scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level Haploid Assembly of *Cannabis sativa* L. cv. Pink Pepper

Byeong-Ryeol Ryu[1,2], Gyeong-Ju Gim[3], Ye-Rim Shin[1], Min-Ji Kang[1], Min-Jun Kim[1], Tae-Hyung Kwon[4], Young-Seok Lim[1,5], Sang-Hyuck Park[2] ✉ & Jung-Dae Lim[1,6] ✉

As molecular research on hemp (*Cannabis sativa* L.) continues to advance, there is a growing need for the accumulation of more diverse genome data and more accurate genome assemblies. In this study, we report the three-way assembly data of a cannabidiol (CBD)-rich cannabis variety, 'Pink Pepper' cultivar using sequencing technology: PacBio Single Molecule Real-Time (SMRT) technology, Illumina sequencing technology, and Oxford Nanopore Technology (ONT). This assembly anchors scaffolds to the ten chromosomes of hemp, and to avoid confusion with previous cannabis genetic research, the chromosomes have been labeled based on an earlier reference genome. The total assembled genome length is 770 Mbp, with a GC content of 34.09% and a repeat region accounting for 77.13% of the genome. This assembly, which incorporates the unique strengths of the three sequencing technologies, demonstrated the highest complete BUSCO scores (97.8%-99.6%) among the reported cannabis genomes, as evaluated using three different BUSCO databases. With annotations for 30,459 protein-coding genes, this dataset can serve as a valuable resource for advancing genetic research on hemp.

Background & Summary

Cannabis sativa L. is a primarily annual, dioecious or monecious herb that has been traditionally cultivated for fiber production, with a history dating back to around 8,000 BC[1,2]. Although many fiber-use cannabis plants are still being cultivated today, there has been a recent increase in interest in the unique chemical components of cannabis called cannabinoids, and the research and medicinal application have been growing[3–8].

Generally, Δ^9-tetrahydrocannabinol (Δ^9-THC) and cannabidiol (CBD) are the most well-known among over 100 cannabinoids, as they are the most abundant[9,10]. These two components mainly exist in the form of Δ^9-tetrahydrocannabinolic acid (Δ^9-THCA) and cannabidiolic acid (CBDA) within the plant, and they are converted to CBD and Δ^9-THC through the process of decarboxylation, in which the carboxyl group is removed upon heating and light exposure, through chemical reactions[11].

These two main cannabinoids are used for different purposes. Δ^9-THC, a representative drug permitted in 25 states of the USA and a few countries such as Canada, the UK, Croatia, and the Czech Republic, is often used for recreational purposes due to its psychoactive properties[12,13]. However, ongoing medical research is being conducted to explore its potential uses. On the other hand, CBD is reported to be effective for medical purposes such as anti-anxiety[14], antioxidant and anti-inflammatory[15], anticonvulsant[16], and synergistic effects with anti-cancer drugs[17]. In the cannabis cultivation industry for medical purposes, there have been active breeding efforts to reduce Δ^9-THC levels and increase CBD levels for several years[18]. Researchers continue to seek a better understanding of the biological and physiological characteristics of medicinal (Type III) cannabis to further advance its breeding[19].

Since the completion of the initial draft genome of the marijuana strain 'Purple Kush' in 2011[20], efforts have been made to establish a comprehensive database and obtain high-quality data for genomes of various strains

[1]Department of Bio-Health Convergence, Kangwon National University, Chuncheon, 24341, Republic of Korea. [2]Institute of Cannabis Research, Colorado State University-Pueblo, 2200 Bonforte Blvd, Pueblo, CO, 81001-4901, USA. [3]National Agrobiodiversity Center, National Academy of Agricultural Science, Rural Development Administration, Jeonju, 54874, Republic of Korea. [4]Institute of Biological Resources, Chuncheon Bioindustry Foundation, Chuncheon, 24232, Republic of Korea. [5]Department of Bio-Health Technology, Kangwon National University, Chuncheon, 24341, Republic of Korea. [6]Department of Bio-Functional Material, Kangwon National University, Samcheok, 25949, Republic of Korea. ✉e-mail: sanghyuck.park@csupueblo.edu; ijdae@kangwon.ac.kr

Assembly name	Modifier	Cannabinoid type	Assembly level	Scaffold count	Sequencing technology
Chemdog91_175268	Chemdog91	Type 1	Scaffold	175,088	Illumina
ASM151000v1	LA Confidential	Type 1	Contig	—	454
ASM186575v1	Cannatonic	Type 2	Contig	—	PacBio
ASM209043v1	Pineapple Banana Bubba Kush	Type 1	Contig	—	PacBio
ASM341772v2	Finola	Type 3	Chromosome	5,303	PacBio
ASM23057v5	Purple Kush	Type 1	Chromosome	12,836	PacBio
Oct15_3.7Mb_N50_Jamaican_Lion_Assembly	Jamaican Lion DASH	Type 3	Contig	—	PacBio
cs10	CBDRx	Type 3	Chromosome	220	Illumina, ONT
JL_Mother	Jamaican Lion ^4	—	Contig	—	PacBio
JL_Father	Jamaican Lion ^4	—	Contig	—	PacBio
JL5	Jamaican Lion ^4	Type 3	Contig	—	PacBio
ASM1303036v1	JL	—	Chromosome	483	PacBio
Cannbio-2	Cannbio-2	Type 3	Chromosome	147	PacBio
Csat_AbacusV2	Abacus	Type 3	Chromosome	160	PacBio
ASM2916894v1	Pink Pepper	Type 3	Chromosome	17	Illumina, PacBio, ONT

Table 1. The list of assemblies of *Cannabis sativa* L. currently available in NCBI GenBank. Modifier refers to the cultivar name or isolate name. The cannabinoid types were classified according to the pharmacological classification proposed by Lewis *et al.* (2018)[73]: Type 1, Δ^9-tetrahydrocannabinol-predominant; Type 2, Cannabis that contains both Δ^9-tetrahydrocannabinol and cannabidiol; Type 3, cannabidiol-predominant.

(Table 1)[21–24]. The current cannabis assemblies lack consistency in terms of total assembly size, and the naming of chromosome numbers and orientations is not standardized[25]. Previously published chromosome-level cannabis assemblies contain at least 147 scaffolds, indicating a need for better continuity (Table 1). Additionally, the average number of N's per 100 kbp is 2,772, reflecting a very high proportion of unknown sequences. Kovalchuk *et al.* (2020) pointed out that the Cannabis genome assembly is incomplete, contains gaps, is poorly aligned with low resolution, and the quality of the consensus sequence obscures the accuracy of annotations[26]. Furthermore, such assemblies create confusion for data users in distinguishing between real genome differences and assembly errors.

With the increasing use of Cannabis for both agricultural and medicinal purposes, it has become essential to establish a comprehensive and high-resolution cannabis genomic database. This resource is crucial for comparative genomics, evolutionary studies, breeding improvements, and understanding the genetic regulation of key agronomic traits, such as cannabinoid production. Recently, there has been growing number of studies examining small-scale variations such as single nucleotide polymorphisms (SNPs) in specific genes, as well as mid-larger scale variations like long terminal repeats (LTRs), using genomic data. The accurate identification of variations relies on the quality of sequencing and genome assembly. Therefore, ensuring high-quality genomic data is critical for the reliable interpretation of genetic variation.

To achieve a high-precision cannabis genome assembly, we utilized three sequencing technologies: Pacbio Single Molecule, Real-Time (SMRT) sequencing, Oxford Nanopore Technologies (ONT), and Illumina high-throughput short-read sequencing to achieve high precision overlap hybrid assembly. We generated two types of 3rd generation primary reads of 'Pink Pepper' based on PacBio SMRT, well-established for its high accuracy[27], and ONT, which is advantageous for its longer read lengths[28]. Then, the accuracy of the genome assembly was then increased by aligning it with the Illumina sequencing data of the same variety, resulting in a chromosome-level genome (Fig. 1a). The assembled genome was classified into 10 chromosomes, with a size of 770 Mb. The GC content was 34.09%, N per 100 kbp was 0.69, complete Benchmarking Universal Single-Copy Orthologs (BUSCO) was 99.6% (viridiplantae_odb10), 97.8% (eudicots_odb10) and 98.6% (embryophyte_odb10). Overall repeats accounted for 77.13% of the entire genome. Based on transcriptome data from leaves, flowers, roots, and stems, and protein sets related to cannabis, 30,459 genes encoding proteins were predicted, accounting for 92.92% of the total 32,779 genes.

In this data, we present the complete genome sequence of the Pink Pepper cultivar, selectively bred for high CBD production. Based on this assembled genome, we can provide more precise fundamental information for not only cannabis breeding but also studies on the biological characteristics, and plant responses through the analysis of Differentially Expressed Genes. Consequently, understanding the cannabinoid and terpene biosynthesis mechanisms in cannabis could ultimately contribute to the development and application of medical cannabis.

Methods

Cannabis variety and cultivation. The variety of *C. sativa* used in this study was 'Pink Pepper,' which is a type 3 cannabis strain with a high content of CBD (open field, 11.404 ± 1.117%·inflorescence dry weight, 3.267 ± 0.335%·leaf dry weight). The cannabis was a cut-clone, and rooting was induced in tap water before being cultivated. To secure rooting space, a large pot (15 L) was filled with bed soil (bio bed soil, Heungnong Jongmyo Co., Pyeongtaek, Korea) for cultivation. The plants were grown in a green house for 90 days (24 ± 4°C), the photoperiod was adjusted to 18 hours/day using shading curtains. Although the strain was auto-flowering, the light

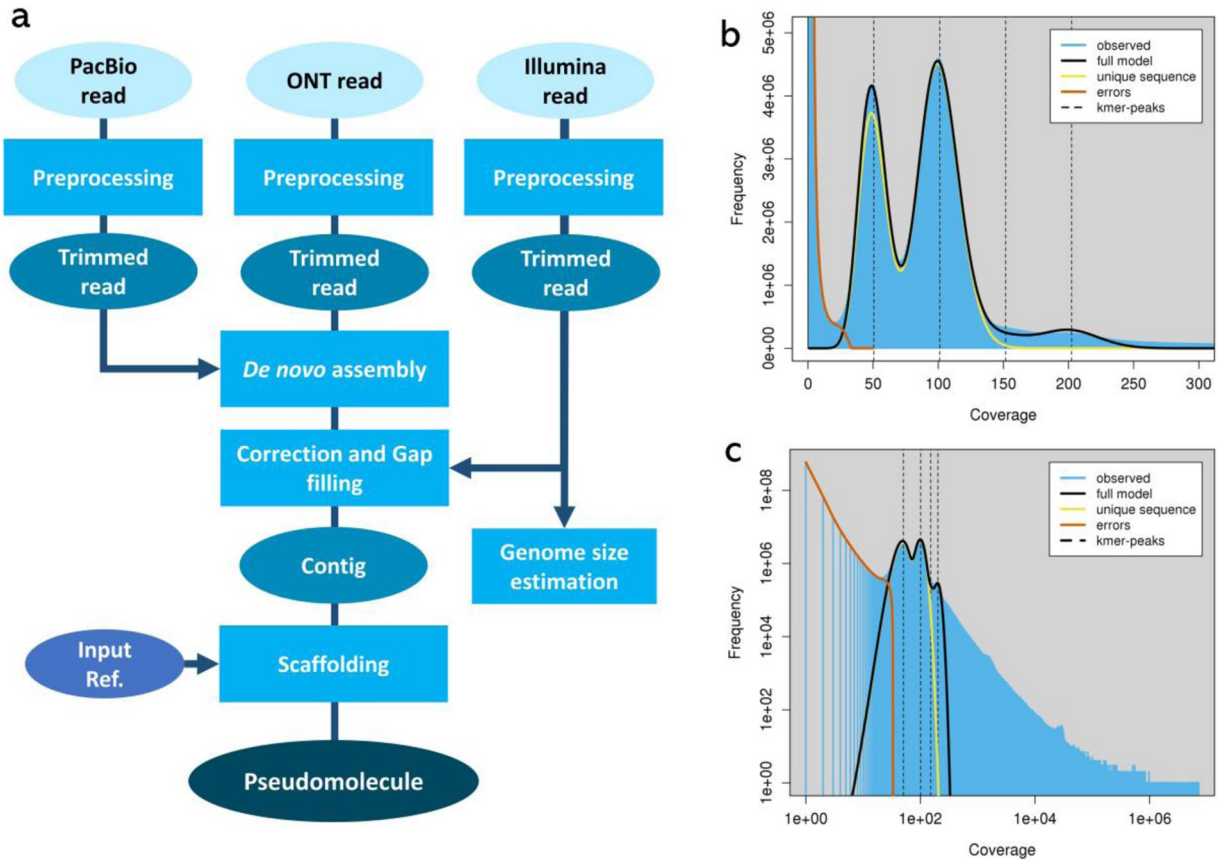

Fig. 1 Schematic diagram of the genome assembly of *Cannabis sativa* L. conducted in this study (**a**). The reference genome used for scaffolding was GCA_900626175.2 of NCBI GenBank database. The distribution of k-mer analysis using GenomeScope 2.0 (kmer: 19). Max k-mer coverage at 300 × (**b**), and 1,000,000 × (**c**). The blue portion in the figure represents the analyzed k-mer frequency, while the orange and yellow lines represent errors and unique sequences, respectively (**b, c**).

was adjusted to 12 hours/day to activate flower differentiation and induce flower development. Throughout the entire growth cycle, the plants were irrigated with 400 mL of tap water once daily.

Nucleic acid extraction. High molecular weight genomic DNA was extracted from fresh leaf tissue during the vegetative growth phase, using the cetrimonium bromide (CTAB)-based extraction method. Total RNA was extracted from three types of plant tissues: flower, leaf, and root, using the Quick-RNA MiniPrep kit (Zymo Research, Irvine, CA, USA) during the flowering stage. To preserve the integrity of the nucleic acids, the sampled plant tissues were immediately submerged in liquid nitrogen and subsequently stored at −80 °C in a deep freezer (DAIHAN Scientific Co., Ltd., Wonju, Korea) until further analysis.

Quality control and library preparation. DNA concentration, quality, quantity, and integrity were assessed using Victor 3 fluorometry (PerkinElmer Inc., Waltham, MA, USA) and gel electrophoresis. A DNA integrity number (DIN) of seven or higher was confirmed. Quality control and normalization of the Illumina library involved quantification according to the Illumina qPCR quantification protocol guide. For nanopore sequencing, library preparation utilized a ligation sequencing kit with quantification performed using Qubit 3.0 (Thermo Fisher Scientific Inc., Waltham, MA, USA). The Pacbio library was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences of California Inc., Menlo Park, CA, USA).

RNA was quantified with the Agilent Technologies 2100 Bioanalyzer (Santa Clara, CA, USA), achieving a RNA integrity number (RIN) of seven or higher, indicative of high quality. RNA integrity was further verified by gel electrophoresis. mRNA purification was conducted using the TruSeq stranded mRNA kit (Illumina, San Diego, CA, USA), followed by cDNA reverse transcription for library preparation. Illumina paired-end sequencing was subsequently performed.

Sequencing and pre-processing. Using the Illumina NovaSeq. 6000 (San Diego, CA, USA), we generated paired-end read data comprising 815,329,552 reads and totaling 123 giga base pair (Gbp). To remove contaminants and adaptors, fastp v0.21.0 (https://github.com/OpenGene/fastp) and BBDuk v38.87 (k = 31, mcf = 0.5;

Parameters	Draft assembly (NextDenovo)	After polishing (NextPolish)	Contig (PurgeHaplotigs)	Psudomolecule (Ragtag)
Contig (Scaffold) number	130	130	70	17
Contig (Scaffold) length (bp)	809,336,498	810,701,602	770,264,337	770,269,637
Min length (bp)	42,660	42,718	144,850	144,850
Max length (bp)	75,022,527	72,195,770	75,195,770	92,210,330
Average length (bp)	6,225,665	6,236,166	11,003,776	45,309,979
N50 (bp)	23,011,228	23,046,559	23,500,774	76,975,026
N90 (bp)	4,190,848	4,201,318	5,583,923	61,573,785
GC Ratio (%)	34.90	32.22	34.09	34.09

Table 2. Genome statistics during assembly and scaffolding process.

https://sourceforge.net/projects/bbmap/) were used. The contaminant databases included viral, rRNA, human, and bacterial sequences. After quality and adaptor trimming, 94 Gbp of read data was obtained, removing 0.06% viral, 2.61% rRNA, 0.03% human, and 0.03% bacterial reads. For long-read sequencing, ONT sequencing was performed using the ONT GridION (Oxford, UK), repeated five times for high reliability. The generated long-read data had adaptors removed using Porechop (v0.2.3, https://github.com/rrwick/Porechop). A pass with a quality score of seven or higher was confirmed. The number of reads was 11,484,123, with a total base pair count of 89 Gbp and an N50 of 26,677. Using the PacBio Sequel II system (Menlo Park, CA, USA), single molecule, real-time (SMRT) sequencing was performed to generate polymerase reads. Using the SMRT Link v11.1 software with the PacBio Sequel II system, adaptors were removed, and subreads were aligned, resulting in 2,039,056 reads with a total base pair count of 21 Gbp.

***De novo* assembly and scaffolding.** To perform statistical analysis on the basic genomic information, Jellyfish v2.2.10[29] and GenomeScope 2.0[30] were utilized to predict the genome size of the Illumina sequence reads. The analysis was conducted with k-mer 17, 19, and 21, with the 19-mer used for the final genome size estimation. As a result, homozygosity ranged from 98.64% to 98.69%, while heterozygosity ranged from 1.31% to 1.36%. The estimated haploid genome length ranged from 776 Mbp (mega base pair) to 779 Mbp, while the repeat length ranged from 554 Mbp to 556 Mbp. The unique length was estimated from 222 Mbp to 223 Mbp (Fig. 1b and c).

In this data, NextDenovo v2.3.1 (https://github.com/Nextomics/NextDenovo) was used to assemble the ONT reads, and then the PacBio reads were mapped[31], resulting in the generation of 130 contigs with a total length of 809 Mbp. Then, the contigs were polished using Illumina short read to generate contigs of 810 Mbp. finally filtered using PurgeHaplotigs, about 4.99% of redundant sequences were removed to derive 70 contigs (770 Mbp). Using the RagTag software (https://github.com/malonge/RagTag) with default parameters, the generated contig was mapped to the previous version of *C. sativa* reference genome (GCA_900626175.2)[32], and the pseudomolecule of a total of 770 Mbp was created. The longest chromosome was chromosome 2 with a length of 92 Mbp, while the shortest chromosome was chromosome 8 with a length of 51 Mbp, and N50 value was 77 Mbp (Table 2).

Repeat annotation. In general, long-read sequencing techniques, such as Pacbio sequencing and ONT approaches, are advantageous for the accurate detection of repeats containing tandem repeats (TRs). These methods can relatively accurately assemble long repeats spanning genes and detect the length, nucleotide composition, and nucleotide variations of TRs[33]. *De novo* repeat families were identified using RepeatModeler software (https://github.com/Dfam-consortium/RepeatModeler), and the distribution of repeats within assembled genomic sequences was analyzed using RepeatMasker v4.1.2 software[34] (https://github.com/rmhubley/RepeatMasker). To enhance readability, the distribution of repeats was categorized into DNA elements, long interspersed nuclear elements (LINEs), LTRs elements, rolling circles (RCs) elements, and short interspersed nuclear elements (SINEs). The overall repeats represented 77.13% of the cannabis assembly, which was consistent with previous research reporting high repeat levels in cannabis cultivars 'Purple kush' and 'Finola' (73.9% and 73.3%, respectively)[23].

The results indicated a slightly higher repeat content in cannabis compared to its taxonomically close relative *Humulus lupulus* (71.46%)[35,36]. Additionally, it was on the higher side compared to other plants such as *Xanthoceras sorbifolium* (56.39%)[37], *Oryza sativa* (51.63% - 54.34%)[38], *Panax ginseng* (56.9%)[39], and *Nicotiana tabacum* (67.05%)[40]. The most abundant repeat regions were LTR-Gypsy retrotransposons and LTR-Copia retrotransposons, comprising 24.45% and 25.81% of the genome, respectively (Table 3).

Gene annotation. Total RNA from plant tissues, including stems, leaves, roots, and flowers, was reverse-transcribed, and paired-end sequencing was performed using Illumina NovaSeq 6000. Subsequently, *de novo* assembly was conducted to obtain transcriptome data[41]. Simultaneously, an evidence dataset was constructed using protein sequences from 10 registered species on NCBI (Table 4), and the first gene prediction was performed using MAKER (v3.01.03)[42]. Among the genes, only those with an annotation edit distance of 0.25 or lower were selected. GeneMark (v4.38)[43], SNAP (v20060728)[44], and AUGUSTUS (v3.3.2)[45] were performed for gene prediction *ab initio* training.

By integrating the results of the first gene prediction and the *ab initio* training dataset, a second gene prediction for gene model prediction was conducted. EvidenceModeler v1.1.1[46] was used to apply different weights to each dataset. The weights were set to 7 for GeneMark data and 10 for the others.

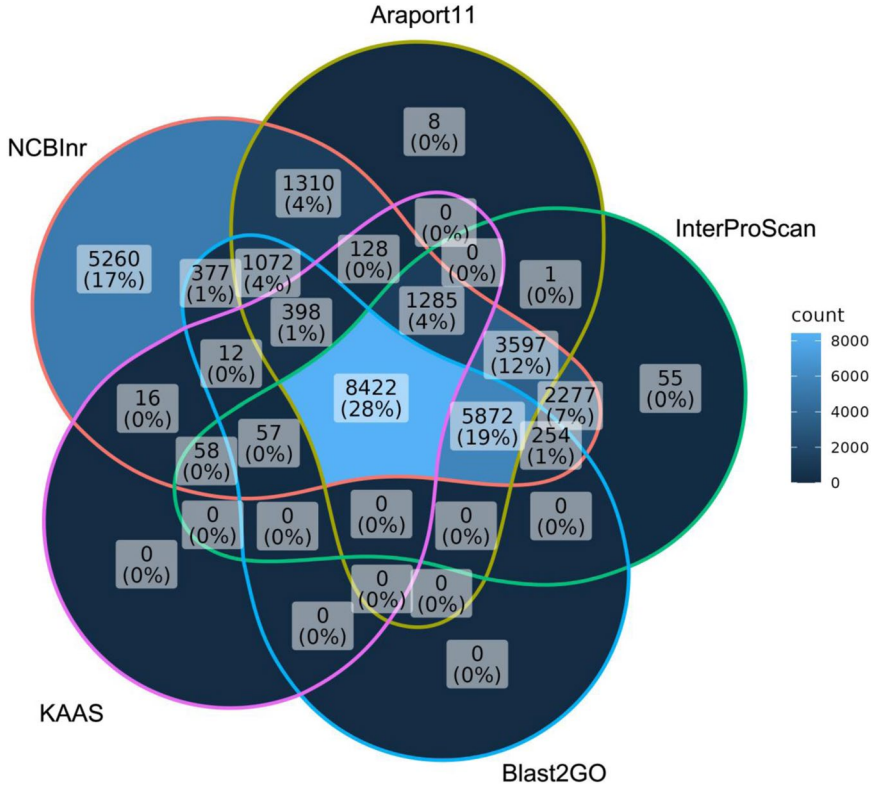

Fig. 2 Number and percentage of annotations by different annotation methods. The data represented in the Venn diagram describes protein IDs that are shared among functional annotation tools: Araport11, annotation database of *Arabidopsis thaliana*; NCBInr, NCBI protein sequence database; InterProScan, InterPro protein sequence database; KAAS, Kyoto encyclopedia of genes and genomes (KEGG) protein sequence database, Blast2GO: Tool for Gene Ontology (GO) analysis and functional annotation.

To predict the function of the identified genes, DIAMOND (v5.34-73.0; maximum target sequence = 20, e-value threshold = 1e-5)[47] was used to analyze the similarity with the non-redundant protein database[48] from NCBI and Araport11[49] from *Arabidopsis thaliana*. Gene ontology (GO) analysis was conducted using BLAST2GO (v5.2.5)[50], protein domains were identified using InterproScan (v5.34-73.0)[51], and KEGG (Kyoto encyclopedia of genes and genomes) pathway analysis was performed using the KAAS web-tool[52]. Annotations were defined as follows: 30,395 (92.73%) for NCBI nr, 22,093 (67.40%) for Araport11, 21,878 (66.74%) for InterProScan, 16,464 (50.23%) for BLAST2GO, and 10,376 (31.65%) for KAAS web-tool. The data from each source were combined and complemented, resulting in 30,459 genes, which accounted for 92.92% of the total cannabis transcriptome (Fig. 2 and Table 5).

Data Records

In the study, the raw data set generated is available in the NCBI SRA database[53]. Specifically, the PacBio sequencing data for the genome is deposited under accession number SRX17887361[54]. The ONT sequencing data is available under accession number SRX17887360[55], and the Illumina data under accession number SRX17887355[56]. The raw mRNA data generated for genome annotation have also been registered in the NCBI SRA database, associated with the following accession numbers: SRX17887359 (stem)[57], SRX17887358 (root)[58], SRX17887357 (leaf)[59], and SRX17887356 (flower)[60].

The assembled genome can be accessed in the GenBank database[61]. Comprehensive gene annotation information, including gene structure, functional predictions, transcriptome and protein data set can be accessed in the Figshare database[62].

Technical Validation

Plant sample validation. The DNA concentration of the leaf sample was 23.616 ng/μl, and 100 μl was extracted (total DNA amount: 3.262 μg). The DIN value was determined to be 7.5, and after passing the quality check, it was used for library preparation. The RNA concentration was 107.024 ng/μl, and 96 μl was extracted (total RNA amount: 10.274 μg). The RIN value was confirmed to be 8.4, and the rRNA ratio was determined to be 2.0.

The RNA concentration of the root sample was 41.937 ng/μl, and 50 μl was extracted (total RNA amount: 2.097 μg). The RIN value was confirmed to be 7.7, and the rRNA ratio was determined to be 4.2.

The RNA concentration of the stem sample was 59.53 ng/μl, and 50 μl was extracted (total RNA amount: 0.281 μg). The RIN value was confirmed to be 7.7, and the rRNA ratio was determined to be 8.3.

Class	Detail class	Count	Masked (bp)	Masked (%)
DNA		41,024	12,608,258	1.64%
	CMC-EnSpm	10,124	9,705,616	1.26%
	MULE-MuDR	12,941	12,653,931	1.64%
	PIF-Harbinger	1,091	459,274	0.06%
	hAT-Ac	1,234	1,022,950	0.13%
	hAT-Tip100	687	415,658	0.05%
LINEs		884	73,654	0.01%
	L1	35,262	33,077,369	4.29%
LTRs		48,295	22,961,302	2.98%
	Caulimovirus	216	334,427	0.04%
	Copia	106,259	188,343,073	24.45%
	Gypsy	101,070	198,792,419	25.81%
RCs		—	—	—
	Helitron	842	733,121	0.10%
SINEs		8,691	3,581,764	0.47%
	tRNA-RTE	368	70,404	0.01%
Unknown		261,254	96,358,396	12.51%
Total interspersed		**630,242**	**581,191,616**	**75.45%**
Low_complexity		29,782	1,598,892	0.21%
Satellite		673	152,804	0.02%
Simple_repeat		142,292	11,143,896	1.45%
Total		**802,989**	**594,087,208**	**77.13%**

Table 3. Result of repeat annotation statistics. Abbreviations: LINEs, Long interspersed nuclear elements; LTRs, Long terminal repeats; SINEs, Short interspersed nuclear elements; RCs, Rolling circles.

Scientific name	Assembly version	ID (GenBank or RefSeq)	Protein No.
Parasponia andersonii	PanWU01x14_asm01	GCA_002914805.1	37,227
Trema orientale	TorRG33x02_asm01	GCA_002914845.1	35,849
Arabidopsis thaliana	TAIR10.1	GCF_000001735.4	48,265
Fragaria vesca	FraVesHawaii_1.0	GCF_000184155.1	31,387
Malus domestica	ASM211411v1	GCF_002114115.1	52,036
Cannabis sativa	cs10	GCF_900626175.2	33,674
Morus notabilis	ASM41409v2	GCF_000414095.2	27,648
Ziziphus jujuba	ASM2079620v1	GCF_020796205.1	42,050
Cannabis sativa	JL_Mother	GCA_012923435.1	27,358
Cannabis sativa	JL_Father	GCA_013030025.1	31,591

Table 4. Used protein database of related species for evidence dataset.

Tools		Number of annotated genes	% of total genes
BLASTP (DIAMOND)	NCBI nr	30,395	92.73%
	Araport11	22,093	67.40%
Protein domains (InterProScan)		21,878	66.74%
Gene Ontology (BLAST2GO)		16,464	50.23%
KEGG pathway (KAAS webtools)		10,376	31.65%
Total		**30,459**	**92.92%**

Table 5. Functional annotation statistics of software for gene prediction.

The RNA concentration of the inflorescence (flower) sample was 952.552 ng/μl, and 50 μl was extracted (total RNA amount: 47.628 μg). The RIN value was confirmed to be 8.3, and the rRNA ratio was determined to be 2.7.

Comparison of read statistics and BUSCO with existing cannabis assemblies. Raw reads from the chromosome-level assembly publicly available on NCBI (Exclude reads from Abacus that are not presented in Sequencing Reads Archive (SRA)) were collected using SRA Toolkit (v3.1.1-ubuntu). Statistics were then generated using SeqKit[63] (v2.8.2, Supplementary Table 1).

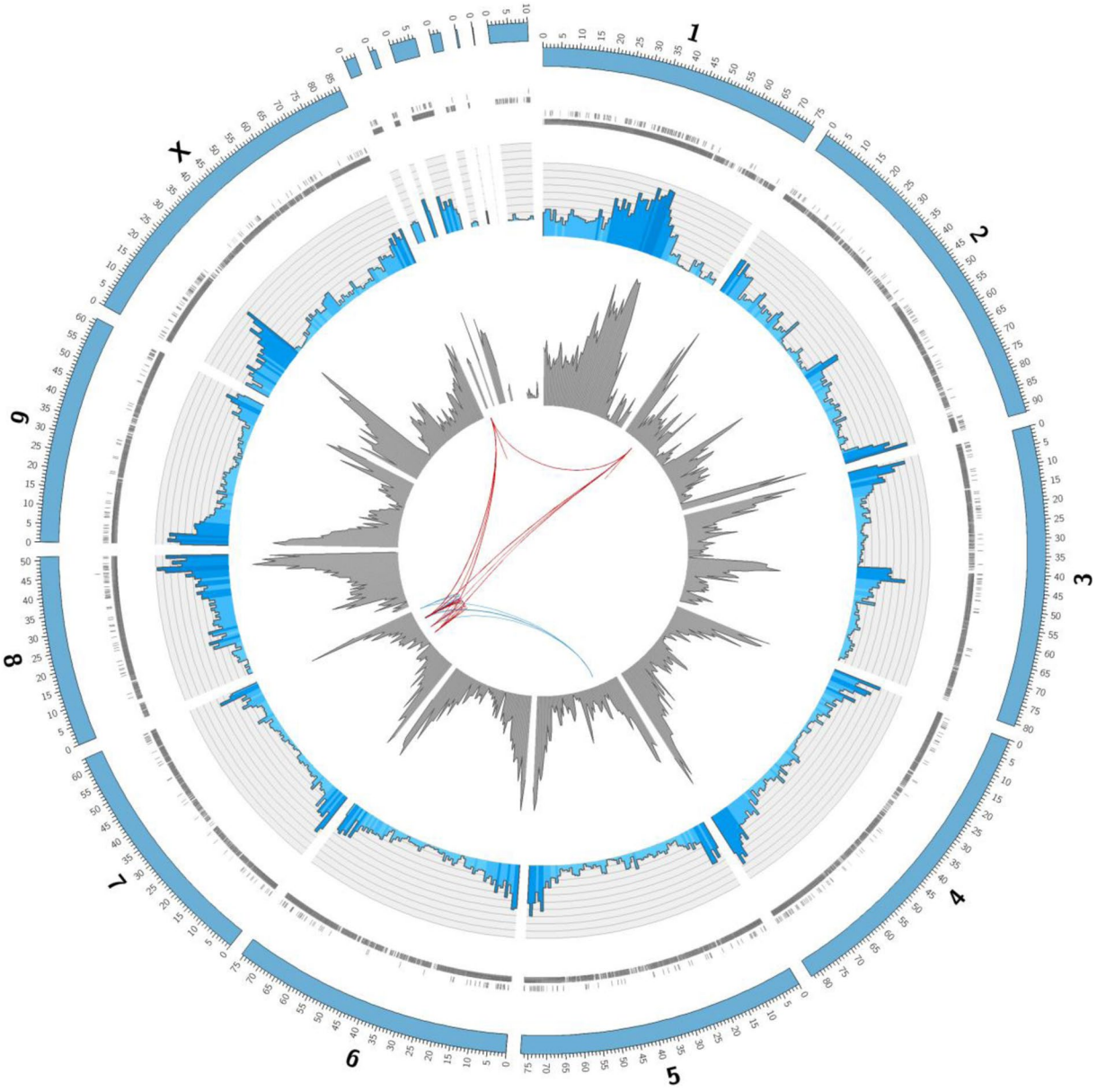

Fig. 3 Circle plot of the *Cannabis sativa* L. cv Pink Pepper genome assembly. From the outermost to innermost layers: Chromosome number, gene, CDS (coding sequence) frequency, mRNA frequency, and the relationship of the main cannabinoid gene. The protruding segments on the chromosomes represent unscaffolded regions. The scale indicating chromosome size is in units of Mbp (mega base pairs). CDS frequency and mRNA frequency are visualized after trimming at the 1 Mbp level. The red links connecting the center represent annotated genes involved in Δ^9-THCA synthesis, while the blue lines represent annotated genes involved in CBDA synthesis (based on the description).

Among them, the Illumina NovaSeq 6000 used for this assembly produced the highest number of reads, generating 815,329,552 paired-end reads totaling 123 Gbp. This result produced 2.7 times more reads than JL's HiSeq X Ten (SRA accession: SRX6757267), which previously held the highest number of reads, with comparable read lengths. The reads produced by ONT GridION had an N50 value of 26,677 and an N60 value of 73,606, which is 1.8 times higher than the N50 value of 14,716 for cs10's ERX3863365 reads, the only other reads produced using ONT. It is also 1.7 times higher than the N50 value of 16,037 for Cannbio-2's PacBio Sequel reads. This indicates a higher overlap proximity of reads, potentially leading to a more contiguous assembly. The reads produced by PacBio Sequel II were evaluated with a Q20 of 98.88% and a Q30 of 97.42%, the highest values next to those of Purple Kush (SRA accession: SRX4178554). These statistics demonstrate the impact of rapidly advancing sequencing technologies on producing high-quality reads. Furthermore, they emphasize the importance of hybrid assembly in offsetting disadvantages and leveraging advantages for downstream analysis.

To compare the completed chromosome-level assembly (Fig. 3) with other assemblies, the final assembly version of the chromosome-level Cannabis genomes registered in NCBI were collected[20,21,24,25] (GenBank

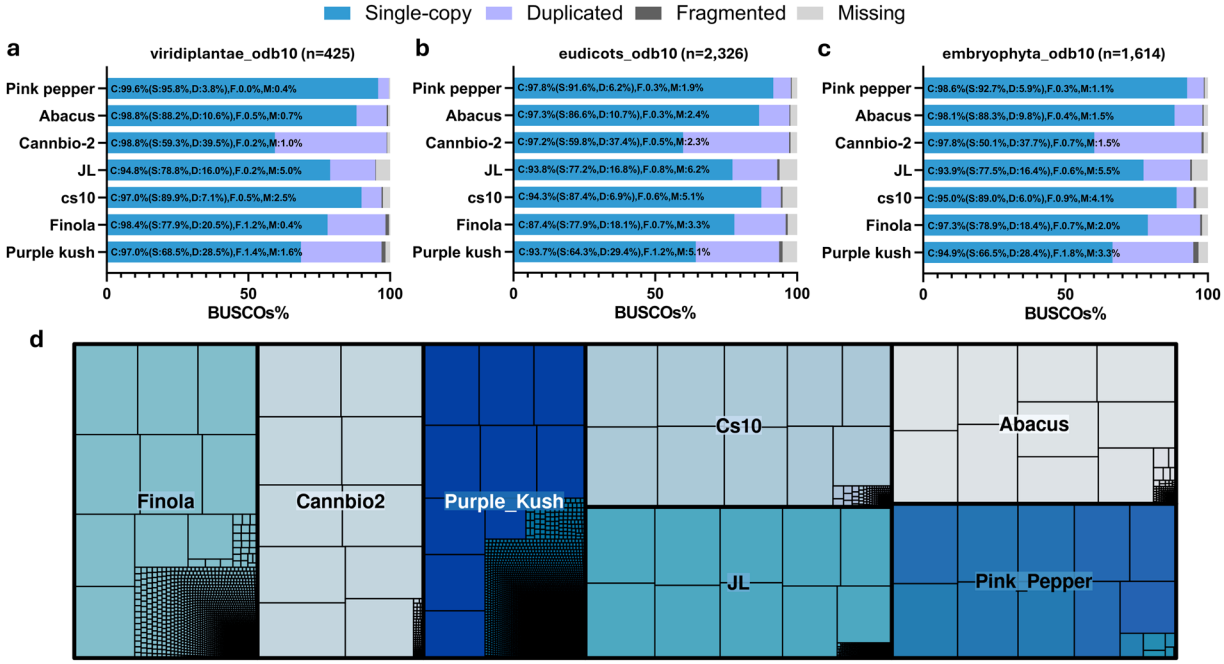

Fig. 4 Assembly completeness evaluation using Benchmark Universal Single-Copy Orthologs (BUSCO) and comparison of assembly continuity using a tree map chart. The evaluations were conducted using viridiplantae_odb10 (**a**), eudicots_odb10 (**b**), and embryophyta_odb10 (**c**). C: complete BUSCOs (S + D), S: Single-copy, D: Duplicated, F: Fragmented, M: Missing. The tree map chart visualizes the continuity of the assembly (**d**). The GenBank accession numbers for the varieties are as follows: Pink Pepper, the assembly data from this study (GCA_029168945.1); Abacus, GCA_025232715.1; Cannbio-2, GCA_016165845.1; JL, GCA_013030365.1; cs10, GCA_900626175.2; Finola, GCA_003417725.2; Purple kush, GCA_000230575.5.

accession: GCA_025232715.1, GCA_013030365.1, GCA_003417725.2, GCA_016165845.1, GCA_000230575.5, GCA_900626175.2), and the collected genome data were validated for integrity using vdb-validate. The BUSCO (v5.2.2) analysis of NCBI's chromosome-level assemblies were conducted using the viridiplantae_odb10, eudicots_odb10, and embryophyte_odb10 databases (Jan 08, 2024 released). Among the registered chromosome-level assemblies, this assembly showed the highest complete BUSCOs% based on all three databases (Fig. 4a-c). Specifically, for the viridiplantae_odb10 database, the complete percentage was 99.6% (single-copy: 95.8%, duplicated: 3.8%), for the eudicots_odb10 database it was 97.8% (single-copy: 91.6%, duplicated: 6.2%), and for the embryophyte_odb10 database, it was 98.6% (single-copy: 92.7%, duplicated: 5.9%). Simultaneously, our assembly data demonstrated a high level of single-copy BUSCOs% (Fig. 4a-c). The treemap, which represents the relative size of the assemblies, highlights the improved continuity of our assembly. Specifically, the number of scaffolds in chromosome-level assemblies is 5,303 for Finola, 147 for Cannbio-2, 12,836 for Purple Kush, 220 for cs10 (CBDRx), 160 for Abacus, and 483 for JL, while this assembly data contains only 17 scaffolds, confirming its superior continuity (Table 1 and Fig. 4d).

Synteny analysis with close genetic relatives of *C. sativa*. Synteny comparison was conducted using protein sequences (protein.fasta) and annotation files (annotation.gff) generated from the annotation through BLASTp (v2.12.0)[64] and MCScanX[65]. Previous studies using *C. sativa* genomes reported synteny comparison results with *Ziziphus jujuba*, which belongs to the same Rosaceae family[24]. In our synteny analysis using between the Pink Pepper genome assembly and the *Z. jujuba* reference genome (RefSeq: GCF_031755915.1), a total of 72,921 genes were identified, with 30,456 classified as collinear. This indicates that *C. sativa* and *Z. jujuba* share 41.77% synteny (Fig. 5a and b).

We further conducted a synteny analysis using the reference genome of *H. lupulus* (RefSeq: GCF_963169125.1), which belongs to the Cannabaceae family, a more specific clade within Rosaceae, and shares significant genetic similarity with *C. sativa*. Out of the 79,354 identified genes, 55,832 were analyzed as collinear genes, revealing a high synteny of 70.36% (Fig. 5a and c). These results further confirm the close genetic relationship between *C. sativa* and *H. lupulus*. The synteny analysis data can be available on Figshare for further analysis and use[66].

Structural comparison between cannabis genomes. To compare the genomic structure using Pink Pepper assembly data, we compared the assembly with the previous reference genome, cs10 (GCA_900626175.2). Whole genome alignment (WGA) was performed using D-GENIES (v1.5.0)[67] with Minimap2 (v2.26; -f = 0.02)[68] as the aligner. The dot plot, with Pink Pepper as the target (reference) and cs10 as the query, revealed significant structural variations, such as gaps, inversions, and repeats, across the chromosomes, despite being from the same species. The comparison showed 19.89% no match, 9.12% matching <25%, 57.40% matching <50%, 13.36%

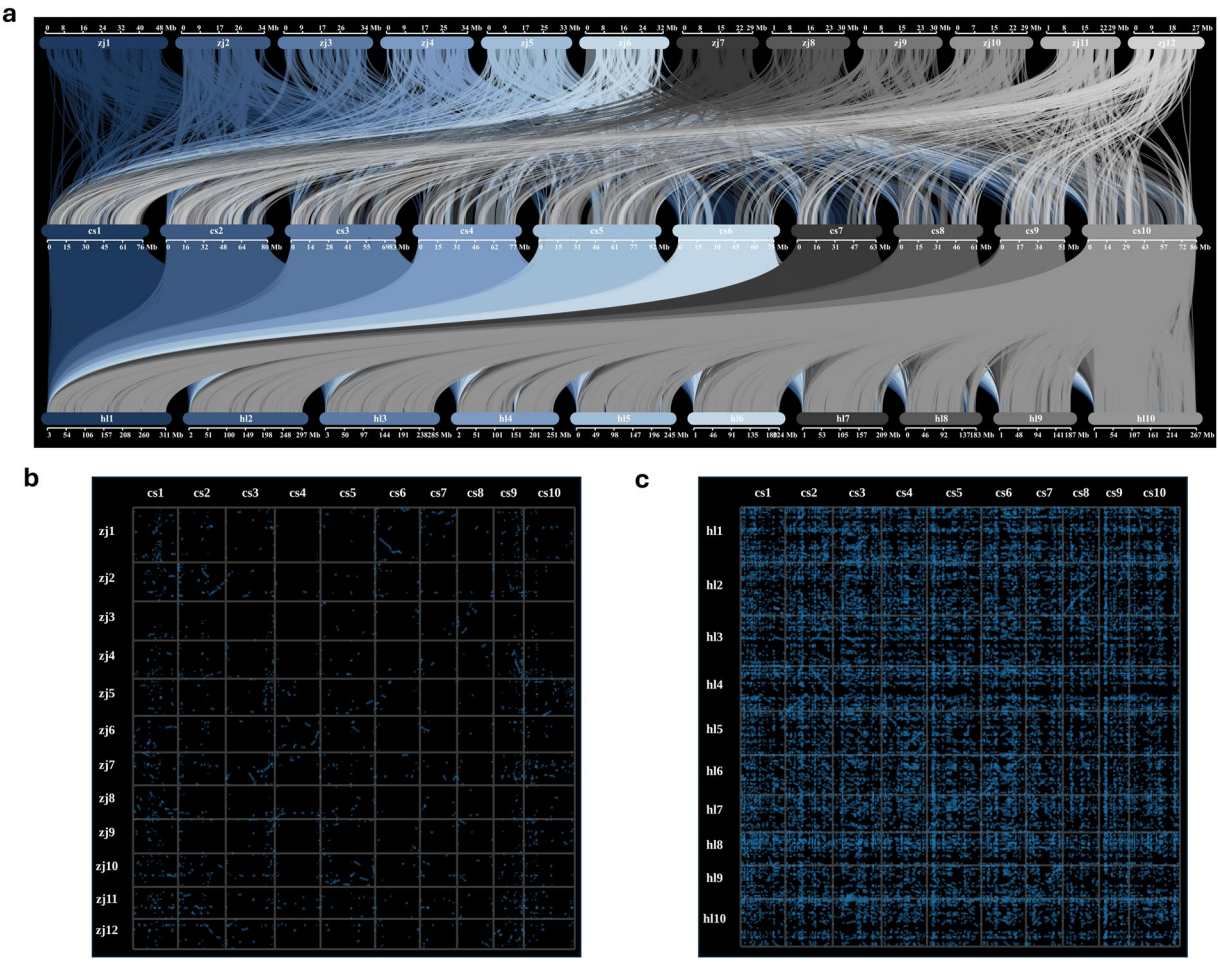

Fig. 5 Synteny analysis between the assembled Pink Pepper genome and the reference genomes of closely related species. The multicolored connecting curves between the chromosomes of the two species represent syntenic blocks, indicating conserved gene blocks between the genomes (**a**). The dot plots generated from the synteny data show the conserved synteny between *Cannabis sativa* L. and other genomes (**b, c**). cs1-10: Chromosome number of *C. sativa* formed by this assembly. zj1-12: chromosome number of the *Ziziphus jujuba* reference genome (RefSeq: GCF_031755915.1). hl1-10: chromosome number of the *Humulus lupulus* reference genome (RefSeq: GCF_963169125.1).

matching <75%, and only 0.23% maching >75% (Fig. 6a). Additionally, distinct structural differences and variations were identified on Chromosome 7, which contains a high density of CBDAS and THCAS (or pseudo- and fragmented) loci in both our assembly and cs10[21,62] (Figs. 3 and 6b).

Figure 6c presents the distribution of structural variations (SVs), categorized by chromosome and interval, using the current assembly as a reference against previously registered genomic datasets (Abacus, Cannbio2, cs10, Finola, JL, and Purple Kush). The analysis was conducted using NUCmer (v3.1; l = 40, g = 90, b = 100, c = 200) and dnadiff (v1.3) with Pink Pepper assembly data as reference. Overall, the number of breakpoints was highest in JL, with 577,769 instances, while Finola exhibited the highest number of relocations (12,979) and translocations (32,620). The most frequent inversions were observed in Purple Kush, totaling 3,158, and Cannbio-2 showed the greatest number of insertions, reaching 220,308. Although visually distinct large-scale structural variations were observed in Fig. 6a, cs10 showed the lowest values across all SV comparisons when compared to other cultivars. This finding suggests significant structural genomic variations among cannabis cultivars bred for diverse purposes and through different ways.

These intra-species WGA results have stem from fragmented assembly, as previously suggested in cannabis genomics[69]. However, they could also be due to phenotypic changes induced by chemical treatments such as silver nitrate and sodium thiosulfate, aimed at inducing male flower through inhibition of ethylene synthesis[70], and repeated inbreeding for strain stabilization[71]. Additionally, these results may be influenced by inbreeding within a limited population to achieve desired chemotypes or phenotypes. Through these differences, the accumulation of multiple high-quality cannabis genome assemblies can significantly enhance the resolution of molecular phylogenetic analyses, enabling the identification of subtle differences in evolutionary relationships and precise elucidation of phylogenetic dynamics. This SVs data can be available on Figshare for further analysis and use[72].

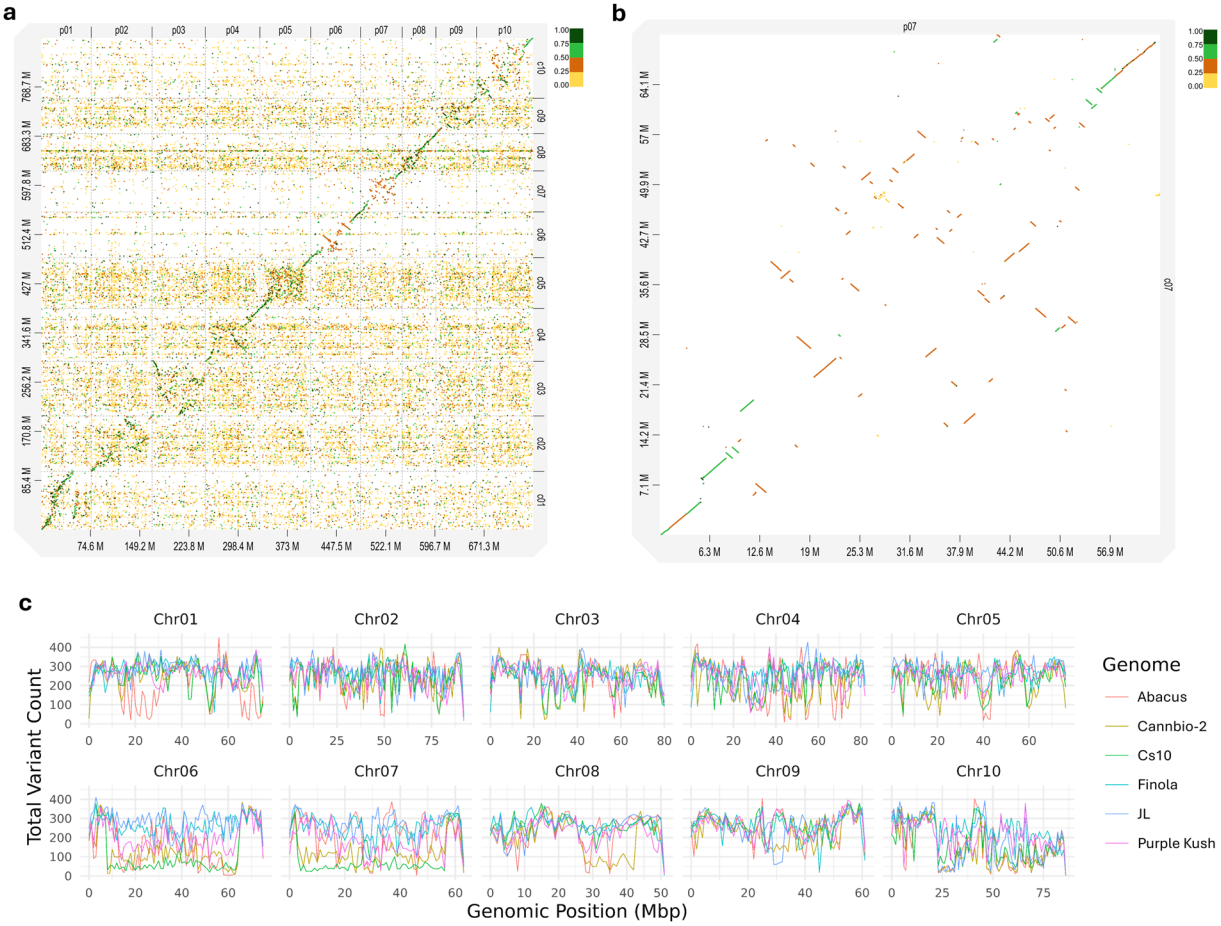

Fig. 6 Whole genome alignment (WGA) dot plot between the assembled Pink Pepper genome and cs10. The dots generated in the plot represent regions of similarity between the two genomes that have been aligned. p01-p10: Chromosome numbers of Pink Pepper, c01-c10: Chromosome numbers of cs10. The WGA excluded unscaffolded contigs (**a**), and the dot plot of chromosome 7, which contains loci related to cannabidiolic acid synthase (CBDAS) and Δ^9-tetrahydrocannabinolic acid synthase (THCAS), shows significant structural differences despite both strains being high-CBD varieties (**b**). Structural variations (SVs) at the chromosome level include breakpoints, duplications, sequence differences, gaps, and jumps, and the variant count was calculated per 10 Mbp (**c**). The GenBank accession numbers for the varieties are as follows: Pink Pepper, the assembly data from this study (GCA_029168945.1); Abacus, GCA_025232715.1; Cannbio-2, GCA_016165845.1; JL, GCA_013030365.1; cs10, GCA_900626175.2; Finola, GCA_003417725.2; Purple kush, GCA_000230575.5.

Chromosome Number	Functional MAKER file	GenBank ID	Size (bp)
1	Cannabis_NC_044371.1_RagTag	CM054919.1	75,645,423
2	Cannabis_NC_044375.1_RagTag	CM054920.1	92,210,330
3	Cannabis_NC_044372.1_RagTag	CM054921.1	80,731,606
4	Cannabis_NC_044373.1_RagTag	CM054922.1	82,824,063
5	Cannabis_NC_044374.1_RagTag	CM054923.1	76,975,026
6	Cannabis_NC_044377.1_RagTag	CM054924.1	75,575,810
7	Cannabis_NC_044378.1_RagTag	CM054925.1	63,199,550
8	Cannabis_NC_044379.1_RagTag	CM054926.1	51,107,535
9	Cannabis_NC_044376.1_RagTag	CM054927.1	61,573,785
X	Cannabis_NC_044370.1_RagTag	CM054928.1	86,057,484
unscaffolded	—	—	24,369,025

Table 6. Chromosome and annotation label of *Cannabis sativa* L. assembly of this data.

Usage Notes

Table 6 provides a summary of the chromosome labels for easier data accessibility.

Code availability

Parameters not mentioned in the main text, excluding threads, were set to default values. No custom code was generated for this work.

Received: 8 July 2024; Accepted: 16 December 2024;
Published: 28 December 2024

References

1. Shahzad, A. Hemp fiber and its composites–a review. *Journal of composite materials* **46**, 973–986 (2012).
2. Attia, Z., Pogoda, C. S., Vergara, D. & Kane, N. C. Variation in mtDNA haplotypes suggests a complex history of reproductive strategy in Cannabis sativa. *bioRxiv* 2020–12 (2020).
3. Leelawat, S. *et al*. Anticancer activity of Δ9-tetrahydrocannabinol and cannabinol *in vitro* and in human lung cancer xenograft. *Asian Pacific Journal of Tropical Biomedicine* **12**, 323–332 (2022).
4. Blaskovich, M. A. *et al*. The antimicrobial potential of cannabidiol. *Communications biology* **4**, 1–18 (2021).
5. Seltzer, E. S., Watters, A. K., MacKenzie, D., Granat, L. M. & Zhang, D. Cannabidiol (CBD) as a promising anti-cancer drug. *Cancers* **12**, 3203 (2020).
6. Gaston, T. E. & Friedman, D. Pharmacology of cannabinoids in the treatment of epilepsy. *Epilepsy & Behavior* **70**, 313–318 (2017).
7. Thapa, D. *et al*. The cannabinoids Δ8THC, CBD, and HU-308 act via distinct receptors to reduce corneal pain and inflammation. *Cannabis and cannabinoid research* **3**, 11–20 (2018).
8. Billakota, S., Devinsky, O. & Marsh, E. Cannabinoid therapy in epilepsy. *Current opinion in neurology* **32**, 220–226 (2019).
9. Radwan, M. M. *et al*. Isolation and pharmacological evaluation of minor cannabinoids from high-potency Cannabis sativa. *Journal of natural products* **78**, 1271–1276 (2015).
10. Borille, B. T. *et al*. Near infrared spectroscopy combined with chemometrics for growth stage classification of cannabis cultivated in a greenhouse from seized seeds. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **173**, 318–323 (2017).
11. Ryu, B. R. *et al*. Conversion characteristics of some major cannabinoids from hemp (Cannabis sativa L.) raw materials by new rapid simultaneous analysis method. *Molecules* **26**, 4113 (2021).
12. Malabadi, R. B., Kolkar, K. & Chalannavar, R. Medical Cannabis sativa (Marijuana or Drug type); The story of discovery of Δ9-Tetrahydrocannabinol (THC). *International Journal of Innovation Scientific Research and Review* **5**, 4134–4143 (2023).
13. National Conference of State Legislatures. *State Medical Cannabis Laws*. https://www.ncsl.org/health/state-medical-cannabis-laws (2024).
14. Blessing, E. M., Steenkamp, M. M., Manzanares, J. & Marmar, C. R. Cannabidiol as a potential treatment for anxiety disorders. *Neurotherapeutics* **12**, 825–836 (2015).
15. Atalay, S., Jarocka-Karpowicz, I. & Skrzydlewska, E. Antioxidative and anti-inflammatory properties of cannabidiol. *Antioxidants* **9**, 21 (2019).
16. Jones, N. A. *et al*. Cannabidiol Displays Antiepileptiform and Antiseizure Properties *In Vitro* and *In Vivo*. *J Pharmacol Exp Ther* **332**, 569–577 (2010).
17. The effects of cannabidiol and its synergism with bortezomib in multiple myeloma cell lines. A role for transient receptor potential vanilloid type-2 - Morelli - 2014 - International Journal of Cancer - Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.28591?casa_token=C9zXgL9ZPwQAAAAA%3AvashIza5lImcCEYUrgsnibwgFqOE_sIkZa6VFroY7yJ9MrKn90kR0cGM_EmqjGOoNdQ2rWC5Q6hzCAk.
18. Fulvio, F., Righetti, L., Minervini, M., Moschella, A. & Paris, R. The B1080/B1192 molecular marker identifies hemp plants with functional *THCA synthase* and total THC content above legal limit. *Gene* **858**, 147198 (2023).
19. The characterization of key physiological traits of medicinal cannabis (Cannabis sativa L.) as a tool for precision breeding | BMC Plant Biology. https://link.springer.com/article/10.1186/s12870-021-03079-2.
20. van Bakel, H. *et al*. The draft genome and transcriptome of Cannabis sativa. *Genome Biol* **12**, R102 (2011).
21. Grassa, C. J. *et al*. A new Cannabis genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytologist* **230**, 1665–1679 (2021).
22. McGarvey, P. *et al*. *De novo* assembly and annotation of transcriptomes from two cultivars of *Cannabis sativa* with different cannabinoid profiles. *Gene* **762**, 145026 (2020).
23. Laverty, K. U. *et al*. A physical and genetic map of Cannabis sativa identifies extensive rearrangements at the THC/CBD acid synthase loci. *Genome Res.* **29**, 146–156 (2019).
24. Gao, S. *et al*. A high-quality reference genome of wild Cannabis sativa. *Horticulture Research* **7**, 73 (2020).
25. Braich, S., Baillie, R. C., Spangenberg, G. C. & Cogan, N. O. A new and improved genome sequence of Cannabis sativa. *Gigabyte* **2020** (2020).
26. Kovalchuk, I. *et al*. The Genomics of Cannabis and Its Close Relatives. *Annual Review of Plant Biology* **71**, 713–739 (2020).
27. Rhoads, A. & Au, K. F. PacBio Sequencing and its Applications. *Genomics, Proteomics & Bioinformatics* **13**, 278–289 (2015).
28. Sun, X. *et al*. Nanopore Sequencing and Its Clinical Applications. in *Precision Medicine* (ed. Huang, T.) 13–32, https://doi.org/10.1007/978-1-0716-0904-0_2 (Springer US, New York, NY, 2020).
29. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
30. 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 (2020).
31. Zhou, Y. *et al*. *De novo* assembly of plant complete genomes. *T* **1**, 1–8 (2022).
32. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:ERS2852417 (2020).
33. De Roeck, A. *et al*. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biol* **20**, 239 (2019).
34. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **5**, 4.10.1–4.10.14 (2004).
35. Fu, X.-G. *et al*. Phylogenomic analysis of the hemp family (Cannabaceae) reveals deep cyto-nuclear discordance and provides new insights into generic relationships. *Journal of Systematics and Evolution* **61**, 806–826 (2023).
36. Padgitt-Cobb, L. K. *et al*. A draft phased assembly of the diploid Cascade hop (Humulus lupulus) genome. *The Plant Genome* **14**, e20072 (2021).
37. Liang, Q. *et al*. The genome assembly and annotation of yellowhorn (Xanthoceras sorbifolium Bunge). *GigaScience* **8**, giz071 (2019).
38. Liang, J., Kong, L., Hu, X., Fu, C. & Bai, S. Chromosomal-level genome assembly of the high-quality Xian/Indica rice (Oryza sativa L.) Xiangyaxiangzhan. *BMC Plant Biol* **23**, 94 (2023).

39. Lee, D.-J. *et al*. Chromosome-Scale Genome Assembly and Triterpenoid Saponin Biosynthesis in Korean Bellflower (Platycodon grandiflorum). *International Journal of Molecular Sciences* **24**, 6534 (2023).
40. Edwards, K. D. *et al*. A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics* **18**, 448 (2017).
41. Li, Z. *et al*. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* **12**, 540 (2011).
42. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
43. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research* **26**, 1107–1115 (1998).
44. Zaharia, M. *et al*. Faster and More Accurate Sequence Alignment with SNAP. Preprint at https://doi.org/10.48550/arXiv.1111.5572 (2011).
45. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
46. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
47. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
48. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**, D501–D504 (2005).
49. Cheng, C.-Y. *et al*. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal* **89**, 789–804 (2017).
50. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
51. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
52. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**, W182–W185 (2007).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP402544 (2023).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX17887361 (2023).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX17887360 (2023).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX17887355 (2023).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX17887359 (2023).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX17887358 (2023).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX17887357 (2023).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX17887356 (2023).
61. Cannabis sativa cultivar Pink pepper isolate KNU-18-1, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JAQSJK000000000 (2023).
62. Cannabis sativa L. (Pink pepper) Annotation Data Set. *figshare* https://doi.org/10.6084/m9.figshare.21391449.v6 (2024).
63. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one* **11**, e0163962 (2016).
64. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1–9 (2009).
65. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40**, e49–e49 (2012).
66. Cannabis sativa, L. (Pink pepper) synteny result data set. *figshare* https://doi.org/10.6084/m9.figshare.27196350.v1 (2024).
67. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
68. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
69. Hurgobin, B. *et al*. Recent advances in Cannabis sativa genomics research. *New Phytologist* **230**, 73–89 (2021).
70. Flajšman, M., Slapnik, M. & Murovec, J. Production of feminized seeds of high CBD Cannabis sativa L. by manipulation of sex expression and its application to breeding. *Frontiers in plant science* **12**, 718092 (2021).
71. Barcaccia, G. *et al*. Potentials and Challenges of Genomics for Breeding Cannabis Cultivars. *Front. Plant Sci.* **11** (2020).
72. Cannabis sativa, L. (Pink pepper) whole genome alignment data set. *figshare* https://doi.org/10.6084/m9.figshare.27198693.v1 (2024).
73. Lewis, M. A., Russo, E. B. & Smith, K. M. Pharmacological Foundations of Cannabis Chemovars. *Planta Med* **84**, 225–233 (2018).

Acknowledgements

This study was supported by the Ministry of Science and ICT (MSIT, Korea) (support program: 2021-DD-UP-0379) and the BK21 FOUR program of the National Research Foundation (NRF, Korea). We thankful the National Institute of Food and Drug Safety Evaluation for granting the Narcotics Academic Researcher approval (Permit Number: Seoul-1806, Seoul, Korea) and the Narcotics Raw Material Handling Approval (Narcotics Policy Division-4789) for the collection, cultivation, analysis, and use of the plant material in this study.

Author contributions

J.-D.L. designed and conceived the experiment. B.-R.R., G.-J.G., Y.-R.S. drafted the manuscript and visualized the data. B.-R.R., G.-J.G., S.-H.P. analyzed the data, B.-R.R., Y.-S.L. breed the plant. M.-J.Kang, M.-J.Kim, T.-H.K. propagated, and managed the plant material and G.-J.G., S.-H.P. corrected the manuscript. S.-H.P., J.-D.L. supervised the study. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.-H.P. or J.-D.L.

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.