# A New and Improved Genome Sequence of *Cannabis sativa*

**Shivraj Braich[1,2], Rebecca C. Baillie[1], German C. Spangenberg[1,2], Noel O.I. Cogan[1,2*]**

[1]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia

[2]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3086, Australia

*Corresponding author

**E-mail: noel.cogan@agriculture.vic.gov.au**

## Abstract

Cannabis is a diploid species (2n = 20), the estimated haploid genome sizes of the female and male plants using flow cytometry are 818 and 843 Mb respectively. Although the genome of Cannabis has been sequenced (from hemp, wild and high-THC strains), all assemblies have significant gaps. In addition, there are inconsistencies in the chromosome numbering which limits their use. A new comprehensive draft genome sequence assembly (~900 Mb) has been generated from the medicinal cannabis strain Cannbio-2, that produces a balanced ratio of cannabidiol and delta-9-tetrahydrocannabinol using long-read sequencing. The assembly was subsequently analysed for completeness by ordering the contigs into chromosome-scale pseudomolecules using a reference genome assembly approach, annotated and compared to other existing reference genome assemblies. The Cannbio-2 genome sequence assembly was found to be the most complete genome sequence available based on nucleotides assembled and BUSCO evaluation in *Cannabis sativa* with a comprehensive genome annotation. The new draft genome sequence is an advancement in Cannabis genomics permitting pan-genome analysis, genomic selection as well as genome editing.

# Main Content

## Context

The legalisation of medicinal cannabis has spread across the globe leading to increased benefits for a range of conditions. *Cannabis sativa* (NCBI:txid3483) is an erect, annual, wind-pollinated herb, that is typically dioecious although monoecious forms can exist. The plant is diploid (2n = 20) with gender driven by a pair of sex chromosomes (X and Y) along with the nine autosomes [1, 2]. The diploid genome sizes of the female and male plants using flow cytometry are 1,636 ± 7.2 and 1,683 ± 13.9 Mbp, respectively [3, 4]. Cannabis plants are best known for cannabinoid biosynthesis, most prominent of these include delta-9-tetrahydrocannabinol ($\Delta^9$-THC, or simply THC) and cannabidiol (CBD). Preparations from medicinal cannabis extract have various pharmacological effects (depending on the cannabinoid composition) for example, CBD has effects as a muscle relaxant, anticonvulsant, neuroprotective, antioxidant, anxiolytic and also has antipsychotic activity; while THC's effects can be utilised as a psychopharmaceutical, as well as an analgesia, appetite stimulation, antiemesis and also for muscle relaxation [5]. Besides CBD and THC, other cannabinoids such as cannabichromene (CBC) [6], cannabigerol (CBG) [7] and delta-9-tetrahydrocannabivarin (THCV) [8] have also been recognised to have pharmacological effects. Moreover, secondary metabolites from cannabis plant tissues, such as flavonoids and terpenes are also known to contribute to psychoactive or therapeutic effects [9]. The biosynthesis of cannabinoids and terpenes with medicinal properties is currently only partly understood and additional genetic and genomic studies will further illuminate the different production mechanisms that the various plant genotypes deliver.

An initial draft genome sequence of cannabis was published in 2011 that generated 534 Mbp of assembled nucleotides available from the drug-type variety, Purple Kush (PK) [10].

Following the generation of an initial draft genome sequence, several chromosome-scale whole genome sequence assemblies were made available in 2018 using long-read sequencing technology from the strains; PK (high THC producing female plant, GenBank-GCA_000230575.5), Finola (hemp, male plant, GenBank-GCA_003417725.2) and CBDRx (high CBD producing plant, genome sequence assembly named cs10 within GenBank-GCA_900626175.2) and recently in 2020 from the strain, JL (wild-type, female plant, GenBank-GCA_013030365.1) with assembled sequence size of 639 Mb, 784 Mb, 714 Mb and 797 Mb, respectively (without Ns) [11-13]. Despite the use of long-read sequencing technology, the published assemblies have significant gaps and inconsistent nomenclature of chromosomes numbering and orientation. The availability of a comprehensive genome sequence from a medicinal strain will add clarity relating to gene characterisation and functional analysis as well as valuable diversity for a pan-genome analysis.

The current study reports the development of an improved comprehensive draft genome sequence for *Cannabis sativa* that integrates the dataset generated from a female genotype which produces a balanced CBD:THC cannabinoid ratio, Cannbio-2 (Cb-2, Figure 1, [14]). The study also provides the genome annotation using the published extensive transcriptome dataset [15] as evidence and evaluation of the generated genome sequence and compares the sequence dataset to available whole genome sequence assemblies.

## Methods

### Plant materials and DNA isolation

All plants were maintained under artificial conditions in controlled environment facilities and all the work undertaken was performed under Medicinal Cannabis Research Licence (RL011/18) and Permit (RL01118P6) issued by the Department of Health (DoH), Office of Drug

Control (ODC) Australia. A variety of seeds were imported from a legal source in Canada and were screened with DNA markers and using comprehensive chemical analysis [14]. Cannbio-2 was identified as a female plant and selected as an optimal strain that produces a balanced CBD:THC cannabinoid ratio [14]. Fresh leaves were sampled from the female cannabis plant, Cannbio-2, and the harvested tissue was stored at -80° C until required. Genomic DNA was isolated with the DNeasy® Plant 96 Kit (QIAGEN, Hilden, Germany) following manufacturer's instructions. Isolated high molecular weight DNA was quantified by fluorometry (Qubit, Thermo Fisher Scientific, Waltham, U.S.A.) and assessed for quality using a 1 % (w/v) pulse-field gel electrophoresis and with genomic ScreenTape on the TapeStation 2200 platform (Agilent Technologies, Santa Clara, CA, USA).

**Pacific Biosciences sequencing and genome assembly**

Single Molecule Real Time (SMRT) bell libraries were prepared from the extracted DNA using the SMRTbell™ Template Prep Kit 1.0-SPv3 according to the protocol "20 kb Template Preparation Using BluePippin Size-Selection System" as recommended by the manufacturer (Pacific Biosciences) with the exception that the initial DNA was not sheared. Incompletely formed or non-SMRTbell DNA was removed by exonuclease treatment. The SMRTbell templates were size-selected using the BluePippin system (Sage Sciences) on a 0.75% (w/v) agarose gel cassette aiming to remove library insert sizes smaller than 15 kb. Size-selected libraries were further cleaned using the AMPure PB beads (Pacific Biosciences). The SMRTbell templates were quantified by a high-sensitivity fluorometric assay (Qubit, Thermo Fisher Scientific, Waltham, U.S.A.) and quality assessed using Genomic DNA ScreenTape on the TapeStation 2200 platform (Agilent Technologies, Santa Clara, CA, USA). The generated SMRT bell templates were sequenced on the PacBio Sequel instrument (PacBio Sequel System, RRID:SCR_017989) with the Sequel™ SMRT® cells 1M v2 Tray as per the manufacturer's

instructions. The raw PacBio reads were error-corrected and assembled using the SMRT Link's Hierarchical Genome Assembly Process (HGAP4) *de novo* assembly application (v5.0.0) with default parameters to generate the *de novo* assembly. RaGOO [16] (v1.1) that uses minimap2 (v2.10, RRID:SCR_018550) [17] was used to reference align, to order and orientate the draft genome assembly contigs of Cannbio-2 to chromosome scale pseudomolecules using reference genomes of cs10, PK and JL. Default parameters with the exception of the "-b" option, to break chimeric contigs and "-g 100" to use gap size of 100 N's for padding in pseudomolecules was used.

**Comparison of genome assemblies**

Available whole genome assemblies of cs10, PK, Finola and JL were compared to the generated genome assembly in the current study. For the comparisons, whole-genome sequence alignments were created using minimap2 [17] (v2.10) with the parameter "-x asm5 –cs" to generate pairwise alignment format (PAF) file using the Cannbio-2 genome sequence assembly as the reference and published genome sequence assemblies as query. The alignments were converted to dot plot using dotPlotly v1.0 [18] in R.

**Genome annotation**

The genome annotation was performed following the GenSAS [19] v6 pipeline on the draft assembly contigs ordered into pseudomolecules. Repeat regions in the genome assembly were masked using RepeatMasker v4.0.7 (RRID:SCR_012954) [20] (with '*Arabidopsis thaliana*, *Oryza sativa* and other dicots' repeat libraries) and *de novo* repeat finding tool RepeatModeler v1.0.11 (RRID:SCR_015027) [21] to create a soft-masked consensus sequence. Transcript alignments were generated using BLASTN (v2.7.1, RRID:SCR_001598), BLAT (v35, RRID:SCR_011919) and PASA (v2.3.3, RRID:SCR_014656) using the Cannbio transcriptome assembly [15] as the database (BioProject: PRJNA560453, BioSample:

SAMN13503240-SAMN13503310, SRA: SRR10600874-SRR10600944). Initial *ab initio* gene predictions were made using Augustus (v3.3.1, RRID:SCR_008417) [22] with species '*Arabidopsis thaliana*'. EVidenceModeler (EVM, v06/25/2012, RRID:SCR_014659) [23] was used to create the consensus gene set by combining gene predictions from Augustus (weight score-1) and results from transcripts alignments (weight score-10). The consensus gene set was further refined using PASA to create the final gene set which was used for functional annotation. Functional analysis of the final gene set was primarily conducted using DIAMOND (v0.9.22, RRID:SCR_016071) [24] analysis to SwissProt database. Putative THCAS/CBDAS genes were identified based on the annotation and plotted across the genome using karyoplyteR (v1.10.0) [25] in R. Other tools were also utilised for the functional analysis including InterProScan (v5.25-68.0, RRID:SCR_005829) [26] and Pfam (v1.6, RRID:SCR_004726) [27]. The results from functional analysis were merged in creating an annotated genome submission in a GFF3 format.

## Results and Discussion

### Generation of genome sequence assembly

Cannbio-2 was sequenced to 86 x genome coverage by generating 70.09 Gbp of sequence data. The draft sequence assembly generated by HGAP4 resulted in 8,477 contigs assembled in 913.5 Mb with maximum contig length of 1,705,170 bp and N50 of 187,352 bp (Table 1). The draft genome sequence assembly of Cannbio-2 was comprehensively analysed through a reference guided assembly approach using the published genome sequence assemblies of PK, cs10 and JL as references to guide the chromosome scale sequence assembly process, resulted in genome assembly sizes (with Ns) of 756.33 Mb, 904.08 Mb and 891.96 Mb respectively (Table 2). Cannbio-2 genome sequence assembly guided using cs10 genome

sequence assembly was found to be the largest based on nucleotides assembled and was used for subsequent analysis to compare the draft genome to the other available references. Furthermore, cs10-guided assembly was also chosen for further analysis due to its chromosome nomenclature (which uses the linkage groups nomenclature from a previous study [28]). The statistical analysis of the new genome assembly generated from the current study and previously published chromosome-scale genome assemblies are summarised in Table 1. The analysis revealed that the generated genome sequence was found to be the most complete with assembly size of 903 Mb when compared to the whole genome assemblies of cs10 (714 Mb), Finola (784 Mb), PK (639 Mb) and JL (797 Mb). The size of the generated genome assembly was found to be larger than the estimated *C. sativa* (Hemp) genome size using the flow-cytometry (818 Mb) [4]. The differences in the genome size could possibly reflect bias introduced due to the use of a different accession to orient and order the contigs to pseudomolecules or potential haplotype duplication or the genome variations (such as insertions, inversions, tandem repeats to name a few) between the hemp and medicinal cannabis strain.

**Comparison of genome assemblies**

The generated genome assembly was found to be consistent in terms of chromosome nomenclature with few structural differences based on the alignment results when compared to the cs10 genome assembly (Figure 2). Despite the larger size of the generated genome assembly, large regions of duplication were not apparent when alignments were visualised as represented in Figure 2, highlighting the contiguity of the generated assembly. Comparisons were also made between Finola, PK and JL to the generated genome based on the alignment results (Figure 3, 4 and 5; Alignment files in GigaDB [29]). The comparisons of the genome sequences revealed large pericentromeric differences and chromosome inversions between

the Cannbio-2 genome sequence and the genome sequences of Finola, PK and JL. Moreover, comparisons of JL, cs10, PK and Finola genome sequences revealed inconsistencies between these genome sequences in terms of orientation and numbering of chromosomes (Alignment files in GigaDB [29]).

**Genome annotation**

The total predicted features from the repeat-masked consensus sequence were found to be 3,419,223. Initial *ab initio* gene predictions that were made using Augustus resulted in prediction of 40,633 genes. The consensus gene set, derived by EVidenceModeler, generated a prediction of 36,758 genes which was further refined using PASA. The total predicted features from the final gene set following PASA refinement were 109,686 with 36,632 genes, 37,107 mRNA and 35,947 proteins. The predicted features per chromosome are as summarised in Table 3. Figure 6 represents the karyoplot of the density of masked repeats and genes across the 10 chromosomes of the Cannbio-2 annotated genome. Functional analysis of the final gene set based on DIAMOND analysis to SwissProt database, resulted in the identification of 16 putative THCAS/CBDAS genes across the Cannbio-2 genome sequence with 12 of these genes coded by chromosome 7 (Figure 6).

## Data validation and quality control

Genomic DNA was extracted from fresh leaves of the Cannbio-2 plant using the DNeasy 96 Plant Kit (QIAGEN, Hilden, Germany), according to the manufacturer's instructions. Whole genome of Cannbio-2 was re-sequenced using an enzymatic MspJI (NEB, MA, United States) shearing method [30] as described previously [31]. The library was assessed using a D1000 ScreenTape on the TapeStation 2200 (Agilent, Santa Clara, CA, USA) and was subjected to paired-end sequencing on a HiSeq 3000 instrument (Illumina Inc., San Diego, CA, USA). The

initial generated fastq sequences were quality trimmed using a custom perl script (available in GigaDB [29]) and adaptor trimmed by cutadapt (v2.6, RRID:SCR_011841) [32]. The trimmed sequence reads were aligned to the generated sequence assembly of the Cannbio-2 strain using the BWA-MEM software package [33] (v0.7.17, RRID:SCR_010910) with default parameters, to evaluate the genome assembly. The alignment results of the sequence reads to the generated genome assembly indicated that out of a total of 178.72 million QC-passed reads, 99.65% sequence reads were found to be mapped with 86.78% of sequence reads being properly paired, suggesting that the generated genome assembly contained comprehensive genomic information.

Benchmarking Universal Single-Copy Orthologs (BUSCO, v4.0.6, RRID:SCR_015008) [34] approach was used with the eudicotyledons_odb10 dataset in genome mode for all the genome assemblies to assess the completeness of the conserved proteins in the published and current genome sequence assemblies. Only pseudomolecules were used in the BUSCO analysis across all the genomes. The Cannbio-2 genome sequence captured 93% of genes as predicted by BUSCO evaluation which was found to be higher than all other published genome assemblies of (cs10-90.3%; Finola-82.6%; JL-86.5%; PK-78.2%; Figure 7). The results from the BUSCO analysis confirms the completeness of the Cannbio-2 genome sequence assembly. Furthermore, a detailed BLASTN analysis (v2.9.0) was performed to search for inadvertent chloroplast (KR184827.1, 153,848 bp and NC_027223.1, 153,854 bp) and mitochondrial (KR059940.1, 414,545 bp) genomes to check for the integration of organellar genomes in the generated assembly. The similarity results of Cannbio-2 genome to the organellar genomes showed incorporation of small fragments with a maximum length of 30 kb for chloroplast genome sequence assembly and 12 kb for mitochondrial genome sequence assembly

(BLASTN results in GigaDB [29]). The similarity results suggest no significant integration of these inadvertent genome sequences into the Cannbio-2 genome sequence assembly.

## Conclusion and future perspective

The results suggest that the Cannbio-2 draft genome is the most comprehensive genome sequence of cannabis published to date. The development of a contiguous cannabis genome sequence will provide novel insights into the identification of genome-wide sequence variants. The research from the current study will also enable genomic selection, genome editing and pan-genome sequence analysis in medicinal cannabis.

## Disclaimer

The genome sequence data generated in this study was not assessed for the presence of potential haplotype duplication and genome heterozygosity.

## Availability of supporting data

Sequence data has been deposited at DDBJ/EMBL/GenBank under the BioProject ID PRJNA667278. The Cannbio-2 sequence reads (short reads and long reads), genome assembly (draft genome assembly sequence and cs10 guided genome assembly sequence), contigs tilling path to chromosomes table, genome annotation and additional files have been deposited in the *GigaScience* GigaDB repository [29].

## Declarations

### List of abbreviations

BUSCO: Benchmarking Universal Single Copy Orthologs; Cb-2: Cannbio-2; CBC: cannabichromene; CBD: cannabidiol; CBG: cannabigerol; DoH: Department of Health; EVM: EVidenceModeler; HGAP4: Hierarchical Genome Assembly Process; ODC: Office of Drug Control; PAF: pairwise alignment format; PK: Purple Kush; SMRT: Single Molecule Real Time; $\Delta^9$-THC or THC: delta-9-tetrahydrocannabinol; THCV: delta-9-tetrahydrocannabivarin.

**Competing Interests**

The authors declare that they have no competing interests.

**Funding**

This work was supported by funding from Agriculture Victoria and Agriculture Victoria Services.

**Author's Contributions**

S.B. and R.C.B. prepared plant materials, performed DNA extraction and sequencing of the libraries. S.B. conducted the data analysis and drafted the manuscript. N.O.I.C. assisted in the experimental design and data analysis. G.C.S. and N.O.I.C. conceptualized the project and assisted with preparation of the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**

The authors would like to thank Doris Ram, Alix L. Malthouse, Melinda C. Quinn and Larry S. Jewell for providing their support in the maintenance of the medicinal cannabis strains.

## References

1. Hirata K. Cytological basis of the sex determination in *Cannabis sativa* L. The Japanese Journal of Genetics. 1924;4:198-201.

2. Sakamoto K, Shimomura K, Komeda Y, Kamada H and Satoh S. A male-associated DNA sequence in a dioecious plant, *Cannabis sativa* L. Plant and Cell Physiology. 1995;36 8:1549-54.

3. Faux A-M, Berhin A, Dauguet N and Bertin P. Sex chromosomes and quantitative sex expression in monoecious hemp (*Cannabis sativa* L.). Euphytica. 2014;196 2:183-97. doi:10.1007/s10681-013-1023-y.

4. Sakamoto K, Akiyama Y, Fukui K, Kamada H and Satoh S. Characterization; Genome Sizes and Morphology of Sex Chromosomes in Hemp (*Cannabis sativa* L.). Cytologia. 1998;63 4:459. doi:10.1508/cytologia.63.459.

5. Russo E and Guy GW. A tale of two cannabinoids: The therapeutic rationale for combining tetrahydrocannabinol and cannabidiol. Medical Hypotheses. 2006;66 2:234-46. doi:10.1016/j.mehy.2005.08.026.

6. Izzo AA, Capasso R, Aviello G, Borrelli F, Romano B, Piscitelli F, et al. Inhibitory effect of cannabichromene, a major non-psychotropic cannabinoid extracted from *Cannabis sativa*, on inflammation-induced hypermotility in mice. British Journal of Pharmacology. 2012;166 4:1444-60. doi:10.1111/j.1476-5381.2012.01879.x.

7. Borrelli F, Pagano E, Romano B, Panzera S, Maiello F, Coppola D, et al. Colon carcinogenesis is inhibited by the TRPM8 antagonist cannabigerol, a Cannabis-derived non-psychotropic cannabinoid. Carcinogenesis. 2014;35 12:2787. doi:10.1093/carcin/bgu205.

8. McPartland JM, Duncan M, Di Marzo V and Pertwee RG. Are cannabidiol and Δ(9) -tetrahydrocannabivarin negative modulators of the endocannabinoid system? A systematic review. British Journal of Pharmacology. 2014;172 3:737. doi:10.1111/bph.12944.

9. Russo EB. Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. British Journal of Pharmacology. 2011;163 7:1344-64. doi:10.1111/j.1476-5381.2011.01238.x.

10. van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, et al. The draft genome and transcriptome of *Cannabis sativa*. Genome Biology. 2011;12 10:R102. doi:10.1186/gb-2011-12-10-r102.

11. Grassa CJ, Wenger JP, Dabney C, Poplawski SG, Motley ST, Michael TP, et al. A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. bioRxiv. 2018:458083. doi:10.1101/458083.

12. Laverty KU, Stout JM, Sullivan MJ, Shah H, Gill N, Holbrook L, et al. A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci. Genome research. 2019;29 1:146-56. doi:10.1101/gr.242594.118.

13. Gao S, Wang B, Xie S, Xu X, Zhang J, Pei L, et al. A high-quality reference genome of wild *Cannabis sativa*. Horticulture Research. 2020;7 1:73. doi:10.1038/s41438-020-0295-3.

14. Plant Breeders Rights. http://pericles.ipaustralia.gov.au/pbr_db/. Accessed 07 September 2020.

15. Braich S, Baillie RC, Jewell LS, Spangenberg GC and Cogan NOI. Generation of a Comprehensive Transcriptome Atlas and Transcriptome Dynamics in Medicinal Cannabis. Scientific Reports. 2019;9 1:16583. doi:10.1038/s41598-019-53023-6.

16. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology. 2019;20 1:224. doi:10.1186/s13059-019-1829-6.

17. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34 18:3094-100. doi:10.1093/bioinformatics/bty191.

18. dotPlotly v1.0 https://github.com/tpoorten/dotPlotly

19. Humann JL, Lee T, Ficklin S and Main D. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York; 2019. p. 29-51.

20. Smit, AFA, Hubley, R & Green, P. RepeatMasker. 2013-2015 http://www.repeatmasker.org

21. Smit, AFA, Hubley, R. RepeatModeler. 2008-2015 http://www.repeatmasker.org/RepeatModeler/

22. Oliver Keller, Martin Kollmar, Mario Stanke, Stephan Waack, A novel hybrid gene prediction method employing protein multiple sequence alignments, Bioinformatics, Volume 27, Issue 6, 15 March 2011, Pages 757–763, https://doi.org/10.1093/bioinformatics/btr010

23. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008 Jan 11;9(1):R7. doi: 10.1186/gb-2008-9-1-r7.

24. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59-60. doi: 10.1038/nmeth.3176.

25. Gel B, Serra E (2017). “karyoploteR : an R / Bioconductor package to plot customizable genomes displaying arbitrary data.” Bioinformatics, 33(19), 3088-3090. doi: 10.1093/bioinformatics/btx346.

26. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May 1;30(9):1236-40. doi: 10.1093/bioinformatics/btu031.

27. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014 Jan;42(Database issue):D222-30. doi: 10.1093/nar/gkt1223.

28. Weiblen GD, Wenger JP, Craft KJ, Elsohly MA, Mehmedic Z, Treiber EL, et al. Gene duplication and divergence affecting drug content in *Cannabis sativa*. New Phytologist. 2015;208 4:1241-50. doi:10.1111/nph.13562.

29. Braich S; Baillie RC; Spangenberg GC; Cogan NOI (2020): Supporting data for "A New and Improved Genome Sequence of Cannabis sativa" GigaScience Database. http://dx.doi.org/10.5524/100821

30. Shinozuka H, Cogan NOI, Shinozuka M, Marshall A, Kay P, Lin Y-H, et al. A simple method for semi-random DNA amplicon fragmentation using the methylation-dependent restriction enzyme MspJI. BMC Biotechnology. 2015;15 1:25. doi:10.1186/s12896-015-0139-7.

31. Malmberg MM, Shi F, Spangenberg GC, Daetwyler HD and Cogan NOI. Diversity and Genome Analysis of Australian and Global Oilseed *Brassica napus* L. Germplasm Using Transcriptomics and Whole Genome Re-sequencing. Frontiers in Plant Science. 2018;9 508 doi:10.3389/fpls.2018.00508.

32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17 doi:10.14806/ej.17.1.200.

356 33. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
357 MEM. arXiv:13033997v1. 2013.
358 34. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO:
359 assessing genome assembly and annotation completeness with single-copy
360 orthologs. Bioinformatics (Oxford, England). 2015;31 19:3210-2.
361 doi:10.1093/bioinformatics/btv351.
362
363

364 **Table 1.** Statistics of Cannbio-2 genome assembly from the current study as compared to
365 published whole genome sequence assemblies.

| Data Type | Cb-2[d] | Cb-2[r] | cs10 | JL | Finola | PK |
|---|---|---|---|---|---|---|
| Number of contigs/scaffolds | 8,477 | 10 | 10 | 10 | 10 | 10 |
| Assembly size with Ns (Mb) | 914 | 904 | 854 | 798 | 785 | 640 |
| Assembly size without Ns (Mb) | 914 | 903 | 714 | 797 | 784 | 639 |
| Largest contig/scaffold (Mb) | 1.7 | 106 | 105 | 93 | 101 | 79 |
| N50 (Mb) | 0.2 | 91 | 92 | 83 | 87 | 72 |
| N90 (Mb) | 0.05 | 72 | 65 | 69 | 50 | 51 |

366 [d] Draft Cb-2 genome assembly

367 [r] RaGOO assigned Cb-2 genome assembly using cs10 as the reference

368 **Table 2.** Number of bases per chromosome of Cannbio-2 genome assembled guided by PK,
369 cs10 and JL genome assembly as the reference.

| Sequence | PK-guided assembly | cs10-guided assembly | JL-guided assembly |
|---|---|---|---|
| Cs_Cb2_01 | 91,352,534 | 86,898,403 | 104,860,357 |
| Cs_Cb2_02 | 84,314,258 | 105,265,154 | 105,786,500 |
| Cs_Cb2_03 | 89,716,256 | 87,707,768 | 91,501,419 |
| Cs_Cb2_04 | 85,532,416 | 100,932,893 | 92,102,208 |
| Cs_Cb2_05 | 84,300,950 | 91,493,340 | 95,601,317 |
| Cs_Cb2_06 | 72,493,431 | 97,797,982 | 89,863,944 |
| Cs_Cb2_07 | 75,583,091 | 85,051,101 | 92,903,079 |
| Cs_Cb2_08 | 72,000,744 | 71,555,044 | 79,110,046 |
| Cs_Cb2_09 | 62,999,750 | 71,141,854 | 76,841,943 |
| Cs_Cb2_10 | 38,036,213 | 106,236,836 | 63,393,006 |
| **Total assembled size (Mb)** | 756,329,643 | 904,080,375 | 891,963,819 |

370

371

372 **Table 3.** Number of predicted features following repeat masking and following genome
373 sequence annotation (protein, mRNA and gene) per chromosome of the Cannbio-2 genome
374 sequence assembly.

| Sequence Name | Predicted Features- Repeats | Predicted Features- Annotation |
|---|---|---|
| Cs_Cb2_01 | 358,706 | 15,722 |
| Cs_Cb2_02 | 385,162 | 11,967 |
| Cs_Cb2_03 | 310,124 | 9,171 |
| Cs_Cb2_04 | 389,512 | 11,461 |
| Cs_Cb2_05 | 335,407 | 9,121 |
| Cs_Cb2_06 | 345,018 | 9,335 |
| Cs_Cb2_07 | 308,329 | 9,693 |
| Cs_Cb2_08 | 286,461 | 10,872 |
| Cs_Cb2_09 | 283,015 | 9,558 |
| Cs_Cb2_10/X | 417,489 | 12,786 |
| **Total** | 3,419,223 | 109,686 |

375

**Figure 1.** Example of Cannbio-2 plant with its leaf characteristics.

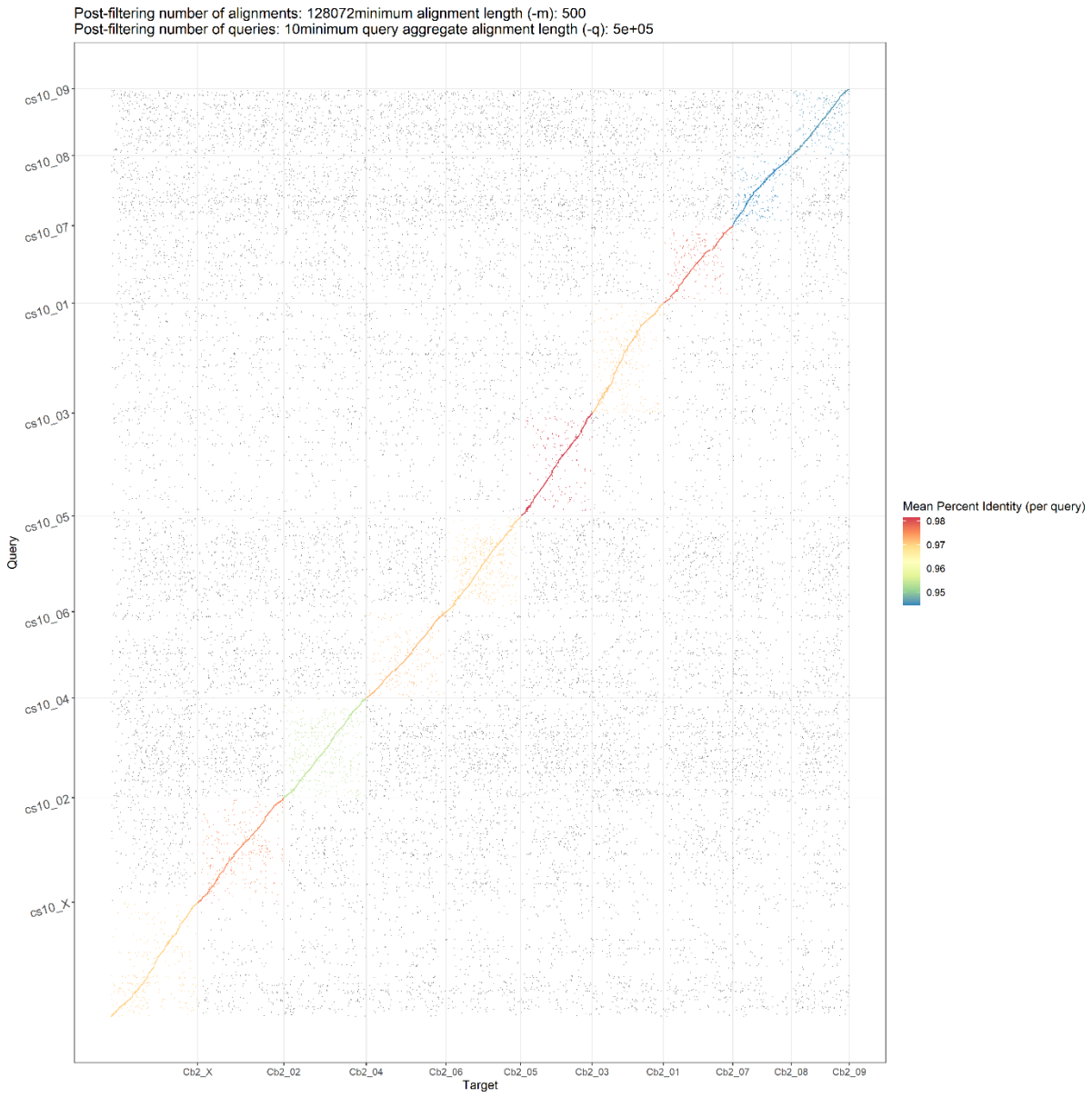



**Figure 2.** Dot plot showing alignments of Cannbio-2 sequence assembly to the whole genome sequence assembly of cs10.

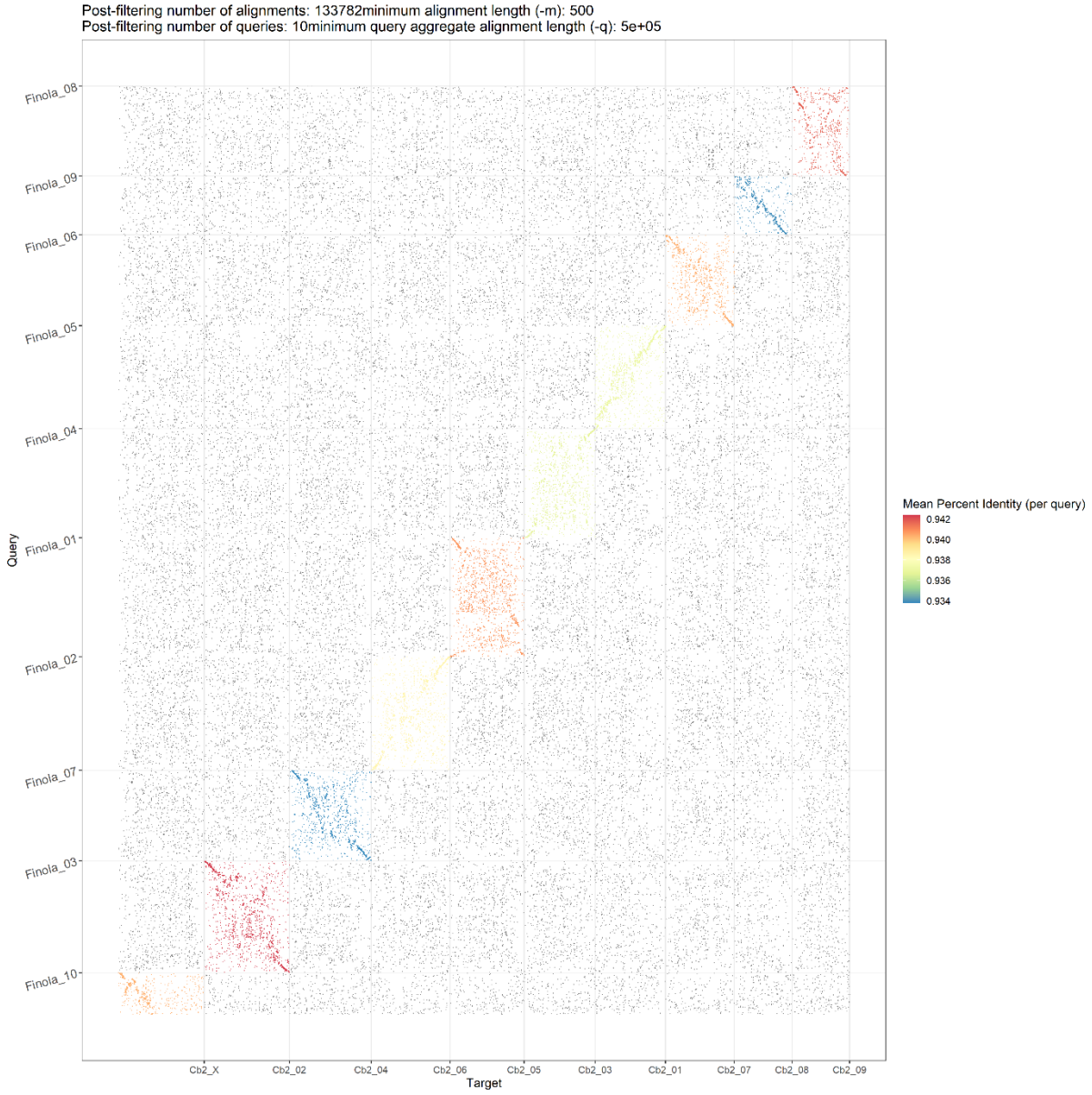



**Figure 3.** Dot plot showing alignments of Cannbio-2 sequence assembly to the whole genome sequence assembly of Finola.

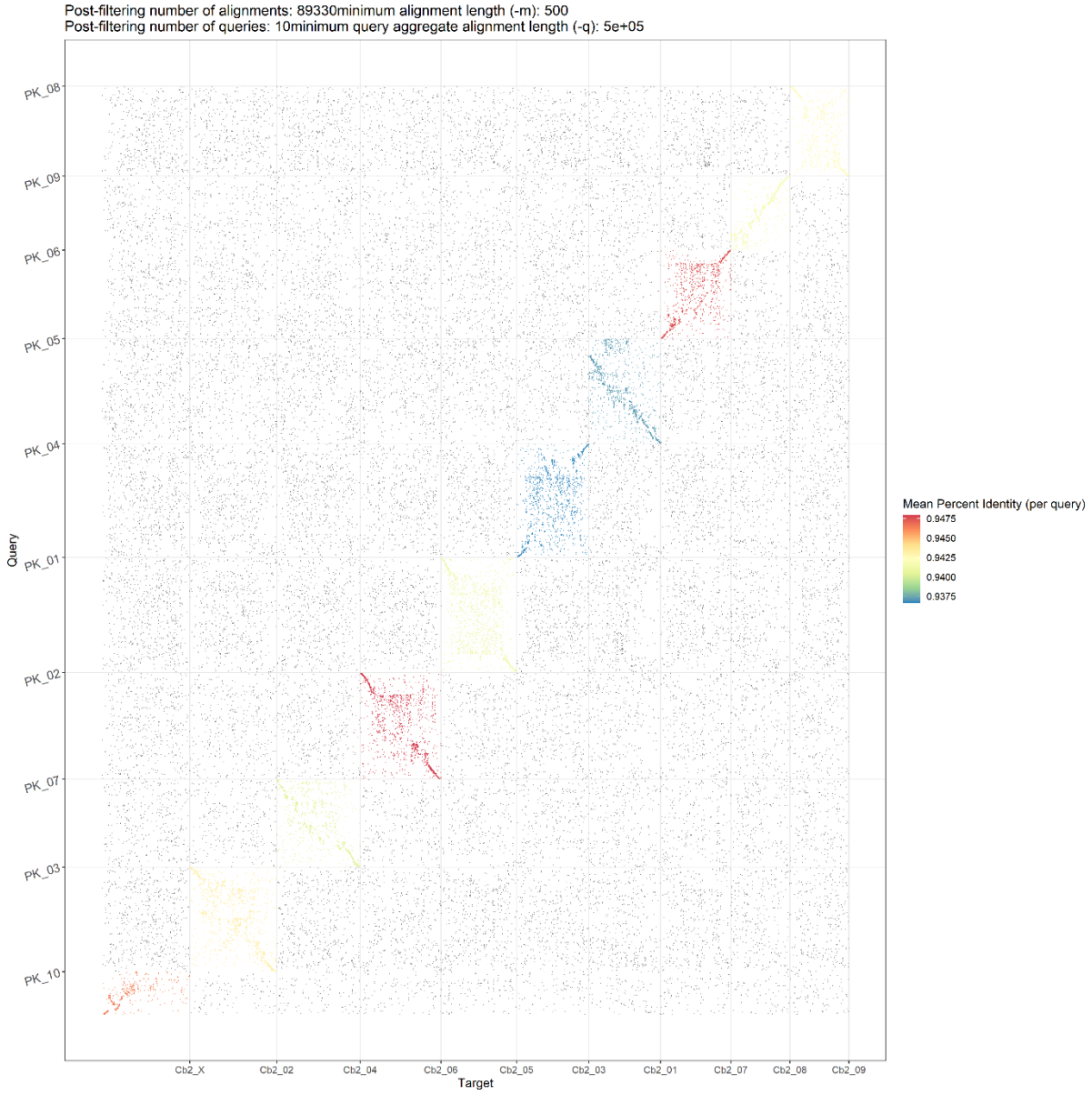
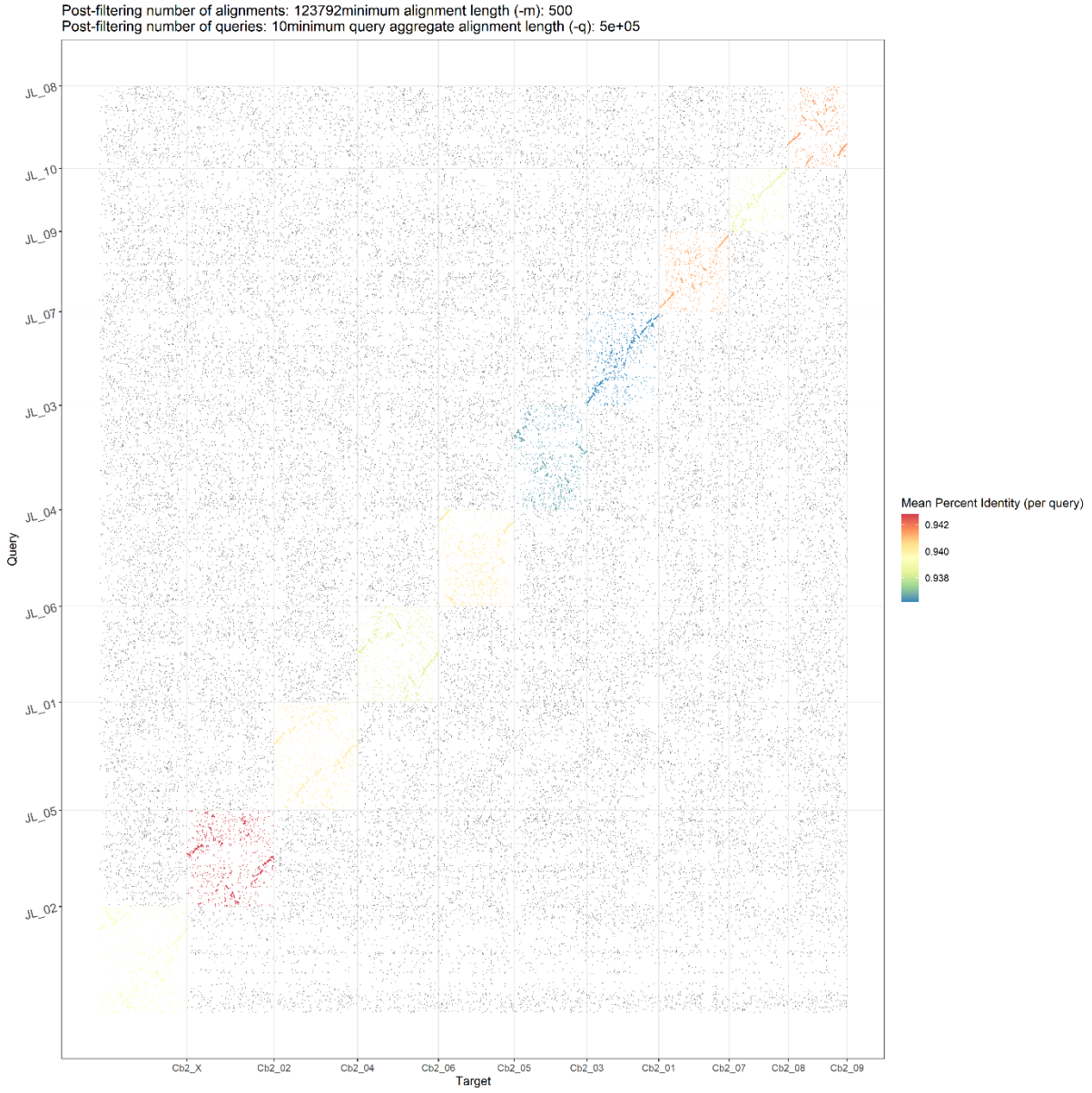



**Figure 4.** Dot plot showing alignments of Cannbio-2 sequence assembly to the whole genome sequence assembly of PK.




391

392 **Figure 5.** Dot plot showing alignments of Cannbio-2 sequence assembly to the whole genome

393 sequence assembly of JL.

394

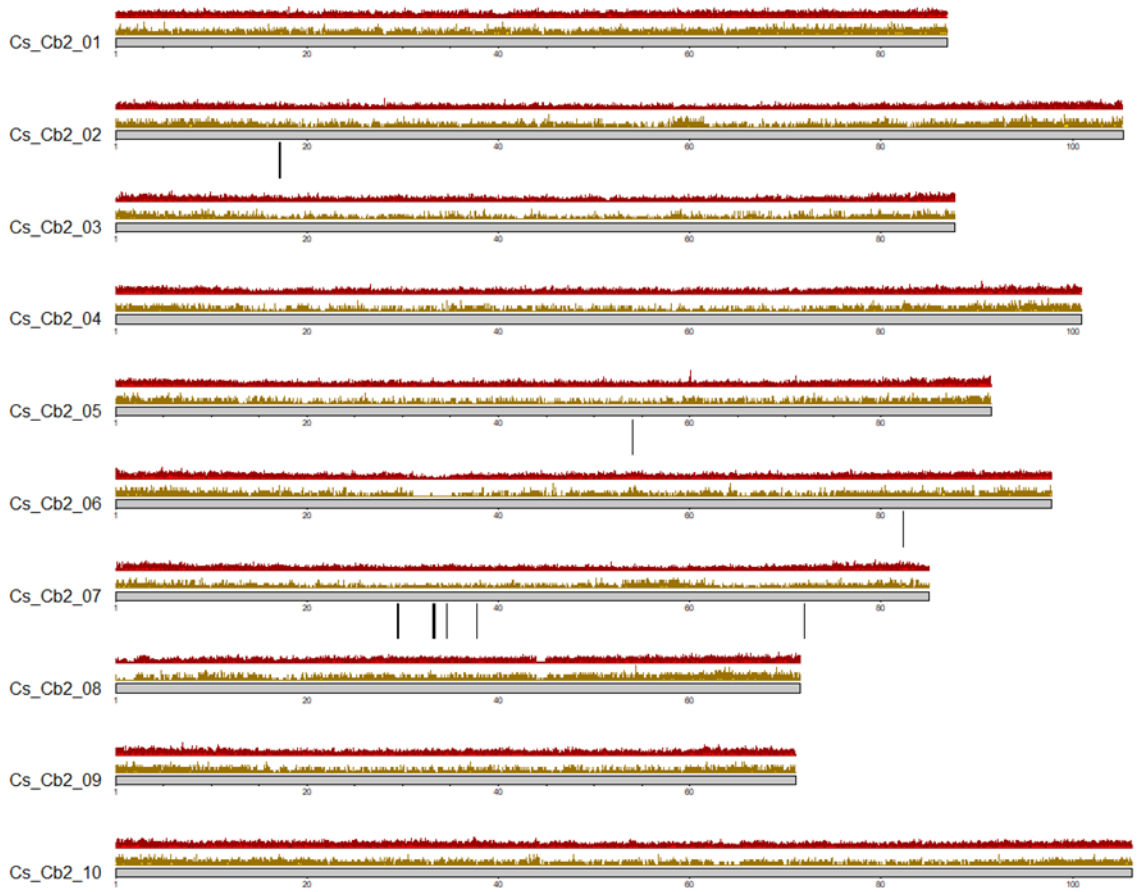
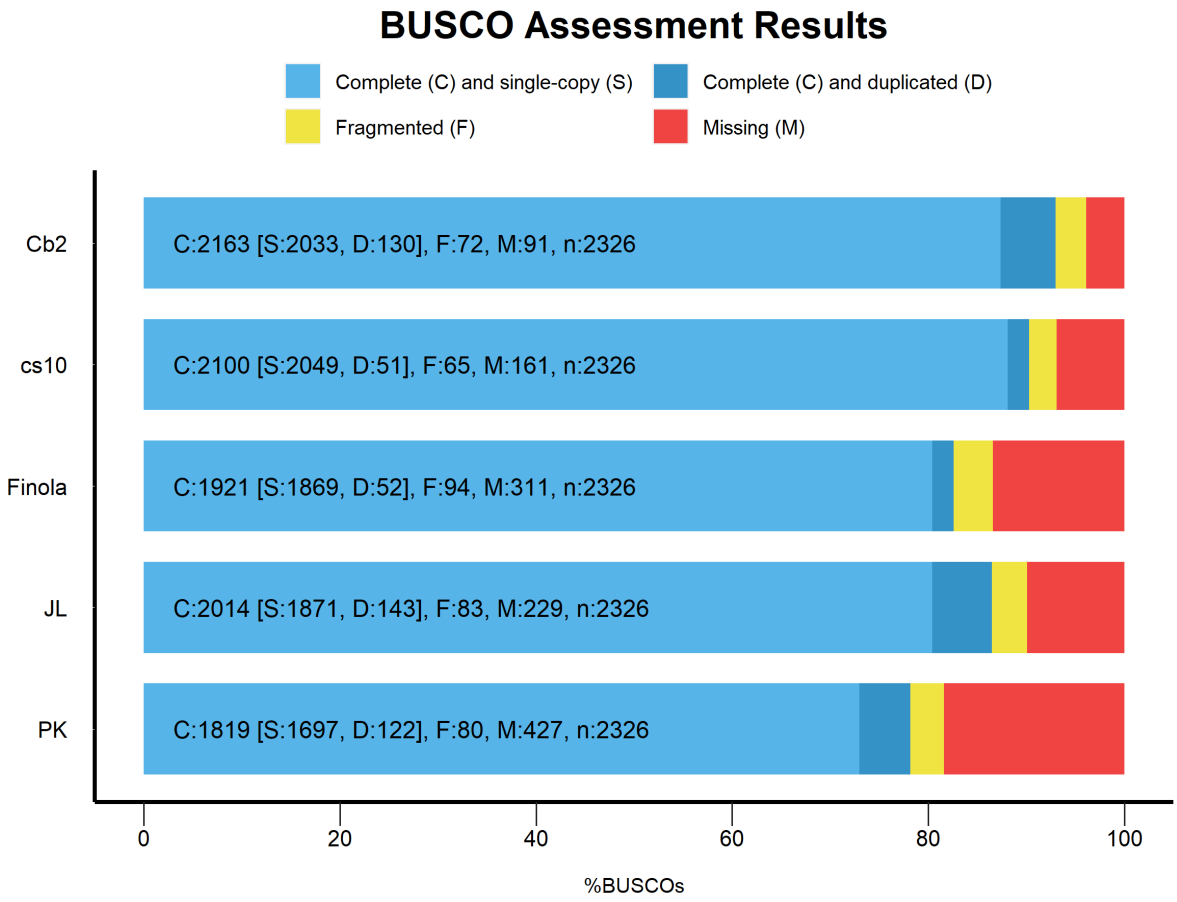



**Figure 6.** Cannbio-2 genome sequence assembly's karyoplot representing genome-wide density of masked repeat regions (gold), gene density (red) and regions of putative THC/CBD synthase genes (black lines).




400

401 **Figure 7.** BUSCO evaluation results of Cannbio-2 genome sequence assembly from the current

402 study as compared to published chromosome-scale whole genome sequence assemblies of

403 cs10, Finola, JL and PK.