio sequencing technology to generate an initial draft genome resource. More than 69 million sequencing reads contributed to an assembled genome of approximately 835 Mbp. The resulting assembly was fragmented, comprising 15,088 scaffolds, with a scaffold N50 of 6.85 kb and a maximum scaffold length of 92.9 kb; therefore, it should be regarded as an initial draft rather than a highly contiguous or chromosome-level reference genome. Nevertheless, the assembly supported extensive downstream genome mining and annotation. A total of 37,342 protein-coding genes were predicted, with 16,920 genes assigned Gene Ontology terms and 9,463 genes mapped to KEGG pathways. Candidate genes associated with important biological and quality-related traits were identified, including 21 flavonoid-related genes, 9 chalcone synthase (CHS) genes, 15 terpene synthase (TPS) genes, and 89 disease-resistance-related genes. Genome-wide analysis also identified 133,814 SSR loci, predominantly dinucleotide repeats, providing a substantial resource for subsequent molecular marker development. Collectively, these results demonstrate the utility of the assembly as an initial genomic resource for gene discovery, functional annotation, candidate-gene mining, and marker development, while recognizing that improved long-read coverage and chromosome-scale scaffolding will be required for a more complete reference genome.

Keywords GC-4, draft genome, SSR marker, *Cuminum cyminum*, NGS, Flavonoid, Chalcone synthase, disease resistance

² ICAR- National Research Centre on Seed Spices, Ajmer, Rajasthan, India	EC	Enzyme Commission
³ ICAR-Central Arid Zone Research Institute, Jodhpur, Rajasthan, India	Gb	Giga base
⁴ ICAR-Directorate of Groundnut Research, Junagadh, Gujarat, India	GC4	Gujarat Cumin 4
⁵ ICAR- Division of Horticultural Science, KAB-II, New Delhi, India	GO	Gene Ontology
	KEGG	Kyoto Encyclopedia of Genes and Genomes
	LRR	Leucine-Rich Repeat
	MAS	Marker-Assisted Selection
	MBP	Megabase pairs
	MCL	Markov Cluster Algorithm
	NBS	Nucleotide-Binding Site
	PacBio	Pacific Biosciences
	QTL	Quantitative Trait Locus
	SSRs	Simple Sequence Repeats
	TIR	Toll/Interleukin-1 receptor
	TPS	Terpene Synthase
Abbreviations:		
BUSCO	Benchmarking Universal S	
CC	Coiled-Coil	
CDS	Coding Sequences	
CTAB	Cetyl Trimethyl Ammonium	

Introduction

Cumin (*Cuminum cyminum* L., 2n = 14), commonly known as "Jeera," is widely used in household kitchens, adding a distinctive flavour to various culinary dishes. This annual herbaceous plant, renowned for its aromatic properties, serves a multifaceted role as a culinary enhancer, a captivating additive, and a therapeutic agent. Owing to its rich repository of essential metabolites, cumin has attracted attention for its applications in food, pharmaceuticals, and essential oils (Heidari and Sadeghi, 2014).

The regions of Asia, the Middle East, and North Africa have emerged as primary custodians of cumin, with India and Iran standing out as the largest producers as well as exporters (Lim, 2013). Its widespread appeal is evident in its incorporation into culinary creations, breweries, liquors, pharmaceuticals, toiletry preparations, and perfumes worldwide (Kumar *et al.*, 2015). Over millennia, the historical narrative of cumin traces its origins to ancient Egypt and then follows its journey across the Middle East, India, China, and the Mediterranean. Among its myriad counterparts, cumin holds the distinction of being India's second most exported spice, following closely behind chilli. In fiscal year 2023-24, India exported 184,000 tons of cumin, generating \$700 million in foreign exchange, with exports projected to reach 192,000 metric tons in 2024-25.

In scientific research, whole-genome sequencing has dismantled traditional barriers, revealing a rich data repository that accelerates plant studies. A pivotal moment in 2000 marked the completion and publication of the genome sequence of *Arabidopsis thaliana*, providing invaluable insights into higher-plant sequencing. Since then, genome sequencing across

species has advanced, revealing the latent potential of crops to combat adversity. Within the extensive family of flowering plants, Apiaceae, which comprises approximately 4000 species across 434 genera, plays a prominent role. However, progress in this plant family is hindered by a lack of complete genome data, which impedes the development of Apiaceae crops. Biomolecular resources for Apiaceae species remain limited, creating a bottleneck for scientific exploration.

The genome sizes of known plants in the Apiaceae family show considerable variations among species, like Coriander's (*Coriandrum sativum* L.) assembled genome size is approximately 2.13 Gb (Song *et al.*, 2020), Fennel (*Foeniculum vulgare* Mill.) is around 1.32-1.43 Gb (Palumbo *et al.*, 2018), Celery (*Apium graveolens* L.) is 2.21 Gb (Li *et al.*, 2020) and Parsley (*Petroselinum crispum* Mill.) is an estimated 1.85 Gb (Liu *et al.*, 2025). The *de novo* draft genome assembly of cumin was the largest, with 13.3 Gb covering 58.6M spots and 17.2G bases submitted to NCBI and DDBJ (Accession Number: DRX467835, Sample ID: Ion Torrent S5 sequencing of SAMD00611418).

The cumin genome is five times larger than the carrot genome and more than twice the size of the fennel genome. The substantial sequencing volume was required to obtain high coverage depth, thereby ensuring reliable representation of repetitive and difficult-to-sequence regions in the approximately 2.66 Gb (2n) cumin genome, as estimated by flow cytometry (Tomar *et al.*, 2021). Genome size variation among Apiaceae species is primarily associated with polyploidy and the differential accumulation of repetitive DNA sequences, especially the expansion of long terminal repeat retrotransposons, a notable phenomenon reported in parsley (Liu *et al.*, 2025).

The first-draft genome of cumin is a significant genomic resource for this economically important seed spice crop. The availability of a reference genome provides a foundation for understanding the genomic architecture of cumin and enables comprehensive gene prediction and functional annotation. Such genomic resources are particularly important for cumin because its narrow genetic base, arising from its monotypic nature, limits the available genetic variation and increases its vulnerability to biotic and abiotic stresses, thereby adversely affecting crop productivity and seed quality. The genome sequence will help identify genes and regulatory networks associated with key agronomic traits, including plant growth and development, stress responses, and secondary metabolite biosynthesis. In addition, genome sequence information has enabled the development of molecular markers, particularly SSRs. A total of 133,814 SSR loci were identified from 69,538 genomic sequences, providing a valuable resource for genetic diversity studies, linkage mapping, and marker-assisted breeding in cumin. Furthermore, the reference genome provides an opportunity to characterize key gene families, including transcription factors, disease resistance genes, and genes involved in the biosynthesis of valuable bioactive compounds and essential oils. These genomic resources will enhance our understanding of the molecular basis of important traits and facilitate the development of improved cultivars with enhanced stress tolerance, superior seed quality, and greater productivity. Overall, the cumin genome sequence provides a valuable platform for functional genomics, comparative genomics, and molecular breeding, promoting sustainable genetic improvement of this important aromatic and medicinal crop.

Materials and Methods

Plant material and DNA extraction

The cumin genotype GC-4, a high-yielding and widely cultivated variety in India, was selected for whole-genome sequencing. Authentic GC-4 seeds were obtained from the ICAR-National Research Centre on Seed Spices (ICAR-NRCSS), Ajmer, Rajasthan. Plants were grown under controlled greenhouse conditions, and young, healthy leaves were harvested for genomic DNA isolation. Large fragments of DNA with high molecular weight were extracted using a modified cetyl trimethyl ammonium bromide (CTAB) protocol optimized for aromatic seed spices. DNA quantity and purity were assessed spectrophotometrically (Nano Drop; A_{260}/A_{280}

ratio), fluorometrically (Qubit 2.0), and by agarose gel electrophoresis to ensure integrity before library preparation. Electrophoresis of the cumin sample: Lane 1 displays the cumin DNA sample; Lane 2 shows a duplicate of the cumin DNA sample; Lane M contains the λ Hind III DNA ladder. Both cumin DNA samples show high-molecular-weight, intact DNA bands with little smearing, indicating good quality and purity for downstream applications (Fig. S1).

Bioanalyzer library loaded in Agilent DNA HS chip.

The lack of low-molecular-weight peaks indicated that the library was devoid of adapter dimers and minor contaminating fragments, confirming effective cleanup. The overall fluorescence intensity and profile shape suggested adequate DNA concentration and fragment uniformity, making the library appropriate for long-read sequencing applications (Fig. S2).

Genome sequencing and library preparation

Genome sequencing was performed using a hybrid strategy combining Ion S5 (Thermo Fisher Scientific) and PacBio RS II (Pacific Biosciences) platforms. For Ion S5 sequencing, we prepared paired-end libraries with insert sizes of approximately 200 bp, 400 bp, and 600 bp following the manufacturer's standard protocols. Sequencing was conducted at the Department of Biotechnology, Junagadh Agricultural University, Junagadh, Gujarat, India.

For long-read sequencing, a high-molecular-weight DNA library (>10 kb) was constructed and sequenced on the PacBio RS II system at Xcelris Labs Ltd, Ahmedabad, to generate long continuous reads for assembly scaffolding and repeat resolution. Combining short- and long-read technologies enabled accurate base calling and improved assembly contiguity for the GC-4 genome.

Read quality assessment and genome assembly

Raw sequence reads generated from both Ion S5 and PacBio platforms were subjected to a comprehensive quality assessment before assembly. Quality parameters, including fragment size distribution, G+C content, base-wise quality scores, and ambiguous base frequency, were evaluated using Fast QC v0.11.9 (Andrews, 2010). Poor quality reads, adapter dimers, and short or ambiguous sequences were removed using Trimmomatic v0.39 with default filtering parameters to ensure high-quality datasets for assembly.

We assembled long-read data from the PacBio RS II platform using Canu v1.5 (Koren et al., 2017), optimised for error correction and long-read scaffolding. Short-read data from the Ion S5 platform were independently assembled using SPAdes v3.15.2 (Antipov et al., 2016) with default parameters. A hybrid genome assembly was subsequently performed by combining PacBio and Ion S5 reads using SPAdes hybrid mode and the de novo assembler in CLC Genomics Workbench v9.5 (Qiagen, Hilden, Germany). Assembly quality was evaluated using the N50 statistic, total assembly length, and contig number. BUSCO v3 (Simão et al., 2015) was used to assess genome completeness based on the Eudicotyledons odb10 ortholog set under default parameters.

Gene prediction and annotation

Identification of potential cumin transcripts within a specific genomic area was performed using Augustus 3.2.2 and SNAP. For similarity-based gene prediction, five sequenced plant species (*Arabidopsis thaliana*, *Mimulus guttatus*, *Solanum tuberosum*, *Solanum lycopersicum*, and *Daucus carota*) were chosen, and their protein sequences were sourced from Phytozome v12 (<http://www.phytozome.net>). The BLAST tool with parameters (identity ≥ 0.95 ; coverage ≥ 0.90) was used to pinpoint ESTs, mRNAs, and proteins that exhibit significant homology to the cumin genome sequence. The thresholds for identity and coverage used in the alignment were established on the basis of the methods applied for the carrot and potato genomes. After filtering and clustering, manual curation and iterative refinement rounds with Exonerate v2.2.0 were executed to realign and refine the sequences.

Functional annotation of protein-coding genes was conducted using BLASTP with an E-value of 1×10^{-4} , targeting multiple databases such as NCBI non-redundant protein sequence (nr), TrEMBL, Swiss-Prot, GO, and KEGG. Blast2GO retrieved the corresponding GO IDs, while WEGO was used to depict the distribution of gene functional profiles visually. This integrated approach ensured a thorough exploration of cumin-specific transcripts and their functional annotations within the genomic context.

Gene family identification and clustering

Gene family analysis was conducted using the OrthoMCL pipeline to identify orthologous and paralogous gene

clusters among cumin and selected reference species. Protein-coding gene sets for (*Mimulus guttatus* v2.0, *Solanum lycopersicum*) iTAG2.4, *Daucus carota* v2.0, and *Arabidopsis thaliana* (TAIR10) were retrieved from the Phytozome v12 database (<http://www.phytozome.net>). For each species, the longest isoform per gene was selected to ensure uniformity in subsequent comparisons.

All-against-all sequence similarity searches were performed using BLASTP (E-value threshold = 1×10^{-5}). Alignments with less than 50% coverage of either sequence were excluded to minimize spurious hits. The filtered similarity matrix was then clustered using the Markov Cluster Algorithm (MCL) implemented in OrthoMCL, applying an inflation parameter of 1.5, to delineate orthologous gene families shared across the studied species.

Functional gene identification and pathway annotation

To identify and characterize key functional genes in the cumin genome, we adopted a domain-based and homology-driven annotation approach. Resistance-related genes were predicted by screening the predicted protein sequences for conserved domains characteristic of plant disease resistance proteins, incorporating structures like CC, TIR, NBS, and LRR motifs. Domain-specific profiles corresponding to these motifs were retrieved from multiple protein family databases, including PFAM, PANTHER, PRINTS, PROSITE, SMART, and SUPERFAMILY. The abundance of CC motifs was specifically verified using PAIRCOIL2.

To identify genes associated with secondary metabolic pathways, we annotated assembled protein-coding genes with EC numbers and mapped them to biochemical pathways in the KEGG database. Genes related to the flavonoid biosynthesis pathway were extracted based on their assigned EC numbers. In addition, putative terpenoid synthase (TPS) proteins were identified using HMMER v3.0 to search PFAM domain models PF03936 (terpene synthase N-terminal domain) and PF01397 (terpene synthase C-terminal domain). Fig. 1 shows the workflow for genome sequencing of cumin genotype GC-4.

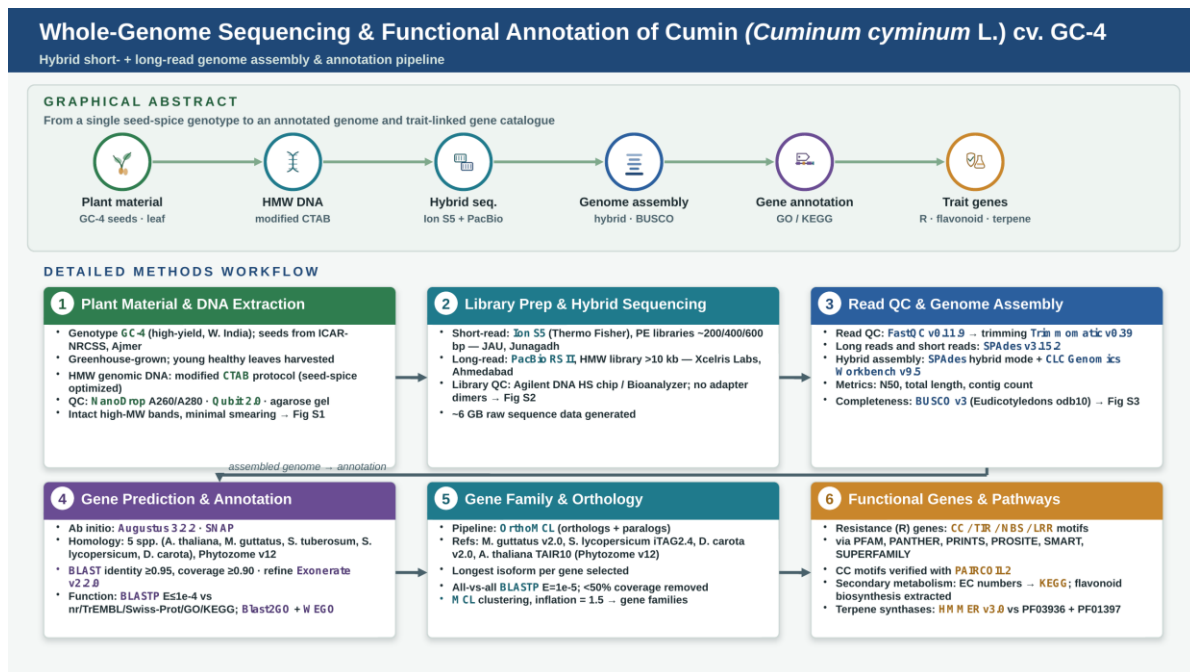


Fig. 1 Work flow of methodology for genome sequencing of cumin genotype GC-4

Identification and development of SSR markers

Simple sequence repeats were identified from the cumin genome assembly using MegaSSR, a freely available software for large-scale SSR identification and high-throughput SSR marker development. The MegaSSR pipeline integrates MISA for identification of microsatellite repeats, Primer3 software for designing PCR primers flanking the identified SSR loci, primer search (EMBOSS) for *in silico* validation of the designed primers, and Bandwagon were used for *in silico* visualization of the expected PCR amplification profiles. The pipeline identified SSR loci across the genome and developed corresponding primer pairs.

PCR Amplification, gel electrophoresis and band analysis for SSR primers

Twenty SSR markers were used for validation in cumin genotype GC-4. PCR amplification was carried out in an Applied Biosystems Veriti Thermal Cycler using the following program: initial denaturation at 95°C for 4 minutes, followed by 35 cycles of denaturation at 95°C for 30 seconds, annealing at optimized temperatures (53–65°C, depending on the primer) for 30 seconds, and extension at 72°C for 1 minute. A final extension step was performed at 72°C for 7 minutes. The amplified PCR products were resolved by electrophoresis on a 1.5% agarose gel. A 100 bp DNA ladder was used as a molecular size marker for estimating band sizes.

Results and discussion

Cumin genome assembly

The results of the genome assembly of cumin described here represent the first genomic resource for this economically important seed spice crop. A total of 69.13 million filtered and trimmed reads were generated from Ion S5 (200–600 bp libraries) and PacBio RS II (>10 kb libraries), with average read lengths ranging from 203 bp to >3,000 bp (Supplementary Table S1).

Table 1 Statistics of cumin genome data and assembled reads

Description	Statistics
Sequencing data	>33.39 Gb
Estimated/assembled draft genome	~835 Mb
Contigs	269,364
Average contig length	3,149 bp
Calculated assembly size	~835 Mb
Maximum contig	96,509 bp
N50	5,638 bp
Total no of scaffolds	15088
Scaffolds N50	6850
Average scaffold length	6702
Maximum scaffold length	92966

Based on the average read lengths, the Ion S5 sequencing generated approximately 18.28 Gb of sequence data, whereas the PacBio RS II platform generated more than

15.10 Gb. Thus, the combined sequencing dataset comprised more than 33.39 Gb of sequence data (Table 1), providing substantial sequence coverage for genome assembly. A hybrid genome assembly was achieved by combining PacBio long reads and Ion S5 short reads using the hybrid mode of SPAdes and the de novo assembler implemented in CLC Genomics Workbench. The resulting draft cumin genome assembly was approximately 835.6 Mbp. The assembly comprises 15,088 scaffolds, indicating reasonable contiguity, further

supported by an N50 scaffold length of 6,850 bases and a maximum scaffold length of 92,966 bases. The overall GC content of the assembled cumin genome was determined to be 37.37%. The CLC-based assembly produced 269,364 contigs totalling ~835 Mbp, with a maximum contig length of 96,509 bp and an N50 of 5,714 bp. In comparison, the SPAdes assembly resulted in 159,839 contigs totalling ~308 Mbp, with a maximum contig length of 22,904 bp and an N50 of 2,727 bp (Table 2).

Table 2 Assembly statistics of the *Cuminum cyminum* GC-4 genome using CLC and SPAdes

Assembly Metric	CLC_Cumin Assembly	SPAdes_Cumin Assembly
Total contigs (≥ 0 bp)	269,364	16,530,727
Contigs $\geq 1,000$ bp	188,402	102,975
Contigs $\geq 5,000$ bp	58,415	10,722
Contigs $\geq 10,000$ bp	11,562	1,027
Contigs $\geq 25,000$ bp	213	0
Contigs $\geq 50,000$ bp	38	0
Total length (≥ 0 bp)	848,092,686 bp	1,467,397,550 bp
Total length ($\geq 1,000$ bp)	803,575,029 bp	267,799,022 bp
Total length ($\geq 5,000$ bp)	478,240,220 bp	76,225,617 bp
Total length ($\geq 10,000$ bp)	154,252,233 bp	12,484,566 bp
Total length ($\geq 25,000$ bp)	8,257,735 bp	0
Total length ($\geq 50,000$ bp)	2,903,897 bp	0
Total number of contigs	231,976	159,839
Largest contig (bp)	96,509	22,904
Total assembly length (bp)	835,573,982	308,477,586
GC content (%)	35.46	34.27
N50 (bp)	5,714	2,727
N90 (bp)	1,655	907
L50 (number of contigs)	47,104	32,136
L90 (number of contigs)	148,858	113,303
N's per 100 kbp	0	0

Differences in assembly statistics between SPAdes and CLC Genomics Workbench may reflect their distinct assembly algorithms and approaches to integrating PacBio long reads with Ion S5 short reads. Therefore, the observed differences reflect assembler-specific reconstruction of the same genomic data rather than differences in the underlying sequencing dataset. The relatively fragmented assembly (N50 = 5.7 kb) suggests that additional high-quality long-read sequencing at adequate depth, along with complementary short-read data and chromosome-level scaffolding such as Hi-C, would be needed to improve genome completeness and contiguity. The final cumin assembly (835 Mb) represented only ~64% of the estimated haploid genome size (~1.33 Gb) and showed low contiguity (N50 = 6,850

bp) despite using PacBio and Ion S5 sequencing technologies. Similar discrepancies between estimated and assembled genome sizes have been reported in several plant species with large and repetitive genomes. In parsley, the assembled genome (~2.12 Gb) was about 10–12% smaller than the estimated genome size (~2.36 Gb), largely because repetitive elements account for nearly 78% of the genome (Liu et al., 2025). Likewise, the tea plant (*Camellia sinensis*) genome (3.02 Gb) possesses highly repetitive sequences (80.9%), resulting in assembly challenges and incomplete recovery of repetitive regions (Xia et al., 2017). The likely abundance of repetitive sequences in cumin may therefore have caused repeat collapse and assembly fragmentation. Furthermore, the absence of Hi-C sequencing data

prevented chromosome-level scaffolding and accurate contig ordering, contributing to the reduced assembly completeness and low contiguity observed in the present study.

We evaluated the cumin genome assembly completeness using BUSCO (Benchmarking Universal Single-Copy Orthologs) version 4.1.1. Assembly metrics and BUSCO results are presented in Fig. S3. A total of 2,326 conserved ortholog groups were searched, of which 1,069 genes were successfully identified in the assembly (Table S2). This indicates that a substantial portion of the expected single-copy orthologs is present. BUSCO analysis indicates relatively low genome completeness (~46%) compared with the ≥90% completeness generally expected for high-quality reference genomes. This reduced BUSCO score likely reflects the fragmented nature of the current assembly and the presence of missing genomic regions. However, BUSCO completeness values can vary considerably among plant genome assemblies and are not always strongly correlated with assembly contiguity. Recent comparative analyses of plant genomes have shown BUSCO completeness ranging from 60% to 99%, with several species exhibiting relatively low completeness despite high assembly contiguity. Species such as *Fraxinus angustifolia*, *Ginkgo biloba*, *Gastrodia elata*, and *Paeonia suffruticosa* have shown low BUSCO scores because of their large, complex, repeat-rich genomes (Cheng et al., 2025). Similarly, cumin's large genome size, probable abundance of repetitive elements, and lack of Hi-C-based chromosome scaffolding may have hindered recovery of conserved orthologs. Therefore, the low BUSCO completeness observed in the present study likely reflects the inherent complexity of the cumin genome and limitations in assembly completeness rather than the complete absence of these essential genes.

Despite these limitations, this assembly provides valuable foundational resources for the cumin research community. The lack of a complete reference genome for cumin has been a major challenge for genome editing and trait improvement efforts. The assembly contains usable gene space (1,069 complete BUSCOs), enabling preliminary investigations of metabolic pathways relevant to cumin's pharmacological properties (e.g., terpenoid and essential oil biosynthesis), disease resistance mechanisms against *Fusarium oxysporum* f. sp. *cumini*, and comparative genomics within Apiaceae.

Gene prediction and annotation

Potential cumin-specific coding sequences (CDS) were identified using *de novo* gene prediction tools Augustus v3.2.2 and SNAP. For homology-based prediction, protein sequences from five sequenced plant species - *A. thaliana*, *M. guttatus*, *S. tuberosum*, *S. lycopersicum*, and *D. carota* were retrieved from Phytozome v12 (<http://www.phytozome.net>). BLAST searches (identity ≥ 95%; coverage ≥ 90%) were performed to assess significant homology in ESTs, mRNAs, and proteins to the cumin genome. The thresholds for sequence identity

and coverage were selected following strategies used in carrot and potato genome studies. Subsequently, manual curation and iterative refinement were performed using Exonerate v2.2.0 for assembly polishing and clustering of the predicted sequences. Functional annotation of protein-encoding genes was carried out using BLASTP (E-value ≤ 1×10^{-4}) against multiple databases, including NCBI non-redundant (nr), TrEMBL, Swiss-Prot, Gene Ontology (GO), and KEGG. GO identifiers were assigned using Blast2GO, and WEGO was employed to visualize the distribution of gene functional classifications.

By combining *de novo* prediction with homology-based methods, a total of 37,342 protein-coding genes were predicted in the cumin genome. These genes have an average length of 3,267 bp, with a mean of 4.83 exons per gene (Table 3).

Table 3 Genome features and gene structure statistics of cumin

Parameter	Value
Genome GC (%)	37.37%
Exon region size (bp)	4,970,297
Exon region (%)	0.58
Exon region GC (%)	42.06
Gene number	37,342
Genes in families	27,549
Unclustered genes	9,763
Family number	15,164
Average genes per family	1.82
mRNA number	37,342
Average mRNA length (bp)	3,267
Exon number	180,591
Average exon length (bp)	243.48
Exon No. per gene	4.83

The predicted gene set provides a comprehensive resource for functional genomics studies and offers insights into pathways underlying important agronomic and nutritional traits in cumin. A total of 37,342 genes were predicted, with an average mRNA length of 3,267 bp and an average of 4.83 exons per gene. Exons collectively span ~0.58% of the genome, with a GC content of 42.06%, while the overall genome GC content is 37.37%. A search was conducted for sequence similarity against public databases to explore potential functions of putative genes. A cumulative of 34,143 genes were annotated utilizing the Nr, InterPro, GO, and KEGG databases (Table S3). Data in table reveal the functional annotation of predicted gene models in the cumin genome. Homology- and domain-based searches against the NCBI Nr, InterPro, GO, and KEGG databases identified annotations for most predicted genes. The number of annotations often exceeds the number of sequences because genes can have multiple functional assignments, providing a comprehensive overview of the cumin genome's functional potential.

Based on the Gene Ontology database, a total of 16,920 predicted genes in the cumin genome were functionally annotated across three main GO categories, namely Biological Process, Cellular Component and Molecular Function, comprising 1,223 distinct functional terms (Fig. 2). Among biological processes, genes associated with oxidation-reduction, protein phosphorylation, and transcriptional regulation were most abundant. For

cellular components, genes were predominantly linked to membranes, nuclei, and ribosomes, while molecular function annotations highlighted protein binding, ATP binding, and kinase activities as the most represented functional categories.

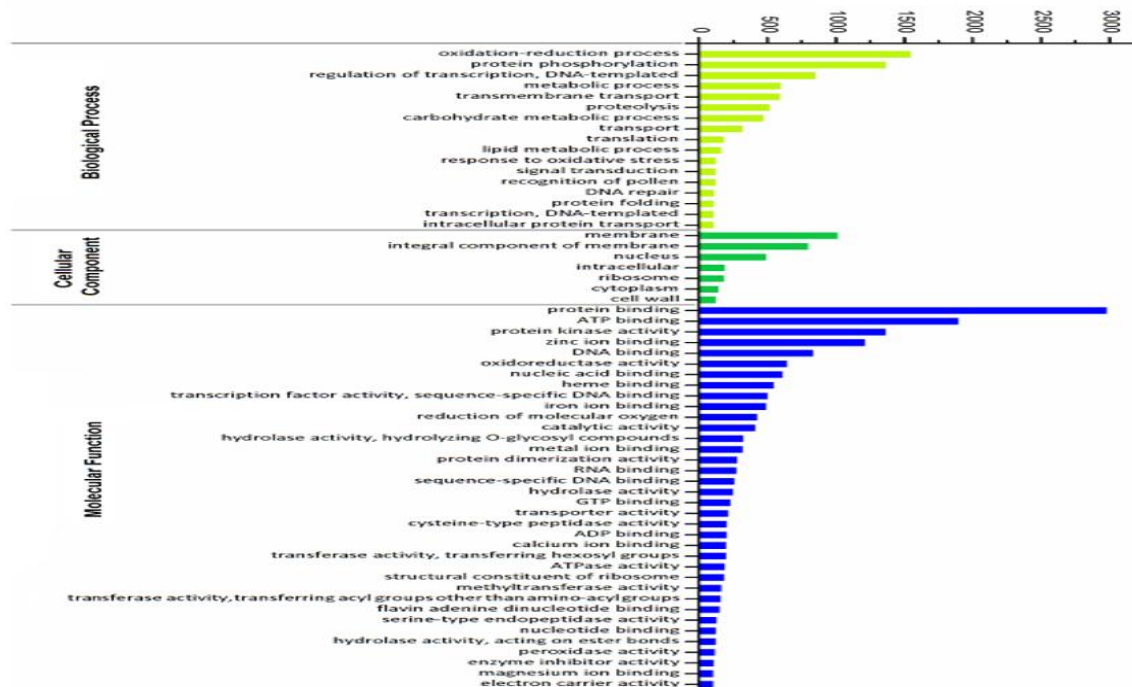


Fig. 2 Functional classification of cumin genes based on Gene Ontology (GO) annotations, showing the distribution across Biological Process (light green), Cellular Component (green), and Molecular Function (blue) categories

Based on pairwise protein sequence similarities, we analysed the cumin genome for gene families. Cumulatively, 27,549 genes were grouped into 15,164 gene family clusters, with an average family size of 1.82 genes, while 9,763 genes remained unclustered (Table 3). Gene family sizes varied considerably, with group 4 containing the largest number of genes (201 members). Gene ontology analysis of these families revealed that the most populous families are predominantly associated with key biological processes, including cell recognition, protein phosphorylation, flavin adenine dinucleotide binding, serine-type endopeptidase activity and zinc ion binding, indicating their potential functional importance in cumin biology. Gene Ontology analysis of these families showed that the largest families are mainly linked to essential biological processes such as cell recognition, serine-type endopeptidase activity, flavin adenine dinucleotide binding, protein phosphorylation, and zinc ion binding, suggesting their possible functional significance in cumin biology.

Functional genes in cumin

Cumin resistance-related genes were identified through the most conserved motif structures of plant resistance proteins, including coiled coil (CC), Toll/interleukin-1 receptor (TIR), nucleotide-binding site (NBS), and leucine-rich repeat (LRR) domains. Conserved motifs were predicted using domain profiles from software tools like PFAM, PANTHER, PRINTS, PROSITE, SMART, and SUPERFAMILY databases, with CC domains specifically predicted using PAIRCOIL2. We annotated genes involved in flavonoid biosynthesis using EC numbers from the KEGG database, and identified TPS genes by screening the cumin genome with HMMER 3.0 using PF03936 and PF01397 domain models.

The assembled genome comprises 37,342 predicted mRNAs and 180,591 exons spanning 4,970,297 bp. Functional annotation was performed using multiple databases, including Nr (39,798 genes), InterPro (39,311 genes), Gene Ontology (GO, 38,141 genes), and KEGG (9,463 genes), enabling comprehensive characterization of gene functions and pathways. Based on gene prediction and functional annotation, cumin-specific genes were classified into several functional categories (FV, Table 4).

Table 4 Predicted functional genes in cumin

Character	Number of genes	Representative genes
Flavonoid	21	U78D2, C75A2, C75A4, C75B3, C93C2, F3PH, FAOMT, FL3H3, MOMT, SOMT2, SOT5, UFOG, UFOG1, UFOG2, UFOG3, UFOG4, UFOG5, UFOG6, UFOG7, UGFGT and Y1103.
Chalcone synthase	9	6DCS, CHS1, CHS2, CHSA, CHSB, CHSD, CHSL1, CHSY, PKS5
Chalcone isomerase	4	CFI, CFI1, CFI2B, CFI3
Flavanone synthase	3	C93C1, FNSI, C93B1,
Terpenoid synthase	15	BAMS, GBIS1, GBIS2, HUMS, TPS01, TPS05, TPS07, TPS08, TPS09, TPS16, TPS18, TPS22, TPS26, TPS29, TPS30
Disease resistance	89	ADR2, CDR1, CHS1, CSA1, DF230, DR206, DRL12, DRL13, DRL14, DRL15, DRL16, DRL17, DRL18, DRL19, DRL2, DRL20, DRL21, DRL23, DRL24, DRL25, DRL26, DRL27, DRL28, DRL29, DRL3, DRL30, DRL31, DRL32, DRL33, DRL34, DRL36, DRL37, DRL38, DRL39, DRL4, DRL40, DRL41, DRL42, DRL43, DRL45, DRL5, DRL7, DRL8, DRL9, DSC1, DSC2, EDR1, EDR2, EDR2L, EDR4, LAZ5, LOV1A, NDR1, R13L1, R13L2, R13L3, R13L4, RFL1, RGA1, RGA2, RGA3, RGA4, RLM1B, RLM3, RP8HA, RP8L2, RP8L3, RP8L4, RPM1, RPP1, RPP13, RPP4, RPP5, RPP8, RPS2, RPS4C, RPS4L, RPS4W, RPS5, RPS6C, RPS6R, RPS6R, SUMM2, TAO1, WR52C, WR52N, WR52R, WR52W, Y4117
Antifungal	4	DEF1, DEF15, DEF2, DEF4
Early flowering	13	ASHH2, EFM, ELF3, ELF6, HD16N, PAF1, PIE1, REF6, RUP1, RUP2, SKIP, SWC6, VIP6
Aromatic	11	5MAT, ANTA, AVT3A, AVT3B, AVT3C, DDC, ISS1, PGL1, PGL2, PGL3, SOT16,
Drought	8	AL7A1, DIS1, ERG14, HDG11, LSM5, SAD2, SDIR1, SSP1A
cold	22	ARF1, CAX1, CO410, CRIM1, CRIM2, CRPK1, CRPM1, CRPM1, CRPM2, CSP1, CSP2, CSP3, DERL1, DRE1A, FBK83, GSTU6, MSI4, PER3, RBG7, RL131, SRC2, ZAT6
Rust	12	ADR2, LRK10, LRL11, LRL12, LRL13, LRL15, LRL21, LRL22, LRL24, LRL25, LRL26, LRL27
Germination	11	CHR10, DOF37, NPG1, NPGR1, NPGR2, P2C37, P2C75, PARN, PBL21, RDL1, RVE3
Nematodes	2	ELF3, HSPR2
Heat	189	1FEH1, ALMT1.....
Transcription factor	3023	ABI4, ABI3, ABI4, ABI5, ABR1, AGL8, AI5L2, AI5L4, AI5L5, AI5L6, AI5L7, AI5L8, AIL1, AIL5, AIL6, AIL7, ALC, AMS, ANL2, ANT, AP2L1, AP2L3, AP2L4, APG, APL, ARFE, AS1, ASIL1, ASIL2, ASR3, ATB12, ATB13, ATB14, ATB15, ATB16, ATB17, ATB21, ATB23, ATB40, ATB51, ATB52, ATBH9, ATHB4, ATHB5, ATHB6, ATHB7, ATML1, BBM, BBM2,

The identification of genes associated with flavonoid biosynthesis (21 genes), including chalcone synthase (9), chalcone isomerase (4), flavanone synthase (3), and terpenoid metabolism (14), indicates a diverse genetic

framework for the biosynthesis of bioactive compounds in cumin. Flavonoids and related phenolic compounds are associated with antioxidant and other potential health-promoting properties. At the same time, terpenes

constitute major components of cumin essential oil and contribute to its characteristic aroma, flavour, and biological activities. The presence of these biosynthetic genes therefore highlights the potential of cumin as a source of bioactive phytochemicals with nutraceutical relevance. It provides candidate targets for further investigation of the genetic basis of its phytochemical composition and health-promoting properties.

A large number of disease-resistance genes (89), including RPS, RPP, RPM, RGA, and NDR1, indicate diverse components of plant immune-signalling and pathogen-response pathways. The presence of four defensin-like antifungal genes further supports potential mechanisms of innate defence against fungal pathogens. Genes associated with drought (8), cold (22), and heat responses (189) suggest a broad genetic repertoire potentially involved in adaptation to environmental stresses. The identification of 13 early-flowering genes indicates the presence of regulatory components involved in flowering-time control and developmental transitions. Genes associated with germination (11) may contribute to seed dormancy release, germination regulation, and early seedling establishment. The detection of 11 aromatic-related genes suggests potential involvement in aromatic compound biosynthesis and modification. Two genes associated with nematode responses, ELF3 and HSPR2, provide candidate loci for investigating nematode-related stress responses. Overall, identifying 3,023 transcription-factor genes highlights extensive regulatory complexity and provides a broad set of candidate genes for studying metabolism, development, stress tolerance, and disease resistance in cumin.

Comparative analysis with available Apiaceae genomes indicates that the functional composition of the cumin genome is largely consistent with other members of the family. Like parsley, celery, coriander, and fennel, cumin has numerous genes associated with stress responses,

secondary metabolite biosynthesis, and disease resistance, reflecting the adaptive and medicinal importance of Apiaceae species (Li et al., 2020; Liu et al., 2025). The identification of genes involved in flavonoid biosynthesis, heat tolerance, and fungal defence in cumin parallels observations in coriander and parsley, where secondary metabolite pathways contribute to aroma and bioactive compound production (Song et al., 2020; Liu et al., 2025). Furthermore, the discovery of 133,814 SSR markers provides a valuable genomic resource comparable to other Apiaceae crops, facilitating genetic diversity analysis, linkage mapping, and marker-assisted breeding. Despite its lower assembly completeness, the cumin genome provides important functional and comparative genomic information. It establishes a foundation for future genomic improvement and breeding applications in this economically important seed spice crop.

SSR identification and mining

Simple sequence repeats in the cumin genome were identified and localized using MISA (<http://pgrc.ipk-gatersleben.de/misa/>). SSRs were classified based on the size and copy number of their tandem repeats as follows: monomer (≥ 8 repeats), dimer (≥ 4 repeats), trimer (≥ 4 repeats), tetramer (≥ 3 repeats), pentamer (≥ 3 repeats), and hexamer (≥ 3 repeats). A total of 269,364 sequences, spanning 835,092,686 bp, were analyzed for SSRs. Overall, 133,814 SSRs were identified within 69,538 sequences. Among these, 29,258 sequences contained more than one SSR, and 29,685 SSRs were present in compound formations (Table 4). The identified SSRs were distributed among different repeat types: dimers (112,728), trimers (15,820), tetramers (4,311), pentamers (440), and hexamers (515), highlighting the predominance of dimeric repeats in the cumin genome (Table S4). This distribution pattern is typical for SSRs in genomic data (Fig. 3) and suggests high mutability of

short repeats and selection against Long Repeats. The dominant form of microsatellites in the analyzed dataset, trend indicates short SSRs (6-9 repeats) are the most while the incidence of longer SSRs is rare.

Table 5 Summary of SSRs in the cumin genome

Parameters	Value
Total number of sequences examined	269,364
Total size of examined sequences (bp)	848,092,686
Total number of identified SSRs	133,814
Number of SSR containing sequences	69,538
Number of sequences containing more than 1 SSR	29,258
Number of SSRs present in compound formation	29,685
Total size of examined sequences (bp)	848,092,686
Total number of identified SSRs	133,814
Number of SSR containing sequences	69,538
Number of sequences containing more than 1 SSR	29,258
Number of SSRs present in compound formation	29,685

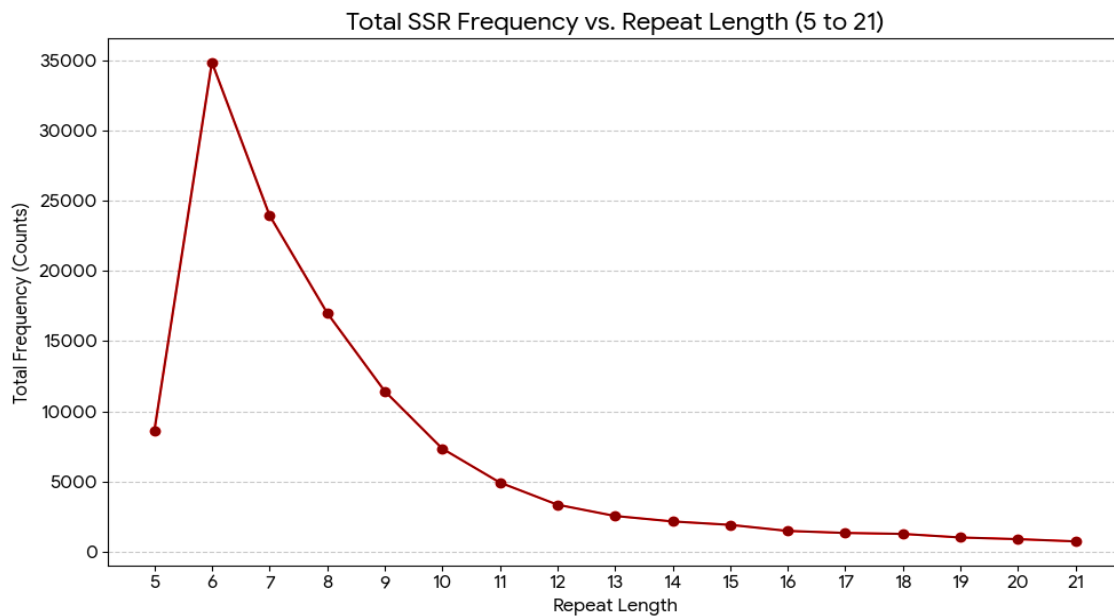


Fig. 3 The graph shows a clear inverse exponential relationship between the Repeat Length (x-axis) and the Total Frequency (y-axis)

Validation of identified SSR Markers

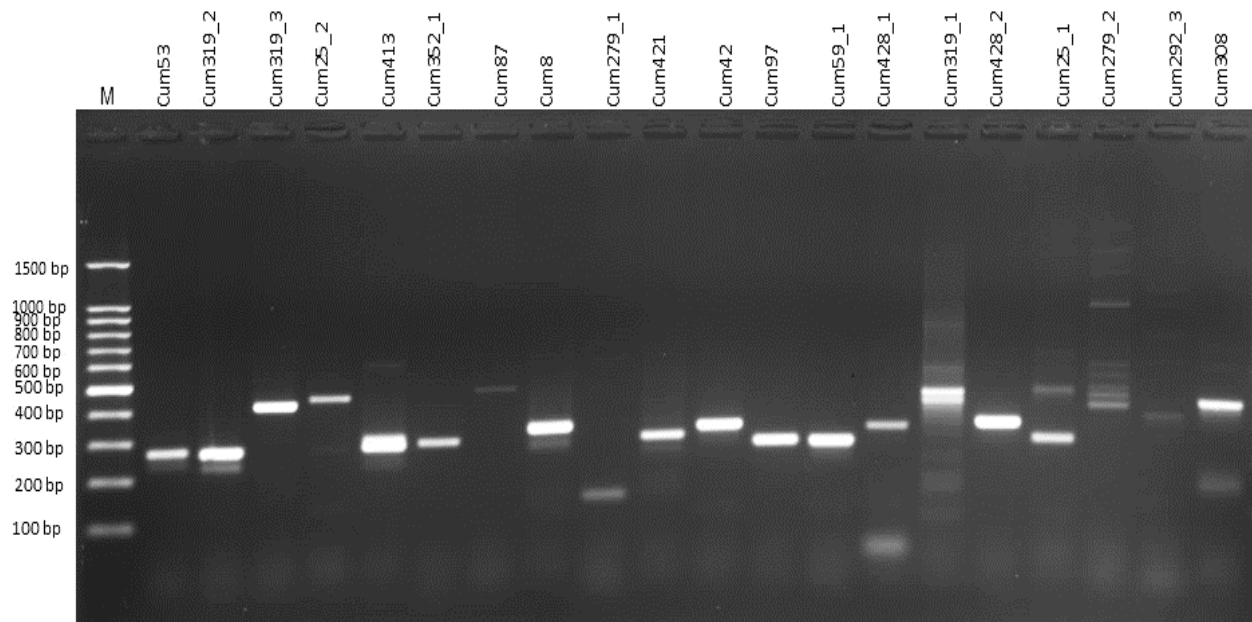
Twenty microsatellite makers were used for validation in cumin genotype GC-4 (Table 5). Gradient PCR was performed to validate 20 SSR markers in cumin genotype

GC-4. The annealing temperature ranged from 53°C to 65°C, and optimal temperatures were determined individually for each primer. Out of the 20 SSR markers analyzed, all markers showed successful amplification. Of these, 16 markers produced clear single bands, with

band sizes ranging from approximately 100 to 400 bp. (Table 6 and Fig 4). These results were confirmed by 4
The remaining 4 markers exhibited double or multiple times with PCR amplification and gel run.
bands, with band sizes ranging from 300 to 1000 bp

Table 6 Microsatellite makers validated for cumin genotype GC-4

Sr. No	Sequence ID	Repeat Type	Repeat Sequence	Repeat Size	Forward Primer	Reverse Primer
1	Cumin_contig_413	Penta	(TATCC)3	15	TGCAATTGACAACCTGA GGCT	GGAGAGCAGCATACC CGAAG
2	Cumin_contig_319_2	Di	(TC)6	12	AAGATCAGAGGTGAAA CAAGGA	GCATTTGCGGCATCC ATGG
3	Cumin_contig_319_3	Hexa	(ATTGTA)3	18	TGGCCACAATCACAAG TCAA	CCTGGTTCCCAAGTCA GGTAG
4	Cumin_contig_25_1	Di	(TG)7	14	TGTGTGTGTCTGTGTGT GTTG	GTCCCCCAAGGTCTG AACTG
5	Cumin_contig_25_2	Compound	(AC)7 a(AC)7	29	GGATAATGAGTTGCGC TACCG	TCCTTGCCTTCTCCTT GGTT
6	Cumin_contig_53	Tri	(AAG)5	15	TGCTCAGGAGGACTCA AACA	GCATTGTCTCCCTTGG CTCT
7	Cumin_contig_352_1	Di	(CA)6	12	GATGCCATGCTTCGTGA TACG	GATGAGTAGTTAAGT TCGGCCA
8	Cumin_contig_87	Di	(AT)7 (AAA)5	14	AGTGTGTCTTGCAACTT GGC	CAAACATTCAGGCCC AGCTC
9	Cumin_contig_279_1	Compound	(AAAAA)3 *...	14	AGGTGGCTGGCAGATA CAAC	AAACACTCTTGAGAGG GCAGG
10	Cumin_contig_279_2	Di	(TT)7	14	AAACACTCTTGAGGG CAGG	AGTGTGGTCCCATGT AGCTTG
11	Cumin_contig_8	Di	(TA)6	12	GTCAGAGAACTTGAGG TAAGCT	GAACATAGCACCGGC CCTAT
12	Cumin_contig_421	Di	(AT)7	14	AACGACGAGAATCTCC GAAAA	ACTGGTTAGGGTTTG ATTTGCT
13	Cumin_contig_42	Di	(AT)6	12	AACCAGGGGTGATTGC ACAT	GCTCCATATGCATCTT TGGCG
14	Cumin_contig_97	Di	(AC)7	14	TGCGCTTGCATGAACA CTTT	ATAGAAACAGGCGTG CGTGT
15	Cumin_contig_59_1	Di	(AT)7	14	AAAGGCATGTAGCTGA TGTGC	CAGACATTGATCTGG CTAGCG
16	Cumin_contig_428_1	Penta	(TAAAT)3	15	GCGATGGATGATGTTTC TGCC	CACGCTTGTTTTGTCT GGCA
17	Cumin_contig_428_2	Penta	(ATCAA)3	15	ATTGATTTACGGCGAG GCGA	GCTTTCCAACCCATTC CAACC
18	Cumin_contig_308	Compound	(TA)8 (TATA)4*	16	CCATCTTTTCTCGTCG ATGCG	GCACAGAAAACCTCGG GAGGA
19	Cumin_contig_292	Compound	(GAGAGA)4 (GA)12 *...	24	TGCCTCTGATCCACAAG CAG	TGGAGGCCTCTCACA TAGGT
20	Cumin_contig_292_3	Compound	(CC)7 (CCCCC)3 *...	15	CTTCCCCAGGAGCCAC AACT	CCTTACCAGCCCCCTC TTTG

Fig. 4. Gel electrophoresis of amplified Microsatellite markers (SSR) for Cumin genotype GC-4

These SSR markers constitute a potential molecular marker resource. Once evaluated across diverse germplasm, the validated markers can serve as a foundation for subsequent genetic diversity studies, linkage mapping, population genetic analysis, and marker-assisted breeding. The success of transferring SSR markers relies on the evolutionary relationships among species. In the Apiaceae family, only a limited number of SSRs have been previously reported, notably in carrot and celery. SSR has been proven to be a game-changer in the development and comprehensive characterization of 300 new markers in carrot (*Daucus carota*) and the development of 11 (expressed sequence tags) dbEST-derived microsatellite markers in celery (*Apium graveolens* L. var. *dulce*). Integrating SSR markers with lncRNA (long noncoding RNA)-based microsatellite marker studies is unprecedented in carrot breeding, opening new avenues for genetic improvement. This study may provide a scientific basis for cumin genotype identification and differentiation, evolutionary genetics, phylogenetic studies, and the breeding and protection of breeders' rights for other Apiaceae species in the future.

The relatively limited genome representation in the present assembly is a major consideration, given the difference between the 835 Mb assembled sequence and the approximately 1.3 Gb haploid genome size estimated from the reported 2C DNA content of 2.66 pg (flow cytometry). Thus, the current assembly represents approximately 65% of the expected haploid genome. Thus, high-coverage long-read sequencing

complemented by accurate short-read data and Hi-C scaffolding would be required to recover the remaining genomic fraction and generate a more complete, contiguous, and representative reference genome for cumin.

Conclusion

The draft-genome sequencing data identified genes and regulatory elements associated with essential oil biosynthesis and important agronomic traits, including resistance to major biotic stresses such as cumin blight caused by *Alternaria burnsii*. Data generated in this study also provide a valuable genomic resource for identifying high-density single nucleotide polymorphisms (SNPs) in cumin. Genome-wide SNP discovery can facilitate the development of high-throughput genotyping platforms such as genotyping-by-sequencing (GBS), which are widely used for genetic diversity analysis, linkage mapping, genome-wide association studies, and marker-assisted breeding. Compared with SSR markers, SNPs are more abundant, uniformly distributed across the genome, and highly amenable to large-scale automated genotyping. In addition, the SSR markers identified in this study constitute useful resources for marker-assisted selection and genetic improvement programs. These genomic resources provide valuable targets for molecular breeding and the development of high-yielding and stress-resilient cultivars. Furthermore, integrating genomic information with transcriptomic and metabolomic datasets will facilitate precision breeding approaches aimed at improving both agronomic performance and the biosynthesis of economically

important secondary metabolites, including cuminaldehyde, cymene, and cumin alcohol. Collectively, these resources establish a strong foundation for accelerating genetic improvement and ensuring sustainable production of this economically important spice crop.

Acknowledgements

This research was supported by the Indian Council of Agricultural Research, Department of Agricultural Research and Education, Government of India, under the Extramural Research Project. The authors are also thankful to Dr B.A. Golakiya, Retd. Professor and Head, Department of Biotechnology, Junagadh Agricultural University, Junagadh, Gujarat, for providing the necessary facilities to conduct this research work.

Declaration of Funding and Competing interests

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest. The authors have no competing interests to declare that are relevant to the content of this article.

Data availability statement

The datasets generated and/or analysed during the current study are given in the manuscript, and additional data are available at the DNA Data Bank of Japan (DDBJ) and NCBI with Accession Number: DRX467835, Sample ID: SAMD00611418.

Funding

Funding received under Extramural Research Project of Indian Council of Agricultural Research, New Delhi.

References

Acquadro A, Magurno F, Portis E, Lanteri S (2006) dbEST-derived microsatellite markers in celery (*Apium graveolens* L. var. dulce). Mol Ecol Notes 6(4): 1080-1082.

Al-Snafi A E (2016) The pharmacological activities of *Cuminum cyminum*: A review. IOSR J Pharm 6(6): 46-65.

Antipov D, Korobeynikov A, McLean J S, Pevzner PA (2016) HybridSPAdes: An algorithm for hybrid assembly of short and long reads. Bioinform 32:1009-1015.

Avatar R, Dashora S L, Sharma R K, Sharma M M (1991) Analysis of genetic divergence in cumin (*Cuminum cyminum* L.). Indian J Gen Plant Breed 51(3) m:289-291.

Bahraminejad A, Mohammadinejad G (2013) Use of microsatellite markers for molecular characterization of cumin (*Cuminum cyminum* L.) ecotypes. Iranian J Gen Plant Breed 2(1): 35-41.

Cavagnaro P F, Chung S M, Manin S, Yildiz M, Ali A, Alessandro M S, Simon P W (2011) Microsatellite isolation and marker development in carrot-genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across Apiaceae. BMC Genom 12(1): 386.

El-Ghorab A H, Nauman M, Anjum F M, Hussain S, Nadeem M (2010) A comparative study on chemical composition and antioxidant activity of ginger (*Zingiber officinale*) and cumin (*Cuminum cyminum*). J Agric Food Chem 58(14): 8231-8237.

Gebhardt Y, Witte S, Forkmann G, Lukačín R, Matern U, Martens S (2005) Molecular evolution of flavonoid dioxygenases in the family Apiaceae. Phytochem 66(11): 1273-1284.

Hundekar A, Sarvamangala S C, Manjunath H, Naika M B, Bapurayagouda P (2024) Characterization of Diverse Carrot (*Daucus carota* L.) Genotypes Using Long Noncoding RNA Based Microsatellite Markers. J Adv Biol Biotechn 27(8): 1316-1324.

Koren S, Walenz B P, Berlin K, Miller J R, Bergman N H, Phillippy A M, Canu (2017) Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722-736.

Kumar S, Saxena S N, Mistry J G, Fougat R S, Solanki R K, Sharma R (2015) Understanding *Cuminum cyminum*: An important seed spice crop of arid and semi-arid regions. Int J Seed Spice 5(2):1-19

Kumar S, Saxena S N, Sakure A A, Patel P, Ravi Y, Khan A, Khandelwal V (2025) Biotechnological advances in cumin (*Cuminum cyminum*). In C. Kole (Ed.), Genetics, Genomics and Breeding of Seed Spices (pp. 69-96). Singapore: Springer Nature

Li M Y, Feng K, Hou X L, Jiang Q, Xu Z S, Wang G L, Xiong, A S (2020) The genome sequence of celery (*Apium graveolens* L.), an important leaf vegetable crop rich in apigenin in the Apiaceae family. Hort Res 7: 1-13.

Li MY, Feng K, Hou XL, Jiang Q, Xu ZS, Wang GL, Liu JX, Wang F, Xiong AS (2020) The genome sequence of celery (*Apium graveolens* L.), an important leaf vegetable crop rich in apigenin in the Apiaceae family. Horticulture research. 2020 Dec 1;7.

Liu H, Zhang JQ, Chen C, Wang YH, Xu ZS, Zhao QZ, Zhang J, Xue JY, Xiong AS (2025) The parsley genome assembly and DNA methylome shed light on apigenin biosynthesis in the Apiaceae. Plant Physiology. 197(3):kiaf077.

Milan K M, Dholakia H, Tikku P K, Vishveshwaraiah P (2008) Enhancement of digestive enzymatic activity by cumin (*Cuminum cyminum* L.) and role of spent cumin as a bionutrient. Food Chem 110(3): 678-683.

- Mnif S, Aifa S (2015) Cumin (*Cuminum cyminum* L.) from traditional uses to potential biomedical applications. *Chem Biodivers* 12(5): 733–742.
- Moghaddam M, Miran S N K, Pirbalouti A G, Mehdizadeh L, Ghaderi Y (2015) Variation in essential oil composition and antioxidant activity of cumin (*Cuminum cyminum* L) fruits during stages of maturity. *Ind Crops Prod* 70: 163–169.
- Morad M, Mokhtar M, Alsamman M, Alsamman A, El Allali A (2023) MegaSSR: A webserver for large scale SSR identification, classification, and marker development. *Front Plant Sci* 14.
- Mortazavian S M M, Arshadi-Bidgoli M, Sadeghi D, Bakhtiarizadeh M R (2024) Identification and validation of EST-SSR in transcriptome sequences by RNA-Seq in cumin (*Cuminum cyminum* L). *Plant Gene* 40.
- Palumbo F, Galla G, Vitulo N, Barcaccia G (2018) First draft genome sequencing of fennel (*Foeniculum vulgare* Mill.): Identification of simple sequence repeats and their application in marker-assisted breeding. *Mol Breed* 38(10):122.
- Palumbo F, Vannozzi A, Barcaccia G (2021) Impact of genomic and transcriptomic resources on Apiaceae crop breeding strategies. *Int J Mol Sci* 22(18): 9713.
- Pei Q, Li N, Yang Q, Wu T, Feng S, Feng X, Song X (2021) Genome-wide identification and comparative analysis of ARF family genes in three Apiaceae species. *Front Genet* 11: 590535.
- Rahman A Y A, Usharraj A O, Misra B B, Thottathil G P, Jayasekaran K, Feng Y, Alam M (2013) Draft genome sequence of the rubber tree (*Hevea brasiliensis*). *BMC Genomics* 14: 75.
- Rebey I B, Jabri-Karoui I, Hamrouni-Sellami I, Bourgou S, Limam F, Marzouk B (2012) Effect of drought on the biochemical composition and antioxidant activities of cumin (*Cuminum cyminum* L) seeds. *Ind Crops Prod* 36(1): 238–245.
- Reyes-Calderón A, Gutiérrez-García C, Urióstegui-Pena A G, Srivastava A, Aguilar-Marcelino L, Iqbal H M, Sharma A (2023) Identification of cumin (*Cuminum cyminum*) microRNAs through deep sequencing and their impact on plant secondary metabolism. *Plants* 12(9): 1756.
- Rostami-Ahmadvandi H, Cheghamirza K, Kahrizi D, Bahraminejad S (2013) Comparison of morpho-agronomic traits versus RAPD and ISSR markers to evaluate genetic diversity among cumin (*Cuminum cyminum* L) accessions. *Aust J Crop Sci* 7(3): 361–367.
- Sharma K, Kumar S (2025) Cumin (*Cuminum cyminum* L): Present breeding status and future prospects. *Ind Crops Prod* 230: 121117.
- Singh R P, Gangadharappa H V, Mruthunjaya K (2017) *Cuminum cyminum* – A popular spice: An updated review. *Pharmacogn J* 9(3): 319–326.
- Song X, Nie F, Chen W, Ma X, Gong K, Yang Q, Wang J, Li N, Sun P, Pei Q, Yu T (2020) Coriander Genomics Database: a genomic, transcriptomic, and metabolic database for coriander. *Hortic Res* 7.
- Sowbhagya H B (2013) Chemistry, technology, and nutraceutical functions of cumin (*Cuminum cyminum* L): An overview. *Crit Rev Food Sci Nutr* 53(1): 1–10.
- Tomar R S, Solanki R K, Padhiyar S M, Kheni J V, Kakani R K (2022) Flow cytometric genome size estimation of major seed spice grown in India. *Pharma Innov J* 11(6): 507–510.
- Uncu A T, Patat A S, Uncu A O (2024) Whole-genome sequencing and identification of antimicrobial peptide coding genes in parsley (*Petroselinum crispum*), an important culinary and medicinal Apiaceae species. *Funct Integr Genomics* 24(5): 142.
- Wang X J, Luo Q, Li T, Meng P H, Pu Y T, Liu J X, Xiong A S (2022) Origin, evolution, breeding, and omics of Apiaceae: A family of vegetables and medicinal plants. *Hortic Res* 9.
- Wanner J, Bail S, Jirovets L, Buchbauer G, Schmidt E, Gochev V, Stoyanova A (2010) Chemical composition and antimicrobial activity of cumin oil (*Cuminum cyminum*). *Nat Prod Commun* 5(9): 1357–1364.
- Zhu Q, Guang Y, Xiao X, Shan N, Wan C, Yang Y (2019) The complete chloroplast genome sequence of celery (*Apium graveolens* L). *Mitochondrial DNA B* 4(1): 463–464.

Supplementary Figure legend

Fig. S1 Quality check on 1% agarose gel (3 µl loaded)

Fig. S2 Bioanalyzer electropherogram

Fig. S3 Assessment of cumin genome assembly completeness via BUSCO