Origin and Evolution of the Cannabinoid Oxidocyclase Gene Family

Robin van Velzen[1,2,*] and M. Eric Schranz[1]

[1]Plant Sciences, Biosystematics Group, Wageningen University, Wageningen, The Netherlands

[2]Bedrocan International, Veendam, The Netherlands

*Corresponding author: E-mail: robin.vanvelzen@wur.nl.

Accepted: 4 June 2021

Abstract

Cannabis is an ancient crop representing a rapidly increasing legal market, especially for medicinal purposes. Medicinal and psychoactive effects of *Cannabis* rely on specific terpenophenolic ligands named cannabinoids. Recent whole-genome sequencing efforts have uncovered variation in multiple genes encoding the final steps in cannabinoid biosynthesis. However, the origin, evolution, and phylogenetic relationships of these cannabinoid oxidocyclase genes remain unclear. To elucidate these aspects, we performed comparative genomic analyses of *Cannabis*, related genera within the Cannabaceae family, and selected outgroup species. Results show that cannabinoid oxidocyclase genes originated in the *Cannabis* lineage from within a larger gene expansion in the Cannabaceae family. Localization and divergence of oxidocyclase genes in the *Cannabis* genome revealed two main syntenic blocks, each comprising tandemly repeated cannabinoid oxidocyclase genes. By comparing these blocks with those in genomes from closely related species, we propose an evolutionary model for the origin, neofunctionalization, duplication, and diversification of cannabinoid oxidocycloase genes. Based on phylogenetic analyses, we propose a comprehensive classification of three main clades and seven subclades that are intended to aid unequivocal referencing and identification of cannabinoid oxidocyclase genes. Our data suggest that cannabinoid phenotype is primarily determined by the presence/absence of single-copy genes. Although wild populations of *Cannabis* are still unknown, increased sampling of landraces and wild/feral populations across its native geographic range is likely to uncover additional cannabinoid oxidocyclase sequence variants.

Key words: biosynthesis pathway, *Cannabis sativa* L, comparative genomics, gene evolution, gene copy number variation, synteny.

Significance statement

Cannabis genome sequencing efforts have revealed extensive cannabinoid oxidocyclase gene variation. However, phylogenetic relationships and evolution of these genes remain unclear. Our comprehensive analysis of currently available data reveals that these genes comprise three main clades and seven subclades that originated through *Cannabis*-specific gene duplication and divergence. Our new conceptual and evolutionary framework serves as a reference for future description and functional analyses of cannabinoid oxidocyclases.

Introduction

The plant *Cannabis sativa* L. (henceforth *Cannabis*) is an ancient yet controversial crop. *Cannabis* cultivars are commonly divided into "Fiber-type" (or hemp) cultivars that are used for the production of fiber and/or seed oil and "drug-type" (or marijuana) cultivars that are used for recreational, ritual, or medicinal purposes (Small and Cronquist 1976; Sawler et al. 2015; Lynch et al. 2016; Vergara et al. 2016; McPartland and Guy 2017). *Cannabis* currently represents a rapidly emerging legal industry with an estimated multibillion global market,

primarily for medicinal purposes. However, many fundamental aspects about the molecular evolution of *Cannabis* remain unknown (Kovalchuk et al. 2020; Hurgobin et al. 2021). In this article, we aim to elucidate the origin and evolution of a unique class of biosynthetic genes found in the *Cannabis* genome.

Many of the medicinal properties of *Cannabis* are due to its production of cannabinoids; a unique class of psychoactive terpenophenolic ligands (Gaoni and Mechoulam 1964; Mechoulam 2005). The two most abundant and well-known cannabinoids are Δ^9-tetrahydrocannabinol (THC) and cannabidiol (CBD), but more than 120 others have been identified in *Cannabis* (ElSohly et al. 2017). It is important to note, however, that cannabinoids are synthesized and stored in the plant as acids that are not medicinally active. Only from exposure to light during storage or heat during processing for consumption (e.g., smoking or baking) these acids are nonenzymatically decarboxylated to their neutral forms that have psychoactive and/or medicinal properties. Some other plant genera such as *Rhododendron* and *Radula* have also been found to make cannabinoids (Iijima et al. 2017; Gülck and Møller 2020). THC is responsible for the psychoactive effect of *Cannabis* through its partial agonist activity at endocannabinoid receptors (Gaoni and Mechoulam 1964; Mechoulam and Parker 2013). This effect is the reason for the large-scale use of *Cannabis* as an intoxicant. But accumulating evidence from clinical trials indicates that moderate doses of THC can be used medicinally to, for example, reduce nausea and vomiting, pain, and improvement of sleep and appetite (van de Donk et al. 2019; Grimison et al. 2020; Suraev et al. 2020). CBD has a weak affinity for endocannabinoid receptors and is not psychoactive (Pertwee 2005). It has been found to modulate the effects of THC and endocannabinoids and may be effective for symptomatic treatment of anxiety and psychosis and in treating some childhood epilepsy syndromes (Gofshteyn et al. 2017; Bhattacharyya et al. 2018; Skelley et al. 2020).

Fiber-type cultivars typically have low content (<0.4%) of THC and intermediate content (2–4%) of CBD while drug-type cultivars typically have high content (14–40%) of THC and low content (<1%) of CBD (Small and Cronquist 1976). However, cultivars exist with alternative chemical profiles such as drug-type cultivars with high levels of CBD and other classifications based on chemotype have been proposed (Hazekamp et al. 2016; Wenger et al. 2020). The most commonly used chemotypes are based on the ratio between CBD and THC, where chemotype I is THC-dominant, chemotype II has similar levels of both THC and CBD, and chemotype III is CBD-dominant. Some *Cannabis* plants synthesize cannabichromene (CBC) or cannabigerol (CBG; CBG-dominant plants are considered chemotype IV) (Morimoto et al. 1997; Fournier et al. 2004). These lesser-known cannabinoids may have anti-inflammatory effects but the evidence is relatively scarce (Brierley et al. 2019; Udoh et al. 2019).

The Cannabinoid Biosynthetic Pathway

Within the *Cannabis* plant, cannabinoids are synthesized in multicellular epidermal glands (glandular trichomes) that are most abundant on the bracts of female inflorescences. The cannabinoid biosynthetic pathway has been largely elucidated, and for many steps in the pathway, the corresponding enzymes have been isolated and characterized (fig. 1). In brief, cannabinoid biosynthesis relies on two precursors from two distinct metabolic pathways: olivetolic acid from the polyketide pathway and geranyl-geranyl pyrophosphate (GPP) from the methylerythritol phosphate (MEP) pathway. Olivetolic acid is c-terminally prenylated with GPP into cannabigerolic acid (CBGA) by CBGA synthase—a transmembrane aromatic prenyltransferase with a plastid localization signal (Fellermeier and Zenk 1998; Luo et al. 2019). CBGA is then secreted into the extracellular storage cavity via an unknown mechanism and further processed by secreted cannabinoid oxidocyclases that perform different types of oxidative cyclizations of its linear prenyl moiety into derived cannabinoid acids such as tetrahydrocannabinolic acid (THCA), cannabidiolic acid (CBDA), and cannabichromenic acid (CBCA) (Taura et al. 1995, 1996; Morimoto et al. 1998; Sirikantaramas et al. 2005; Rodziewicz et al. 2019). *Cannabis* plants accumulating CBGA are assumed to have nonfunctional cannabinoid oxidocyclases (De Meijer and Hammond 2005; Onofri et al. 2015).

The three currently known cannabinoid oxidocyclase enzymes THCA synthase (THCAS), CBDA synthase (CBDAS), and CBC synthase (CBCAS) are highly similar in their biochemical properties and sequence characteristics (Taura, Sirikantaramas, Shoyama, Shoyama, et al. 2007; Laverty et al. 2019). The amino acid sequences are also highly similar, with THCAS and CBCAS being 92% identical to each other and respectively 84% and 83% identical to CBDAS (Sirikantaramas et al. 2004; Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; Laverty et al. 2019).

THCAS, CBDAS, and CBCAS are members of the berberine bridge enzyme (BBE)-like gene family (PF08031) (Sirikantaramas et al. 2004; Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007). This family is named after an oxidocyclase from *Eschscholzia californica* involved in alkaloid biosynthesis and part of the larger oxygen-dependent FAD-linked oxidoreductase family (PF02913) (Hauschild et al. 1998; Winkler et al. 2008). Like other BBE-like synthases, THCAS, CBDAS, and CBCAS contain an N-terminal signal peptide, a flavin adenine dinucleotide (FAD)-binding domain, a substrate-binding domain, and a BBE-like specific C-terminus that is part of the FAD-binding module (Sirikantaramas et al. 2004; Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007). In accordance with this domain structure, THCAS and CBDAS have been found to be catalytically active in the extracellular storage cavity of the glandular trichome and rely on covalently bound FAD and O_2 for their activity

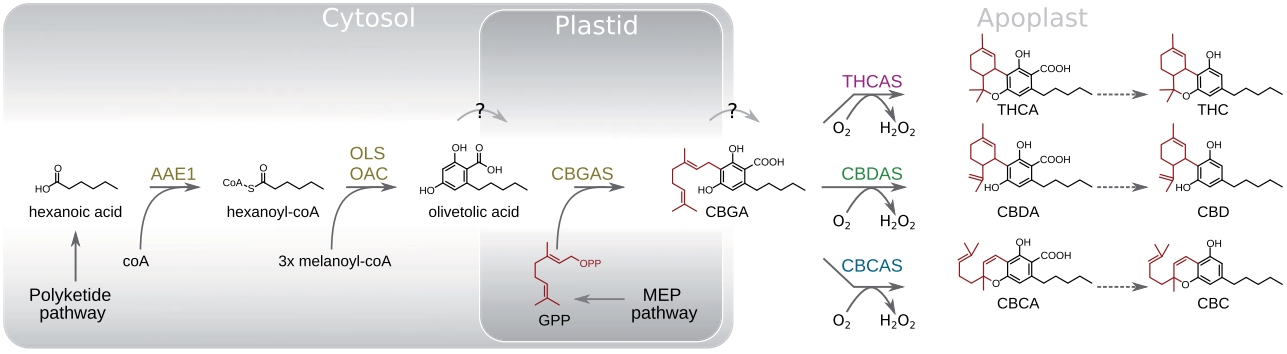

Fig. 1.—Cannabinoid biosynthesis pathway. Dotted arrows indicate nonenzymatic decarboxylations; solid arrows indicate enzymatic reactions; enzyme names are shown in blue, while resulting compounds are shown in black. Compound (sub)structures depicted in red signify those that represent a single unit of GPP. AAE1, acyl-activating enzyme 1; CBC, cannabichromene; CBCA(S),cannabichromenic acid (synthase); CBD, cannabidiol; CBDA(S), cannabidiolic acid (synthase); CBGA(S),cannabigerolic acid (synthase); GPP(S), geranyl-pyrophosphate (synthase); MEP, methylerythritol phosphate; OAC, olivetolic acid cyclase; OLS, olivetol synthase; THC, tetrahydrocannabinol; THCA(S), tetrahydrocannabinolic acid (synthase).

(Sirikantaramas et al. 2005; Rodziewicz et al. 2019). CBCAS is less extensively studied, but considering its high sequence similarity with THCAS, probably shares these biochemical activities (Morimoto et al. 1997; Taura, Sirikantaramas, Shoyama, Shoyama, et al. 2007; Gülck and Møller 2020). However, the latest phylogenetic classification of plant BBE-like genes was based on Arabidopsis sequences only (Brassicaceae) (Daniel et al. 2016) and consequently lacks genes from *Cannabis* and related genera. Even though some BBE-like enzymes related to cannabinoid oxidocyclases have been identified (Aryal et al. 2019), it still remains unclear exactly how the various described cannabinoid oxidocyclase genes are related to each other and to other BBE-like enzymes. Therefore, a comprehensive phylogenetic analysis of BBE-like enzymes including cannabinoid oxidocyclase genes is warranted.

Cannabinoid Oxidocyclase Gene Evolution

Although environmental factors play a role in determining the amount of cannabinoids present in different parts and stages of the plant (Rustichelli et al. 1998), in most populations the ratio between THCA and CBDA has been found to be under genetic control (Mandolino et al. 2003; Weiblen et al. 2015; Toth et al. 2020; Wenger et al. 2020). Codominant inheritance of CBDA and THCA chemotypes is consistent with a Mendelian single-locus (de Meijer et al. 2003; Onofri et al. 2015; Weiblen et al. 2015). This led to the model in which THCAS and CBDAS are encoded by alternate alleles of the same gene (B_T and B_D, respectively) (de Meijer et al. 2003). However, later genome sequencing revealed that they are encoded by different genes (rather than alleles) within a large polymorphic genomic region with low levels of recombination (Kojoma et al. 2006; van Bakel et al. 2011; McKernan et al. 2015; Onofri et al. 2015; Weiblen et al. 2015; Laverty et al. 2019; Grassa et al. 2021). Thus, they are treated as separate genes below.

The genes encoding THCAS, CBDAS, and CBCAS have been identified (Sirikantaramas et al. 2004; Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; Laverty et al. 2019). The *THCAS* gene comprises a 1638 bp intronless open reading frame that is found in all drug-type cultivars (Sirikantaramas et al. 2004; Kojoma et al. 2006; van Bakel et al. 2011; McKernan et al. 2015; Onofri et al. 2015; Weiblen et al. 2015; Vergara et al. 2019). For this reason, the gene has been used as a diagnostic marker for detecting psychoactive cultivars for crop breeding and forensics (Kojoma et al. 2006; Kitamura et al. 2016). It should be noted, however, that a nonpsychoactive cultivar from Malawi has a *THCAS* gene but accumulates the cannabinoid precursor CBGA instead of THCA (chemotype IV). This is probably due to a single amino acid mutation leading to a defective (B_{T0}) variant (Onofri et al. 2015). Recently, a different CBGA-dominant cultivar was found to have another single amino acid mutation in THCAS (Garfinkel et al. 2021). Gene copy number variation has been suggested based on amplicon sequencing of the *THCAS* locus (McKernan et al. 2015; Weiblen et al. 2015; Vergara et al. 2019). But amplicons may have included closely related genes such as *CBCAS* (see below). Thus, it remains unclear if *THCAS* occurs in multiple copies and, consequently, if copy number variation could be a target for the breeding of cultivars.

The *CBDAS* gene comprises a 1635 bp intronless open reading frame that is found in all CBDA-producing cultivars (Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007). However, different missense mutations have been described from CBGA-dominant (i.e., chemotype IV) hybrid cultivars from Italy and Ukraine that are considered B_{D01} and B_{D02} variants, respectively (Onofri et al. 2015). Another missense mutation was described from a cultivar from Turkey that is considered a weak B_{Dw} variant resulting in a partial accumulation of CBGA due to partly impaired activity of the encoded CBDAS (Onofri et al. 2015). In THCA-dominant cultivars,

fragments have been found that are 93% identical to functional *CBDAS* and share a four base pair deletion that results in a truncated and most probably nonfunctional protein (Weiblen et al. 2015; Cascini et al. 2019; Vergara et al. 2019). This deletion showed strict association with THCA-producing (chemotypes I and II) cultivars, suggesting tight genetic linkage with *THCAS*. Indeed, it could be used to discriminate between fiber and drug-type cultivars as well as accurately predict chemotype in feral and cultivated plants (Cascini et al. 2019; Wenger et al. 2020). Notably, up to three different variants of such putative pseudogenes were detected in single cultivars (van Bakel et al. 2011; Weiblen et al. 2015; Laverty et al. 2019; Vergara et al. 2019) suggesting multiple duplicated loci. Moreover, given their sequence divergence from *CBDAS* and close linkage with *THCAS* it is unclear if they should be considered variants of *CBDAS* or as separate loci.

The *CBCAS* gene comprises a 1638 bp intronless open reading frame. It was recently identified and described based on genome sequencing of the cultivar "Finola" (Laverty et al. 2019). Based on sequence comparisons, "inactive THCA" synthase sequences described from European fiber-type cultivars more than a decade earlier were also considered *CBCAS* (Kojoma et al. 2006; Laverty et al. 2019). Other reported amplified fragments may also represent the same gene. For example, lowly expressed fragments from CBDA-dominant (chemotype III) cultivars such "Carmen" and "Canna Tsu" have been reported to be similar to those "inactive THCA" synthases (McKernan et al. 2015; Onofri et al. 2015; Weiblen et al. 2015). In order to confirm whether these indeed represent *CBCAS*, a comprehensive analysis is required.

In addition to *THCAS*, *CBDAS*, and *CBCAS*, other yet uncharacterized sequences have been described. These may encode enzymes for other cannabinoids, but their copy numbers and sequence properties are not well described or cataloged (Hurgobin et al. 2021). Besides the functionally characterized *CBDAS* gene, Taura et al. amplified two other gene fragments both of which contain a 1635 bp intronless open reading frame from genomic DNA and named these *CBDAS2* and *CBDAS3*. They share 84% identity with *CBDAS* but did not encode enzymes with CBDAS activity (Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007). Weiblen et al. (2015) amplified a fragment from chemotype III fiber-type cultivar "Carmen" with 95% identity to *THCAS* and two identical fragments from cultivars "Skunk#1" and "Carmen" with 92% identity to *THCAS*. A pseudogene was described from the genome of the cultivar "Purple Kush" with 92% identity to *THCAS* but appeared phylogenetically separate from the previously mentioned fragments (van Bakel et al. 2011; Weiblen et al. 2015). In cultivar "Finola," a putative pseudogene with 93% identity to *THCAS* was reported (Laverty et al. 2019). In various other fiber-type cultivars, "mutated *THCAS*" fragments were reported, some of which were pseudogenized (Cascini et al. 2019). A recent phylogenetic analysis also identified a set of lineages representing functional and nonfunctional "unknown cannabinoid synthases" (Vergara et al. 2019). But, it remains unclear how these relate to the gene fragments listed above and to each other.

The total number of cannabinoid oxidocyclase genes varies considerably across cultivars. Onofri et al. (2015) amplified up to 5 (in cultivar "Haze") different full-length fragments in chemotype I drug-type cultivars and up to 3 (in cultivars from Yunnan and Northern Russia and an inbred Afghan hashish landrace) different full-length fragments in chemotype III fiber-type cultivars. Inbred individuals of cultivars "Carmen" and "Skunk #1" are expected to be homozygous but yielded four and five cannabinoid synthase fragments, respectively (Weiblen et al. 2015). McKernan et al. (2015) detected up to six different fragments (including pseudogenes) of *THCAS* and related sequences. A recent study on copy number variation in cannabinoid oxidocyclase genes estimated that some of the analyzed cultivars could have up to 10 different fragments (Vergara et al. 2019). Based on these results it is clear that cannabinoid oxidocyclase genes can be considered a unique gene family that stems from a recent expansion and includes genes with unknown function (Onofri et al. 2015; Weiblen et al. 2015; Vergara et al. 2019; Hurgobin et al. 2021). However, due to differences in 1) primers used for amplification, 2) reference genomes used for copy number estimation, and 3) level of homozygosity, these numbers are not directly comparable and may not be accurate assessments of gene copy number. There is also no appropriate classification to facilitate the unequivocal naming and referencing of cannabinoid oxidocyclase genes.

Finally, it remains unclear whether these genes are specific to *Cannabis*. A phylogenetic analysis sampling cannabinoid oxidocyclase genes from cultivars "Skunk#1," "Carmen," and "Purple Kush" a priori considered cannabinoid oxidocyclase genes to comprise a clade (Weiblen et al. 2015). A more recent phylogenetic analysis based on genomic data applied the same a priori assumption (Hurgobin et al. 2021). Another study sampling cultivars "Pineapple Banana Bubble Kush" and "Purple Kush" suggested that all cannabinoid oxidocyclase genes may comprise a clade but did not include functional *CBDAS* sequences nor homologs from *Cannabis'* most closely related genus *Humulus* (Vergara et al. 2019). Therefore, it remains unknown whether cannabinoid oxidocyclase genes are specific to *Cannabis* or represent more ancient gene duplications in, for example, an ancestor of *Cannabis* and related genera within the Cannabaceae family such as *Humulus* and *Trema* (Padgitt-Cobb et al. 2019; Vergara et al. 2019). In addition, the genomic localization of many described gene sequences remains unknown and, consequently, we lack a clear overview of the patterns of gene duplication and divergence across the *Cannabis* genome (Weiblen et al. 2015).

Aims

We present a comparative analysis of cannabinoid oxidocyclase genes in the genomes of *Cannabis*, closely related genera and informative outgroup species. This was greatly aided by the recent release and publication of several diverse *Cannabis* genome assemblies based on long-read sequencing technologies (McKernan et al. 2018; Laverty et al. 2019; S. Gao et al. 2020; Grassa et al. 2021). In addition, genomic information is available for other genera in the Cannabaceae family. Recent species-level phylogenetic analyses of the Cannabaceae family based on plastome sequences suggest that the genera *Parasponia* and *Trema* together are sister to *Cannabis* and *Humulus* (Jin et al. 2020). Draft genome assemblies have recently become available for *Humulus*, *Parasponia*, and *Trema*, that can be used for comparative analyses of *Cannabis* genes (van Velzen et al. 2018; Padgitt-Cobb et al. 2019; Kovalchuk et al. 2020). This provides an excellent opportunity to perform a comprehensive reconstruction of the evolution of cannabinoid oxidocyclase genes. In addition, *Morus notabilis* (Moraceae), *Medicago truncatula* (Fabaceae), and *Arabidopsis thaliana* (Brassicaceae) were included as outgroups, allowing us to place our results within a broader phylogenetic perspective and in the context of the existing BBE-like gene family classification (Daniel et al. 2016). Based on phylogenetic and synteny analysis, we elucidate the evolution of these genes and address the following questions:

1. How are cannabinoid oxidocyclases related to other berberine bridge enzymes?
2. Are cannabinoid oxidocyclase genes specific to *Cannabis* or do they represent more ancient duplications in, for example, an ancestor of *Cannabis* and related genera within the Cannabaceae family?
3. What are the phylogenetic relationships of *THCAS*, *CBDAS*, and *CBCAS* with other closely related genes?
4. What are the patterns of duplication and divergence of cannabinoid oxidocyclase genes across *Cannabis* genomes?

We also present a comprehensive clade-based classification of all cannabinoid oxidocyclase genes to resolve current confusion due to inconsistencies in naming and aid their future referencing and identification.

Results

Cannabinoid Oxidocyclase Genes Are Specific for *Cannabis*

To place cannabinoid oxidocyclase genes within the context of the BBE-like gene family we performed a phylogenetic analysis of BBE-like protein sequences from selected Eurosid genomes (supplementary table S1, Supplementary Material online). These include genomes from *C. sativa* cultivar "CBDRx," *Humulus lupulus* cultivar "Cascade," *Parasponia andersonii*, and *Trema orientalis* from the Cannabacae family. Genomes from *Morus notabilis* (Moraceae), *Medicago truncatula* (Fabaceae), and *Arabidopsis thaliana* (Brassicaceae) were included as outgroups (fig. 2*A*). The resulting gene tree recovered 11 clades, including groups 1–7 earlier described (Daniel et al. 2016) based on Brassicaceae sequences (all seven groups were monophyletic, except that group 3 was confirmed to be nested within group 4). *Cannabis* BBE-like sequences were found in groups 2, 5.1, 5.2, 6, and 7.1. In addition, *Cannabis* sequence accession XP_030480925.1 represented an undescribed clade which we named group 8; *Cannabis* sequence accession XP_030480615.1 represented an undescribed clade including berberine bridge enzyme originally described from *E. californica* which we named group 9. *THCAS*, *CBDAS*, and *CBCAS* are members of a newly defined group 10. Within this group, a Cannabaceae-specific gene expansion can be identified within which all three known cannabinoid oxidocyclase occur in a *Cannabis*-specific clade, which we name the cannabinoid oxidocyclase clade. This suggests that *THCAS*, *CBDAS*, and *CBCAS* originated from a single ancestral cannabinoid oxidocyclase gene within the *Cannabis* lineage.

We also found that BBE-like genes often occurred near each other in the *Cannabis* CBDRx genome. We therefore retrieved genomic locations of all BBE-like genes in other genomes including *T. orientalis*, *M. notabilis*, and *A. thaliana*. This revealed that BBE-like genes from different clades are commonly colocalized in these genomes (fig. 2*B*). This suggests that selection favors BBE-like genes to remain in close genomic proximity. It is known that genes involved in the same pathway have the tendency to cluster in plant genomes (Liu et al. 2020). However, it is not clear if and how the various BBE-like genes share pathways and we therefore have no conclusive explanation for this intriguing pattern.

Phylogenetic Classification of Cannabinoid Oxidocyclase Genes

To elucidate the phylogenetic relationships of *THCAS*, *CBDAS*, and *CBCAS* with other highly homologous genes, we performed an extensive phylogenetic analysis of cannabinoid oxidocyclase genes from genome assemblies of *Cannabis* cultivars "CBDRx", "Jamaican Lion," "Finola," "Purple Kush," and a putatively wild *Cannabis* plant from Jilong, Tibet (McKernan et al. 2018; Laverty et al. 2019; S. Gao et al. 2020; McKernan et al. 2020; Grassa et al. 2021). Additional analyses based on sequences from NCBI and from more fragmented *Cannabis* genomes are shown in supplementary figures S1 and S2, Supplementary Material online, respectively. Based on the resulting gene trees we consistently recovered the same three main clades (A–C; fig. 3) that we describe below.

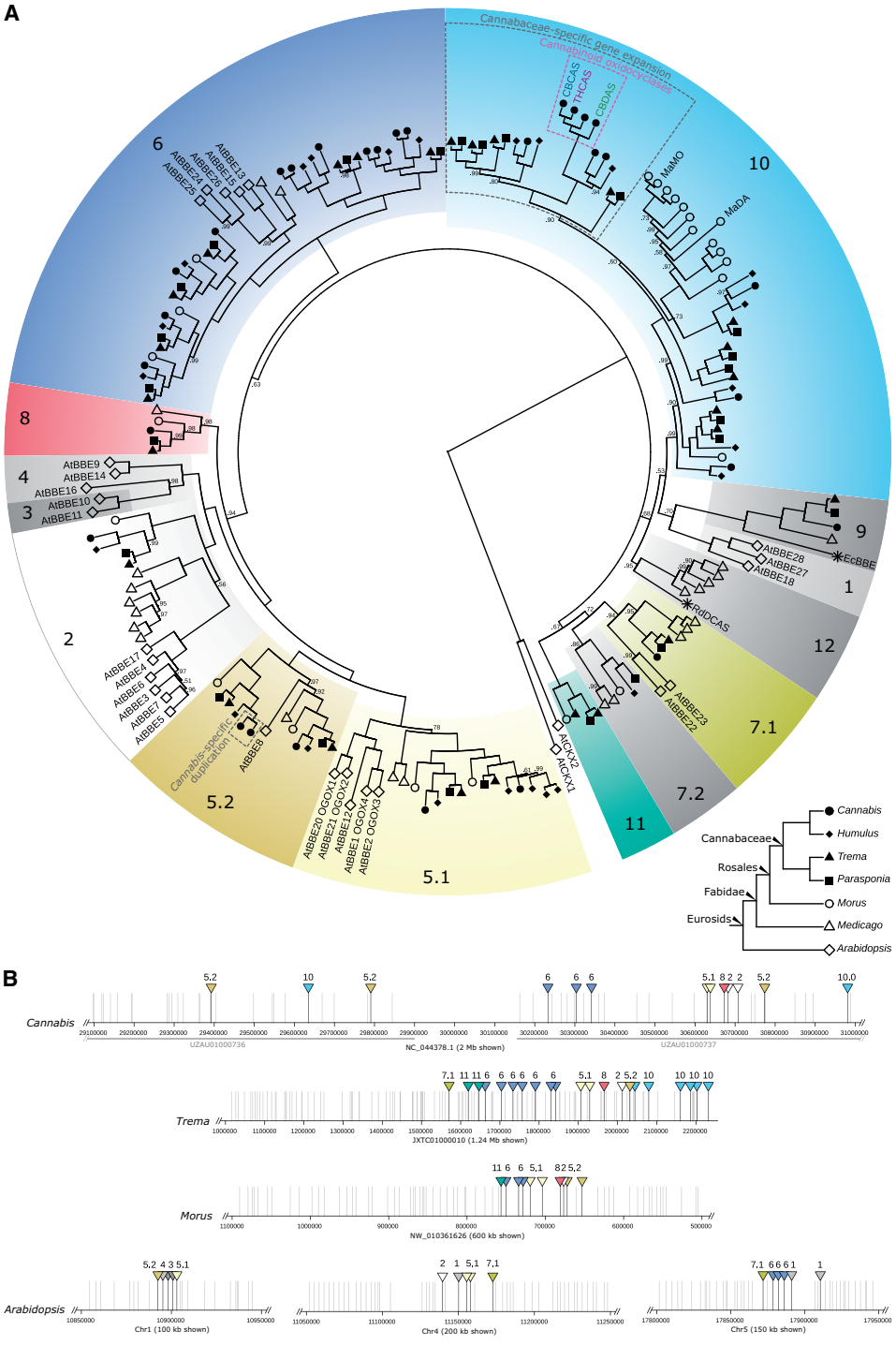

Fig. 2.—Eurosid berberine bridge enzyme gene family analysis. (A) Gene tree based on protein sequences showing that cannabinoid oxidocyclases comprise a *Cannabis*-specific clade. Shapes indicate sampled species *Cannabis sativa* (solid circles; $N = 30$), *Humulus lupulus* (solid diamonds; $N = 24$), *Parasponia andersonii* (solid squares; $N = 25$), *Trema orientalis* (solid triangles; $N = 26$), *Morus notabilis* (open circles; $N = 24$), *Medicago truncatula* (open triangles; $N = 23$), and *Arabidopsis thaliana* (open diamonds; $N = 29$), stars indicate additional sequences from *Rhododendron dauricum* and *Eschscholzia californica*. Colored blocks indicate the identified groups 1–12; node labels indicate posterior probabilities below 1.0. Bottom right inset shows known relationships among sampled species. (B) genomic colocalization of berberine bridge enzymes in *C. sativa* cultivar "CBDRx", *T. orientalis*, *M. notabilis*, and *A. thaliana*. Grey horizontal bars indicate contigs in the *Cannabis* chromosomal scaffold shown in figure 4A. Vertical lines indicate locations of annotated genes; berberine bridge enzymes are indicated with triangles in color and numbering consistent with those in (A). For displaying purposes, genomic scaffolds are not shown in the same scale (size shown is indicated).

Clade A

Clade A includes *THCAS* and *CBCAS*. It comprises four subclades and can be characterized by three unique nonsynonymous substitutions (supplementary table S2, Supplementary Material online). Subclade A1_THCAS comprises full-length coding sequences from THCA-producing plants such "Purple Kush," "Skunk#1," and "Chemdog91," including functionally characterized *THCAS* (Sirikantaramas et al. 2004), "drug-type *THCAS*" sequences (Kojoma et al. 2006), "active *THCAS*" sequences (McKernan et al. 2015), and fully functional (B_T) as well as nearly defective (B_{T0}) coding sequences (Onofri et al. 2015) (fig. 3, supplementary figs. S1 and S2, Supplementary Material online; table 1 and supplementary table S5, Supplementary Material online). Subclade A1_THCAS sequences can be characterized by three unique nonsynonymous substitutions and may be further divided into 2 groups and 18 types of which 9 were previously described by Onofri et al. (2015) and 9 are new (supplementary table S2, Supplementary Material online). Group 1 comprises six types (1/1–1/6) that are identical or similar to the *THCAS* reference (Sirikantaramas et al. 2004; Onofri et al. 2015). Type 1/1 (Onofri et al. 2015) comprises sequences from various cultivars that are identical to the functionally characterized *THCAS* described by (Sirikantaramas et al. 2004). Type 1/2 (Onofri et al. 2015) comprises accession KP970849.1 which differs by a single synonymous substitution from type 1/1 and can therefore be considered functionally equivalent. Type 1/3 (Onofri et al. 2015) comprises the defective B_{T0} allele from Malawi that differs only by 706 C (Gln). Type 1/4 (Onofri et al. 2015) comprises sequences from cultivars "Skunk #1," "AK47," "Chemdog91," "CannaTsu," "Black84," and a hashish landrace from Afghanistan that share 749 A (Asp). Type 1/5 comprises sequences from cultivars "Purple Kush," "Blueberry Essence," and "C4xCannaTsu" that share the unique amino acid (aa) substitution 998G (Arg). Type 1/6 comprises accession KT876046.1 from cultivar "Otto" that differs by only one aa substitution. Group 2 comprises six types sharing the nonsynonymous substitution 373C (Leu). Type 2/1 (Onofri et al. 2015) comprises sequences from cultivars "Haze," "Alaskan ice," and "Otto" that share two nonsynonymous substitutions. Type 2/2 (Onofri et al. 2015) comprises accession KP970853.1 from cultivar "Haze" that has one nonsynonymous substitution. Type 2/3 (Onofri et al. 2015) comprises accession which differs by two synonymous substitutions from type 2/1 and can therefore be considered functionally equivalent. Type 2/4 comprises sequences from Boseung province in Korea described by Doh et al. (2019) that share three nonsynonymous substitutions. Type 2/5 comprises accession MN422091.1 from Jecheon province in Korea that has five nonsynonymous substitutions. Type 2/6 comprises sequences from low-THCA cultivar "Cheungsam" described by Doh et al. (2019) and can be characterized by ten nonsynonymous substitutions. Other types remain ungrouped. Type 3 (Onofri et al. 2015) comprises accession KP970851.1 from a hashish landrace from Afghanistan that has the unique aa substitution 187C (Leu). Type 4 (Onofri et al. 2015) comprises accession KP970855.1 from cultivar "Haze" and has aa substitutions 794 G (Gly) and 1229 A (Glu). Type 5 comprises partial sequences from various regions in Pakistan described by Ali et al. (2019) of which at least one is a pseudogene. They can be characterized by two unique aa substitutions: 851T (Val) and 883C (Pro). Type 6 comprises accession MT338560.1 from Oregon CBD that shares 998 G (Arg) with type 1/5 but has an additional unique aa substitution 1064A (Asn). Type 7 comprises accession LC120319.1 from cultivar "Big Bud" described by Kitamura et al. (2016) that shares 749 A (Asp) with type 1/4 but has one additional nonsynonymous substitution (1018 G; Ala). Type 8 comprises the putative *THCAS* sequence of a putatively wild plant from Jilong, Tibet that can be characterized by six unique aa substitutions.

Subclade A2_CBCAS comprises full-length coding as well as nonfunctional (pseudo)gene sequences from drug-type cultivars such as "Purple Kush" and "Jamaican Lion" as well as fiber-type cultivars such as "Finola" and "Carmen". It includes the functionally characterized *CBCAS* (Laverty et al. 2019), "mutated *THCAS*" (Cascini *et al.* 2019), "fiber-type *THCAS*" sequences (Kojoma et al. 2006), and "inactive *THCAS*" sequences (Kojoma et al. 2006; McKernan et al. 2015; Cascini et al. 2019; McKernan et al. 2020). Subclade A2_CBCAS sequences can be characterized by 12 unique aa substitutions (supplementary table S2, Supplementary Material online).

Subclade A3 comprises at least one full-length coding sequence from cultivar "Finola" and at least two nonfunctional (pseudo)gene copies from drug-type cultivars "Purple Kush" and "Jamaican Lion." These sequences can be characterized by a duplication of the 3rd codon (TAC; Tyr) and 13 unique nonsynonymous substitutions (supplementary table S2, Supplementary Material online). They have not yet been functionally assessed but given that at least one variant comprises a full-length coding sequence it is expected to have some functional relevance; probably as a cannabinoid oxidocyclase.

Subclade A4 comprises three nonfunctional (pseudo)gene sequences from whole-genome assemblies of cultivars "Purple Kush," "Finola," and "Jamaican Lion." They share four nonsense mutations and 14 unique aa substitutions and can be divided into two types (supplementary table S2, Supplementary Material online).

Clade B

Clade B comprises two subclades and can be characterized by 16 unique aa substitutions (supplementary table S2, Supplementary Material online). Subclade B1_CBDAS comprises full-length coding *CBDAS* sequences from chemotype III cultivars such as "Finola," "Carmen," and "CBDRx" (fig. 3,

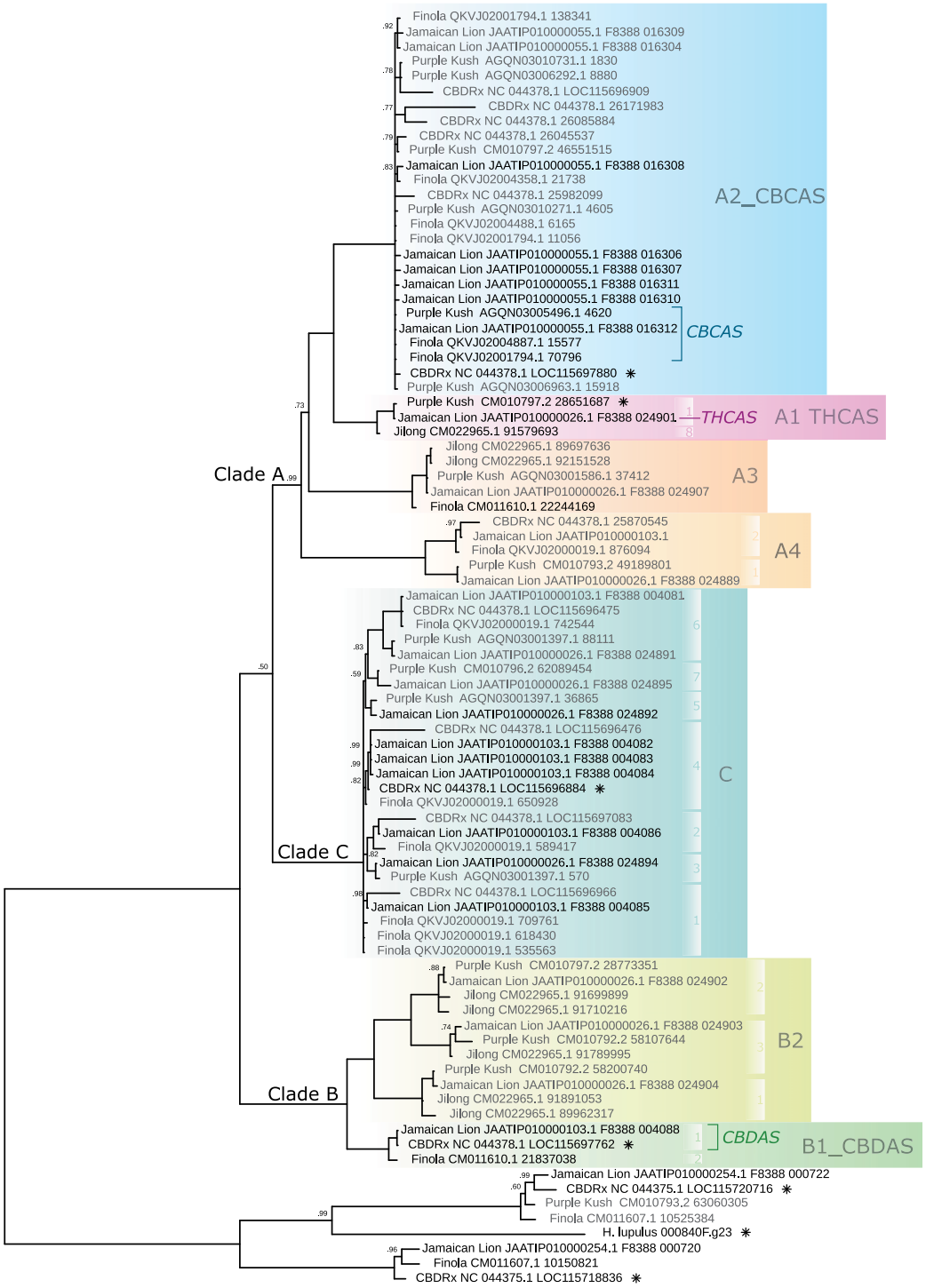

Fig. 3.—Cannabinoid oxidocyclase gene tree. Based on nucleotide sequences from whole-genome assemblies of *Cannabis sativa* cultivars "CBDRx" ($N = 13$), "Finola" ($N = 15$), "Jamaican Lion" (mother; $N = 26$), "Purple Kush" ($N = 16$), and a putatively wild plant from Jilong, Tibet ($N = 8$) and *Humulus lupulus* ($N = 1$); including closely related berberine bridge enzymes as outgroups for a total of 78 gene sequences used in this phylogenetic reconstruction (McKernan et al. 2018; Laverty et al. 2019; S. Gao et al. 2020; Grassa et al. 2021). Labels indicate genbank accession of genomic contig and locus tag (when available) or start position. Putative nonfunctional (pseudo)genes are in gray; functionally characterized *THCAS*, *CBDAS*, and *CBCAS* genes (Sirikantaramas et al. 2004; Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; Laverty et al. 2019) are labeled. Sequences included in the BBE-like analysis shown in figure 2A are marked with an asterisk. Colored blocks indicate the identified clades; white blocks indicate sequence types. Node labels indicate posterior probabilities below 1.0.

Table 1
Number of Full-Length Coding (cd) and Nonfunctional (nf) Cannabinoid Oxidocyclase Gene Sequences in High-Quality Genome Assemblies

Cultivar	Finola		Purple Kush		CBDRx		Jamaican Lion (Mother)		Jamaican Lion (Father)		Jilong	
Chemotype	III		I		III		II		III		Unknown	
Gene status	cd	nf	cd	nf	cd	nf	cd	nf	cd	nf	cd	nf
A1_THCAS	0	0	1	0	0	0	1	0	0	0	1	0
A2_CBCAS	2	4	2	3	1	5	6	2	3	1	0	0
A3	1	0	0	1	0	0	0	1	0	0	0	2
A4	0	1	0	1	0	1	0	2	0	1	0	0
B1_CBDAS	1	0	0	0	1	0	1	0	1	0	0	0
B2	0	0	0	3	0	0	0	3	0	0	0	5
C	0	5	0	4	1	4	7	3	4	1	0	0
Sums	4	10	3	12	3	10	15	11	8	3	1	7
Total	14		15		13		26		11		8	

supplementary figs. S1 and S2, Supplementary Material online; table 1). These sequences can be characterized by a 3 bp deletion at position 755 and 14 unique aa substitutions (supplementary table S2, Supplementary Material online). They can be further divided into two types that correspond with groups 5 and 6 described by Onofri et al. (2015). Type 1 is characterized by 1423A (Lys) and found in fiber-type cultivars such as "Ermo" and "C.S.," drug-type cultivars "Jamaican Lion" and "CBDRx," and landraces from China and Japan (Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; Onofri et al. 2015; Cascini et al. 2019; Grassa et al. 2021). It includes the CBDA synthase gene *CBDAS1* and a defective sequence B_{D01} coding sequence (Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; Onofri et al. 2015). Type 2 can be characterized by three unique Serines (supplementary table S2, Supplementary Material online) and is found in chemotype III fiber-type cultivars such as "Finola," "Carmen," "Ermes," "Futura 75," "Tygra," and "Uso31" (Onofri et al. 2015; Weiblen et al. 2015; Cascini et al. 2019; Laverty et al. 2019). It includes fully functional (B_D), weakly functional (B_{DW}), and defective (B_{D01} and B_{D02}) coding sequences (Onofri et al. 2015). We note that some sequences described by Cascini et al. are ambiguous at type-specific positions and so probably represent a mix of both types (see supplementary table S3, Supplementary Material online).

Subclade B2 comprises the nonfunctional (pseudo)gene sequences from THCA-producing cultivars such as "Purple Kush," "Jamaican Lion," "Skunk#1," "Chocolope," and "Northern Light"; including "mutated *CBDAS*" sequences described by Cascini et al. and "marijuana-type CBDA synthase" sequences described by Weiblen et al. (2015) (Cascini et al. 2019) (fig. 3, supplementary figs. S1 and S2, Supplementary Material online). They can be characterized by a 4 or 6 bp frame-shift deletion at position 153 and by two unique nonsynonymous substitutions (supplementary table S2, Supplementary Material online). They can be further divided into three types based on additional missense and aa mutations (we note, however, that because of the shared frame-shift deletions these mutations probably did not have any significance in terms of actual coding sequence and can therefore be considered "secondary"). The first type can be characterized by four secondary nonsense mutations and eight secondary unique aa substitutions (supplementary table S2, Supplementary Material online). The second type can be characterized by two secondary unique aa substitutions. The third type can be characterized by nine secondary unique aa substitutions (supplementary table S2, Supplementary Material online). Accession LKUA01006620.1 from cultivar "LA confidential" may be a chimera of types 2 and 3.

Clade C

Clade C comprises full-length coding as well as nonfunctional (pseudo)gene sequences from cultivars "Purple Kush," "Finola," "CBDRx," and "Jamaican Lion" as well as *CBDAS2* and *CBDAS3* from a chemotype III "domestic" cultivar from Japan described as having no CBDA synthase activity by Taura, Sirikantaramas, Shoyama, Yoshikai et al. (2007) (fig. 3, supplementary figs. S1 and S2, Supplementary Material online; table 1). They share 19 unique aa substitutions and can be divided into seven types (fig. 3; supplementary table S2, Supplementary Material online).

Patterns of Cannabinoid Oxidocyclase Gene Duplication and Divergence

To reconstruct patterns of gene duplication and divergence, we assessed microsynteny across genomes of *Cannabis* cultivars "CBDRx," "Jamaican Lion," (mother) "Finola," "Purple Kush," and a putatively wild *Cannabis* plant from Jilong, Tibet. Based on nucleotide alignments and protein comparisons, we found that all cannabinoid oxidocyclase genes occur in two main syntenic clusters, together with other BBE-like genes. The first main syntenic cluster comprises a tandemly repeated array of genes from clade C in the genome

assemblies of cultivars "Finola," "CBDRx," and "Jamaican Lion" (fig. 4A). The array is flanked at the 3′ end by a group 5.2 BBE-like gene, a receptor-like protein, a Patellin protein, a TWINKLE DNA primase-helicase, and a caseinolytic protease. In the chemotype II Jamaican Lion genome, there are two putative allelic variants; the first comprising two full-length coding sequences and two nonfunctional (pseudo)gene copies and the second comprising five full-length coding sequences and one nonfunctional (pseudo)gene copy. In the assembly of chemotype III cultivar "Finola," it comprises six nonfunctional (pseudo)gene copies. In the assembly of chemotype III cultivar "CBDRx," it comprises one full-length coding sequence and four nonfunctional (pseudo)gene copies. The array is flanked at the 5′ end by one of three variants of a large genomic region with very little nucleotide-level identity (supplementary fig. S3, Supplementary Material online). All variants comprise another copy of a group 5.2 BBE-like gene. The first variant comprises a single copy of *THCAS*, a tandemly repeated array of subclade B2 nonfunctional (pseudo)genes, and a nonfunctional (pseudo)gene from the A4 subclade. It is present in cultivars "Jamaican Lion," "Purple Kush," and the putatively wild plant from Jilong, Tibet. The second variant comprises a single copy of a type 1 *CBDAS* and can be found in cultivars "Jamaican Lion" and "CBDRx." The third variant comprises only a type 2 *CBDAS* and can be found in cultivar "Finola." We found no nucleotide-level alignments between these variants except for the context around the group 5.2 BBE-like in the second and third variants (fig. 4A and supplementary fig. S3, Supplementary Material online). This suggests high levels of divergence across this large genomic region.

The second cluster comprises a tandemly repeated array of genes from subclade A2_CBCAS in the genome assemblies of cultivars "CBDRx" and "Jamaican Lion" (fig. 4B and supplementary fig. S4, Supplementary Material online). The array is flanked at the 5′ end by a RING/FYVE/PHD-type zinc finger family protein and a receptor-like kinase; and at the 3′ end by an ankyrin repeat family protein and an NRT1/PTR family protein in both assemblies. Some of these flanking genes are considered pathogen response genes (McKernan et al. 2020). In the chemotype II "Jamaican Lion" genome, there are two putative allelic variants, the first of which comprises six full-length *CBCAS* coding sequences and two nonfunctional (pseudo)gene copies and the second of which is only partially assembled. In the "CBDRx" genome, it comprises one full-length *CBCAS* coding sequence and five nonfunctional (pseudo)gene copies. Interestingly, it also includes a nonfunctional (pseudo)gene from subclade A3 but given the lack of additional A3 copies within the array, this appears to be a relatively recent insertion. In the genome assemblies of cultivars "Finola" and "Purple Kush," subclade A2_CBCAS gene copies appear in several unplaced scaffolds probably representing the same array (Hurgobin et al. 2021). Subclade A2_CBCAS gene copies are absent in the assembly of the putatively wild plant from Tibet. No further synteny was found with *Humulus*, *Parasponia*, or *Trema*; suggesting that this syntenic cluster is specific to *Cannabis*.

To assess the direction of evolution, we then assessed protein-level microsynteny in genomes from closely related Cannabaceae species *H. lupulus*, *P. andersonii*, and *T. orientalis*. This revealed that each comprises a tandemly repeated array of group 10 BBE-like genes that are closely related to the known cannabinoid oxidocyclase genes, as well as a single copy of the group 5.2 BBE-like gene (not found in *Humulus*) and the receptor-like protein, the Patellin protein, the TWINKLE DNA primase-helicase, and the caseinolytic protease listed above (fig. 3A). This suggests that cannabinoid oxidocyclases originated within an ancestral syntenic block and experienced a series of tandem gene duplications, translocations, and divergence.

Discussion

Origin of Cannabinoid Oxidocyclases from within the BBE-like Gene Family

Since the cannabinoid oxidocyclase genes were first discovered and described, it has been known that they are members of the BBE-like gene family (Sirikantaramas et al. 2004; Daniel et al. 2017). However, the BBE-like family is large and the most recent classification of plant BBE-like genes was based only on analysis of genes from *Arabidopsis* in the Brassicaceae family (Daniel et al. 2016). Even though some BBE-like enzymes related to cannabinoid oxidocyclases have been identified (Aryal et al. 2019), it remained unclear how the various described cannabinoid oxidocyclase genes are related to each other and to other berberine bridge enzymes.

Our results show that cannabinoid oxidocyclase genes from *Cannabis* originated from a newly defined clade (Group 10) within the BBE-like gene family (fig. 2A). *Rhododendron dauricum* daurichromenic acid synthase (*RdDCAS*), another plant cannabinoid oxidocyclase (Iijima et al. 2017), originates from another clade (Group 12). Within Group 10 gene expansions occurred independently in Moraceae and Cannabaceae (fig. 2A). The expansion in Moraceae includes the recently described *Morus alba* Diels–Alderase (*MaDA*) and moracin C oxidase (*MaMO*) genes that are responsible for the production of the medicinal compound chalcomoracin (Han et al. 2018; L. Gao et al. 2020). The expansion in Cannabaceae eventually led to the origin of cannabinoid oxidocyclases. Such gene diversification and enzymatic versatility confirm that BBE-like enzymes play important roles in generating biochemical novelty (Daniel et al. 2017). Because most plant BBE-like genes (including those in Group 10) have a secretory signal peptide they may be considered to be "preadapted" for a role in the extracellular space. Within the Cannabaceae-specific gene expansion, *THCAS*, *CBDAS*, and *CBCAS* form a clade that is sister to

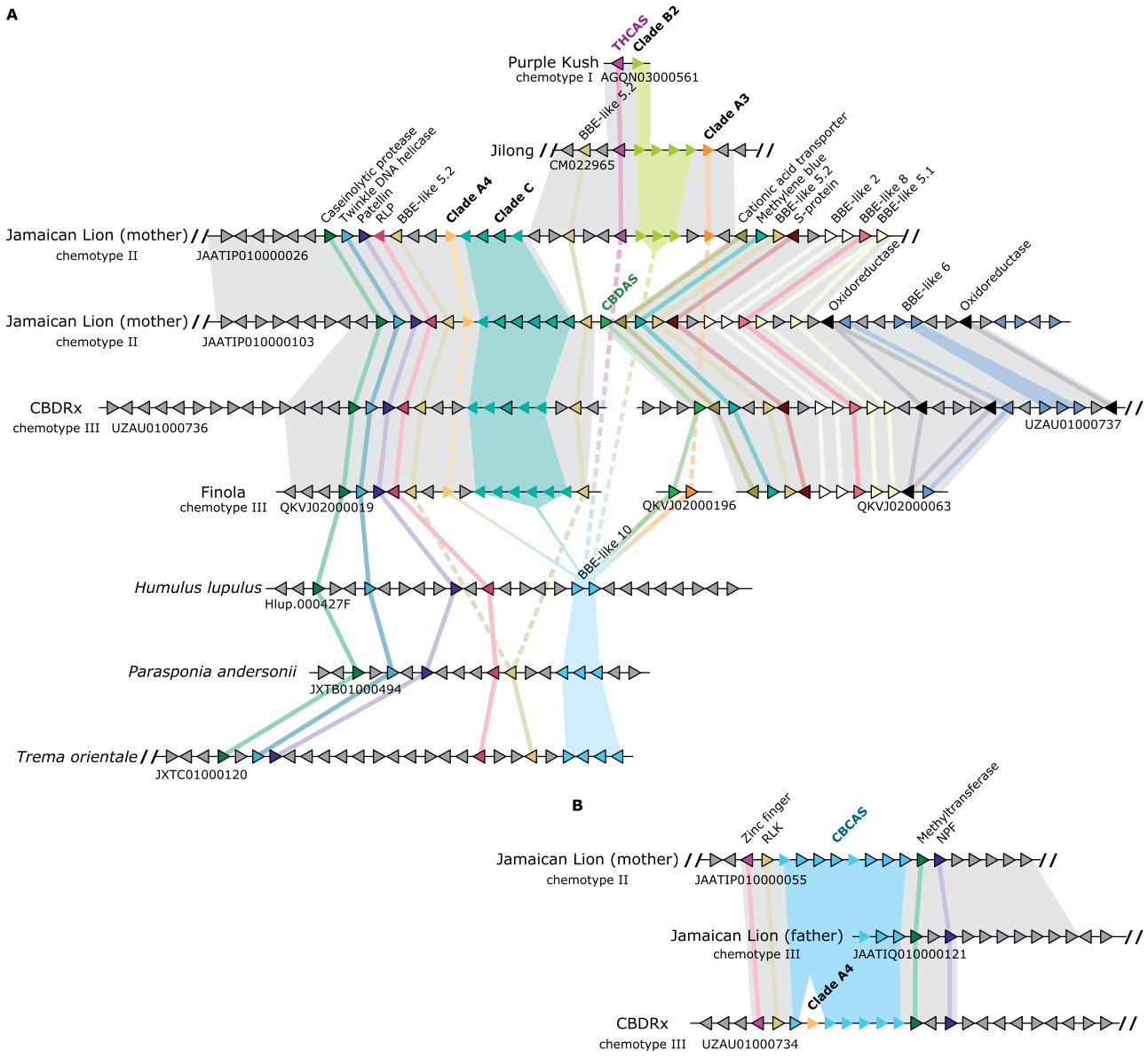

Fig. 4.—Cannabinoid oxidocyclase microsynteny assessments. (A) Syntenic block comprising *CBDAS*, *THCAS*, and related genes. (B) Syntenic block comprising *CBCAS* tandemly repeated array. Triangles indicate genes (not to scale) colored according to their homology and putative orthologs are connected with colored lines. Nonfunctional (pseudo)genes are shown without black outlines. Gray backgrounds indicate LASTZ nucleotide alignments based on results shown in supplementary figures S3 and S4, Supplementary Material online. Cannabinoid oxidocyclase genes are members of BBE-like group 10 (see fig. 2) and labeled in boldface. BBE, berberine bridge enzyme; NPF, NRT1/PTR family protein; RLK, receptor-like kinase; RLP, receptor-like protein.

homologous genes from *Cannabis* and *Humulus* (fig. 2*A*). These results show unequivocally and for the first time that cannabinoid oxidocyclase genes did not originate from more ancient duplications within the Cannabaceae but are specific to *Cannabis*.

The central cannabinoid precursor CBGA is the common substrate for THCAS, CBDAS, and CBCAS (fig. 1). We therefore hypothesize that the CBGA biosynthetic pathway existed before the origin of cannabinoid oxidocyclases. Thus, other *Cannabis* cannabinoid biosynthesis genes such as those encoding CBGAS, OAC, OLS, and acyl-activating enzyme 1 (Raharjo et al. 2004; Taura et al. 2009; Gagne et al. 2012; Stout et al. 2012) may have originated from more ancient duplications in an ancestor of *Cannabis* and related genera within the Cannabaceae family such as *Humulus*, *Parasponia*, and *Trema*. The comparative approach that we leveraged here can help elucidate the order in which these pathway genes evolved and, thus, reconstruct the origin of a novel and societally relevant biosynthetic pathway.

A Novel Classification of Cannabinoid Oxidocyclase Genes

Based on our phylogenetic analysis, we classified the cannabinoid oxidocyclase genes into three main clades (A–C) comprising a total of seven (sub)clades (fig. 3). *THCAS* and *CBCAS* are most closely related and occur in subclades A1 and A2, respectively, while *CBDAS* occurs in subclade B1. In addition to these three subclades representing functionally characterized cannabinoid oxidocyclase genes, we identified four previously unrecognized subclades. Based on current sampling, two of these clades contain only pseudogenes. Within subclade A4, two types can be recognized that share four nonsense mutations. Similarly, within subclade B2, three types can be recognized that share a frame-shift mutation (supplementary table S2, Supplementary Material online). Therefore, it seems that within each of these two subclades, the most recent common ancestor was likely already nonfunctional. Contrastingly, clade C and subclade A3 each include full-length coding sequences that are most likely functional enzymes. Taura *et al.* (2007) prepared recombinant enzymes based on two clade C full-length sequences (accession numbers AB292683.1 and AB292683.1) and reported they did not exhibit CBDA synthase activity but did not show the underlying experimental data. The subclade A3 sequence from cultivar "Finola" was reported as a pseudogene (Laverty et al. 2019) but based on our assessment of the genome sequence deposited on genbank it encodes a full-length protein. It has not yet been experimentally tested. Consequently, there is potential that the products of clade C and subclade A3-encoded enzymes are of biochemical and potential medical importance.

Our clade-based classification is intended to aid unequivocal referencing and identification of cannabinoid oxidocyclase genes. For example, based on our analyses we were able to confirm that sequences variously named "Fiber-type," "inactive," or "obscure" *THCAS* (McKernan et al. 2015; Onofri et al. 2015; Weiblen et al. 2015; McKernan et al. 2020) can be classified as variants of *CBCAS* (Laverty et al. 2019). Similarly, we reclassify sequences variously described as "marijuana-type" or "mutated" *CBDAS* (Weiblen et al. 2015; Cascini et al. 2019) as representing subclade B2. In retrospect, much of the confusion about gene identity stems from the general tendency to name sequences in accordance with the primers used for their amplification. For example, *CBCAS*-like and clade B2 pseudogenes were probably erroneously classified because they were generally amplified with primers that were considered specific for *THCAS* or *CBDAS*, respectively. These coamplifications are undoubtedly due to the high levels of sequence similarity between these genes. Sequences representing clade C have been often neglected in amplicon-based studies because they did not amplify using *THCAS* or *CBDAS* primers (Onofri et al. 2015). Moreover, these sequences were variously considered either a variant of *CBDAS* or *THCAS*, leading to additional confusion (Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; Weiblen et al. 2015). We anticipate that our classification will help avoid such confusion about the identity and relationships of cannabinoid oxidocyclase genes in the future. Our comprehensive analyses sampling all currently available sequences consistently recovered the same clades (see fig. 3, supplementary figs. S1 and S2, Supplementary Material online), suggesting that this classification is robust. New cannabinoid oxidocyclase sequences can be associated with the corresponding clade by phylogenetic analysis or based on the clade-specific missense mutations listed in supplementary table S2, Supplementary Material online. Our sequence alignments and phylogenetic trees are available for analysis via https://doi.org/10.4121/13414694 (last accessed June 2021). In case sequences fall outside any of our described clades, new (sub)clades can be defined in accordance with our system.

Localization and Divergence of Oxidocyclase Genes in the *Cannabis* Genome

The genetic basis underlying the ratio of THCA and CBDA is relatively well known. Genome sequencing of the chemotype I drug-type cultivar "Purple Kush" and chemotype III fiber-type cultivar "Finola" revealed that *THCAS* and *CBDAS* genes are located at different loci within a single large polymorphic genomic region with low levels of recombination (Laverty et al. 2019). However, the genomic locations of most other oxidocyclase genes have remained unknown. Consequently, a comprehensive overview of the patterns of gene duplication and divergence across the *Cannabis* genome has been lacking (Weiblen et al. 2015). Our assessment of microsynteny based on nucleotide alignments and protein comparisons revealed that cannabinoid oxidocyclase genes occur in two large syntenic blocks. The first block comprises a conserved region including a tandemly repeated array of clade C genes and a divergent region including either *THCAS* and subclade B2 pseudogenes; *CBDAS* and a subclade A3 gene; or only *CBDAS* (fig. 4*A*, supplementary fig. S3, Supplementary Material online). Close linkage of clade B2 pseudogenes with *THCAS* explains why they are considered markers for drug-type cultivars (Cascini et al. 2019). It also explains why a "*CBDAS* genotype assay" differentiating between subclades B1 and B2 can be used to accurately predict levels of THCA versus CBDA (Wenger et al. 2020). The second block comprises a conserved region including a tandem repeat of *CBCAS*-like genes (fig. 4*B*). This explains why, even though *CBCAS*-like genes have been considered to be associated with fiber-type cultivars and genomes of some drug-type cultivars indeed lack any *CBCAS*-like gene, genomes of other chemotype I drug-type cultivars such as "Purple Kush," "Skunk #1," and "Pineapple Banana Bubble Kush" do have full-length *CBCAS* genes (supplementary tables S4 and S5, Supplementary Material online). We have not found any nonfunctional *THCAS* gene closely linked to *CBDAS* as predicted

by (Wenger et al. 2020). However, given the deep divergence between *THCAS* and *CBDAS* it seems unlikely that they comprise orthologous genes (see figs. 3 and 4). Similarly, it is yet unclear if *CBDAS* and the subclade B2 pseudogenes are orthologous or comprise different paralogous loci. Long-read whole-genome sequencing of additional cultivars or wild plants may uncover haplotypes including *THCAS*, *CBDAS*, and clade B2 pseudogenes and help elucidate these aspects.

An Evolutionary Model for the Origin and Diversification of Cannabinoid Oxidocycloase Genes

Our protein-level microsynteny analysis including genomes from closely related species *H. lupulus*, *P. andersonii*, and *T. orientalis* revealed an ancestral syntenic block including several Group 10 BBE-like genes (fig. 4*A*). Below, we propose our most parsimonious evolutionary interpretation of cannabinoid oxidocyclase gene duplication and divergence (fig. 5). First, a group 10 BBE-like gene in an ancestral *Cannabis* neofunctionalized to use CBGA as substrate. Subsequent gene duplication and divergence lead to a set of ancestral genes representing the three main extant clades A, B, and C. Next, tandem duplication of this set together with the closely linked group 5.2 BBE-like gene resulted in two blocks. Block 1 retained ancestral genes representing clades A (diverging into an ancestor of the extant clade A4 pseudogenes) and C; clade B was apparently lost. Block 2 retained ancestral genes representing clades A (an ancestral clade A3 gene originated through duplication and divergence) and B; clade C was apparently lost. Support for this hypothetical tandem duplication event can be found in the BBE-like gene tree where the two corresponding group 5.2 BBE-like genes also comprise a *Cannabis*-specific duplication (fig. 2*A*). Finally, large-scale divergence of the second block led to the three variants of the divergent region described above. The tandemly repeated array of subclade A2_CBCAS genes most likely originated from a duplication of an ancestor of *THCAS* and translocation to another chromosomal region. Whole-genome sequencing studies revealed that lack of synteny between contigs comprising either *THCAS* or *CBDAS* genes is due to differential expansions of highly repetitive LTR retrotransposon elements (Laverty et al. 2019; Grassa et al. 2021). Therefore, flanking repetitive elements may have facilitated tandem duplication and/or translocation of cannabinoid oxidocyclase genes (Pisupati et al. 2018; Romero et al. 2020; Grassa et al. 2021).

The nature of the hypothesized ancestral cannabinoid oxidocyclase remains unknown (Vergara et al. 2019). Based on the observation that sequence variation was higher among "*CBDAS*"-like than among "*THCAS*"-like sequences, Onofri et al. (2015) considered *CBDAS* the ancestral type. We now know, however, that this perceived variation was due to the existence of additional gene lineages (fig. 3). Based on our gene tree, it is not possible to conclusively reconstruct whether the ancestral function was similar to that of THCAS, CBDAS, CBCAS, or another yet unidentified synthase (Vergara et al. 2019). In vitro functional studies suggest that these enzymes can produce multiple products and, therefore, perhaps the ancestral enzyme was promiscuous (Zirpel et al. 2018). Nevertheless, given the relative recent divergence of cannabinoid oxidocyclase genes it may be possible to reconstruct an ancestral protein sequence for functional testing with reasonable accuracy.

Re-evaluating Gene Copy Number Variation

Earlier studies have assessed cannabinoid oxidocyclase gene copy number with the aim to link this to chemical variation (McKernan et al. 2015; Vergara et al. 2019). We argue that, in light of our phylogenetic classification, these earlier results need to be reevaluated/reassessed. Claims have been made with regard to copy number variation in *THCAS* (McKernan et al. 2015; Vergara et al. 2019). However, we have found no instance of multiple copies of the subclade A1 functional THCAS gene (supplementary table S5, Supplementary Material online). This is in line with recent findings based on a similar comparative genomics approach (Hurgobin et al. 2021). Instead, copies counted toward "*THCAS*" copy number variation were due to tandem repeat copies of subclade A2. Similarly, although multiple types can be recognized within subclade B1_CBDAS they have been recovered exclusively as a single copy. Instead, we found copy number variation reported for *CBDAS* to be due to amplification of tandem repeat copies of subclade B2 pseudogenes in THCA-producing cultivars (McKernan et al. 2015; Vergara et al. 2019). Tight genetic linkage between *THCAS* and the subclade B2 tandemly repeated array (fig. 4; table 1) explains why "*CBDAS* copy number" was found to be positively correlated with the production of THCA and negatively with that of CBDA (Vergara et al. 2019).

Based on our classification, claims of *THCAS* and *CBDAS* copy number variation can be attributed to off-target amplification by the primers used (McKernan et al. 2015; Weiblen et al. 2015). Particularly illustrative in this regard is that sequences from clade C have been considered variants of *THCAS* or *CBDAS* in different studies, depending on the primers used for amplification (Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; Weiblen et al. 2015). Thus, based on currently available data, we consider *THCAS* and *CBDAS* each as single-copy genes and gene copy number variation to be restricted to the tandem duplications of subclades A2_CBCAS and B2, and clade C. Given that these (sub)clades may include a variable number of pseudogenes (fig. 4; table 1), it is not apparent how their copy number would have important functional relevance. For example, no correlation was found between the copy number of *CBCAS*-like genes and the production of CBCA (Vergara et al. 2019). Thus, we conclude that biosynthesis of the two major cannabinoids THCA and

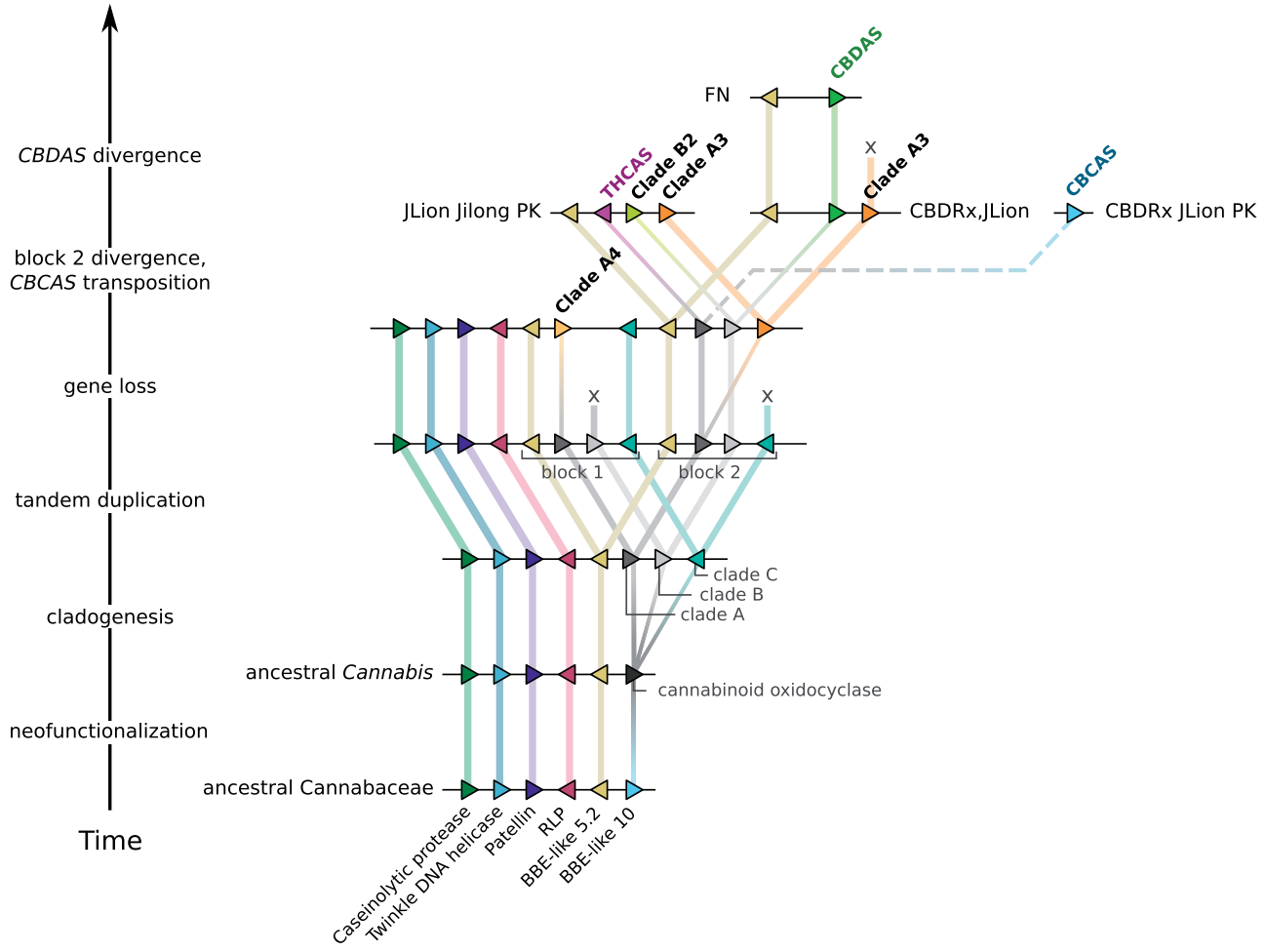

Fig. 5.—Evolutionary model of cannabinoid oxidocyclase gene duplication and diversification. Most parsimious hypothesis based on microsynteny patterns shown in figure 4 and phylogenetic reconstruction shown in figure 3. Triangles indicate genes (not to scale) colored according to their homology and putative orthologs are connected with colored lines. Dashed line indicates transposition of *CBCAS* to another syntenic block. Cannabinoid oxidocyclase genes are labeled in boldface. BBE, berberine bridge enzyme; RLP, receptor-like protein.

CBDA are the result of presence/absence, sequence variation, and expression of single-copy genes (Weiblen et al. 2015; McKernan et al. 2020; Wenger et al. 2020; Grassa et al. 2021).

Gene Sequence Variation and Potential Geographic Origins

Besides the divergence at the genomic level mentioned above, sequence variation within cannabinoid oxidocyclase gene sequences may help shed light on their evolutionary history. For example, subclade A4 pseudogenes are generally single copy but can be divided into two divergent types (supplementary table S2, Supplementary Material online; fig. 3, supplementary figs. S1 and S2, Supplementary Material online). Type 1 can be found in CBDA-dominant cultivars "Finola," "CBDRx," and on the *CBDAS*-containing haplotype of chemotype II cultivar "Jamaican Lion" (mother). Type 2 on the other hand can be found in chemotype I drug-type cultivars "Purple Kush", "LA confidential," and on the *THCAS*-containing haplotype of chemotype II cultivar "Jamaican Lion" (mother). Similarly, the full-length subclade A3 gene that is closely linked to *CBDAS* in the genome of chemotype III fiber-type cultivar "Finola" is sister to the subclade A3 pseudogenes closely linked to *THCAS* in drug-type cultivars "Purple Kush," "Jamaican Lion," and the plant from Jilong (fig. 3). These findings further corroborate our evolutionary interpretation of cannabinoid oxidocyclase gene duplication and divergence shown in figure 5 and suggest significant and consistent divergence between haplotypes containing *CBDAS* and *THCAS*. Thus, genomic divergence described above correlates with the prevalence of THCA and CBDA and, hence, perhaps with genetic origins of drug- versus fiber-type cultivars. Drug-type cultivars are considered to have originated in two different regions of the Himalayan foothills, while fiber-type cultivars are considered to have been developed

independently in Europe and in East Asia (Clarke and Merlin 2013; Clarke and Merlin 2016). The observed genetic variation may therefore be a consequence of divergence between these different geographic regions. Similarly, variation within *THCAS* sequences may also reflect geographic origin. Level of sequence divergence between *THCAS* from different geographic areas in South Korea is relatively high (supplementary table S3, Supplementary Material online; supplementary fig. S1, Supplementary Material online). Accessions from Boseung province share three unique aa substitutions, the accession from Jecheon province in Korea that has five unique aa substitutions, and sequences from Cheungsam share ten unique aa substitutions (Doh et al. 2019). Moreover, the *THCAS* sequence from the putatively wild accession from Jilong, Tibet is also relatively different from canonical *THCAS* and placed as a sister to all other clade A1 sequences (S. Gao et al. 2020). This suggests that additional sampling throughout the native range of *Cannabis* is likely to reveal additional genetic variation. However, germplasm from regions of origin is scarce, especially when restricting samples to those that are compliant with international regulations such as the Nagoya-protocol. We therefore strongly support earlier calls for increased efforts to develop well-curated public germplasm banks covering *Cannabis*' entire natural variation (Welling et al. 2016; Small 2018; Kovalchuk et al. 2020; McPartland and Small 2020).

Materials and Methods

Sequence Sampling

Sampling of Berberine Bridge Protein Sequences

We sampled full-length predicted berberine bridge protein sequences from the Eurosid clade based on whole-genome assemblies of *C. sativa* cultivar "CBDRx" (GCF_900626175.1; $N = 29$), *Humulus japonicus* cultivar "Cascade" (from http://hopbase.cgrb.oregonstate.edu; last accessed June 2021; $N = 24$), *Parasponia andersonii* (GCA_002914805.1; $N = 25$), *Trema orientalis*(GCA_002914845.1; $N = 26$), *Morus notabilis* (GCF_000414095.1; $N = 24$), *Medicago truncatula* (JCVI MedtrA17_4.0; $N = 23$), and *Arabidopsis thaliana* (TAIR10; $N = 29$) (supplementary table S1, Supplementary Material online) (Tang et al. 2014; Berardini et al. 2015; van Velzen et al. 2018; Grassa et al. 2021). Some *Cannabis* and *Humulus* genes were found to be misannotated or lacking an annotation. These were manually corrected based on alignment with a closely related and correctly annotated genome sequence. Because the CBDRx genome does not include *THCAS*, we included accession Q8GTB6.1 (Sirikantaramas et al. 2004). We indicated the putative orthologs of *Morus alba* Diels–Alderase (MaDA) and moracin C oxidase (MaMO) in *Morus notabilis* (L. Gao et al. 2020). In addition, we included daurichromenic acid synthase from *Rhododendron dauricum* (accession BAZ95780.1) and berberine bridge enzyme from Eschscholzia *californica* (accession AAC39358.1) (Hauschild et al. 1998; Iijima et al. 2017). Sequences of *A. thaliana* CYTOKININ OXIDASE 1 and 2 were used as outgroups.

Sampling of Cannabinoid Oxidocyclase Nucleotide Sequences

We mined available near chromosome-level genome assemblies of *C. sativa* for homologs of characterized cannabinoid oxidocyclase sequences (i.e., *THCAS*, *CBDAS*, and *CBCAS*) using BLASTP and BLASTN implemented in Geneious Prime 2019 (https://www.geneious.com; last accessed June 2021). This resulted in 13 cannabinoid-related genes from cultivar "CBDRx" (GCF_900626175.1), 26 from "Jamaican Lion" (mother: GCA_012923435.1) (McKernan et al. 2018; McKernan et al. 2020), 15 from "Finola" (GCA_003417725.2), 16 from "Purple Kush" (GCA_000230575.4) (van Bakel et al. 2011; Laverty et al. 2019), and 8 from a putatively wild plant from Jilong, Tibet (GCA_013030365.1) (S. Gao et al. 2020) We similarly mined sequences from additional genome assemblies of cultivars "Cannatonic" (GCA_001865755.1; $N = 11$), "Chemdog91" (GCA_001509995.1; $N = 5$), "Jamaican Lion" (father) (GCA_013030025.1; $N = 11$), "LA confidential" (GCA_001510005.1; $N = 8$), and "Pineapple Banana Bubble Kush" (GCA_002090435.1; $N = 11$) (supplementary table S4, Supplementary Material online) (McKernan et al. 2018; Vergara et al. 2019; McKernan et al. 2020). When necessary, structural annotations were manually modified based on nucleotide alignments with annotated genes with the highest identity. When genes comprised putative pseudogenes (i.e., coding sequence was fragmented due to premature stop codons and/or frame-shifts), they were annotated manually such that CDS remained homologous and all nonsense and frameshift mutations were indicated.

In addition, we compiled homologous nucleotide sequences available from ncbi databases the majority of which came from published studies (supplementary table S3, Supplementary Material online) (Sirikantaramas et al. 2004; Kojoma et al. 2006; Taura, Sirikantaramas, Shoyama, Yoshikai, et al. 2007; El Alaoui et al. 2013; McKernan et al. 2015; Onofri et al. 2015; Weiblen et al. 2015; Cascini et al. 2019; Doh et al. 2019). Based on preliminary analyses, some sequences described by Cascini et al. (2019) were found to have relatively high levels of ambiguous nucleotides, probably due to unspecific amplification of multiple genes or gene variants. Some sequences amplified from Moroccan hashish samples described by (El Alaoui et al. 2013) were suspected to be chimeric (probably due to differential specificity between forward and reverse sequencing primers). These ambiguous and suspected chimeric sequences were excluded from final analyses. The most closely related BBE-like sequences from *Cannabis* and *Humulus* were used as outgroups.

Phylogenetic Analyses

Multiple sequence alignment was performed with MAFFT v7.450 with automatic selection of appropriate algorithm, a gap open penalty of 1.26, and an offset value 0.123. For aligning the protein and nucleotide sequence data sets, we used the BLOSUM62 and 100PAM scoring matrix, respectively. Optimal models of sequence evolution as determined using Modeltest-NG v.0.1.5 on XSEDE via the CIPRES gateway (Miller et al. 2010; Darriba et al. 2020) were WAG+I+G4 for the protein data set and GTR+I+G4 for all nucleotide data sets. Gene trees were reconstructed in a Bayesian framework using MrBayes v 3.2.6 (Ronquist and Huelsenbeck 2003) implemented in Geneious Prime with a chain length of 2.2 million generations; sampling every 1,000th generation; 4 heated chains with a temperature of 0.2 and applying the optimal model of sequence evolution. The first 200,000 generations were discarded as burnin.

Within the berberine bridge gene family tree, clades were numbered in accordance with earlier classification (Daniel et al. 2016). Within the cannabinoid oxidocyclase gene tree, clades and types were characterized based on unique non-synonymous substitutions (i.e., substitutions resulting in a change to a specific amino acid that, based on current sampling, were unique for as well as constant within that clade) where possible. All site numbers were projected to the *THCAS* reference sequence described by Sirikantaramas et al. (2004) (accession AB057805.1) and, within the *THCAS* clade, type names were kept in accordance with those from Onofri et al. (2015).

Microsynteny Assessment

Nucleotide microsynteny was assessed for the cultivars "Jamaican Lion" (mother), "CBDRx," "Finola," "Purple Kush," and a putatively wild specimen from Jilong, Tibet. Because we found inconsistencies between the different genome assemblies in the ordering and orientation of sequences into scaffolds we considered genomic contigs only. Nucleotide-level alignments were generated by performing gapped extensions of high-scoring segment pairs using Lastz version 1.04.03. To avoid seeding in repetitive sequence we indexed unique words with single alignments only (–maxwordcount = 1; –masking = 1). To reduce runtime we set –notransition and –step = 20. To keep tandem repeats we set –nochain. To filter short and/or dissimilar alignments we set –hspthresh = 100,000, –filter = identity: 95, and –filter = nmatch: 2000. The –rdotplot output was used to generate alignment dotplots in R. For nucleotide level microsyntenic blocks of interest we further assessed microsynteny with genomic sequences from *H. lupulus* cultivar "Cascade," *P. andersonii*, and *T. orientalis* based on predicted protein sequences (van Velzen et al. 2018; Padgitt-Cobb et al. 2019).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Bob Harris (Pennsylvania State University, USA) for advice on generating dotplots based on Lastz output. Marleen Botermans (NVWA, The Netherlands) provided valuable comments on an early draft of the manuscript. Bastian Daniel (Graz University, Austria) kindly helped with BBE-like gene family classification. Suggestions from John McPartland (University of Vermont, USA) and two anonymous reviewers helped improve the text after submission.

Data Availability

The sequence data underlying this article are available in the GenBank Nucleotide Database at www.ncbi.nlm.nih.gov (last accessed June 2021) and can be accessed with the accession codes listed in supplementary tables S1, S3, and S4, Supplementary Material online. The multiple-sequence alignments and associated gene trees are available in the 4TU.ResearchData repository at https://doi.org/10.4121/13414694 (last accessed June 2021).

Literature cited

Ali S, Mufti M, Khan M, Aziz I. 2019. The identification of SNPs in *THCA synthase* gene from Pakistani *Cannabis*. Asia Pac J Mol Biol Biotechnol. 27:1–9.

Aryal N, Orellana DF, Bouie J. 2019. Distribution of cannabinoid synthase genes in non-*Cannabis* organisms. J Cannabis Res. 1(1):8.

van Bakel H, et al. 2011. The draft genome and transcriptome of *Cannabis sativa*. Genome Biol. 12(10):R102.

Berardini TZ, et al. 2015. The *Arabidopsis* information resource: making and mining the "gold standard" annotated reference plant genome. Genesis 53(8):474–485.

Bhattacharyya S, et al. 2018. Effect of cannabidiol on medial temporal, midbrain, and striatal dysfunction in people at clinical high risk of psychosis: a randomized clinical trial. JAMA Psychiatry. 75(11):1107–1117.

Brierley DI, et al. 2019. Chemotherapy-induced cachexia dysregulates hypothalamic and systemic lipoamines and is attenuated by cannabigerol. J Cachexia Sarcopenia Muscle. 10(4):844–859.

Cascini F, et al. 2019. Highly predictive genetic markers distinguish drug-type from fiber-type *Cannabis sativa* L. Plants 8(11):496.

Clarke RC, Merlin MD. 2013. *Cannabis*: evolution and ethnobotany. Berkeley (CA): University of California Press.

Clarke RC, Merlin MD. 2016. *Cannabis* domestication, breeding history, present-day genetic diversity, and future prospects. CRC Crit Rev Plant Sci. 35(5-6):293–327.

Daniel B, et al. 2017. The family of berberine bridge enzyme-like enzymes: a treasure-trove of oxidative reactions. Arch Biochem Biophys. 632:88–103.

Daniel B, et al. 2016. Structure of a berberine bridge enzyme-like enzyme with an active site specific to the plant family Brassicaceae. PLoS One. 11(6):e0156892.

Darriba D, et al. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol. 37(1):291–294.
De Meijer EPM, Hammond KM. 2005. The inheritance of chemical phenotype in *Cannabis sativa* L. (II): cannabigerol predominant plants. Euphytica 145(1-2):189–198.
Doh EJ, et al. 2019. DNA markers to discriminate *Cannabis sativa* L. "Cheungsam" with low tetrahydrocannabinol (THC) content from other South Korea cultivars based on the nucleotide sequences of tetrahydrocannabinolic acid synthase and putative 3-ketoacyl-CoA synthase genes. Evid Based Complement Alternat Med. 2019:8121796.
van de Donk T, et al. 2019. An experimental randomized study on the analgesic effects of pharmaceutical-grade cannabis in chronic pain patients with fibromyalgia. Pain 160(4):860–869.
El Alaoui MA, et al. 2013. Extraction of high quality DNA from seized Moroccan *Cannabis* resin (Hashish). PLoS One. 8(10):e74714.
ElSohly MA, Radwan MM, Gul W, Chandra S, Galal A. 2017. Phytochemistry of *Cannabis sativa* L. In: Kinghorn A, Falk H, Gibbons S, Kobayashi J, editors. Phytocannabinoids. Progress in the Chemistry of Organic Natural Products, vol 103. Cham: Springer International Publishing. p. 1–36.
Fellermeier M, Zenk MH. 1998. Prenylation of olivetolate by a hemp transferase yields cannabigerolic acid, the precursor of tetrahydrocannabinol. FEBS Lett. 427(2):283–285.
Fournier G, Beherec O, Bertucelli S. 2004. Santhica 23 et 27: deux variétés de chanvre (*Cannabis sativa* L.) sans Δ-9-THC. Ann Toxicol Anal. 16(2):128–132.
Gagne SJ, et al. 2012. Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique catalytic route to plant polyketides. Proc Natl Acad Sci U S A. 109(31):12811–12816.
Gao L, et al. 2020. FAD-dependent enzyme-catalysed intermolecular [4 + 2] cycloaddition in natural product biosynthesis. Nat Chem. 12(7):620–628.
Gaoni Y, Mechoulam R. 1964. Isolation, structure, and partial synthesis of an active constituent of hashish. J Am Chem Soc. 86(8):1646–1647.
Gao S, et al. 2020. A high-quality reference genome of wild *Cannabis sativa*. Hortic Res. 7(1):73.
Garfinkel AR, Otten M, Crawford S. 2021. SNP in potentially defunct tetrahydrocannabinolic acid synthase is a marker for cannabigerolic acid dominance in *Cannabis sativa* L. Genes 12(2):228.
Gofshteyn JS, et al. 2017. Cannabidiol as a potential treatment for febrile infection-related epilepsy syndrome (FIRES) in the acute and chronic phases. J Child Neurol. 32(1):35–40.
Grassa CJ, et al. 2021. A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. New Phytol. 230(4):1665–1679.
Grimison P, et al. 2020. Oral THC: CBD cannabis extract for refractory chemotherapy-induced nausea and vomiting: a randomised, placebo-controlled, phase II crossover trial. Ann. Oncol. 31(11):1553–1560. doi: 10.1016/j.annonc.2020.07.020.
Gülck T, Møller BL. 2020. Phytocannabinoids: origins and biosynthesis. Trends Plant Sci. 25(210):985–1004. doi: 10.1016/j.tplants.2020.05.005.
Han H, et al. 2018. Chalcomoracin is a potent anticancer agent acting through triggering oxidative stress via a mitophagy- and paraptosis-dependent mechanism. Sci. Rep. 8(9566):1–14.
Hauschild K, Pauli HH, Kutchan TM. 1998. Isolation and analysis of a gene bbe1 encoding the berberine bridge enzyme from the California poppy *Eschscholzia californica*. Plant Mol Biol. 36(3):473–478.
Hazekamp A, Tejkalová K, Papadimitriou S. 2016. *Cannabis*: from cultivar to chemovar II—a metabolomics approach to cannabis classification. Cannabis Cannabinoid Res. 1(1):202–215.
Hurgobin B, et al. 2021. Recent advances in *Cannabis sativa* genomics research. New Phytol. 230(1):73–89.
Iijima M, et al. 2017. Identification and characterization of daurichromenic acid synthase active in anti-HIV biosynthesis. Plant Physiol. 174(4):2213–2230.
Jin J, et al. 2020. Born migrators: historical biogeography of the cosmopolitan family Cannabaceae. J Syst Evol. 58(4):461–473.
Kitamura M, Aragane M, Nakamura K, Watanabe K, Sasaki Y. 2016. Development of loop-mediated isothermal amplification (LAMP) assay for rapid detection of *Cannabis sativa*. Biol Pharm Bull. 39(7):1144–1149.
Kojoma M, Seki H, Yoshida S, Muranaka T. 2006. DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type" *Cannabis sativa* L. Forensic Sci Int. 159(2-3):132–140.
Kovalchuk I, et al. 2020. The genomics of *Cannabis* and its close relatives. Annu Rev Plant Biol. 71(1):713–739.
Laverty KU, et al. 2019. A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci. Genome Res. 29(1):146–156.
Liu Z, et al. 2020. Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. New Phytol. 227(4):1109–1123.
Luo X, et al. 2019. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. Nature. 567(7746):123–126.
Lynch RC, et al. 2016. Genomic and chemical diversity in cannabis. CRC Crit Rev Plant Sci. 35(5-6):349–363.
Mandolino G, Bagatta M, Carboni A, Ranalli P, de Meijer E. 2003. Qualitative and quantitative aspects of the inheritance of chemical phenotype in *Cannabis*. J Ind Hemp. 8(2):51–72.
McKernan K, et al. 2018. Cryptocurrencies and Zero Mode Wave guides: an unclouded path to a more contiguous *Cannabis sativa* L. genome assembly. Open Science Framework. Available from: http://dx.doi.org/10.31219/osf.io/7d968. Accessed June 2021.
McKernan KJ, et al. 2020. Sequence and annotation of 42 *Cannabis* genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes. bioRxiv: 2020.01.03.894428. Available from: https://doi.org/10.1101/2020.01.03.894428. Accessed June 2021.
McKernan KJ, et al. 2015. Single molecule sequencing of *THCA synthase* reveals copy number variation in modern drug-type *Cannabis sativa* L. bioRxiv: 028654. Available from: https://doi.org/10.1101/028654. Accessed June 2021.
McPartland JM, Guy GW. 2017. Models of *Cannabis* taxonomy, cultural bias, and conflicts between scientific and vernacular names. Bot Rev. 83(4):327–381.
McPartland JM, Small E. 2020. A classification of endangered high-THC cannabis (*Cannabis sativa* subsp. indica) domesticates and their wild relatives. PhytoKeys 144:81–112.
Mechoulam R. 2005. Plant cannabinoids: a neglected pharmacological treasure trove. Br J Pharmacol. 146(7):913–915.
Mechoulam R, Parker LA. 2013. The endocannabinoid system and the brain. Annu Rev Psychol. 64:21–47.
de Meijer EPM, et al. 2003. The inheritance of chemical phenotype in *Cannabis sativa* L. Genetics 163(1):335–346.
Miller MA, Pfeiffer W, Schwartz T. 2010. 2010 Gateway Computing Environments Workshop (GCE) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: ieeexplore.ieee.org. p. 1–8.
Morimoto S, Komatsu K, Taura F, Shoyama Y. 1997. Enzymological evidence for cannabichromenic acid biosynthesis. J Nat Prod. 60(8):854–857.
Morimoto S, Komatsu K, Taura F, Shoyama Y. 1998. Purification and characterization of cannabichromenic acid synthase from *Cannabis sativa*. Phytochemistry 49(6):1525–1529.

Onofri C, de Meijer EPM, Mandolino G. 2015. Sequence heterogeneity of cannabidiolic- and tetrahydrocannabinolic acid-synthase in *Cannabis sativa* L. and its relationship with chemical phenotype. Phytochemistry 116:57–68.
Padgitt-Cobb LK, et al. 2019. A phased, diploid assembly of the Cascade hop (*Humulus lupulus*) genome reveals patterns of selection and haplotype variation. bioRxiv: 786145. Available from: 10.1101/786145. Accessed June 2021.
Pertwee RG. 2005. Inverse agonism and neutral antagonism at cannabinoid CB1 receptors. Life Sci. 76(12):1307–1324.
Pisupati R, Vergara D, Kane NC. 2018. Diversity and evolution of the repetitive genomic content in *Cannabis sativa*. BMC Genomics. 19(1):156.
Raharjo TJ, Chang W-T, Choi YH, Peltenburg-Looman AMG, Verpoorte R. 2004. Olivetol as product of a polyketide synthase in *Cannabis sativa* L. Plant Sci. 166(2):381–385.
Rodziewicz P, Loroch S, Marczak Ł, Sickmann A, Kayser O. 2019. Cannabinoid synthases and osmoprotective metabolites accumulate in the exudates of *Cannabis sativa* L. glandular trichomes. Plant Sci. 284:108–116.
Romero P, Peris A, Vergara K, Matus JT. 2020. Comprehending and improving cannabis specialized metabolism in the systems biology era. Plant Sci. 298:110571.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Rustichelli C, Ferioli V, Baraldi M, Zanoli P, Gamberini G. 1998. Analysis of cannabinoids in fiber hemp plant varieties (*Cannabis sativa* L.) by high-performance liquid chromatography. Chromatographia 48(3-4):215–222.
Sawler J, et al. 2015. The genetic structure of marijuana and hemp. PLoS One 10:e0133292.
Sirikantaramas S, et al. 2004. The gene controlling marijuana psychoactivity: molecular cloning and heterologous expression of Delta-1-tetrahydrocnnabinolic acid synthasefrom *Cannabis sativa* L. J Biol Chem. 279(38):39767–39774.
Sirikantaramas S, et al. 2005. Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes. Plant Cell Physiol. 46(9):1578–1582.
Skelley JW, Deas CM, Curren Z, Ennis J. 2020. Use of cannabidiol in anxiety and anxiety-related disorders. J Am Pharm Assoc. 60(1):253–261.
Small E. 2018. Dwarf germplasm: the key to giant Cannabis hempseed and cannabinoid crops. Genet Resour Crop Evol. 65(4):1071–1107.
Small E, Cronquist A. 1976. A practical and natural taxonomy for *Cannabis*. Taxon 25(4):405–435.
Stout JM, Boubakir Z, Ambrose SJ, Purves RW, Page JE. 2012. The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in *Cannabis sativa* trichomes. Plant J. 71(3):353–365.
Suraev AS, et al. 2020. Cannabinoid therapies in the management of sleep disorders: A systematic review of preclinical and clinical studies. Sleep Med Rev. 53:101339.
Tang H, et al. 2014. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics. 15:312.
Taura F, Morimoto S, Shoyama Y. 1996. Purification and characterization of cannabidiolic-acid synthase from *Cannabis sativa* L. biochemical analysis of a novel enzyme that catalyzes the oxidocyclization of cannabigerolic acid to cannabidiolic acid. J Biol Chem. 271(29):17411–17416.
Taura F, Morimoto S, Shoyama Y, Mechoulam R. 1995. First direct evidence for the mechanism of.DELTA.1-tetrahydrocannabinolic acid biosynthesis. J Am Chem Soc. 117(38):9766–9767.
Taura F, Sirikantaramas S, Shoyama Y, Shoyama Y, Morimoto S. 2007. Phytocannabinoids in *Cannabis sativa*: recent studies on biosynthetic enzymes. Chem Biodivers. 4(8):1649–1663.
Taura F, et al. 2007. Cannabidiolic-acid synthase, the chemotype-determining enzyme in the fiber-type *Cannabis sativa*. FEBS Lett. 581(16):2929–2934.
Taura F, et al. 2009. Characterization of olivetol synthase, a polyketide synthase putatively involved in cannabinoid biosynthetic pathway. FEBS Lett. 583(12):2061–2066.
Toth JA, et al. 2020. Development and validation of genetic markers for sex and cannabinoid chemotype in *Cannabis sativa* L. GCB Bioenergy. 12(3):213–222.
Udoh M, Santiago M, Devenish S, McGregor IS, Connor M. 2019. Cannabichromene is a cannabinoid CB2 receptor agonist. Br J Pharmacol. 176(23):4537–4547.
van Velzen R, et al. 2018. Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. Proc Natl Acad Sci U S A. 115(20):E4700–E4709. doi: 10.1073/pnas.1721395115.
Vergara D, et al. 2016. Genetic and genomic tools for *Cannabis sativa*. CRC Crit Rev Plant Sci. 35(5-6):364–377.
Vergara D, et al. 2019. Gene copy number is associated with phytochemistry in *Cannabis sativa*. AoB Plants. 11(6):plz074.
Weiblen GD, et al. 2015. Gene duplication and divergence affecting drug content in *Cannabis sativa*. New Phytol. 208(4):1241–1250.
Welling MT, et al. 2016. A belated green revolution for cannabis: virtual genetic resources to fast-track cultivar development. Front Plant Sci. 7:1113.
Wenger JP, et al. 2020. Validating a predictive model of cannabinoid inheritance with feral, clinical, and industrial *Cannabis sativa*. Am J Bot. 107(10):1423–1432.
Winkler A, et al. 2008. A concerted mechanism for berberine bridge enzyme. Nat Chem Biol. 4(12):739–741.
Zirpel B, Kayser O, Stehle F. 2018. Elucidation of structure-function relationship of THCA and CBDA synthase from *Cannabis sativa* L. J Biotechnol. 284:17–26.

Associate editor: Yves Van De Peer