

1 **Genomic description of critical upstream cannabinoid biosynthesis genes**

2

3 Peter A. Innes¹, Daniela Vergara¹

4

5 1 Department of Ecology and Evolutionary Biology, University of Colorado Boulder,
6 Boulder, Colorado

7

8 *Correspondence: peter.innes@colorado.edu, daniela.vergara@colorado.edu

9

10 **Keywords:** biochemical pathway, cannabinoids, gene copy number, hemp, marijuana

11

12

13

14

15

16

17

18

19

20

21

22

23

Abstract

Cannabinoid production is one of the key attributes of the plant *Cannabis sativa* and the characterization of the genes involved is an essential first step to develop tools for their optimization. We used bioinformatic approaches to annotate and explore variation in the coding genes for critical enzymes comprising the cannabinoid pathway: Olivetol Synthase (OLS), Olivetolic Acid Cyclase (OAC), and Cannabigerolic Acid Synthase (CBGAS), in multiple *C. sativa* genomes. These upstream genes of the Cannabinoid Oxidocyclase Genes THCAS, CBDAS, and CBCAS generate the necessary precursor molecules to produce the cannabinoids THC and CBD. We found that these genes vary in copy number and confirm that OLS, OAC, CBGAS, and the Cannabinoid Oxidocyclases are on separate chromosomes, while homologs are found in proximity. CBGAS, located on Chromosome X, suggests potential dosage effects in female plants. Except for the Cannabinoid Oxidocyclase genes, the other genes have multiple exons, up to 10 in CBGAS. Through differential exon usage explorations in CBGAS we found evidence for potential regulatory differences. This study provides valuable insight on the genomic identity and variation of cannabinoid biosynthesis genes that will benefit future research on the origin and evolution of this pathway, driver of economic, social, and medicinal value.

Introduction

Cannabis sativa L. (marijuana, hemp) is a flowering plant from the family Cannabaceae (Bell et al. 2010). **Error! Bookmark not defined.** *C. sativa* likely diverged from its closest extant relative, hops (*Humulus* spp.), approximately 25-27 million years ago in central Asia (McPartland et al 2019; Ren et al 2021). It has been used by humans for more than 10,000 years, making it one of the oldest domesticated plants (Li 1973; Ren 2021; Russo 2007). It is unclear if distinct wild populations of *C. sativa* still exist, though humans have helped spread the cultivated form across the globe (McPartland et al. 2019; Kovalchuk et al. 2020). Worldwide, this versatile plant has been cultivated for multiple uses. Hemp-type *C. sativa* produces fiber that can be used in paper, rope, or clothing, and grain that can be used to extract oil for cooking, personal hygiene and beauty products, while marijuana-

type *C. sativa* produces compounds used for medicinal and recreational purposes (Ahmed et al. 2022).

The *Cannabis* genus is considered to have a single species, *Cannabis sativa* L (Watts 2006), but it comprises at least two marijuana-type lineages and one hemp lineage (Sawler et al. 2015; Lynch et al. 2016; Vergara et al. 2016; Kovalchuk et al. 2020; Vergara et al. 2021). *Cannabis sativa* has a mixed mating system with either dioecious or monoecious (with individuals that produce male and female flowers) lineages. Dioecious lineages have sex chromosomes where females are homogametic (XX), and males are heterogametic (XY). Monoecious lineages appear to have two X chromosomes as well (Peil et al. 2003; Hillig 2005; Punja and Holmes 2020).

Cannabis sativa is renowned for its production of secondary metabolites, foremost cannabinoids and terpenes. Cannabinoids are a class of compounds that interact with the endocannabinoid system (Gertsch et al. 2008) and many have medicinal (Russo 2011; Swift et al. 2013; Volkow et al. 2014) or psychoactive (Russo and McPartland 2003; ElSohly and Slade 2005) properties. Two of the most widely known cannabinoids are Δ -9-tetrahydrocannabinolic acid (THCA), and Cannabidiolic acid (CBDA), which are converted to the neutral forms Δ -9-tetrahydrocannabinol (THC) and cannabidiol (CBD), respectively, once heated (Hart et al. 2001).

'Marijuana-type' lineages used for human consumption has focused on female domestication, since most of the cannabinoid production is found in the trichomes of female flowers (Sirikantaramas et al. 2005; Gagne et al. 2012). There has been strong human selection against males and monoecious individuals. On the other hand, hemp cultivated for their stalks or seeds for fiber or grain production, respectively, produce lower cannabinoid or terpenes (Schafroth et al. 2021).

Biochemical pathway

The last four steps of the biochemical pathway (Figure 1) from which the cannabinoids are produced begins with steps 1 and 2: the conversion of the compound hexanoyl-CoA to Olivetolic Acid via the action of two enzymes, Olivetol Synthase (OLS) and Olivetolic

82 Acid Cyclase (OAC). In step 3 (figure 1) Olivetolic Acid is then converted to another
 83 compound, Cannabigerolic Acid (CBGA), via the enzyme Cannabigerolic Acid Synthase
 84 (CBGAS). Cannabigerolic Acid, CBGA, has been referred to as the “mother cannabinoid,”
 85 since it is the precursor to the cannabinoids produced in step 4: THCA, CBDA, and a third
 86 cannabinoid produced in low quantities, Cannabichromenic Acid (CBCA; Page and Stout
 87 2017; Vergara et al. 2019; van Velzen and Schranz 2020; Smith et al. 2021). The enzymes
 88 that produce these final cannabinoids are referred to as cannabinoid oxidocyclases (van
 89 Velzen and Schranz 2020) because they oxidize the CBGA compound while adding an
 90 additional ring.

91 Like THCA and CBDA, decarboxylation is responsible for the formation of
 92 Cannabigerol (CBG), Cannabichromene (CBC), and other neutral cannabinoids (Valliere et
 93 al. 2019). Due to their low abundance, these “minor cannabinoids” have generally been
 94 less well-studied than THC and CBD, although they display a range of interesting
 95 pharmacological properties and potential medicinal value (Allen et al. 2022).

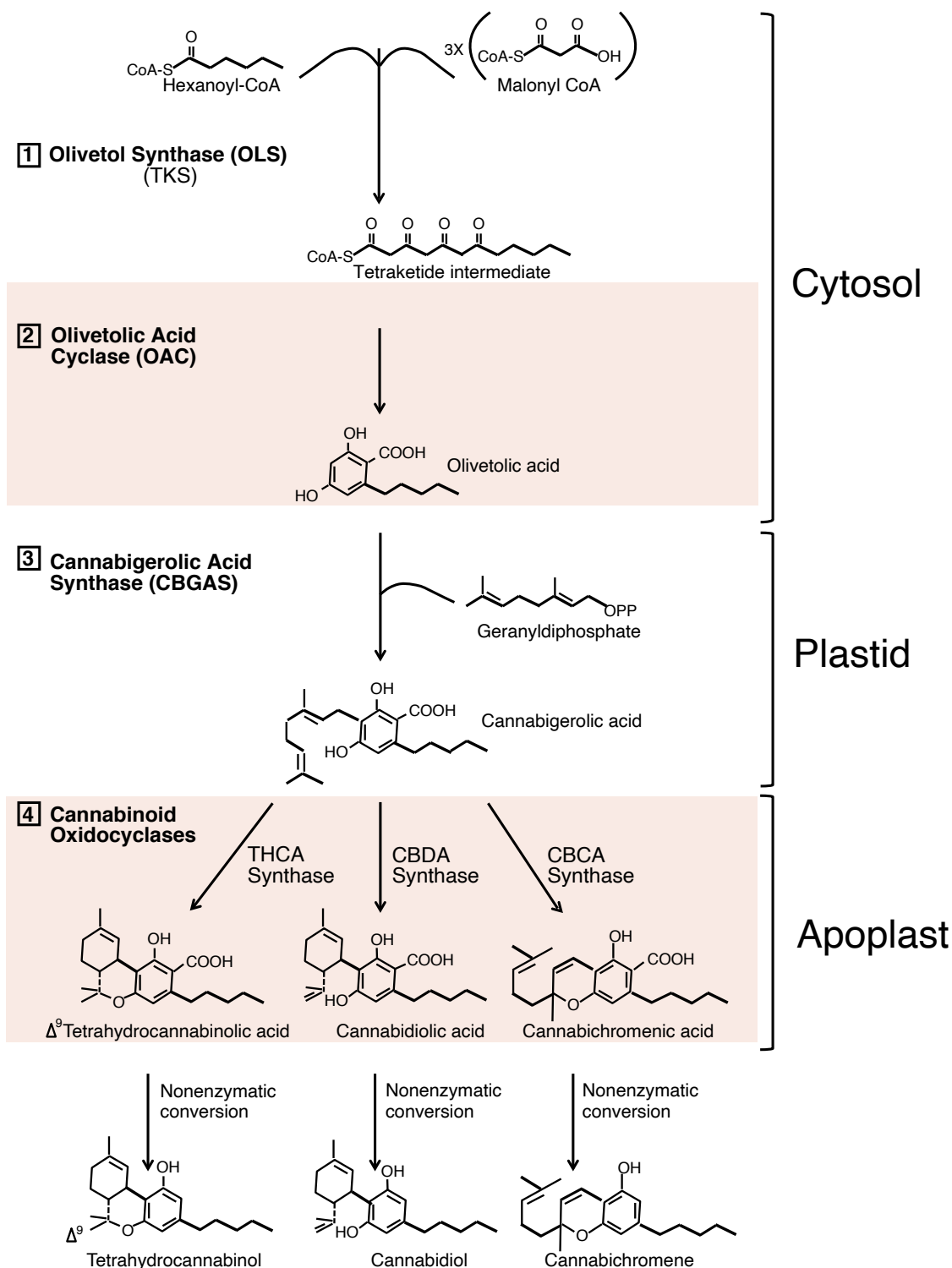


Figure 1 Biochemical Pathway. Last four steps of the biochemical pathway responsible for production of cannabinoids THCA and CBDA. Each step of the pathway has its proposed location in the cell. Figure modified from (Page and Stout 2017; Vergara et al. 2019; Gülck and Møller 2020).

Enzyme-coding genes in the biochemical pathway

Compared to the cannabinoid oxidocyclase genes, which act in the last step of the cannabinoid biosynthesis pathway, less is known about the genes acting upstream: OLS, OAC, and CBGAS. OLS (also known as Tetraketide Synthase, TKS) was cloned and characterized as a plant type III polyketide synthase that produces olivetol (Taura et al. 2009), which is then cyclized by OAC (Gagne et al. 2012) to produce olivetolic acid. OLS and OAC are located on separate chromosomes (Lavery et al. 2019).

The enzymatic activity of CBGAS was described as an aromatic prenyltransferase by which alkylation of olivetolic acid with geranylpyro-phosphate yields cannabigerolic acid (CBGA) (Fellermeier and Zenk 1998). Following an initial genetic sequence description for CBGAS (*CsPT1*; Page and Boubakir 2014) two other similar prenyltransferases (*CsPT4* and *CsPT7*) were described, together forming a cluster (Luo et al. 2019; Rea et al. 2019) that nests within a larger group of aromatic prenyltransferases (Rea et al. 2019; Gülck et al. 2020; Gülck and Møller 2020). Transcriptomic and biochemical analyses suggest that between these three different genes —*CsPT1*, *CsPT4*, and *CsPT7*— the former two are more likely to produce CBGA *in vivo* (Luo et al. 2019; Rea et al. 2019) as they are highly expressed in glandular trichomes along with OAC and THCAS. On the other hand, *CsPT7* is moderately expressed in trichomes and co-expressed with earlier cannabinoid biosynthesis genes, including OLS (Gülck et al. 2020). Additionally, *CsPT4*, but not *CsPT1*, served as a functional CBGAS in both yeast and tobacco (Luo et al. 2019; Gülck et al. 2020). Whether or precisely how *CsPT1* and *CsPT4* differ in function remains difficult to decipher, as they both have plastid localization signal (Luo et al. 2019; Gülck et al. 2020), and are in close genomic proximity on the X chromosome (Lavery et al. 2019). These findings lead to hypotheses of their possible redundancy (Gülck et al. 2020), yet to be experimentally determined. Differences in expression of *CsPT1* have been found between hemp and marijuana (van Bakel et al. 2011), raising questions about how this gene is regulated. Due to the multiexonic structure of the CBGAS genes, alternative splicing may play a role (Lavery et al. 2019).

The diversity, evolutionary origin, function, and genomic context of the Cannabinoid oxidocyclase genes, the last step of the pathway, have been characterized previously (Laverty et al. 2019; van Velzen and Schranz 2020; Grassa et al. 2021). These genes are highly similar at the structural and genetic sequence levels and the enzymes they code for may be classified as “promiscuous” as they act on the same substrate (Onofri et al. 2015; Vergara et al. 2019; Vergara et al. 2020). In their function as synthases, they may also exhibit “sloppiness” (Auldridge et al. 2006; Franco 2011; Chakraborty et al. 2013; Vergara et al. 2020): *in vitro*, they can produce up to eight different products, including each other’s compounds (Zirpel et al. 2018). The genes encoding these synthases are likely derived from the same ancestral gene via duplication events (Onofri et al. 2015; Padgitt-Cobb et al. 2019; van Velzen and Schranz 2020) and are located in close proximity in the genome (Laverty et al. 2019; van Velzen and Schranz 2020; Grassa et al. 2021).

Here we review the genomic locations of the genes coding for the enzymes Olivetol Synthase (OLS), Olivetolic Acid Cyclase (OAC), Cannabigerolic Acid Synthase (CBGAS), and the cannabinoid oxidocyclases (THCAS, CBDAS, CBCAS) and use phylogenetic-based analyses to describe their sequence variation among nine different reference genome assemblies of diverse *C. sativa* lineages. Of these assemblies, four have been assembled to the chromosome level while the remaining five vary in the degree of completion (Gao et al. 2020; Grassa et al. 2021). These upstream genes exhibit duplications and potential copy number variation among assemblies, like the cannabinoid oxidocyclases (Laverty et al. 2019; Vergara et al. 2019; van Velzen and Schranz 2020). Additionally, homologous sequences/copies are in close proximity while the different enzymes are spread throughout the genome on multiple chromosomes (Laverty et al. 2019; Grassa et al. 2021).

We also confirm that these upstream genes have multiple exons, two in the case of OLS and OAC, but up to ten for CBGAS. It is plausible that the multiple exons of the CBGAS-like genes are alternatively spliced, which could influence the genes’ expression or function. Therefore, we examined alternative splicing and found small but significant differences between varieties. Lastly, the location of these CBGAS genes on the X

chromosome may have potential implications for cannabinoid production in female plants, which we discuss.

Methods

Assemblies

We examined nine different *C. sativa* genome assemblies. These include three hemp types, USO31 (Woods et al. 2021), Carmagnola (Woods et al. 2021), and the previously published Finola assembly (BioProject: PRJNA73819; Lavery et al. 2019). We include the hybrid CBDRx assembly ‘cs10’ (PRJEB29284; Grassa et al. 2021); four marijuana-type assemblies: PBBK (Pineapple Banana Bubba Kush NCBI BioProject PRJNA378470; Vergara et al. 2019), Purple Kush (Purple Kush NCBI BioProject: PRJNA73819) improved from (van Bakel et al. 2011), Jamaican Lion (NCBI BioProject: PRJNA486541) and Cannatonic (NCBI BioProject: PRJNA350523); and the purportedly wild cannabis assembly from Jilong, Tibet (Gao et al. 2020).

Mining Cannabinoid-related gene sequences

We used BLASTN (Altschul et al. 1990; Gish and States 1993) to find the cannabinoid pathway genes in the aforementioned genome assemblies. For Olivetolic Acid Synthase and Olivetolic Acid Cyclase, steps 1 and 2 on Figure 1, we used the NCBI GenBank cDNA sequences AB164375.1 (Taura et al. 2009), and JN679224.1 (Gagne et al. 2012), respectively.

For Cannabigerolic Acid Synthase which is step 3 on Figure 1, we used previously reported *C. sativa* DNA sequences with putative CBGAS function: CsPT1, CsPT4, and CsPT7 (BK010678, BK010648, BK010683), respectively (Luo et al. 2019). CsPT1 (*C. sativa* prenyl-transferase 1) corresponds to the sequence first reported to be related to CBGAS activity (Page and Boubakir 2014). Initial alignments of these CBGAS-like genes against *C. sativa* assemblies revealed all three have a multi-exonic structure with 10 exons and multiple large introns. To accurately annotate exon boundaries of these genes across assemblies, we searched the cs10 assembly annotation, and found that all three genes were present.

We used the cs10 exons of each gene as BLASTN queries against each of the *C. sativa* genome assemblies. We used the subsequent BLASTN results to extract nucleotide sequences of the exons with BEDTOOLS (Quinlan and Hall 2010) , and lastly we manually verified canonical splice site sequences (GT donor and AG acceptor).

For the BLAST queries for the oxidocyclase genes in the last step of the pathway we used the THCAS accessions KP970852.1 (Onofri et al. 2015) and JQ437488.1, the CBDAS accession AB292682.1, and the cDNA for CBCAS (Page and Stout 2017). We used a similar methodology as in (Vergara et al. 2019) where we only recovered sequences with an alignment length of more than 1100 bp.

ML Trees

We constructed maximum likelihood (ML) gene trees using the program MEGA v.11 (Tamura et al. 2021) for the coding or partial coding regions of OLS, OAC, and CBGAS from each assembly. For all topologies node support was estimated using 500 bootstrap samples and branches with less than 75% support were collapsed. We note that some exons include internal stop codons suggesting they are pseudogenes.

As outgroups, we used BLAST to find similar genes in *Humulus sp.* for both OLS and OAC, which were found for the former but not for the latter gene. For CBGAS, we used two aromatic prenyltransferase genes from *Humulus* (HIPT1-I and HIPT2) as an outgroup, previously shown to cluster with CsPT1, CsPT4, and CsPT7 amidst a broader prenyltransferase gene tree (CITE Gulck et al 2020). Finally, for the cannabinoid oxidocyclase genes we included two cannabinoid-like genes from *Humulus*, 14 sequences from three species from the order Rosales, two of them also from the family Cannabaceae – *Trema orientale* and *Parasponia andersonii* with four and three sequences respectively – and a more distantly related species from the family Moraceae as an outgroup, *Morus notabilis*, with seven sequences.

Visualization of gene locations and structure, and analysis of alternative splicing

We used the program Chromomap (Anand and Lopez 2020) to map each gene from the cs10 assembly to its respective location. Since CBGA has multiple exons, we used the R package ‘Gviz’ (Hahne and Ivanek 2016) to illustrate the location and size of each exon/intron.

Lastly, for the CBGAS-like genes, we investigated patterns of alternative splicing inferred through differential exon usage within two *C. sativa* varieties, to understand more about their complex multi-exon structure (Supplementary materials).

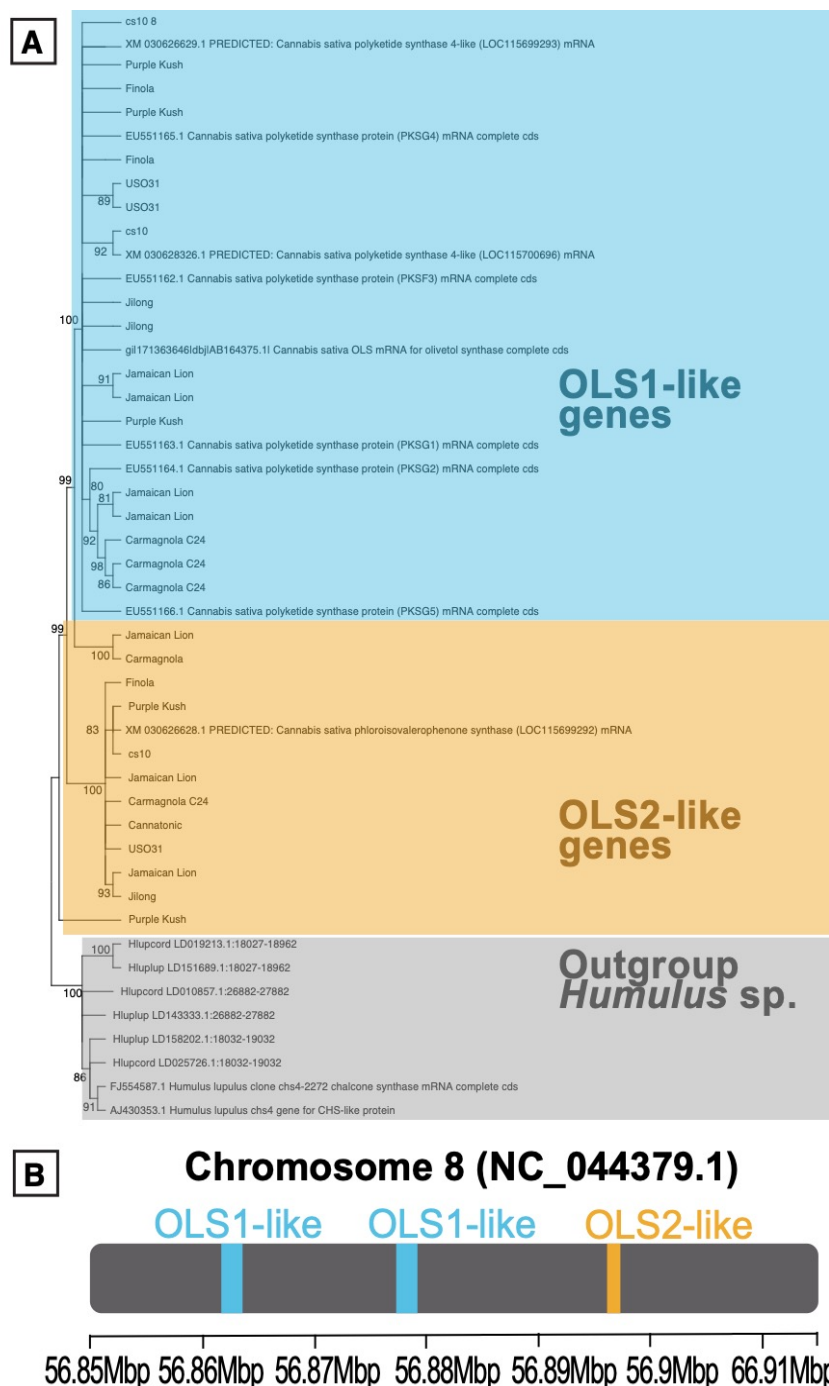
Results

Step 1: Olivetol Synthase

We found multiple copies of OLS-related genes, which include a complete OLS gene, denominated OLS1, as well as a partial OLS gene, which we call OLS2. The nine assemblies differ in copy number of OLS1 (Table S1). Jamaican Lion had four copies, while six other assemblies had two copies, and Cannatonic and PBBK had none. OLS1 is 1,158 bp, with two exons. The first exon is 159bp, the second exon is 999bp, and the intron varies between 103-245bp. OLS2 is in close proximity to OLS1 with 78-80% BLAST percent identity to the second exon of OLS1 and also of approximately 1000 bp. OLS2 is truncated in most assemblies and also varies in copies, with one, two, or up to three copies in Jamaican Lion (table S1).

Although the OLS1 genes from all assemblies cluster together in similarity (Figure 2A), there are two OLS2 genes that cluster with the OLS1 genes instead of the other OLS2 copies. All OLS1 and OLS2 genes are found in chromosome 8 in the cs10 assembly and cs10 has two copies of OLS1-like genes and one copy of OLS2-like genes (Figure 2B). These three genes in cs10 have been annotated, the two copies of OLS1 (LOC115699293 and LOC115700696) and OLS2 (LOC115699292) which was described as a phloroisovalerophenone synthase. The assemblies Finola, Jilong, and Carmagnola have truncated OLS1 genes, and indels causing the premature stop are found within the second

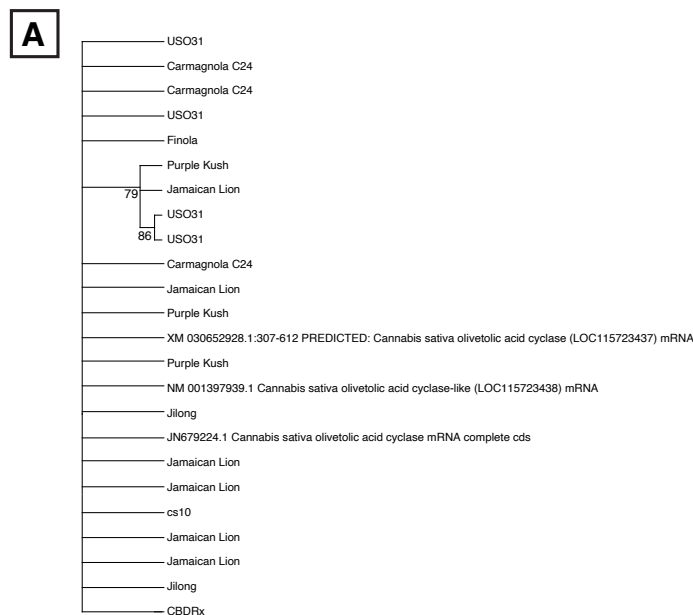
243 exon but all seem to have complete OLS2 genes. We found similar genes in *H. lupulus* that
244 cluster as an outgroup suggesting OLS may be specific to *C. sativa*.



245
246 **Figure 2. Olivetolic Acid Synthase. A)** OLS gene tree obtained from seven genome assemblies using
247 Maximum Likelihood approach genome assemblies. OLS was missing from two of the original nine
248 assemblies. Numbers in the branches indicate the percent of 500 bootstraps supporting the topology **B)**
249 Location of the two OLS1 gene copies and the unique OLS2 gene copy in the cs10 assembly.

Step 2: Olivetolic Acid Cyclase

Like Olivetol Synthase, the nine assemblies differ in number of gene copies of OAC (Table S2), with six copies in Jamaican Lion, four in USO31, three in Carmagnola and PK, two in cs10 and Jilong, one in Finola, and none in Cannatonic and PBBK. OAC also has two exons, the first one of 126bp and the second one of 180bp. The intron is small, approximately 100bp although it varies in the different assemblies (Table S2). USO31 and Purple Kush have truncated genes with indels in the second exon not close to the boundary with the intron. The gene is very similar between varieties/assemblies, and the ML tree reveals substantial polytomy (Figure 3A) that precludes clear relationships between genes from different assemblies. The two OAC copies in the cs10 assembly are found in close proximity in chromosome 9 and have been annotated (LOC115723437 and LOC115723438).



B Chromosome 9 (NC_044376.1)

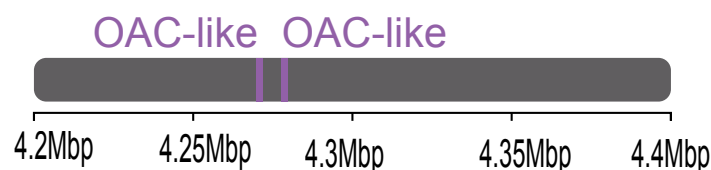


Figure 3. Olivetolic Acid Cyclase. A) Relationship via Maximum Likelihood approach between the OAC genes from seven, as it was absent from two of the genome assemblies. Numbers in the branches indicate the percent of 500 bootstraps supporting the topology **B)** Location of the two OAC gene copies in the cs10 assembly.

Step 3: Cannabigerolic acid synthase

Our initial search for CsPT1, CsPT4, and CsPT7 in the cs10 reference genome annotations revealed near identical matches for all three genes, each with 10 coding exons and annotated as “2-acylphloroglucinol 4-prenyltransferase.” CsPT1 and CsPT7 both had a single top hit (LOC115713215 and LOC115713205, respectively), while CsPT4 closely matched two separately annotated genes: LOC115713171 and LOC115713185. The latter of these two CsPT4 homologs comprised a complete alignment, while the former had a premature stop codon near the end of the seventh exon (our BLAST search identified all 10 exon regions, but only 7 were annotated as ‘CDS’ in the cs10 reference genome annotations). CsPT4 (LOC115713171 but not LOC115713185) and CsPT7 also had multiple isoforms annotated, indicating alternative splicing may play a role in the regulation of these genes. Notably, we found that all CBGAS-like genes are located within the same 193Kb window on the X chromosome (Figure 4).

Subsequent BLAST searches of the CBGAS genes revealed one copy of both CsPT1 and CsPT7 in each of the nine *C. sativa* genome assemblies. Of these, Cannatonic, USO31, and Finola all had internal stop codons in CsPT1; these three assemblies plus PBBK and Carmagnola also had internal stops in CsPT7.

For CsPT4, we found 12 closely homologous sequences across the nine assemblies, indicating potential copy number variation. In cs10, in addition to the two annotated copies of CsPT4, we found a third, previously unannotated copy of this gene. To distinguish these three homologs, we numbered them CsPT4.1-4.3 according to their genomic position in the cs10 assembly. CsPT4.3 corresponds to LOC115713185 in the cs10 reference genome and is a complete ‘canonical’ copy. CsPT4.1 corresponds to LOC115713171, with the premature stop near the end of exon 7, caused by a TGG->TAG point mutation at position 1019. CsPT4.2, the unannotated copy from cs10, was missing

likely because it has indel mutations prior to the same TGG->TAG point mutation in exon 7, leading to a more truncated gene. Another distinguishing feature between the three CsPT4 copies is that CsPT4.1 and CsPT4.2 have tandem replications of exon 6, with three and two copies, respectively, although the replications had partial alignments with a total length of 79 out of the total 93 bp of exon 6. Because CsPT4.1 is annotated as a protein coding gene in the cs10 assembly without the additional copy of exon 6, we decided to label the replications of exon 6 as non-coding intronic sequences and did not include them in sequences used for maximum likelihood gene tree analysis.

Among the other assemblies, Cannatonic had CsPT4.1 and CsPT4.2, both with the internal stop codon TGG → TAG at end of exon 7, and Jamaican Lion had three copies (i.e. CsPT4.1, CsPT4.2 and CsPT4.3) only one of which, CsPT4.2, had this same internal stop codon. Purple Kush, Carmagnola, Jilong, and Finola all had a single copy of CsPT4; of these, all were complete except for Finola, which had a 1bp deletion in exon 3 causing an internal stop. Our BLAST search against the USO31 assembly returned hits only for the first two exons of CsPT4. Likewise, PBBK had poor hits for various CsPT4 exons but lacked a complete copy (Table S3).

Phylogenetic-based analysis of the CBGAS gene sequences from all assemblies revealed three well supported clades, corresponding to CsPT1, CsPT4, and CsPT7 (Figure 4). CsPT1 and CsPT7 are more similar to each other than either one is to CsPT4. Within the CsPT4 clade, there is one well-supported split that separates complete/functional from truncated and/or non-functional copies, except for CsPT4_Finola (which has a 1bp deletion in exon 3 causing an internal stop), grouping with the complete/functional CsPT4 copies from other assemblies.

The CBGAS-like genes (CsPT1, CsPT4, CsPT7) contain multiple large introns (Table S3). In cs10, intron 2 of CsPT1 is 3,553bp long. CsPT4 has an even longer second intron compared to CsPT1: in cs10 it is 5,386bp long (CsPT4.3), and even larger in Purple Kush, at 5,946 bp. Intron 5 of CsPT4 is even longer than its intron 2, ranging from 6,344 bp in cs10 to 8541 in Purple Kush, for the ‘canonical’ CsPT4 copies (i.e. those without premature stops codons).

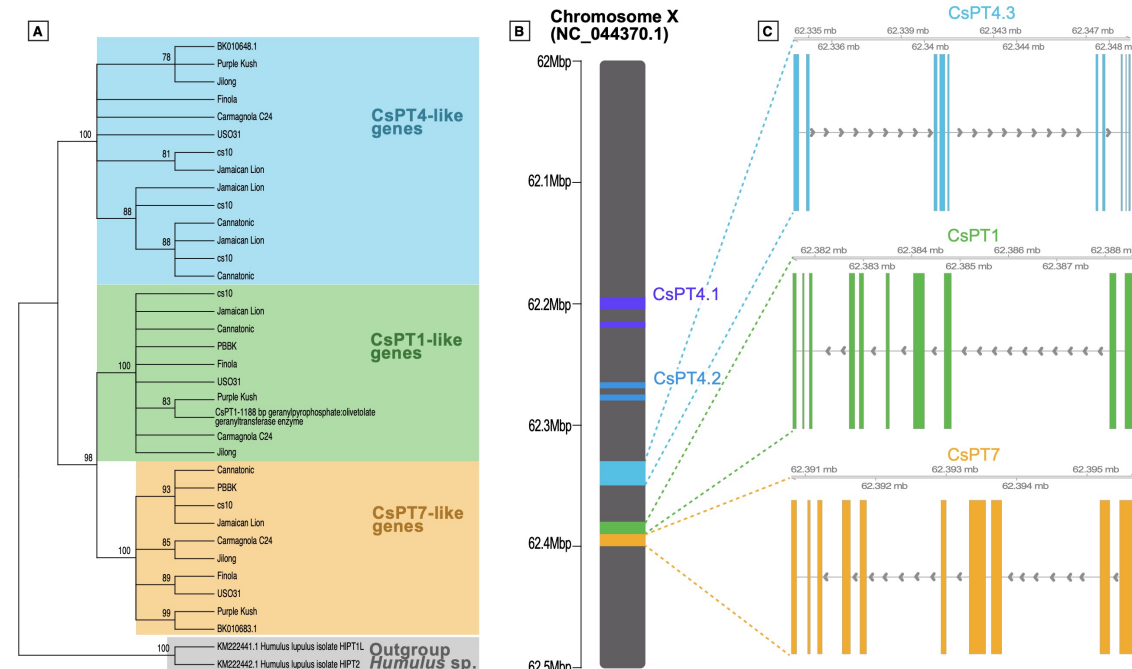


Figure 4. Cannabigerolic Acid Synthase. A) Relationship via Maximum Likelihood approach between the CsPT genes from the nine assemblies showing the clear split between CsPT4 from CsPT1 and 7. Numbers in the branches indicate the percent of 500 bootstraps supporting the topology. **B)** Location of the five CsPT gene copies in the cs10 assembly. **C)** Gene structure of three complete copies of CsPT4.3, CsPT1, and CsPT7 showing exons in blue, green, and yellow, respectively. Introns are the uncolored gaps. Arrows indicate the transcription direction.

Step 4: Cannabinoid oxidocyclase genes

We found 143 genes in the nine different assemblies related to the cannabinoid oxidocyclase genes. These same 143 genes were recovered regardless of whether THCAS, CBDAS, or CBCAS was used as the BLAST query. Jamaican Lion was the assembly that contained the greatest number of oxidocyclase genes (32) and Jilong the fewest (8).

Our ML gene tree reconstruction revealed relationships both between different oxidocyclase gene copies and homologs in different assemblies (Figure 5A). We found 44 genes related to CBCAS, seven to THCAS, and 17 to CBDAS. We also found five clusters of genes with unclear function, labeled in Figure 5A as 'Unknown' groups 1, 2, and 3, 'Additional oxidocyclase-like genes', and 'Additional uncharacterized genes'. Unknown groups 1 and 2 contains 13 genes, six and seven genes, respectively. Unknown group 3 has 45 genes from all nine assemblies which is 32% of all genes. In other words, each genome assembly contains genes from this unknown 3 cluster. The six genes labeled 'Additional oxidocyclase-like genes' form a polytomy alongside the CBDAS and the 'Unknown 3' clusters, making it unclear which of these sets of genes is sister to the clade comprising THCAS, CBCAS, 'Unknown 1', and 'Unknown 2'.

The cannabinoid oxidocyclase genes CBCAS and THCAS are very similar and are sister to each other in our gene tree. Every assembly except for USO31 has both complete and truncated CBCAS-like genes. However, only marijuana-type individuals contain THCAS-like genes. CBDAS forms a group with two clusters, one of them with only truncated genes (Figure 5A).

Considered by genome assembly, the Carmagnola assembly lacks THCAS and CBDAS genes but has one complete CBCAS gene, which is the only complete cannabinoid oxidocyclase gene we found from this assembly. USO31 lacks THCAS and CBCAS genes but has complete CBDAS. The last hemp-type genome assembly, Finola, has a complete CBDAS and CBCAS genes, but no THCAS-like genes.

The hybrid individual cs10 does not have THCAS genes but does have complete CBDAS and CBCAS genes.

371 All marijuana-type genome assemblies contain one THCAS-like gene which is
372 complete for all individuals except Cannatonic. Interestingly, Cannatonic is usually a 1:1
373 CBD:THC or high CBD variety. The only marijuana-type genome that has a
374 complete/functional CBDAS-like gene is Jamaican Lion.

375 The 'Additional uncharacterized genes' from the different assemblies have a
376 BLAST identity of less than 80% to the oxidocyclase genes THCA, CBDA, and CBCA (Table
377 S4). These cluster into two groups of eight genes for 16 genes total, are found outside the
378 oxidocyclase clade, and have the same relationship to the genes from *Humulus* sp as to
379 the oxidocyclase genes in the ML tree (Figure 5A). All assemblies except for USO31 had
380 one of these genes, some of them complete (Table S4).

381 All cannabinoid oxidocyclase genes are found in chromosome 7 in the cs10
382 assembly (Figure 5B), but those 'Additional uncharacterized genes' are found in
383 chromosome 2.

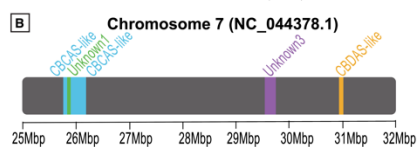
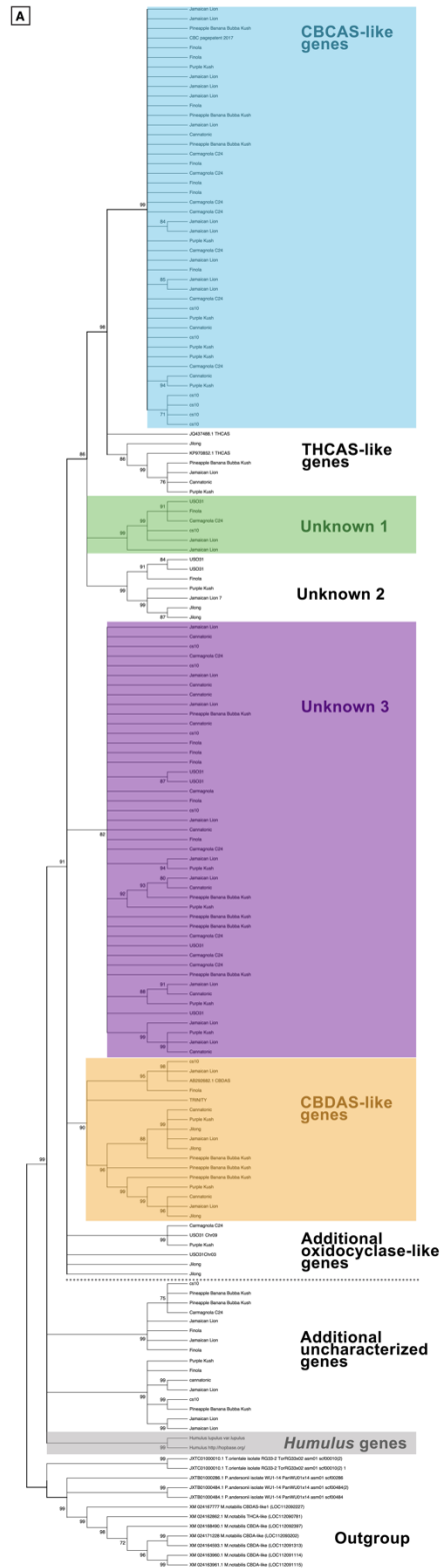


Figure 5. Cannabinoid oxidocyclase genes. A) Gene tree of cannabinoid oxidocyclases and related sequences via Maximum Likelihood approach from the nine assemblies, showing the clear split between CBCAS in blue and THCAS uncolored, three unknown clusters in green, uncolored, and purple, respectively, CBDAS in yellow and two groups with ‘additional oxidocyclase-like genes’ and ‘Additional uncharacterized genes, both uncolored. The *Humulus* sp. genes are colored in gray. Numbers in the branches indicate the percent of 500 bootstraps supporting the topology. The dashed line separates the primary cannabinoid oxidocyclase clade from more dissimilar sequences **B)** Location of the 12 oxidocyclase genes in the cs10 assembly (tandem repeats are collapsed into a single color block).

Discussion

These cannabinoid biosynthesis genes described in this study have both societal and legislative implications. The compounds they produce are used for setting legal definitions for different categories of cannabis products, as well as for ‘chemotaxonomic’ purposes to classify different *C. sativa* varieties (Henry et al. 2018; Reimann-Philipp et al. 2019; Smith et al. 2021). For example, the legal definition of hemp in the United States is any *C. sativa* plant containing 0.3% THC or less (Small and Marcus 2003).

We describe these cannabinoid biosynthesis genes and provide new insights about their sequence and potential copy number variation, and genomic location across diverse *C. sativa* lineages and genome assemblies. Our results suggest that independent of the lineage, whether a hemp-type or a marijuana-type *C. sativa*, these genes related to the cannabinoid pathway can be found in multiple copies and are scattered throughout the genome.

Steps 1 and 2: OLS and OAC

Our results strengthen reports of duplicate copies of OLS in Purple Kush and Finola (van Bakel et al. 2011; Laverty et al. 2019) and of OAC in Tachlita Till, a marijuana-type variety (McGarvey et al. 2020). We find tandem duplications of both these enzyme genes in the genome assemblies of Carmagnola, Finola, cs10, Purple Kush, Jamaican Lion, and Jilong. The fact that both duplications are present in the Jilong assembly, which is a putative wild genotype, suggests the duplication events could have arisen before strong artificial selection pressures from humans. Future research should investigate how prevalent the

OLS and OAC duplications are, and whether the gene copies are functionally redundant, or have any impact on gene dosage. Additionally, for OLS, future research could determine whether the OLS2-like gene, similar to the second exon of OLS1 is functional and what products it yields.

Step 3: CBGAS

We find evidence for potential copy number variation in CBGAS: in the cs10 assembly, arguably the most complete and well-assembled of the nine, the CsPT4 gene is repeated three times in close proximity, although just two of the copies, CsPT4.3 (complete, 10 exons) and CsPT4.1 (truncated, seven exons), are annotated as protein coding genes. Two other assemblies, Jamaican Lion and Cannatonic, also have additional copies of CsPT4. The replications of CsPT4 in cs10 are intriguing and deserve further investigation, for instance, do they exist in other genotypes not included here?

Based on the close homology and proximity of CsPT1, CsPT4, and CsPT7, there has apparently been an extended history of gene duplication at this locus, though the functional consequences, if any, remain unclear. Finola (hemp) and Purple Kush (marijuana-type) do not differ in number of CBGAS (neither CsPT1 or CsPT4) gene copies, yet findings report more than five-fold greater expression of CsPT1 (Aromatic Prenyltransferase, referred by van Bakel as the 'AP' locus) in Purple Kush compared to Finola (van Bakel et al. 2011). It was later suggested that the presence of repeat sequences and indels within introns of CsPT1 might regulate its transcription or alternative splicing (Lavery et al. 2019). We indeed find evidence for differential splicing of CsPT1 and multiple other CBGAS-related genes (Figure S1) between a high THC variety (White Cookies) and a high CBD variety (Canna Tsu), but with a small magnitude of difference as well as sample size, and therefore should be confirmed by additional studies. However, in support of our findings, a recent study reported alternative splicing events at the CBGAS locus, as well as others in the cannabinoid biosynthesis pathway, though tests for differential splicing between varieties were not performed (Wu et al. 2021).

Previous studies have reported varying lengths of intron 2 in CsPT1 (McGarvey et al. 2020), in agreement with our results. We found that CsPT4 has an even longer second and fifth introns compared to those previously reported for CsPT1. The implications of these very large intronic regions, although unclear, should motivate future studies.

The presence of CBGAS genes (both CsPT1 and CsPT4) on the X Chromosome (Chromosome 10) is also something that has not been investigated closely and has potential implications of higher dosage in females who contain two X chromosomes. Females produce higher concentration of cannabinoids due to the abundance of trichomes on their flowers (Sirikantaramas et al. 2005; Gagne et al. 2012). Perhaps that these genes are found on the X chromosome could be another of the underlying reasons of the sex difference in cannabinoid production.

Step 4: Cannabinoid Oxidocyclase genes

Our results support previous findings that the cannabinoid oxidocyclase genes are numerous, in close proximity in the genome (Lavery et al. 2019; Grassa et al. 2021), and vary in copies (van Velzen and Schranz 2020; Vergara et al. 2021) between assemblies (Table S4). Through this study we found that THCA is only found in marijuana-types but not in hemp. This raises the question or whether or not these are derived genes only found in some of the *C. sativa* lineages. Although the idea of some of the cannabinoid oxidocyclase genes being ancestral has been proposed before (Onofri et al. 2015), we cannot conclusively support or disprove it given that no other species is known to share this biosynthetic pathway (Vergara et al. 2019). However, the recent sequencing of the putative wild specimen from Jilong, Tibet, with very few cannabinoid oxidocyclase genes, suggests that increased gene duplication events may be associated with domestication (Gao et al 2020).

We also found that the assembly of Cannatonic, which is a variety sold as a 1:1 CBD:THC, has truncated sequences for both the canonical THCA and two CBDA synthase genes, making it unclear which of the other 13 genes (Cannatonic has 15 cannabinoid

oxidocyclase genes total) are responsible for the production of CBDA and THCA. Therefore, although we only have one genome assembly for Cannatonic, our results suggest that despite the truncated genes, these compounds may still be produced. If this is the case, and other oxidocyclase genes are producing the THCA and or CBDA compounds, it will make it hard for farmers to produce compliant hemp below the 0.3% THC regulated limit. This result supports the idea that the enzymes codified by these oxidocyclase genes are both promiscuous (Onofri et al. 2015; Vergara et al. 2020) and sloppy (Auldrige et al. 2006; Franco 2011; Chakraborty et al. 2013; Zirpel et al. 2018; Vergara et al. 2020).

The large number of CBCAS-like genes is also an intriguing finding given the low production quantities of this cannabinoid, which is the reason why it is considered a ‘minor’ cannabinoid. However, despite low productions of CBCA, expression levels of this gene varies between *C. sativa* varieties (Allen et al. 2022). As previously suggested, perhaps CBCAS is the sloppiest of all synthases in the pathway (Vergara et al. 2020). Similarly, the genes whose products remain to be established are interesting due to their large number. These two results regarding the numerous CBCAS-like genes despite low production of CBCA, and those with still unknown products support the notion that all these genes are acting in unison as a network to produce the complex cannabinoid phenotype (Zager et al. 2019). Our results also suggest that the genes whose products remain unknown are either producing one of the three known cannabinoids THCA, CBDA, and or CBCA, or that they have not been directly associated to the cannabinoid compound—from one of the numerous cannabinoids found in the plant—they are producing, or that there may still be cannabinoid compounds yet to be discovered.

With different phylogenetics-based approaches, the relationship between the oxidocyclase genes remain the same as previously determined (van Velzen and Schranz 2020): CBCAS and THCAS are sister clusters to each other, various clusters of genes with unknown function exist, and some genes that fall outside the oxidocyclase-like clade that could be closer to *Humulus* sp. genes (van Velzen and Schranz 2020).

Entire pathway

All the enzymes from different steps of the pathway are found in different chromosomes, far apart from each other. This suggests that through recombination, individuals may frequently inherit novel combinations of cannabinoid biosynthesis-related genes and alleles that will work together to produce the cannabinoid phenotype. There may be particular alleles of each gene, or combinations of genes in the entire pathway that, when together, could enhance or diminish the production of these compounds. Because these genes are unlinked, this could facilitate artificial selection in eventually combining the best performing genes and alleles throughout the entire pathway to enhance cannabinoid production.

Through our phylogenetic-based gene tree analyses, we found that some of these genes harbor more sequence variation across different *C. sativa* varieties than others. OAC has the least amount of sequence variation, producing a large polytomy in the gene tree, while the CBGAS-like genes and the cannabinoid oxidocyclases show more distinct clustering and branching, suggestive of sequence divergence.

Given that the cannabinoid compounds discussed here are unique to *C. sativa*, our findings, in agreement with previous studies (van Velzen and Schranz 2020), suggest that the enzymes producing these compounds acquired a new function after the speciation event with *Humulus* species approximately 25 mya (Bell et al. 2010). Because all of the genes in the pathway (OLS, OAC, CBGAS, and the oxidocyclases) appear to have a significant history of duplication, each of these genes could be going through their own process of either concerted evolution, in which the copies maintain similar sequence and function, neofunctionalization, where one of the copies acquires a novel function, or subfunctionalization, where the original function of the gene becomes split among the copies (Lynch 2007; Zhang 2013). The apparently major role of gene duplication in the *C. sativa* cannabinoid biosynthesis pathway is consistent with the evolution of plant secondary metabolism pathways more broadly (Ober 2005; Hartmann 2007).

The lack of clustering of cannabinoid biosynthesis genes in the *C. sativa* genome is similar to the genomic organization of some other plant specialized metabolism

pathways, in which genes are spread across multiple chromosomes, for example anthocyanin and saponin production in carrot (Yildiz et al. 2013) and *Nicotiana benthamiana* (Khakimov et al. 2015), respectively, and glucosynilate biosynthesis in *Arabidopsis thaliana* (Kliebenstein and Osbourn 2012; Hofberger et al. 2013). However, a growing number of studies have reported the opposite pattern, with genomic clustering of genes that interact in the same secondary metabolism pathway (Nützmann and Osbourn 2014). Whether or not cannabinoid biosynthesis pathways in other species (e.g. *Radula marginata*, *Rhododendron dauricum*) form genomic clusters is unknown, due to lack of genomic resources (Taura et al. 2016; Hussain et al. 2018; Saeki et al. 2018; Hussain et al. 2019). The evolutionary consequences of having linked versus unlinked represent an intriguing avenue of future comparative genomics research, especially in the context of *C. sativa* crop improvement. Independent of the location of the genes in the cannabinoid pathway, our study provides insights on the future possibilities of modifying the genes upstream in the pathway to silence the entire cannabinoid production downstream, which would be useful for hemp cultivators to assure the maintenance of plants with less than 0.3% THC.

Caveats

Our study comes with limitations, one is the sample size and the other one is the genome assembly quality. Although the individuals included in this study comprise the variation found in *C. sativa* as they are part of different lineages within the species, these are only nine genome assemblies. Therefore, all the results reported here belong to a small subsample within the species.

Our survey highlights the difference in assembly quality among currently available *C. sativa* genomes. The known genome size of *C. sativa* is approximately 800Mb, and the cs10 assembly is currently the representative genome on NCBI GenBank (Grassa et al. 2021). However, among the included genome assemblies in our study, size varied from ~500Mb (PBBK and Cannatonic) to ~1Gb (Jamaicon Lion). Given these substantial size differences, it is likely that the PBBK and Cannatonic assemblies are both missing genomic

regions and that the Jamaican Lion assembly has redundancies. Indeed, we found that PBBK and Cannatonic lacked crucial genes (e.g. OLS, OAC), while Jamaican Lion often had more gene copies compared to other assemblies. Thus, the putative copy number variation among PBBK, Cannatonic, and Jamaican Lion compared to other assemblies could be attributed to sequencing or assembly issues rather than real biological copy number variation. However, for Jamaican Lion, the abundance of additional cannabinoid biosynthesis gene copies might represent important allelic variation and should not be completely disregarded. Due to this potential allelic diversity from the Jamaican Lion assembly, and the fact that we cannot distinguish between alleles or gene copies, we decided to retain all sequences from this assembly in our gene trees. All analyzed assemblies in our study provide valuable resources for *C. sativa* genomic research, but future studies should carefully consider differences in assembly completeness if making comparisons between varieties/assemblies.

Acknowledgments: The authors would like to thank the graduate and undergraduate students of the 2020 Comparative Cannabis Genomics course at the University of Colorado, Boulder, from which this study developed. We also thank Nolan Kane, Rick E. Miller, Liam Friar, and other Kane Lab members for helpful comments and discussion of earlier versions of the manuscript.

Author contributions: PAI and DV performed all analyses and wrote the manuscript.

Competing financial interests: DV is the founder and president of the non-profit Agricultural Genomics Foundation and the sole owner of the company CGRI, LLC.

References:

- Ahmed, A. F., M. Z. Islam, M. S. Mahmud, M. E. Sarker, and M. R. Islam. 2022. Hemp as a potential raw material toward a sustainable world: A review. *Heliyon*:e08753.
- Allen, K. D., A. Torres, K. De Cesare, and R. Gaudino. 2022. Evolution, Expansion and Characterization of Cannabinoid Synthase Gene Family in *Cannabis Sativa*. *bioRxiv*:2022.2011.2018.517131.
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403-410.

597 Anand, L. and C. M. R. Lopez. 2020. chromoMap: an R package for interactive
598 visualization and annotation of chromosomes. *Biorxiv*:605600.

599 Anders, S., A. Reyes, and W. Huber. 2012. Detecting differential usage of exons from
600 RNA-seq data. *Nature Precedings*:1-1.

601 Auldridge, M. E., D. R. McCarty, and H. J. Klee. 2006. Plant carotenoid cleavage
602 oxygenases and their apocarotenoid products. *Current opinion in plant biology*
603 9:315-321.

604 Bell, C. D., D. E. Soltis, and P. S. Soltis. 2010. The age and diversification of the
605 angiosperms re-revisited. *American Journal of Botany* 97:1296-1303.

606 Chakraborty, S., R. Minda, L. Salaye, A. M. Dandekar, S. K. Bhattacharjee, and B. J. Rao.
607 2013. Promiscuity-based enzyme selection for rational directed evolution
608 experiments. *Enzyme Engineering: Methods and Protocols*:205-216.

609 Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson,
610 and T. R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner.
611 *Bioinformatics* 29:15-21.

612 ElSohly, M. A. and D. Slade. 2005. Chemical constituents of marijuana: the complex
613 mixture of natural cannabinoids. *Life sciences* 78:539-548.

614 Fellermeier, M. and M. H. Zenk. 1998. Prenylation of olivetolate by a hemp transferase
615 yields cannabigerolic acid, the precursor of tetrahydrocannabinol. *FEBS letters*
616 427:283-285.

617 Franco, O. L. 2011. Peptide promiscuity: an evolutionary concept for plant defense. *FEBS*
618 *letters* 585:995-1000.

619 Gagne, S. J., J. M. Stout, E. Liu, Z. Boubakir, S. M. Clark, and J. E. Page. 2012.
620 Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique
621 catalytic route to plant polyketides. *Proceedings of the National Academy of*
622 *Sciences* 109:12811-12816.

623 Gao, S., B. Wang, S. Xie, X. Xu, J. Zhang, L. Pei, Y. Yu, W. Yang, and Y. Zhang. 2020. A high-
624 quality reference genome of wild *Cannabis sativa*. *Horticulture research* 7:1-11.

625 Gertsch, J., M. Leonti, S. Raduner, I. Racz, J.-Z. Chen, X.-Q. Xie, K.-H. Altmann, M. Karsak,
626 and A. Zimmer. 2008. Beta-caryophyllene is a dietary cannabinoid. *Proceedings*
627 *of the National Academy of Sciences* 105:9099-9104.

628 Gish, W. and D. J. States. 1993. Identification of protein coding regions by database
629 similarity search. *Nature genetics* 3:266-272.

630 Grassa, C. J., G. D. Weiblen, J. P. Wenger, C. Dabney, S. G. Poplawski, S. Timothy Motley,
631 T. P. Michael, and C. Schwartz. 2021. A new *Cannabis* genome assembly
632 associates elevated cannabidiol (CBD) with hemp introgressed into marijuana.
633 *New Phytologist* 230:1665-1679.

634 Gülck, T., J. Booth, Â. Carvalho, B. Khakimov, C. Crocoll, M. Motawia, B. Møller, J.
635 Bohlmann, and N. Gallage. 2020. Synthetic biology of Cannabinoids and
636 Cannabinoid Glucosides in *Nicotiana benthamiana* and *Saccharomyces*
637 *cerevisiae*. *Journal of Natural Products* 83:2877-2893.

638 Gülck, T. and B. L. Møller. 2020. Phytocannabinoids: origins and biosynthesis. *Trends in*
639 *plant science* 25:985-1004.

640 Hahne, F. and R. Ivanek. 2016. Visualizing genomic data using Gviz and bioconductor. Pp.
641 335-351. Statistical genomics. Springer.

642 Hart, C. L., W. Van Gorp, M. Haney, R. W. Foltin, and M. W. Fischman. 2001. Effects of
643 acute smoked marijuana on complex cognitive performance.
644 Neuropsychopharmacology 25:757-765.

645 Hartmann, T. 2007. From waste products to ecochemicals: fifty years research of plant
646 secondary metabolism. Phytochemistry 68:2831-2846.

647 Henry, P., A. Hilyard, S. Johnson, and C. Orser. 2018. Predicting chemovar cluster and
648 variety verification in vegetative cannabis accessions using targeted single
649 nucleotide polymorphisms. PeerJ Preprints 6:e27442v27441.

650 Hillig, K. W. 2005. Genetic evidence for speciation in Cannabis (Cannabaceae). Genetic
651 Resources and Crop Evolution 52:161-180.

652 Hofberger, J. A., E. Lyons, P. P. Edger, J. Chris Pires, and M. Eric Schranz. 2013. Whole
653 genome and tandem duplicate retention facilitated glucosinolate pathway
654 diversification in the mustard family. Genome biology and evolution 5:2155-
655 2173.

656 Hussain, T., R. V. Espley, J. Gertsch, T. Whare, F. Stehle, and O. Kayser. 2019.
657 Demystifying the liverwort *Radula marginata*, a critical review on its taxonomy,
658 genetics, cannabinoid phytochemistry and pharmacology. Phytochemistry
659 Reviews 18:953-965.

660 Hussain, T., B. Plunkett, M. Ejaz, R. V. Espley, and O. Kayser. 2018. Identification of
661 putative precursor genes for the biosynthesis of cannabinoid-like compound in
662 *Radula marginata*. Frontiers in plant science 9:537.

663 Khakimov, B., V. Kuzina, P. Ø. Erthmann, E. O. Fukushima, J. M. Augustin, C. E. Olsen, J.
664 Scholtalbers, H. Volpin, S. B. Andersen, and T. P. Hauser. 2015. Identification and
665 genome organization of saponin pathway genes from a wild crucifer, and their
666 use for transient production of saponins in *Nicotiana benthamiana*. The Plant
667 Journal 84:478-490.

668 Kliebenstein, D. J. and A. Osbourn. 2012. Making new molecules—evolution of pathways
669 for novel metabolites in plants. Current opinion in plant biology 15:415-423.

670 Kovalchuk, I., M. Pellino, P. Rigault, R. van Velzen, J. Ebersbach, J. R. Ashnest, M. Mau,
671 M. Schranz, J. Alcorn, and R. Laprairie. 2020. The Genomics of Cannabis and Its
672 Close Relatives. Annual Review of Plant Biology 71.

673 Laverty, K. U., J. M. Stout, M. J. Sullivan, H. Shah, N. Gill, L. Holbrook, G. Deikus, R. Sebra,
674 T. R. Hughes, and J. E. Page. 2019. A physical and genetic map of *Cannabis sativa*
675 identifies extensive rearrangements at the THC/CBD acid synthase loci. Genome
676 research 29:146-156.

677 Li, H. L. 1973. An archaeological and historical account of cannabis in China. Economic
678 Botany 28:437-448.

679 Liao, Y., G. K. Smyth, and W. Shi. 2014. featureCounts: an efficient general purpose
680 program for assigning sequence reads to genomic features. Bioinformatics
681 30:923-930.

682 Luo, X., M. A. Reiter, L. d'Espaux, J. Wong, C. M. Denby, A. Lechner, Y. Zhang, A. T.
683 Grzybowski, S. Harth, and W. Lin. 2019. Complete biosynthesis of cannabinoids
684 and their unnatural analogues in yeast. *Nature* 567:123-126.

685 Lynch, M. 2007. Origins of Genome Architecture. *Origins of Genome Architecture*.
686 Lynch, R. C., D. Vergara, S. Tittes, K. White, C. J. Schwartz, M. J. Gibbs, T. C. Ruthenburg,
687 K. deCesare, D. P. Land, and N. C. Kane. 2016. Genomic and Chemical Diversity in
688 Cannabis. *Critical Reviews in Plant Sciences* 35:349-363.

689 McGarvey, P., J. Huang, M. McCoy, J. Orvis, Y. Katsir, N. Lotringer, I. Nesher, M.
690 Kavarana, M. Sun, and R. Peet. 2020. De novo assembly and annotation of
691 transcriptomes from two cultivars of *Cannabis sativa* with different cannabinoid
692 profiles. *Gene* 762:145026.

693 McPartland, J. M., W. Hegman, and T. Long. 2019. Cannabis in Asia: its center of origin
694 and early cultivation, based on a synthesis of subfossil pollen and
695 archaeobotanical studies. *Vegetation History and Archaeobotany*:1-12.

696 Nützmann, H.-W. and A. Osbourn. 2014. Gene clustering in plant specialized
697 metabolism. *Current opinion in biotechnology* 26:91-99.

698 Ober, D. 2005. Seeing double: gene duplication and diversification in plant secondary
699 metabolism. *Trends in plant science* 10:444-449.

700 Onofri, C., E. P. M. de Meijer, and G. Mandolino. 2015. Sequence heterogeneity of
701 cannabidiolic-and tetrahydrocannabinolic acid-synthase in *Cannabis sativa* L. and
702 its relationship with chemical phenotype. *Phytochemistry*.

703 Padgitt-Cobb, L. K., S. B. Kingan, J. Wells, J. Elser, B. Kronmiller, D. Moore, G.
704 Concepcion, P. Peluso, D. Rank, P. Jaiswal, J. Henning, and D. A. Hendrix. 2019. A
705 phased, diploid assembly of the Cascade hop (*Humulus lupulus*)
706 genome reveals patterns of selection and haplotype variation. *bioRxiv*:786145.

707 Page, J. E. and Z. Boubakir. 2014. Aromatic prenyltransferase from Cannabis. Google
708 Patents.

709 Page, J. E. and J. M. Stout. 2017. Cannabichromenic acid synthase from Cannabis sativa.
710 Google Patents.

711 Peil, A., H. Flachowsky, E. Schumann, and W. E. Weber. 2003. Sex-linked AFLP markers
712 indicate a pseudoautosomal region in hemp (*Cannabis sativa* L.). *Theoretical and*
713 *applied genetics* 107:102-109.

714 Punja, Z. K. and J. E. Holmes. 2020. Hermaphroditism in marijuana (*Cannabis sativa* L.)
715 inflorescences—impact on floral morphology, seed formation, progeny sex ratios,
716 and genetic variation. *Frontiers in Plant Science* 11:718.

717 Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing
718 genomic features. *Bioinformatics* 26:841-842.

719 Rea, K. A., J. A. Casaretto, M. S. Al-Abdul-Wahid, A. Sukumaran, J. Geddes-McAlister, S. J.
720 Rothstein, and T. A. Akhtar. 2019. Biosynthesis of cannflavins A and B from
721 Cannabis sativa L. *Phytochemistry* 164:162-171.

722 Reimann-Philipp, U., M. Speck, C. Orser, S. Johnson, A. Hilyard, H. Turner, A. J. Stokes,
723 and A. L. Small-Howard. 2019. Cannabis Chemovar Nomenclature Misrepresents
724 Chemical and Genetic Diversity; Survey of Variations in Chemical Profiles and

Genetic Markers in Nevada Medical Cannabis Samples. Cannabis and Cannabinoid Research.

Ren, G., X. Zhang, Y. Li, K. Ridout, M. L. Serrano-Serrano, Y. Yang, A. Liu, G. Ravikanth, M. A. Nawaz, and A. S. Mumtaz. 2021. Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. Science Advances 7:eabg2286.

Russo, E. B. 2007. History of cannabis and its preparations in saga, science, and sobriquet. Chemistry & Biodiversity 4:1614-1648.

Russo, E. B. 2011. Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. British Journal of Pharmacology 163:1344-1364.

Russo, E. B. and J. M. McPartland. 2003. Cannabis is more than simply Δ9-tetrahydrocannabinol. Psychopharmacology 165:431-432.

Saeki, H., R. Hara, H. Takahashi, M. Iijima, R. Munakata, H. Kenmoku, K. Fuku, A. Sekihara, Y. Yasuno, and T. Shinada. 2018. An aromatic farnesyltransferase functions in biosynthesis of the anti-HIV meroterpenoid daurichromenic acid. Plant physiology 178:535-551.

Sawler, J., J. M. Stout, K. M. Gardner, D. Hudson, J. Vidmar, L. Butler, J. E. Page, and S. Myles. 2015. The Genetic Structure of Marijuana and Hemp. PloS one 10:e0133292.

Schafroth, M. A., G. Mazzocanti, I. Reynoso-Moreno, R. Erni, F. Pollastro, D. Caprioglio, B. Botta, G. Allegrone, G. Grassi, and A. Chicca. 2021. Δ9-cis-Tetrahydrocannabinol: Natural occurrence, chirality, and pharmacology. Journal of natural products 84:2502-2510.

Sirikantaramas, S., F. Taura, Y. Tanaka, Y. Ishikawa, S. Morimoto, and Y. Shoyama. 2005. Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes. Plant and Cell Physiology 46:1578-1582.

Small, E. and D. Marcus. 2003. Tetrahydrocannabinol levels in hemp (Cannabis sativa) germplasm resources. Economic botany 57:545-558.

Smith, C. J., D. Vergara, B. Keegan, and N. Jikomes. 2021. The Phytochemical Diversity of Commercial Cannabis in the United States. bioRxiv.

Swift, W., A. Wong, K. M. Li, J. C. Arnold, and I. S. McGregor. 2013. Analysis of cannabis seizures in NSW, Australia: cannabis potency and cannabinoid profile. PloS one 8:e70052.

Tamura, K., G. Stecher, and S. Kumar. 2021. MEGA11: molecular evolutionary genetics analysis version 11. Molecular biology and evolution 38:3022-3027.

Taura, F., M. Iijima, E. Yamanaka, H. Takahashi, H. Kenmoku, H. Saeki, S. Morimoto, Y. Asakawa, F. Kurosaki, and H. Morita. 2016. A novel class of plant type III polyketide synthase involved in orsellinic acid biosynthesis from Rhododendron dauricum. Frontiers in plant science 7:1452.

Taura, F., S. Tanaka, C. Taguchi, T. Fukamizu, H. Tanaka, Y. Shoyama, and S. Morimoto. 2009. Characterization of olivetol synthase, a polyketide synthase putatively involved in cannabinoid biosynthetic pathway. FEBS letters 583:2061-2066.

768 Valliere, M. A., T. P. Korman, N. B. Woodall, G. A. Khitrov, R. E. Taylor, D. Baker, and J. U.
769 Bowie. 2019. A cell-free platform for the prenylation of natural products and
770 application to cannabinoid production. *Nature communications* 10:565.
771 van Bakel, H., J. M. Stout, A. G. Cote, C. M. Tallon, A. G. Sharpe, T. R. Hughes, and J. E.
772 Page. 2011. The draft genome and transcriptome of *Cannabis sativa*. *Genome*
773 *Biology* 12.
774 van Velzen, R. and M. E. Schranz. 2020. Origin and evolution of the cannabinoid
775 oxidocyclase gene family. *bioRxiv*.
776 Vergara, D., H. Baker, K. Clancy, K. G. Keepers, J. P. Mendieta, C. S. Pauli, S. B. Tittes, K.
777 H. White, and N. C. Kane. 2016. Genetic and Genomic Tools for *Cannabis sativa*.
778 *Critical Reviews in Plant Sciences* 35:364-377.
779 Vergara, D., R. Gaudino, T. Blank, and B. Keegan. 2020. Modeling cannabinoids from a
780 large-scale sample of *Cannabis sativa* chemotypes. *PloS one* 15:e0236878.
781 Vergara, D., E. L. Huscher, K. G. Keepers, R. M. Givens, C. G. Cizek, A. Torres, R. Gaudino,
782 and N. C. Kane. 2019. Gene copy number is associated with phytochemistry in
783 *Cannabis sativa*. *AoB PLANTS* 11:plz074.
784 Vergara, D., E. L. Huscher, K. G. Keepers, R. Pisupati, A. L. Schwabe, M. E. McGlaughlin,
785 and N. C. Kane. 2021. Genomic evidence that governmentally produced *Cannabis*
786 *sativa* poorly represents genetic variation available in state markets. *Frontiers in*
787 *plant science*:1502.
788 Volkow, N. D., R. D. Baler, W. M. Compton, and S. R. B. Weiss. 2014. Adverse Health
789 Effects of Marijuana Use. *New England Journal of Medicine* 370:2219-2227.
790 Watts, G. 2006. Science commentary: Cannabis confusions. *BMJ: British Medical Journal*
791 332:175.
792 Woods, P., B. J. Campbell, T. J. Nicodemus, E. B. Cahoon, J. L. Mullen, and J. K. McKay.
793 2021. Quantitative trait loci controlling agronomic and biochemical traits in
794 *Cannabis sativa*. *Genetics* 219:iyab099.
795 Wu, B., Y. Li, J. Li, Z. Xie, M. Luan, C. Gao, Y. Shi, and S. Chen. 2021. Genome-Wide
796 Analysis of Alternative Splicing and Non-Coding RNAs Reveal Complicated
797 Transcriptional Regulation in *Cannabis sativa* L. *International journal of*
798 *molecular sciences* 22:11989.
799 Yildiz, M., D. K. Willis, P. F. Cavagnaro, M. Iorizzo, K. Abak, and P. W. Simon. 2013.
800 Expression and mapping of anthocyanin biosynthesis genes in carrot. *Theoretical*
801 *and applied genetics* 126:1689-1702.
802 Zager, J. J., I. Lange, N. Srividya, A. Smith, and B. M. Lange. 2019. Gene networks
803 underlying cannabinoid and terpenoid accumulation in cannabis. *Plant*
804 *physiology* 180:1877-1897.
805 Zhang, J. 2013. Gene duplication. *Princeton Guide to Evolution*:397-405.
806 Zirpel, B., O. Kayser, and F. Stehle. 2018. Elucidation of structure-function relationship of
807 THCA and CBDA synthase from *Cannabis sativa* L. *Journal of biotechnology*
808 284:17-26.
809
810

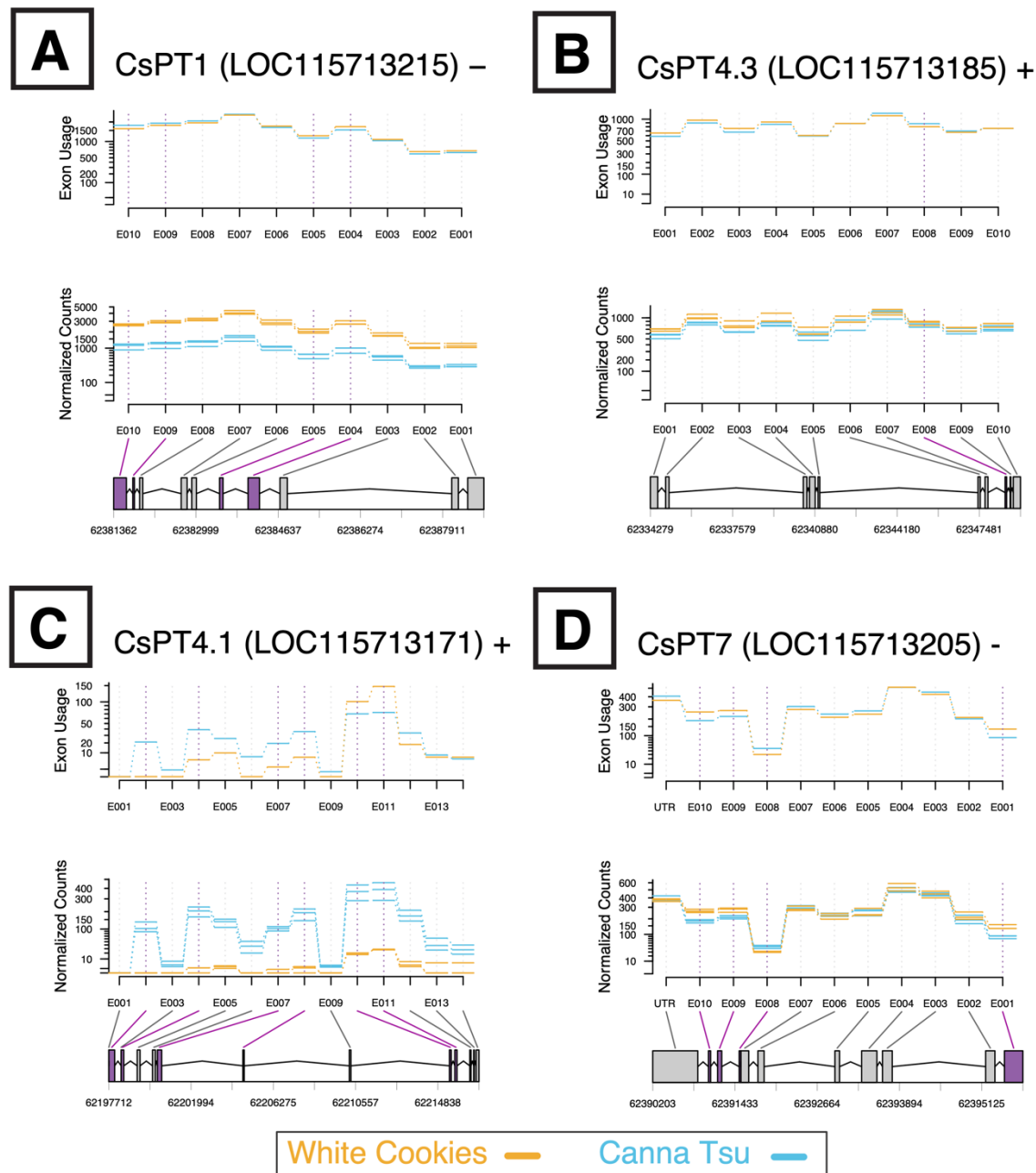
811 **Supplemental materials**—Analysis of differential exon usage in CBGAS genes between
812 two *C. sativa* varieties

813

814 **Methods**

815 CBGAS-like genes differential expression

816 We downloaded publicly available transcriptomic data from floral tissue of two *C. sativa*
817 varieties, White Cookies (high THC) and Canna Tsu (high CBD), which were grown in a
818 common environment (Zager et al. 2019). We aligned three RNA-seq libraries from both
819 varieties to the cs10 reference genome (Grassa et al. 2021) using STAR in two-pass mode
820 (Dobin et al. 2013). We then counted reads at the (exon) level using featureCounts (Liao
821 et al. 2014) of only uniquely mapping reads. We tested for differences in relative exon
822 usage between varieties for CsPT1, CsPT4.3, CsPT4.1, and CsPT7, using DEXSeq (Anders et
823 al. 2012). Due to the absence of CsPT4.2 in the cs10 annotation, we did not include this
824 gene in our differential exon usage analysis. The used approach controls for differences
825 in overall gene expression level between conditions and is an indicator of certain types of
826 alternative splicing, for instance exon skipping or mutually exclusive exons.



827

828 **Figure S1** Differential exon usage diagrams for different CBGAS-like genes. The high-THC variety ‘White
829 Cookies’ is shown in yellow, and the high-CBD variety ‘Canna Tsu’ is blue. Exons shaded in purple have
830 significant differences in relative expression (i.e. ‘usage’) between varieties (FDR < .05), independent of
831 any gene-level differences in expression. Normalized read counts for each sample are shown in the middle
832 section, while bottom sections show gene structure and location on chromosome X (NC_044370.1) of four
833 of the five CBDAS-like genes from the cs10 reference assembly: **A**) CsPT1; **B**) CsPT4.3; **C**) CsPT4.1; and **D**)
834 CsPT7