




Leaf pigmentation in *Cannabis sativa*: Characterization of anthocyanin biosynthesis in colorful Cannabis varieties

Kristina K. Gagalova^{1,2}  | Yifan Yan³ | Shumin Wang⁴ | Till Matzat³ |
 Simone D. Castellarin³ | Inanc Birol^{2,5} | David Edwards⁶  | Mathias Schuetz^{4,7} 

¹Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Perth, WA, Australia

²Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada

³Wine Research Centre, University of British Columbia, Vancouver, BC, Canada

⁴Department of Botany, University of British Columbia, Vancouver, BC, Canada

⁵Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

⁶School of Biological Sciences and Institute of Agriculture, University of Western Australia, Crawley, Western Australia, Australia

⁷Department of Biology, Kwantlen Polytechnic University, Surrey, BC, Canada

Correspondence

Mathias Schuetz, Department of Biology, Kwantlen Polytechnic University, 12666 72 Ave, Surrey, BC V3W2M8, Canada.
 Email: mathias.schuetz@kpu.ca

Funding information

MITACS ACCELERATE grant (#IT21175) to K.K.G., I.B.; Canada Foundation for Innovation/Infrastructure Operating Fund and British Columbia Knowledge Development Fund (BCKDF) F15-03053 to S. C., Natural Sciences and Engineering Research Council of Canada (NSERC) College and Community Innovation Program (CCIP) Innovation Enhancement (IE) Grant (CCIP 555874-20) to M.S.

Abstract

Cannabis plants produce a spectrum of secondary metabolites, encompassing cannabinoids and more than 300 non-cannabinoid compounds. Among these, anthocyanins have important functions in plants and also have well documented health benefits. Anthocyanins are largely responsible for the red/purple color phenotypes in plants. Although some well-known Cannabis varieties display a wide range of red/purple pigmentation, the genetic underpinnings of anthocyanin biosynthesis have not been well characterized in Cannabis. This study unveils the genetic diversity of anthocyanin biosynthesis genes found in Cannabis, and we characterize the diversity of anthocyanins and related phenolics found in four differently pigmented Cannabis varieties. Our investigation revealed that the genes *4CL*, *CHS*, *F3H*, *F3'H*, *FLS*, *DFR*, *ANS*, and *OMT* exhibited the strongest correlation with anthocyanin accumulation in Cannabis leaves. The results of this study enhance our understanding of the anthocyanin biosynthetic pathway and shed light on the molecular mechanisms governing Cannabis leaf pigmentation.

KEYWORDS

2-ODD enzymes, anthocyanins, Cannabis, flavonols, leaf pigmentation

1 | INTRODUCTION

Cannabis sativa is among the earliest cultivated plants and has been used for food and fiber production as well as for medicinal and recreational use (McPartland et al., 2019; Rull, 2022; Small, 2015). Cannabis

plants develop glandular trichomes on their leaves and flowers, which accumulate a diverse array of secondary metabolites. Cannabinoids such as tetrahydrocannabinolic acid (THCA) and cannabidiolic acid (CBDA) are the most well-known Cannabis-derived metabolites (EISOhly & Gul, 2014; Turner et al., 1980). Due to its intoxicating

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Plant Direct* published by American Society of Plant Biologists and the Society for Experimental Biology and John Wiley & Sons Ltd.



effects, the THCA/THC content in Cannabis is a key factor in its classification, with higher levels resulting in stricter legal controls (e.g., as a drug type). In contrast to drug-type Cannabis, hemp-type Cannabis is required to contain less than .3% per dry weight of THCA/THC but is permitted to contain higher levels of other non-intoxicating cannabinoids such as CBDA/CBD.

In addition to cannabinoids, Cannabis plants produce a wide range of other non-cannabinoid molecules, including a diverse array of terpenoids and flavonoids (ElSohly & Slade, 2005). While cannabinoids like THC and CBD often take the spotlight, terpenoids and flavonoids are crucial players in what is referred to as the “entourage effect” (Bautista et al., 2021; Ferber et al., 2020; Koltai & Namdar, 2020; Russo, 2011). This phenomenon highlights the synergistic interactions between cannabinoids and these other metabolites, collectively enhancing and modulating the activity of THC and CBD.

Cannabis accumulates a diverse array of terpenoids, which have been well characterized in previous studies (reviewed in Booth & Bohlmann, 2019; Sommano et al., 2020). However, the diversity of flavonoids found in Cannabis remains relatively understudied until recently (Bassolino et al., 2023; Cerrato et al., 2021, 2023; Izzo et al., 2020). Flavonoids are a group of phenolic molecules that have important functions in plants, including defense against pathogen infection and countering the damage caused by UV radiation (Mierziak et al., 2014; Verdan et al., 2011). Different flavonoids are grouped into different subtypes based on their chemical structure, specifically the arrangement of their carbon atoms and the presence of certain functional groups. Anthocyanins and flavonols are two main flavonoid subtypes and are well known for their antioxidant and anti-inflammatory properties as part of the human diet (Khoo et al., 2017; Kumar & Pandey, 2013). The accumulation of anthocyanins is responsible for the red, purple, and blue colors found in many plants, such as blueberries, rose flowers, and chard leaves. Although Cannabis plants have diverse pigmentation patterns that are often reflected in the naming schemes of some varieties (e.g., Purple Kush, Purple Haze, Pink Kush), the diversity of anthocyanins found in Cannabis have not been studied in detail until recently. Analysis of Cannabis leaves and reproductive tissues identified cyanidin-3-rutinoside as the major anthocyanin in hemp-type Cannabis (Bassolino et al., 2023). Moreover, three LC–MS studies have identified flavonols and other phenolic compounds present in the flowers and leaves of hemp-type Cannabis (Izzo et al., 2020; Cerrato et al., 2021, 2023). These previous studies analyzed hemp-type Cannabis, likely due to the regulatory hurdles of working with drug-type Cannabis, which we examine in this study.

The biosynthesis of flavonoids such as anthocyanins and flavonols commences via the general phenylpropanoid biosynthetic pathway, orchestrated by the enzyme phenylalanine ammonia lyase (PAL) (Vogt, 2010). This biosynthetic step involves the deamination of the amino acid (aa) phenylalanine. Subsequent activity of cinnamate 4-hydroxylase (C4H) and 4-coumarate-CoA ligase (4CL) leads to the formation of *p*-coumaroyl CoA, which is a precursor molecule for the biosynthesis of various phenolic compounds, including flavonoids, lignin, and stilbenes (Vogt, 2010). Chalcone synthase (CHS) operates as a gatekeeper for flavonoid formation by directing *p*-coumaroyl CoA

toward the synthesis of flavonoids. The branch point within the flavonoid biosynthesis pathway that channels metabolic intermediates toward flavonol and anthocyanin production occurs via the action of flavonol synthase (FLS) and dihydroflavonol 4-reductase (DFR), respectively (Martens et al., 2010). FLS and DFR compete for dihydroflavonol, which is a key intermediate molecule for both flavonol and anthocyanin formation. Flavonols and anthocyanins can be further modified via methylation, acetylation, prenylation, or through conjugation to various sugar molecules (Castañeda-Ovando et al., 2009; Dias et al., 2021). For example, anthocyanidins are typically conjugated to various sugar moieties to form anthocyanins, which are colorful pigments that typically accumulate in plant vacuoles (Castañeda-Ovando et al., 2009; Passeri et al., 2016). The enzymes involved in attaching sugars to flavonols and anthocyanidins are broadly categorized as uridine diphosphate-glucosyltransferases (UGT) (Louveau & Osbourn, 2019). Among the UGT family, UDP-glucose: flavonoid 3-O-glucosyltransferases (UGFT) (Zhao et al., 2012) are a distinct subgroup with a specialized function, specifically conjugating glucose to the 3-position of flavonols.

Recent studies have begun characterizing flavonoids in *C. sativa* in terms of metabolites (Bassolino et al., 2023; Cerrato et al., 2021, 2023; Izzo et al., 2020), biosynthesis genes (Bassolino et al., 2020), and transcriptional regulation (Bassolino et al., 2020; Kundan et al., 2022). These studies are noteworthy considering the extensive range of physiological effects that flavonols and anthocyanins can exert on plant physiology, as well as their direct impact on the aesthetic qualities of Cannabis plants. Indeed, Cannabis exhibits a wide spectrum of observable phenotypes, with perhaps one of the most prominent being the coloration of leaves and flowers (Aardema & DeSalle, 2021) (Figure 1). The diversity of colors was likely selected for by clandestine Cannabis breeders during the long period of Cannabis prohibition. In this study, we provide a comprehensive inventory of flavonoid biosynthesis genes in the Cannabis cs10 reference genome. We identified these genes using a knowledge-based identification of pathway enzymes (KIPes) approach that evaluates the functionally significant aa residues and domains within the peptide sequences to accurately identify biosynthesis enzymes (Pucker et al., 2020). We document the biochemical phenotypes of four drug-type Cannabis varieties and evaluate the phenylpropanoid/flavonoid biosynthesis gene expression in the leaves of these varieties using RNAseq. We extend our analysis to encompass the functional characterization of three 2-oxoglutarate and oxygen-dependent dioxygenases (2-ODD) genes (anthocyanidin synthase [ANS], FLS, and flavanone 3-hydroxylase [F3H]), which are responsible for facilitating key oxidation reactions in the flavonoid biosynthesis pathway.

2 | MATERIALS AND METHODS

2.1 | Plant material

Four different Cannabis varieties (CA19210, CK19206, Cali Kush, and Willow-alpha) were cultivated under uniform conditions by a licensed

Willow-alpha

CA19210

CK19206

Cali Kush

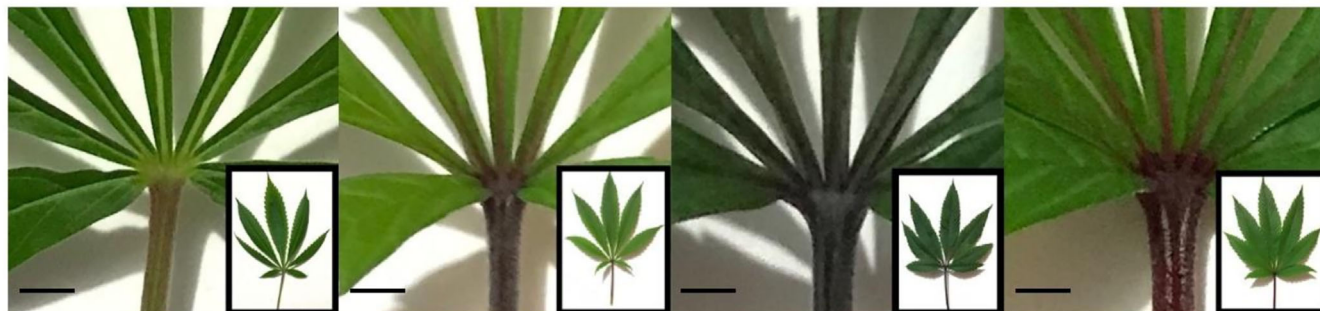


FIGURE 1 Leaf pigmentation phenotypes in four different Cannabis varieties. Petiole and leaf pigmentation phenotypes are shown for Willow-alpha (unpigmented phenotype), CA19210, CK19206, and Cali Kush (pigmented phenotypes). A magnified view of the petiole and primary leaf veins is shown for each variety as well as the overall leaf morphology (inset). Scale bar = .5 cm.

Cannabis producer in British Columbia, Canada. The four plant varieties were selected by visually assessing their leaf pigmentation (Figure 1). Among them, the Cali Kush and CK19206 varieties exhibit pronounced leaf pigmentation with the presence of red/purple coloration. The CA19210 variety exhibited a moderate level of leaf pigmentation, while the Willow-alpha strain lacked any discernible red/purple pigmentation. Plant clones were generated via cuttings and using Stim-Root #1 rooting powder (Plant Products Co. Ltd., Leamington, Canada), followed by planting into reconstituted Jiffy #7 peat pellets (Jiffy Growing Solutions, Zwijndrecht, The Netherlands). Plants were grown at 100% humidity under a plastic dome until roots became visible. The clones were individually potted into 3-gal plastic nursery pots using Sunshine Mix #4 professional growing mix (Sun Gro Horticulture, Agawam, USA). Plants were grown at temperatures between 22 and 24°C, and using NextLight Veg8 240 W full spectrum LED lights (NextLight, Cincinnati, USA) using an 18 h/6 h cycle (hours of light/hours of dark) during clone establishment and vegetative development. After 4 weeks of vegetative development, the plants were transferred to a neighboring grow room equipped with Next Light Mega Pro 640 W full spectrum LED lights (Nextlight, Cincinnati, USA) using a 12 h/12 h cycle (hours of light/hours of dark) to facilitate reproductive development (e.g., flowering). Plants were watered as needed and fertilized using Miracle-Gro All Purpose (24-8-16) fertilizer during vegetative growth and Miracle-Gro Ultra Bloom (15-30-15) during flowering. Biological controls consisting of *Amblyseius californicus* (Biobest, Westerlo, Belgium) were applied at regular intervals to the plants to proactively mitigate spider mite pests. No visible signs of pests or disease were observed throughout the experiment. Leaf collection took place 3 weeks after transfer to the flowering conditions coinciding with the shift to a 12-h light/dark photoperiod. At the time of leaf collection, Cannabis flowers were observed but not yet mature. For each of the four varieties, one fully expanded fan leaf was collected from three individual plant clones and flash frozen in liquid nitrogen prior to further processing for RNA and metabolite extraction. Moreover, the apical flower bud for each plant of the four varieties was harvested after 7 weeks post-transfer

to the flowering conditions. These apical flower buds were flash frozen in liquid nitrogen and subsequently used to determine the cannabinoid content of the four Cannabis plant varieties.

2.2 | Identification of genes involved in the general phenylpropanoid, flavonoid, and anthocyanin pathways

Genes responsible for enzymes within the flavonoid, anthocyanin, and phenylpropanoid pathways were identified as follows. Initially, potential candidates were identified using the KIPes pipeline v0.38 (Pucker et al., 2020), which was executed in its default mode using protein sequences from the cs10 Cannabis genome reference (Grassa et al., 2021). Subsequently, the chosen sequences were subjected to validation as described below, via manual curation using assessments of alignments, phylogenetic trees, and the presence of conserved binding or catalytic residues. Research into the annotation of flavonoid-related genes has been highly successful in *Arabidopsis* (Saito et al., 2013), making this system ideal for characterizing flavonoid biosynthesis genes in other plants. All candidate proteins with a sequence identity greater than 40% with their corresponding *Arabidopsis* homologs were retained. Additionally, proteins lacking more than 80% of their conserved residues according to KIPes were discarded from the selection. Further validation was applied to the enzymes classified as 2-ODD, including FLS, ANS, and F3H that were aligned with multiple sequence alignment (MSA) in Mafft v7.453a (Nakamura et al., 2018) using the argument *-auto*. Following the alignment, each enzyme was individually examined for its catalytic and binding sites.

The annotated enzymes' functional residues (such as catalytic, binding, or structural aminoacids) were identified according to peer-reviewed studies or manual assertion based on sequence similarity to entries in the UniProt database (www.uniprot.org) (The UniProt Consortium et al., 2021) as follows. To annotate the functional residues, BLASTP (Camacho et al., 2009) was employed

to align the protein sequences of the Cannabis predicted protein sequences with those of Arabidopsis. In doing so, residues at specified positions were compared to ascertain their correspondence between the two species, and the active residues of these annotated enzymes were examined, specifically focusing on the catalytic and binding sites.

A sequence clustering approach was employed to categorize the UGT enzymes, utilizing Arabidopsis UGTs as a reference dataset, sourced from the P450 database (<http://www.P450.kvl.dk>) (Paquette et al., 2009). The MSA was conducted using Clustal Omega v1.2.4 (Sievers & Higgins, 2014) with default parameters. Subsequently, a phylogenetic tree was constructed through a neighbor-joining clustering method using FastTree2 (Price et al., 2010), where the local support values are used to score the reliability of the final tree. Cannabis candidate UGTs are assigned to the subclasses based on similarity with Arabidopsis UGTs.

All the identified enzyme sequences were compared with the Arabidopsis reference genome TAIR10/Ararport11 (Cheng et al., 2017) using BLASTP. The percent identity and percent positive scoring residues (residues with similar biochemical properties) were then used to determine the degree of conservation between the two species.

2.3 | Gene expression analysis

One fully expanded leaf was collected from three individual plant clones for each genotype and was used for RNA extraction and analysis. Frozen leaf tissue was ground in liquid nitrogen using a mortar and pestle. RNA was isolated using PureLink Plant RNA reagent (Thermo Fisher Scientific, Waltham, Massachusetts) according to the manufacturer's instructions. RNA sequencing was performed by Genome Quebec using an Illumina NovaSeq 6000 PE (2 × 100 bp). A total of 25 million reads were generated per sample and screened for possible contamination with BioBloomTools v2.3.3 (Chu et al., 2014). A series of Bloom filters were built using a low negative rate ($kmer_size = 25$, $desired_false_positive_rate = .001$, $number_of_hash_functions = 9$) based on the genomes from viruses, archaea, protozoa, bacteria, fungi, aphids (superfamily Aphidoidea), mites (subclass Acari), thrips (order Thysanoptera), and Univec build #10, downloaded in September 2020. The reads without a match to any Bloom filter were kept for quantification.

The cs10 genome reference (Grassa et al., 2021) was downloaded from NCBI, accession ID GCF900626175, and used for alignment. The quantification was conducted through pseudoalignment with Kallisto v0.46 (Bray et al., 2016). The transcript abundances estimated by Kallisto were imported in R and analyzed with differential expression analysis in DESeq2 v1.38.2 (Love et al., 2014). *p*-value statistical significance is adjusted with Benjamini–Hochberg for multiple-testing correction at false discovery rate (FDR .1). Genes with an absolute log₂ fold change higher than 1.5 at *p*-value .05 statistical significance are considered differentially expressed. The RNAseq data generated in this study have been deposited in the NCBI's Sequence Read Archive.

2.4 | Metabolite profiling

Ground leaf powder matching the samples from the transcriptome profiling was used for metabolite analysis. Total phenolics were extracted from ground-frozen leaf tissues using acidified methanol (v/v, methanol:water:formic acid, 49.5:49.5:1). A total of 100 mg of ground leaf tissue per ml of acidified methanol was incubated at room temperature for 2 h with gentle shaking (100 RPM) on an orbital shaker. Samples were centrifuged at room temperature for 10 min to pellet the leaf material, and 1 mL of extraction buffer was aspirated and subsequently used for analysis using high-performance liquid chromatography and mass spectrometry as described in Yan et al. (2020). Five microliters of each extract were injected into an Agilent 1100 Series LC coupled to an MSD Trap XCT Plus System (Agilent Technologies, Mississauga, Canada). Chromatographic separation was performed using an Agilent ZORBAX SB-C18 Column (1.8 μm, 4.6 × 50 mm)—(Agilent Technologies, Mississauga, Canada), with the temperature set to 67.0°C. The mobile phases consisted of aqueous formic acid (98:2, v/v; Solvent A) and acetonitrile/formic acid (98:2, v/v; Solvent B). The LC separation employed a binary solvent gradient at a flow rate of 1.20 mL/min. The gradient conditions were as follows: .20 min, 5.0% Solvent B; 6.00 min, 20.0% Solvent B; 9.00 min, 80.0% Solvent B; 10.00 min, 90.0% Solvent B; 10.10 min, 90.0% Solvent B; 11.00 min, 5.0% Solvent B, stopped at 11.50 min. Mass spectra were generated via electrospray ionization (ESI) in both positive and negative modes. Anthocyanins and flavonols were identified by the following criteria: (i) comparing the retention time and elution order of the identified peaks with authentic standards when available (i.e., cyanidin 3-O-glucoside, cyanidin 3-O-rutinoside, peonidin 3-O-glucoside, and peonidin 3-O-rutinoside); (ii) matching the mass spectra of the identified peaks with the published data (Radwan et al., 2021; Tulio et al., 2008; Wu et al., 2004). The identities of selected compounds (i.e., cyanidin 3-O-glucoside, cyanidin 3-O-rutinoside, peonidin 3-O-glucoside, and peonidin 3-O-rutinoside) were further confirmed using LC Q-TOF (Agilent 1200 UHPLC/6530B coupled with accurate Mass Q-TOF). Quantification of anthocyanins was based on a standard curve of cyanidin 3-O-glucoside (Cat. # 09155) with malvidin 3,5-diglucoside (Cat. # 09305) as an internal standard. The maximal absorption for anthocyanins is reported to be between 512 and 528 nm (Lee et al., 2005). The quantification results from LC-QTOF were corroborated using HPLC-UV/VIS by monitoring 512–528 nm wavelengths (Figure S1). Flavonol quantification was based on a standard curve of quercetin 3-O-glucoside (Cat. #1099) with baicalein (Cat. #14005) as an internal standard. Cyanidin 3-O-rutinoside and peonidin 3-O-rutinoside were verified using LC-QTOF (Agilent) analysis against authentic standards (Cat #0914S and 0945, respectively). All standards were purchased from Extrasynthese (Genay, France).

The cannabinoid content of leaves and flowers was determined as follows. A total of 200 mg of ground frozen leaf and flower material was weighed into 15 mL centrifuge tubes and combined with 4 mL of extraction solvent (9:1 MeOH:chloroform). Samples were vortexed at maximum for 2 min, followed by gently shaking (100 RPM) on an orbital shaker at room temperature for 15 min. The sample tubes were



centrifuged (4000 RPM) using a tabletop centrifuge to pelletize the plant material. A total of 1 mL of extraction solvent was aspirated, diluted tenfold with fresh extraction solvent, and used for ultra-high-performance liquid chromatography UV analysis. Chromatographic separation was performed using a Vanquish Duo UHPLC system with an Accucore Vanquish C18+ Column (1.5 μ m, 2.1 \times 50 mm) from Thermo Fisher Scientific, Waltham, Massachusetts. Column temperature set to 50.0°C, and the mobile phases consisted of aqueous formic acid (.1% formic acid in deionized water, v/v; Solvent A) and methanol/acetonitrile/formic acid (.1% formic acid: 75% methanol: 25% acetonitrile, v/v; Solvent B). LC separation employed a binary solvent gradient at a flow rate of .6 mL/min. The gradient conditions were as follows: .00 min, 70% Solvent B; .30 min, 70.0% Solvent B; 1.33 min, 79.0% Solvent B; 3.30 min, 95.0% Solvent B; 3.50 min, 70% Solvent B; ending at 3.80 min. Cannabinoids were detected using a Vanquish Diode Array Detector (VF-D11-A-01; Thermo Fisher Scientific, Waltham, Massachusetts) and quantified via standard curves generated from authentic standards purchased from Cerilliant Corporation (Round Rock, Texas), CBDA (Cat. #C-144), CBGA (Cat. #C-142), and THCA (Cat. #T-093).

2.5 | Correlation analysis of transcriptome and metabolome

Gene expression quantities and metabolite concentrations from the Cannabis varieties were employed to calculate their non-parametric correlation (Spearman correlation). The gene counts, normalized using the estimated size factor in DESeq2, were compared with the normalized quantities of metabolites, which were adjusted by the corresponding internal standard. The ρ (rho) scores for each individual gene expression and metabolite concentration pair were derived from vectors containing values where each value corresponded to the quantity and concentration from individual samples. This method captures the correlation by accounting for the contributions of samples, yielding a ρ score that reflects the relationship based on these replicates.

Subsequently, the obtained correlation matrix was utilized for clustering metabolic compounds and genes in R and heatmap.2. This clustering was performed using Ward D2 hierarchical clustering coupled with the Euclidean distance metric, resulting in the grouping of metabolites and genes based on their correlation patterns.

2.6 | Strain specific genotyping

RNAseq reads were employed for genotyping the four Cannabis varieties, aiming to investigate their genetic variation. The analysis was performed using the nvar pipeline (Ewels et al., 2020), commit #ccd9d82c, incorporating established genotyping tools. The reference genome, cs10, was indexed with the `--runMode genomegenerate`, and genome annotations were provided via the `--sjdbGTFfile` parameter. Reads from each sample were mapped using STAR v2.7.9a (Dobin et al., 2013) in 2-pass mode, enhancing the accuracy of read mapping

by utilizing junctions detected in the first pass as “annotated” junctions for the second pass. Reads from each sample are uniquely mapped to the reference genome at an average rate of 87%.

The RNA analysis followed the best practices of the Genome Analysis Toolkit (GATK) Variant Discovery v4.2.6.1 (DePristo et al., 2011). This included marking duplicate reads using GATK4's *MarkDuplicates* and separating reads with *SplitNCigarReads*. Variant calling was conducted using GATK *HaplotypeCaller* with a default minimum Phred-scaled confidence threshold of 20 for diploid genomes. The resulting variants were filtered with criteria: FisherStrand (FS) > 30.0, Quality by Depth (QD) < 2.0, and single nucleotide polymorphism (SNP) clusters within a 35 bp window size and cluster size of 3. SNPs were further filtered for QUAL < 60 and retained if present in all three biological replicates of each strain, excluding indels and polymorphic sites. The overall number of high-quality SNPs annotated averaged 220,000. The final SNP dataset was annotated with SnpEff v 5.2c (Cingolani et al., 2012) to predict functional effects using cs10 database annotation.

2.7 | Analysis of protein structures and enzyme-substrate docking of 2-ODDs enzyme family

Cannabis protein 3D structures were predicted for ANS (XP030501512), FLS (XP030502221, XP030492734), and F3H (XP030486031, XP030497321) using the AlphaFold2 algorithm through the ColabLab server notebook (Mirdita et al., 2022). The highest-ranked protein structure model, evaluated using the Local Distance Difference Test (IDDT) score (Mariani et al., 2013), was selected for further analysis. The majority of residues showed a high-quality backbone prediction, with an IDDP score exceeding 70%. Structural alignment with Arabidopsis protein homologs, AtLDOX/ANS (1GP4 (Wilmouth et al., 2002), AtF3H (F4J3A5), and AtFLS1 (B1GV57), was used to infer protein homology, as protein structure is more conserved than aa sequence (Illergård et al., 2009). For this comparison, the structures were aligned using TM-align (Zhang, 2005), with the shortest protein sequence serving as a reference for normalizing the alignment score.

The substrates (2R,3S,4S)-leucocyanidin, (+)-dihydroquercetin, (+)-dihydrokaempferol, and (S)-naringenin described in Table S7 were used in protein docking simulations with the predicted Cannabis proteins. CB-Dock2 (Liu et al., 2022) was used to detect the protein cavities and also calculate the binding affinity scores based on contact residues. The template-based docking engine was used for ANS, specifically employing the 1GP4 protein structure, which CB-Dock2 automatically selected to improve the prediction accuracy. For all other protein structures, the structure-based engine was used without any template. In addition, COACH-D (Wu et al., 2018) was used to identify other ligands with high affinity binding scores, apart from the tested substrates. The protein structures and representations of the protein-ligand complexes were generated respectively with Open-Source PyMOL™ v2.3.0 and LigPlot+ v2.2.8 (Laskowski & Swindells, 2011).

Missense variants identified from RNAseq were analyzed using PremPS v1.0 (Chen et al., 2020). This web-based tool, accessed on July 8, 2024, was employed to estimate the protein unfolding free energy and its effect on protein stability.

2.8 | Computational infrastructure

Data analysis was performed using the computational resources of the Genome Sciences Centre (GSC) in Vancouver, Canada, and the Pawsey Setonix Supercomputer at the Pawsey Supercomputing Research Centre (2023) in Perth, Australia. These facilities provided access to high-performance computing (HPC) cluster resources essential for the comprehensive analysis of genomic data.

3 | RESULTS

3.1 | Pigmentation and anthocyanin assessment in Cannabis varieties

The degree of red/purple leaf pigmentation can be visually assessed during Cannabis growth and development. We utilized four Cannabis varieties, namely, Willow-alpha, CA19210, CK19210, and Cali Kush. The Willow-alpha variety did not show any obvious red/purple pigmentation, even in the petiole and primary leaf vein regions that are often pigmented in other Cannabis varieties (Figure 1). In contrast, the CA19210, CK19206, and Cali Kush varieties all have variably higher red/purple leaf pigmentation (Figure 1). The abundance of anthocyanin-based leaf pigments can be quantified by analyzing leaf extracts for anthocyanin diagnostic absorption spectra between 512 and 528 nm (Lee et al., 2005). Analysis of methanol leaf extracts identified the CA19210 variety as having the highest amounts of anthocyanins, followed by CK19206 and Cali Kush (Figure S1). Willow-alpha was found to have the lowest accumulation of anthocyanins based on UV absorption.

3.2 | Identification of phenylpropanoid, flavonoid, and anthocyanin-producing enzymes in Cannabis

Utilizing the annotated proteins derived from the cs10 Cannabis reference whole-genome prediction, the KIPeS homology search (Pucker et al., 2020) successfully identified over 60 putative enzymes associated with the production of anthocyanins and their precursors (Table 1). The KIPeS search included enzymes involved in the general phenylpropanoid biosynthesis as well as flavonoid biosynthesis. The corresponding transcripts and Cannabis gene models were also retrieved and are listed in Table 1. Similar to other plants, most of the general phenylpropanoid pathway enzymes were identified as small enzyme families in Cannabis. Multiple copies of PAL and 4CL genes were found (Table 1), which is in line with observations in other plant species (Fraser & Chapple, 2011; Hamberger et al., 2007). Overall, we

identified 13 genes involved in the general phenylpropanoid pathway and 14 in the flavonoid pathway; the remaining 30 genes are specifically involved in the production of anthocyanins. The largest enzyme family of the latter is glycosyltransferases, which are responsible for conjugating glycans to the anthocyanidins and represent the final step of the anthocyanin biosynthesis (Table 1).

The general phenylpropanoid pathway genes (PAL, C4H, and 4CL) found in Cannabis show a high level of sequence identity with their Arabidopsis homologs, with the top hits ranging between 71% and 87% aa identity. CHS was found to be a single gene copy with high protein sequence identity with Arabidopsis (87%). We identified several copies encoding for genes in the flavonoid biosynthesis pathway. Flavanone 3-hydroxylase (F3H) was found in two copies and shows high aa conservation with 78% and 80% identity. Seven genes encoding for flavonol 3'-hydroxylase (F3'H) were found with aa conservation ranging between 49% and 66%. Two copies of FLS were found with aa conservation with 51% and 66% identity. Dihydroflavonol 4-reductase (DFR) and ANS enzymes have a high sequence identity with their Arabidopsis homologs, with 70% and 78% aa identity, respectively. Other enzymes from the flavonoid and anthocyanin pathways show lower homology with Arabidopsis homologs, ranging from 52% to 70% aa identity.

Flavone synthase 1 and 2 (FNS1 and FNS2, respectively) perform the same enzymatic reaction although being completely different enzymes (Martens & Mithöfer, 2005). We identified four gene copies of FNS2, but we did not identify any homologs to Arabidopsis FNS1. The anthocyanidin-producing enzymes, DFR and ANS, were found as single gene copies in the cs10 genome. We identified three genes encoding O-methyltransferase (OMT) that putatively facilitate the formation of peonidin from cyanidin and 25 genes encoding flavonoid glycosyl transferases, with a high similarity to the Arabidopsis subclasses UGT72B, UGT75, UGT78D, UGT79B, and UGT80 (Figure S2).

We performed a comparison of the catalytic and substrate binding sites to those found in their respective Arabidopsis homologs to interrogate the functionality of the enzymes identified in the KIPeS homology search. All enzymes shared conserved functional aa residues with Arabidopsis homologs, except for FLS (XP030492734) and ANS (XP030501512) (Table S1). FLS contains three functional residue variants when compared with Arabidopsis: Glu329Ala, Tyr205Met, and His132Tyr, where the first two are involved in the enzymes' catalytic activity and the latter binds to the 2-ODD cofactor (Chua et al., 2008; Welford et al., 2005). In ANS, the residue at position 299, which is crucial for binding the cofactor 2-ODD (Welford et al., 2005; Wilmouth et al., 2002), is a hydrophobic chain aa, similar to its counterpart in Arabidopsis (Ile299Val), indicating a similar substitution.

3.3 | Flavonoid and anthocyanin gene expression

Among the Cannabis varieties, Willow-alpha exhibited the lowest accumulation of anthocyanins in leaves, based on visual inspection and LC-UV analysis of leaf extracts (Figure 1; Figure S1). Hence, this variety served as the baseline for quantifying gene expression in



TABLE 1 Cannabis genes and associated proteins that are putatively involved in the production of phenylpropanoids (white), flavonoids (gray), and anthocyanins (light blue) in Cannabis. Each entry, shown with the appropriate RefSeq ID, for locus, transcript, and protein sequence is displayed with the corresponding Arabidopsis homolog ID. The homolog similarity is shown with % BLASTP identity and % positive scoring residues (percent of amino acids that are either identical between the query and the subject sequence or have similar chemical properties). Different isoforms are included for LOC115709845 (F3'H), LOC115712971 (OMT), LOC115724965 (UGT 79 superfamily), and LOC115720702 (UGT80). FNS2 enzyme is absent in the *Arabidopsis thaliana* genome but found in *Cannabis sativa*.

Locus ID	Transcript ID	Protein ID	Protein length (aa)	Arabidopsis homolog	Gene name	% identity	% positive scoring
LOC115706668	XM030637350	XP030493210	707	AT3G53260	PAL2	82	92
LOC115709282	XM030634382	XP030490242	723	AT2G37040	PAL1	81	90
LOC115709383	XM030637471	XP030493331	709	AT3G53260	PAL2	82	90
LOC115709862	XM030638093	XP030493953	707	AT2G37040	PAL1	83	91
LOC115710186	XM030638528	XP030494388	711	AT3G53260	PAL2	83	91
LOC115710187	XM030638529	XP030494389	710	AT3G53260	PAL2	83	91
LOC115709609	XM030637745	XP030493605	530	AT2G30490	C4H	65	78
LOC115719463	XM030648524	XP030504384	505	AT2G30490	C4H	87	95
LOC115699364	XM030626724	XP030482584	550	AT3G21240	4CL2	71	84
LOC115717276	XM030646244	XP030502104	572	AT1G65060	4CL3	69	83
LOC115725019	XM030654419	XP030510279	553	AT1G51680	4CL1	71	83
LOC115724170	XM030653640	XP030509500	389	AT5G13930	CHS	87	93
LOC115716382	XM030645165	XP030501025	266	AT3G55120	CHI1	62	77
LOC115709313	XM030637390	XP030493250	514	AT5G07990	F3'H	66	81
LOC115709845	XM030638082	XP030493942	514	AT5G07990	F3'H	64	81
LOC115709845	XM030638083	XP030493943	514	AT5G07990	F3'H	64	81
LOC115709933	XM030638196	XP030494056	513	AT5G07990	F3'H	66	82
LOC115714670	XM030643420	XP030499280	519	AT1G58807	F3'H	52	64
LOC115715257	XM030644098	XP030499958	515	AT1G58807	F3'H	52	64
LOC115725467	XM030654999	XP030510859	492	AT3G26180	F3'H	49	60
LOC115702709	XM030630171	XP030486031	384	AT3G51240	F3H	80	89
LOC115712997	XM030641461	XP030497321	373	AT3G51240	F3H	78	89
LOC115708739	XM030636743	XP030492603	514	Na	FNS2	Na	Na
LOC115709046	XM030637075	XP030492935	512	Na	FNS2	Na	Na
LOC115709089	XM030637124	XP030492984	512	Na	FNS2	Na	Na
LOC115721328	XM030650599	XP030506459	544	Na	FNS2	Na	Na
LOC115708857	XM030636874	XP030492734	322	AT5G08640	FLS1	51	68
LOC115717395	XM030646361	XP030502221	337	AT5G08640	FLS1	63	77
LOC115710150	XM030638485	XP030494345	356	AT5G42800	DFR	70	84
LOC115716756	XM030645652	XP030501512	354	AT4G22880	LDOX/ANS	78	89
LOC115695175	XM030622265	XP030478125	467	AT5G54010	UGT79B6	45	67
LOC115695811	XM030622894	XP030478754	465	AT5G54060	UGT79B1	44	67
LOC115697078	XM030623988	XP030479848	460	AT4G27570	UGT superfamily	48	67
LOC115698903	XM030626102	XP030481962	462	AT5G54010	UGT79B6	45	67
LOC115716562	XM030645379	XP030501239	463	AT5G54010	UGT79B6	49	68
LOC115724965	XM030654351	XP030510211	470	AT5G53990	UGT superfamily	43	62
LOC115724965	XM030654352	XP030510212	470	AT5G53990	UGT superfamily	43	62
LOC115724966	XM030654353	XP030510213	466	AT2G36970	UGT superfamily	45	64
LOC115709176	XM030637224	XP030493084	640	AT1G43620	UGT80B1	82	90
LOC115712970	XM030641414	XP030497274	627	AT3G07020	UGT80B2	79	87
LOC115720702	XM030649869	XP030505729	606	AT3G07020	UGT80B2	74	85

(Continues)

TABLE 1 (Continued)

Locus ID	Transcript ID	Protein ID	Protein length (aa)	Arabidopsis homolog	Gene name	% identity	% positive scoring
LOC115720702	XM030649870	XP030505730	605	AT3G07020	UGT80B2	74	85
LOC115703173	XM030630692	XP030486552	488	AT4G01070	UGT72B1	44	62
LOC115705536	XM030632889	XP030488749	478	AT4G01070	UGT72B1	61	78
LOC115705626	XM030633020	XP030488880	478	AT4G01070	UGT72B1	61	78
LOC115705627	XM030633021	XP030488881	469	AT4G01070	UGT72B1	55	75
LOC115706084	XM030633600	XP030489460	484	AT4G01070	UGT72B1	50	66
LOC115718128	XM030647088	XP030502948	477	AT4G01070	UGT72B1	54	71
LOC115718147	XM030647108	XP030502968	476	AT4G01070	UGT72B1	48	64
LOC115720703	XM030649871	XP030505731	508	AT4G01070	UGT72B1	42	61
LOC115725407	XM030654924	XP030510784	479	AT4G01070	UGT72B1	47	65
LOC115716326	XM030645093	XP030500953	460	AT5G17050	UGT78D2	58	74
LOC115722389	XM030651591	XP030507451	503	AT1G05560	UGT75B2	42	60
LOC115722393	XM030651596	XP030507456	492	AT1G05560	UGT75B2	42	60
LOC115722394	XM030651597	XP030507457	490	AT1G05560	UGT75B2	45	62
LOC115722400	XM030651603	XP030507463	473	AT1G05560	UGT75B2	42	61
LOC115722403	XM030651605	XP030507465	459	AT1G05560	UGT75B2	42	62
LOC115712602	XM030640905	XP030496765	241	AT1G67980	OMT (CCOMT)	61	74
LOC115712603	XM030640907	XP030496767	240	AT1G67980	OMT (CCOMT)	60	44
LOC115712971	XM030641415	XP030497275	241	AT1G67980	OMT (CCOMT)	61	74
LOC115712971	XM030641416	XP030497276	241	AT1G67980	OMT (CCOMT)	61	61

comparison to the other three Cannabis varieties. An overview of the phenylpropanoid and flavonoid biosynthesis genes listed in Table 1, as well as their corresponding substrates and products, is provided in Figure 2a. The differential gene expression of these genes is shown as log₂ fold-change in Figure 2b with adjusted *p*-value significance across the pair of genotypes; for example, Willow-alpha versus CA19210; Willow-alpha versus CK19206; Willow-alpha versus Cali Kush. The RNAseq read counts and the comprehensive analysis of differential gene expression can be found in Data S1.

The genes/transcripts encoding for 4CL/LOC115717276 and CHS/LOC115724170 show elevated gene expression for the varieties that accumulated more anthocyanins compared with Willow-alpha (Figure 2b). The genes encoding enzymes involved in flavonol (F3H, F3'H, and FLS) and anthocyanin (DFR, ANS, and OMT) biosynthesis are significantly upregulated in the pigmented Cannabis varieties (Figure 2b). Notably, both gene copies encoding for F3H and FLS, responsible for dihydrokaempferol/dihydroquercetin and kaempferol/quercetin formation, respectively, are significantly upregulated. Moreover, two identified genes copy for F3'H, LOC115709933 and LOC115709845, are upregulated in the pigmented varieties compared with Willow-alpha. On the other hand, the gene responsible for FNS2, involved in the production of apigenin and luteolin from the naringenin and eriodyctiol, respectively, does not exhibit a clear differential gene expression pattern among the evaluated varieties in this study (Figure 2b). The only OMT/LOC115712971 identified that putatively catalyzes the production of peonidin showed increased

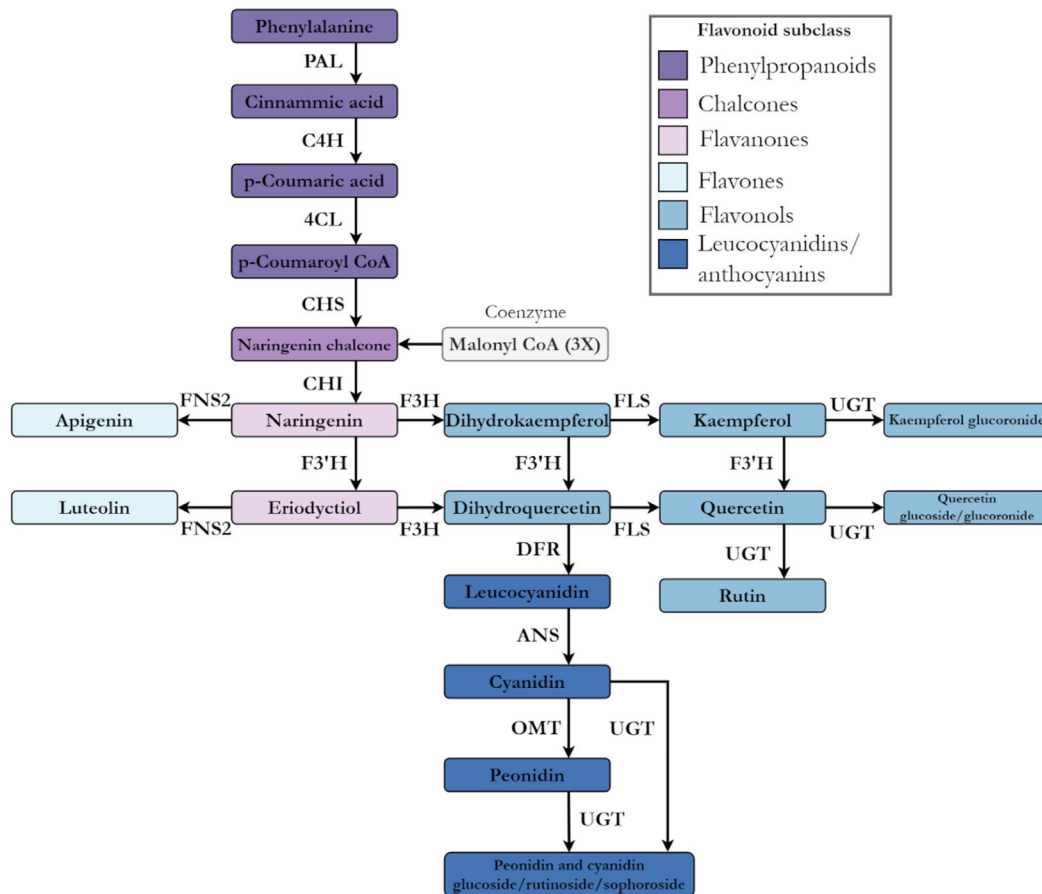
differential gene expression in the pigmented Cannabis varieties compared with Willow-alpha. Finally, at least one enzyme from UGT families 72, 75, 78, and 79 (excluding UGT80) was putatively responsible for the production of flavonoid glycosides and showed differential expression (Figure 2b).

3.4 | Flavonoid, anthocyanin, and cannabinoid metabolite profiling

To better understand the diversity of anthocyanins, flavonols, and flavones found in the Cannabis varieties used in this study, we performed metabolite analysis of leaf extracts. We identified and quantified the abundance of six anthocyanins (Figure 3a,d), four flavonols (Figure 3b,d), and two major flavones (Figure 3c,d) encompassing five of their different glycosides (Tables S2 and S3) in the Cannabis varieties here examined.

The most abundant anthocyanin found in all Cannabis varieties in this study was cyanidin 3-O-rutinoside (Figure 3a,d). Moreover, cyanidin 3-O-rutinoside was much more abundant in Cannabis varieties showing a more pigmented leaf phenotype, such as CA19210, CK19206, and Cali Kush. Peonidin 3-O-rutinoside, an O-methylated derivative of cyanidin 3-O-rutinoside, is the second most abundant anthocyanin found in CA19210, CK19206, and Cali Kush but was nearly undetectable in Willow-alpha (Figure 3a,d). CA19210 is the only Cannabis variety that accumulated cyanidin sophoroside, and this

(a)



(b)

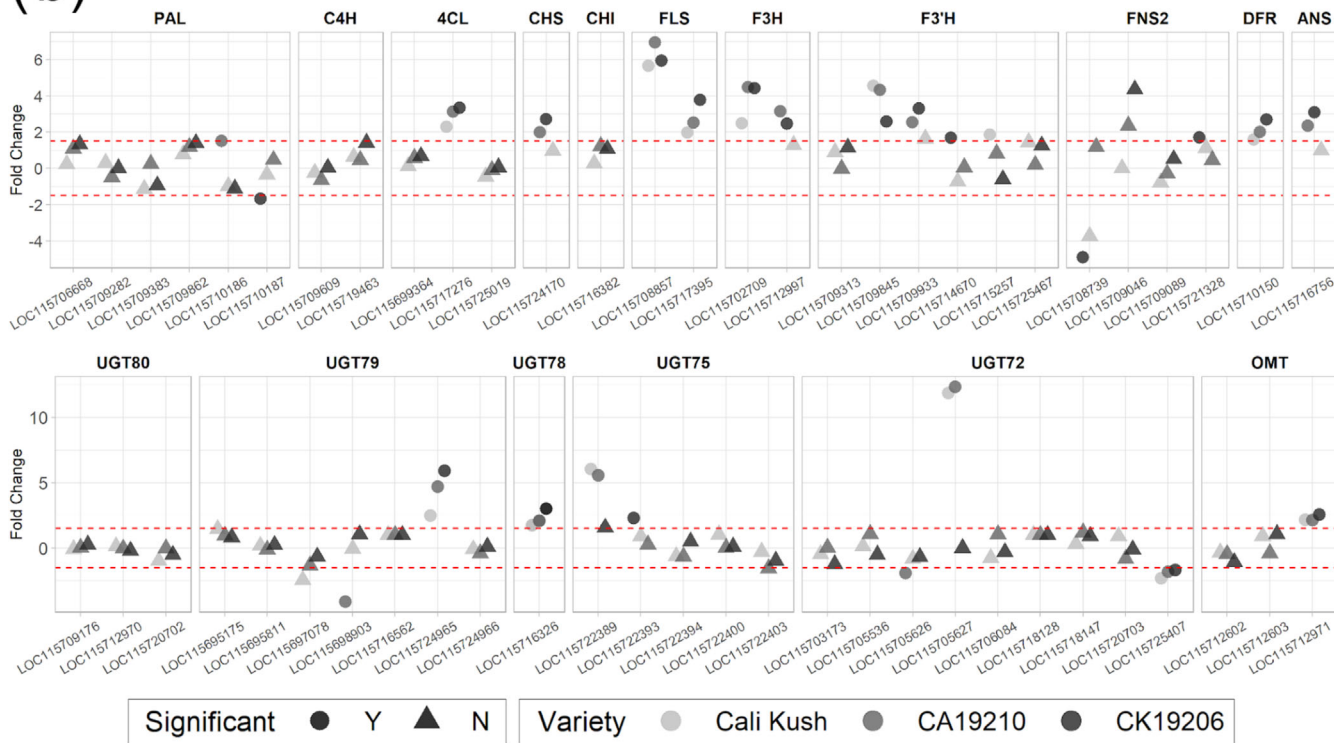


FIGURE 2 Legend on next page.

FIGURE 2 Flavonoid biosynthesis gene expression analysis. (a) Overview of phenylpropanoid and flavonoid (flavanone, flavone, flavonol, and anthocyanin) biosynthesis pathways along with the main substrates and products for each metabolic step. (b) Differential gene expression presented as log₂ fold-change between the pigmented Cannabis varieties (Cali Kush—light gray; CA19210—medium gray; CK19206—black) and Willow-alpha. Differential gene expression analysis for the six glycosyltransferases gene families (UGT80, 79, 78, 78, 75, and 72) identified in the KIPes homology search strategy is also presented. ± 1.5 log₂ fold change threshold is indicated in dashed red lines. Differential gene expression results with statistical significance ($p < .05$ and absolute fold-change > 1.5) are indicated as circles. Non-statistically significant results are indicated as triangles.

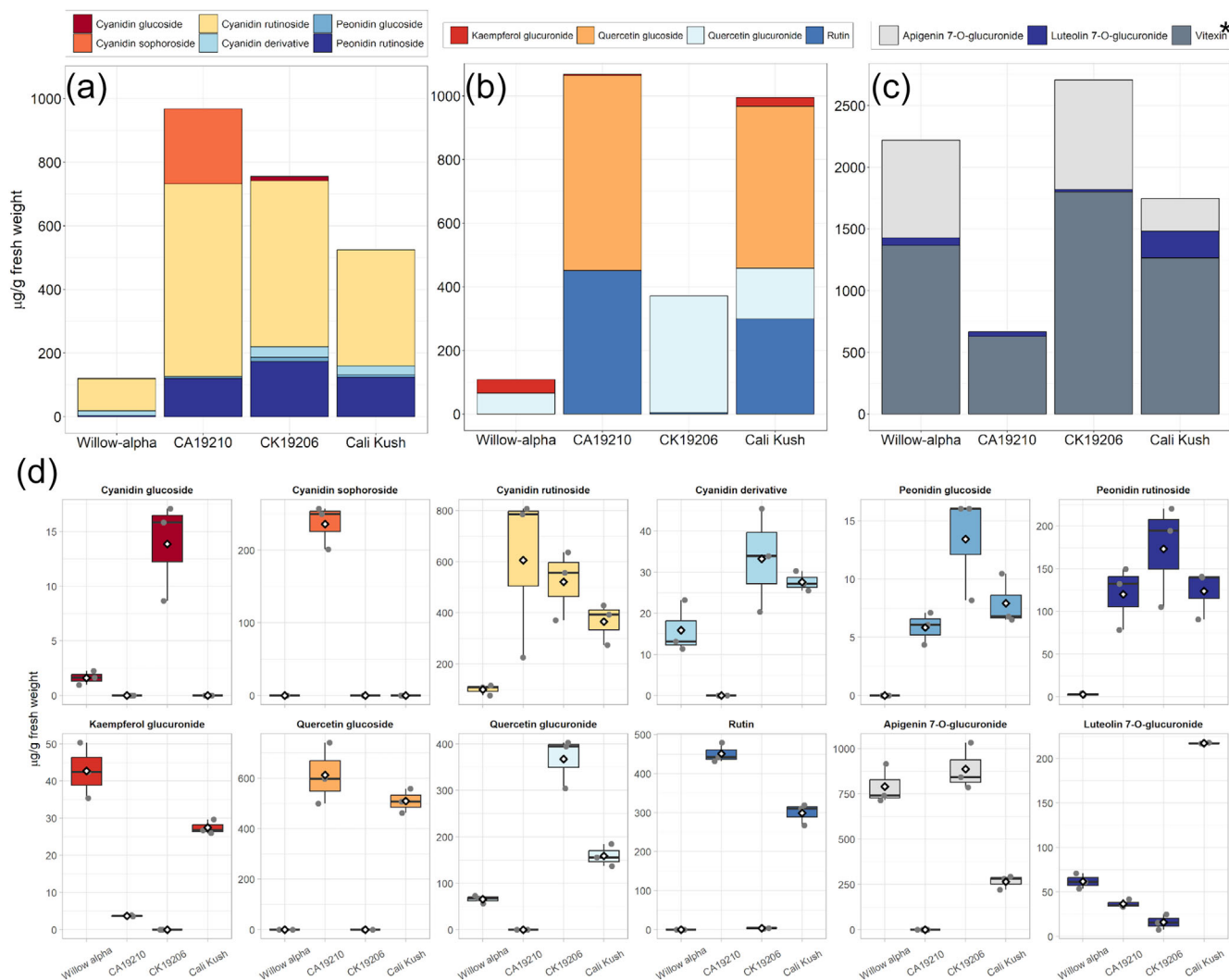


FIGURE 3 Flavonoid metabolite analysis. Cumulative overview of (a) anthocyanin, (b) flavonol, and (c) flavone content in leaf samples from Willow-alpha and three pigmented varieties, CA19210, CK19206, and Cali Kush. Average quantities are shown as normalized scores, resulting in micrograms of metabolites for each gram of fresh leaf weight. The quantities for vitexin (*) are shown as the sum of vitexin 3-O-glucoside, vitexin 7-O-glucoside, and isovitexin-O-glucoside average values (c). This is due to the inability to distinguish between these specific compounds because of their identical molecular masses. (d) Metabolite quantities, across three replicates; quantities are shown as individual normalized scores, with median values represented by the bar in the boxplot and average values as empty diamond shapes.

variety had the highest total amount of anthocyanins (Figure 3a,d). In addition to anthocyanins, we detected several flavonol glycosides, including kaempferol glucuronide, quercetin glucoside, quercetin glucuronide, and rutin (Figure 3b,d). Rutin and quercetin glucoside were the most abundant flavonol glycosides found in CA19210 and Cali

Kush, whereas quercetin glucuronide was the most abundant flavonol glycoside in CK19206 and Willow-alpha (Figure 3b,d). Rutin and quercetin glucoside were the most abundant flavonol glycosides in CA19210 and Cali Kush, whereas quercetin glucuronide was the most abundant in Willow-alpha and CK19206 (Figure 3b,d). Overall,



Willow-alpha and CK19206 accumulated the lowest amount of flavonols, while CA19210 accumulated the highest. Flavone glycosides such as apigenin, luteolin, and vitexin/isovitexin glycosides were also identified (Figure 3c and Tables S3 and S4). Notably, Willow-alpha and CK19206, which had the lowest amounts of anthocyanins + flavonols, exhibited the highest average concentrations of flavones (Figure 3c). Among these, apigenin-7-O-glucuronide and vitexin/isovitexin-glycosides were the most abundant flavones that we identified (Figure 3c).

The cannabinoid content in leaves and flowers was also documented for the four Cannabis varieties used in this study. Leaf samples did not accumulate high levels of CBGA, CBDA, or THCA compared with mature flowers (Figure S3). Willow-alpha accumulated the highest amount of THCA at .43 mg per 100 mg of leaf fresh weight (FW). CA19210 accumulated the lowest amount of THCA (.01 mg/100 mg) but accumulated the highest levels of CBDA (.35mg/100mg) per leaf FW (Figure S3). Willow-alpha flowers accumulated the highest amounts of THCA at 3.8 mg/100 mg per FW (Figure S3). CA19210 accumulated the lowest amount of THCA and highest CBDA at .09 mg/100 g and 2.54 mg/100 g per flower FW respectively. Taken collectively, all the cultivars here used are defined as drug-type Cannabis, with Willow-alpha able to produce a high THCA/low CBDA accumulation chemotype, Cali Kush and CK10206 producing a medium THCA/medium CBDA accumulation chemotype, and CA19210 producing a low THCA/high CBDA accumulation chemotype.

3.5 | Gene expression/metabolite correlation analysis

We correlated gene expression and flavonoid metabolite abundance across the four Cannabis varieties analyzed in this study in order to identify the genes whose expression was most correlated with the production or accumulation of specific metabolites. Figure S4 presents the correlation scores alongside the corresponding genes and clusters of metabolites. The accumulation of peonidin glucoside/rutinoside and cyanidin rutinoside exhibits a robust correlation ($p > .7$) with the gene expression of ANS/LOC115716756, DFR/LOC115710150, F3'H/LOC115709933, CHS/LOC115724170, and 4CL/LOC115717276. Similarly, the production of cyanidin sophoroside demonstrates a strong positive correlation with the gene expression of UGT72B1/LOC115705626. On the other hand, the accumulation of Kaempferol glucuronide negatively correlates with genes involved in the anthocyanin and anthocyanidin biosynthesis, such as DFR/LOC115710150, ANS/LOC115716756, and OMT/LOC115712971.

3.6 | Flavonoid biosynthesis genotype variation analysis

We utilized RNAseq reads to genotype the flavonoid biosynthesis genes in the four Cannabis cultivars. The analysis included all genes

involved in the flavonoid biosynthesis pathway that are listed in Table 1. The list of SNPs and their presence in each variety can be found in Data S2. The most common genetic modifications identified were low-impact variants, followed by SNPs affecting non-coding regions and upstream/downstream regions, and moderate-impact variants resulting in aa sequence changes. A single high-impact event was detected in CA19210, where a premature stop codon was found in UGT72/LOC115706084.

The most prevalent type of variant was the synonymous variant (Data S2). This was followed by downstream gene variants and then missense variants, the latter of which accounted for about one-third the number of synonymous variants. CK19206 exhibited the highest number of synonymous and missense variants, totaling 217 and 104, respectively. In contrast, Willow-alpha had the fewest synonymous SNPs, 189, and the least number of missense variants, 65. Enzymes from large gene families shown in Figure S5, including PAL/LOC115706668, UGT75/LOC115722394, and UGT72/LOC115720703, showed the highest variability in synonymous SNPs in at least one of the varieties. F3'H/LOC115715257 and UGT72/LOC115715257 were among the genes with the highest number of missense variants.

3.7 | Structural and protein-substrate docking analysis of Cannabis 2-ODD enzymes

To gain a more comprehensive understanding of the putative function of the 2-ODD enzymes identified in this study, we conducted protein structural prediction of FLS, ANS, and F3H using AlphaFold. This study aims to enhance our characterization of these enzymes and contribute to a deeper insight into the enzyme family's properties and functions. Following structural prediction, the application of pairwise structure alignment (Table S5) revealed that ANS, produced by a single copy gene within Cannabis, displays the most notable structural similarity score (98.2%) with its corresponding Arabidopsis homolog, AtLDOX/ANS (Table S5). FLS (XP030502221) and F3H (XP030486031), which belong to multicopy gene families, demonstrate the highest structural similarity with their corresponding Arabidopsis homologs, with 97.6% and 94.5% structural similarity, respectively (Table S5).

Through the utilization of protein databank (PDB) templates within the COACH-D protein ligand binding site prediction pipeline (Wu et al., 2018), we identified the binding sites of the crucial cofactors, 2-oxoglutarate (2-OG or α -ketoglutarate AKG) and non-heme iron (Fe). CB-Dock2 (Liu et al., 2022) was employed to predict the substrate binding sites and reveal the docking arrangements of ANS (a), FLS (b), and F3H (c), with leucocyanidin, dihydroquercetin, and narigenin, respectively (Figure 4). The enzyme pocket is situated within the hydrophobic beta-sheet arrangement, referred to as the beta-jelly roll (Wilmouth et al., 2002), and shown in yellow in Figure 4. Additionally, the active site is enclosed by alpha-helices, enveloping the ligand's placement as shown by the highlighted regions. ANS and FLS exhibit a catalytic pocket with a more enclosed structure (Figure 4a,b),

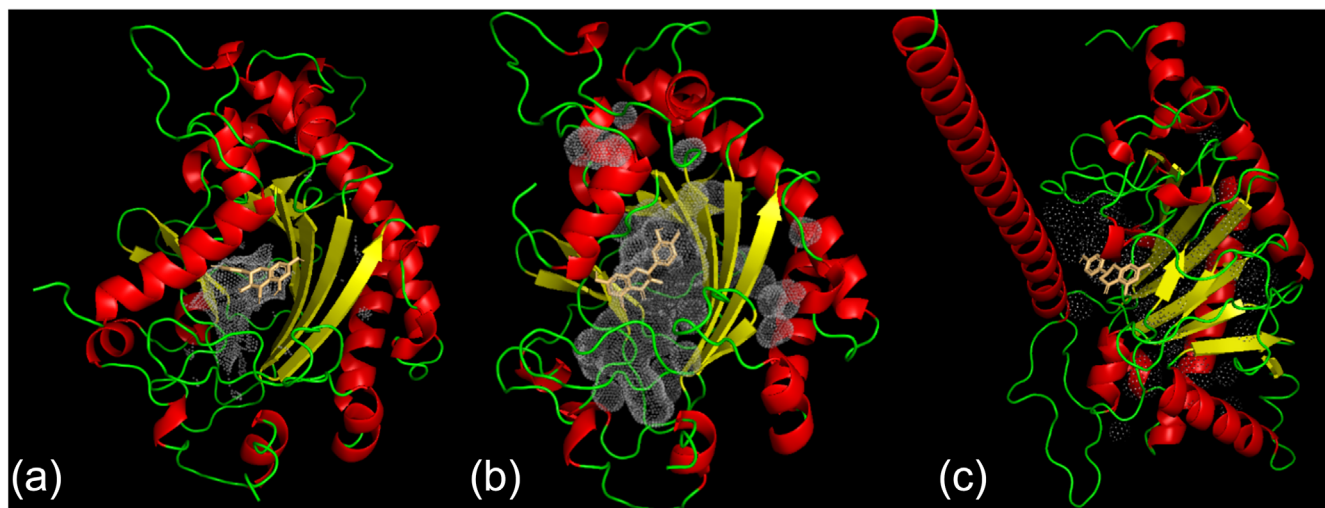


FIGURE 4 Predicted structures for (a) ANS (XP030501512), (b) FLS (XP030492734) and (c) F3H (XP030486031), where beta-sheets are shown in yellow, alpha-helices in red, and unstructured regions in green. The density cloud represents the predicted binding pocket where each substrate is binding. The metabolites shown are (a) leucocyanidin, (b) dihydroquercetin, and (c) naringenin, which are the primary substrates for the three enzymes.

TABLE 2 Binding affinity scores for the Cannabis 2-ODD class enzymes and flavonoid pathway substrates. The lowest affinity score for each enzyme, in the columns, is shown in bold.

	Anthocyanidin synthase (ANS) XP030501512	Flavonoid synthase (FLS) XP030492734	Flavanone 3-hydroxylase (F3H) XP030486031
Leucocyanidin	-6.6	-8.2	-8.0
Dihydroquercetin	-6.0	-8.4	-7.7
Dihydrokaempferol	-5.8	-7.9	-7.7
Naringenin	-5.7	-8.1	-8.1

whereas F3H reveals a more open conformation with an extended N-terminal alpha-helix structure (Figure 4c).

The CBDock2 protein affinity analysis evaluates how flavonoid ligands bind to the examined enzymes. Table 2 reveals the lowest binding affinity scores, indicative of the most stable interactions, which are attributed to the native substrates for each enzyme: leucocyanidin for ANS, dihydroquercetin for FLS, and naringenin for F3H. Remarkably, leucocyanidin and dihydroquercetin also exhibit the second lowest binding affinity scores for ANS and FLS, respectively. F3H demonstrates a notable affinity for leucocyanidin as well, representing its second highest affinity for the enzyme. The full list of tested ligands is shown in Table S6, and the complete list of binding affinity scores and predicted interacting residues is shown in Table S7. Moreover, we report the aa residues predicted to interact with each substrate for the reported enzymes and corresponding ligands (Figures S6–S10).

The protein structures resulting from the observed SNPs were compared with their unaltered equivalent protein in *C. sativa* reference genome, cs10, to determine if changes in the aa sequence could lead to significant structural alterations (Table S8). ANS/XM030645652 had a Pro319Ser missense variant present in all four varieties, which was not identified as a strongly destabilizing

variant and did not affect relevant binding sites as confirmed by protein docking experiments (Table S6). F3H/XM030630171 had the highest number of annotated missense variants, including Ile370Asn, Leu342Ile, and Ser164Ala, found, respectively, in Willow-alpha/CK19206/Cali Kush, CK19206/Cali Kush, and all varieties (Willow-alpha/CA19210/CK19206/Cali Kush). All these variants involve surface aas with non-significant destabilizing effects, similar to Arg53Gly in CA19210, which has a non-significant stabilizing effect on the protein (Table S8). All the annotated aa changes in F3H/XM030630171 are located on the surface of the protein and thus not predicted to interfere with predicted binding ligand sites (Table S8). FLS/XP030492734 did not show any aa changes in the studied varieties.

4 | DISCUSSION

This study investigated the genetic and chemical underpinnings of diverse pigmentation patterns that are observed in drug-type *C. sativa*. We focused on four distinct Cannabis varieties displaying varying degrees of leaf pigmentation. Our analysis encompassed identifying Cannabis genes involved in the phenylpropanoid and flavonoid



biosynthetic pathways, followed by characterization of gene expression and metabolite analysis. We identified ~60 general phenylpropanoid and flavonoid biosynthesis genes using a KIPES-enabled genome-wide homology-based approach (Pucker et al., 2020). KIPES can identify proteins that share significant sequence similarity with known functional proteins and thus increase the likelihood of accurately predicting the function of proteins in other genomes. A previous study identified 20 flavonoid biosynthesis-related genes in the *cs10 Cannabis* reference genome using BLAST and representative sequences mined from *Solanum* species (Bassolino et al., 2020). We identified all but 7 of these 20 sequences and expanded further on this data set by identifying several additional gene models for previously identified genes as well as new gene models for genes encoding F3'H and FNS2 enzymes. Moreover, we used the *Arabidopsis thaliana* genome as our reference dataset because many of the phenylpropanoid/flavonoid biosynthesis genes identified in this study were first characterized in *Arabidopsis*. Moreover, *A. thaliana* remains the best-characterized model system for flavonoid biosynthesis (Saito et al., 2013).

The results show that the significant upregulation of key phenylpropanoid/flavonoid biosynthetic genes is highly correlated with the pigmentation phenotype in *Cannabis* plants. The high differential expression of the 4CL/LOC115717276 and CHS/LOC115724170 genes in the pigmented *Cannabis* varieties suggests increased metabolic flux through the general phenylpropanoid biosynthetic pathway and toward flavonoid biosynthesis, respectively. This pattern of elevated differential gene expression in pigmented *Cannabis* varieties continues for several other key flavonol and anthocyanin biosynthetic genes identified in this study. The differential accumulation of a diverse range of anthocyanins and flavonols in the pigmented *Cannabis* varieties used here largely supports the gene expression findings. Taken together, these results support the hypothesis that differential activation of anthocyanin biosynthetic pathway genes results in the accumulation of anthocyanins and hence the pigmented phenotypes observed in *C. sativa*.

Flavonol glycosides like those found in our study (e.g., kaempferol glucuronide, quercetin glucoside, and rutin), are often colorless or have a pale yellow color (Castañeda-Ovando et al., 2009; Dias et al., 2021). In contrast, anthocyanins such as cyanidin and peonidin 3-O-rutinoside impart red/purple phenotypes (Castañeda-Ovando et al., 2009; Dias et al., 2021). We observed the accumulation of kaempferol and quercetin-derived flavonols along with cyanidin and peonidin-derived anthocyanins in the pigmented *Cannabis* varieties. These metabolites are derived from dihydrokaempferol or dihydroquercetin, which are intermediates leading to subsequent flavonol and anthocyanin biosynthesis. The observation that the most abundant anthocyanins and flavonols found in this study were derived from dihydroquercetin (Figure 3a,b,d) is also supported by the high expression of F3'H, which catalyzes the formation of dihydroquercetin using dihydrokaempferol as a substrate. However, the significant differential gene expression of two FLS genes (FLS/LOC115708857 and FLS/LOC115717395), as well as dihydroflavonol 4-reductase (DFR/LOC115710150), suggests that there is strong competition for

dihydroquercetin and dihydrokaempferol toward subsequent flavonol and anthocyanin biosynthesis (Figure 2a). Notably, the genes encoding flavonoid 3',5'-hydroxylase (F3'5'H) enzymes, which use dihydrokaempferol for the downstream biosynthesis of delphinidin-derived anthocyanins, were not identified in our study. This is further supported by our metabolomic analysis, which also did not identify any delphinidin-derived anthocyanins.

The final phase of anthocyanin and flavonol biosynthesis involves the conjugation of glycosyl moieties to the aglycones (Yoshihara et al., 2005). Our investigation revealed two UGT genes, specifically UGT78D/LOC115716326 and UGT79B/LOC115724965 (Figure 2b), that exhibit a strong correlation with the pigmentation accumulation in colored varieties. The *Arabidopsis* homologs to these two genes have been implicated in controlling pigmentation in *Arabidopsis* (Jones et al., 2003; Knoch et al., 2018; Li et al., 2017). Our results suggest that the two genes are good candidates for the formation of anthocyanins and flavonol-glycosides in *Cannabis*. However, the high number of UGTs found in plants and the tens of thousands of potential substrates make the prediction of substrates of any particular UGT difficult (Osmani et al., 2009). Gene co-expression studies, like those performed in this work, can provide UGT enzymes that are candidates for anthocyanin and flavonol accumulation, but future functional studies would provide more insight into their substrate specificity.

We identified cyanidin 3-O-rutinoside and peonidin 3-O-rutinoside as the most abundant anthocyanins in the four drug-type *Cannabis* varieties. Our findings corroborate a previous report derived from six different hemp-type *Cannabis* varieties (Bassolino et al., 2023). Taken collectively, these results demonstrate that cyanidin 3-O-rutinoside and peonidin 3-O-rutinoside are the dominant anthocyanins found in both drug-type and hemp-type *Cannabis*. This result was somewhat surprising considering the significant genome-wide differentiation between the drug-type and hemp-type *Cannabis* as was previously reported after the genomic analysis of 81 drug-type and 43 hemp-type *Cannabis* varieties (Sawler et al., 2015). Thus, the differences in cannabinoid accumulation, plant growth architecture, disease/pest resilience, and so forth observed in hemp-type and drug-type *Cannabis* do not impact the diversity of anthocyanin accumulation in these different *Cannabis* types. However, the relative abundance of the anthocyanins identified thus far (e.g., cyanidin 3-O-rutinoside, peonidin 3-O-rutinoside, or cyanidin sophoroside) in any particular *Cannabis* variety can vary dramatically (Figure 3a; Bassolino et al., 2023).

In addition to gene expression, other factors such as enzyme turnover, metabolon formation, catalytic activity, and substrate promiscuity are important factors regulating the metabolic flow through a biosynthetic pathway toward specific endpoints. We therefore conducted a more comprehensive examination of the *FLS*, *ANS*, and *F3H* genes, which all belong to the 2-ODD enzyme family, an evolutionarily related group of enzymes that oversee crucial reactions within flavonoid biosynthesis. Among these three, *ANS* emerges as the main enzyme responsible for initiating anthocyanin production in plants (Pelletier et al., 1997). Notably, the primary substrate of this enzyme,

leucocyanidin, is produced by the upstream DFR enzyme, which competes with FLS for dihydroquercetin.

2-ODD enzymes are recognized for their multifunctional catalytic activity, which extends to the utilization of leucocyanidin by ANS, naringenin by F3H, and dihydroquercetin/dihydrokaempferol by FLS (Turnbull et al., 2004). The functional role of FLS has undergone exploration across various plant species (Chen et al., 2023; Xu et al., 2012). Multiple copies of the FLS enzyme have been identified in several studies, including *Arabidopsis* (Owens et al., 2008), grapevine (Fujita et al., 2006), and canola (Schilbert et al., 2021). In *Arabidopsis*, the gene family responsible for encoding FLS enzymes emerged from a recent gene duplication event, subsequently followed by pseudogenization to eliminate redundant gene copies (Preuß et al., 2009). Among these, *FLS1* remains the sole functional enzyme-encoding gene product, with the additional five gene copies yielding non-functional outcomes (Owens et al., 2008; Wisman et al., 1998). The analysis of *FLS* and *ANS* across multiple plant species by Wang et al. (2021) reveals an interesting observation that some enzymes annotated as ANS cluster with FLS. This suggests functional redundancy or incomplete evolution of substrate selectivity among these proteins. The presence of similar activities in this branch enriches the functional diversity of the 2-ODD enzyme family. The enzymatic activity of FLS in *Cannabis* has been characterized by the study conducted by Zhu et al. (2022), which also includes a heterologous assessment of FLS's enzymatic activity in *Escherichia coli*. This study indicates FLS2/XP030492734 as the enzyme demonstrating the highest catalytic activity toward substrates such as dihydrokaempferol and dihydroquercetin. The FLS enzyme, XP030502221, which was also identified in our study, did not show functional activity toward its primary substrates (Zhu et al., 2022). Furthermore, Zhu et al. (2022) present a functional analysis of FLS3/XP030501512, which displays a minor degree of catalytic activity for the same two substrates. In this work, we present comprehensive annotations of enzymes contributing to the anthocyanin pathway, designating XP030501512 as ANS (Table 1), based on homology analysis, representative catalytic sites, and protein structure evaluation. Informed by protein docking findings (Table 2), our attention was drawn to the substrate exhibiting the most robust binding affinity—leucocyanidin—which coincidentally serves as the primary substrate of the ANS enzyme. We therefore hypothesize that the ANS/LOC115716756 is the functional anthocyanin synthase in *Cannabis* based on differential gene expression analysis (Figure 2b) and the protein structure and molecular docking studies (Figure 4; Table 2). The loss of function phenotype of the *Arabidopsis* *Leucoanthocyanidin Dioxygenase* (LDOX/At4g22880), which is the *Arabidopsis* ortholog to ANS/LOC115716756, displays obvious deficiencies in anthocyanin accumulation in leaves, stems, and other tissues (Abrahams et al., 2003). *Cannabis* currently lacks the genetic resources for interrogating enzyme function based on mutant analysis. However, the recent development of CRISPR/Cas9-mediated (clustered regularly interspaced short palindromic repeats / CRISPR-associated protein 9) mutagenesis in hemp-type *Cannabis* will enable these future functional studies (Zhang et al., 2021).

In conclusion, our findings identify the *Cannabis* genes whose expression is associated with *Cannabis* leaf pigmentation, while also identifying the diversity of anthocyanins, flavonols, and flavones found in four drug-type *Cannabis* varieties. Together, these data contribute to a deeper understanding of the underlying mechanisms governing anthocyanin production in *Cannabis* and lay the foundation for future functional characterization of the biosynthesis genes identified in this study.

AUTHOR CONTRIBUTIONS

Conceptualization: Kristina K. Gagalova, Mathias Schuetz, Till Matzat, and Shumin Wang. **Methodology:** Kristina K. Gagalova, Yifan Yan, and Shumin Wang. **Formal analysis:** Kristina K. Gagalova and Mathias Schuetz. **Investigation:** Kristina K. Gagalova and Mathias Schuetz. **Data curation:** Kristina K. Gagalova and Mathias Schuetz. **Writing—original draft preparation:** Kristina K. Gagalova. **Writing—review and editing:** Mathias Schuetz, Simone D. Castellarin, Till Matzat, David Edwards, and Inanc Birol. **Data visualization:** Kristina K. Gagalova. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

We would like to thank Dr. Matthew Workentine for his support with data analysis and annotation during the early stages of this work.

CONFLICT OF INTEREST STATEMENT

M.S., K.K.G., T.M., and S.W. were employed by Willow Analytics during the data collection stage of this study. Willow Analytics was a for-profit analytical testing and *Cannabis* research lab located in Burnaby, British Columbia.

DATA AVAILABILITY STATEMENT

Data will be made available on request.

ORCID

Kristina K. Gagalova  <https://orcid.org/0000-0002-5975-0805>

David Edwards  <https://orcid.org/0000-0001-7599-6760>

Mathias Schuetz  <https://orcid.org/0000-0002-2413-800X>

REFERENCES

- Aardema, M. L., & DeSalle, R. (2021). Can public online databases serve as a source of phenotypic information for *Cannabis* genetic association studies? *PLOS One*, 16(2), e0247607. <https://doi.org/10.1371/journal.pone.0247607>
- Abrahams, S., Lee, E., Walker, A. R., Tanner, G. J., Larkin, P. J., & Ashton, A. R. (2003). The *Arabidopsis* *TDS4* gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. *The Plant Journal*, 35(5), 624–636. <https://doi.org/10.1046/j.1365-313X.2003.01834.x>
- Bassolino, L., Buti, M., Fulvio, F., Pennesi, A., Mandolino, G., Milc, J., Francia, E., & Paris, R. (2020). In silico identification of MYB and bHLH families reveals candidate transcription factors for secondary metabolic pathways in *Cannabis sativa* L. *Plants (Basel, Switzerland)*, 9(11), 1540. <https://doi.org/10.3390/plants9111540>
- Bassolino, L., Fulvio, F., Pastore, C., Pasini, F., Toschi, T. G., Filippetti, I., & Paris, R. (2023). When *Cannabis sativa* L. turns purple: Biosynthesis



- and accumulation of anthocyanins. *Antioxidants*, 12(7), 1393. <https://doi.org/10.3390/antiox12071393>
- Bautista, J. L., Yu, S., & Tian, L. (2021). Flavonoids in *Cannabis sativa*: Biosynthesis, bioactivities, and biotechnology. *ACS Omega*, 6(8), 5119–5123. <https://doi.org/10.1021/acsomega.1c00318>
- Booth, J. K., & Bohlmann, J. (2019). Terpenes in *Cannabis sativa*—From plant genome to humans. *Plant Science*, 284, 67–72. <https://doi.org/10.1016/j.plantsci.2019.03.022>
- Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525–527. <https://doi.org/10.1038/nbt.3519>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1), 421. <https://doi.org/10.1186/1471-2105-10-421>
- Castañeda-Ovando, A., Pacheco-Hernández, M. D. L., Páez-Hernández, M. E., Rodríguez, J. A., & Galán-Vidal, C. A. (2009). Chemical studies of anthocyanins: A review. *Food Chemistry*, 113(4), 859–871. <https://doi.org/10.1016/j.foodchem.2008.09.001>
- Cerrato, A., Biancolillo, A., Cannazza, G., Cavaliere, C., Citti, C., Laganà, A., Marini, F., Montanari, M., Montone, C. M., Paris, R., Virzi, N., & Capriotti, A. L. (2023). Untargeted cannabinomics reveals the chemical differentiation of industrial hemp based on the cultivar and the geographical field location. *Analytica Chimica Acta*, 1278, 341716. <https://doi.org/10.1016/j.aca.2023.341716>
- Cerrato, A., Citti, C., Cannazza, G., Capriotti, A. L., Cavaliere, C., Grassi, G., Marini, F., Montone, C. M., Paris, R., Piovesana, S., & Laganà, A. (2021). Phytocannabinomics: Untargeted metabolomics as a tool for cannabis chemovar differentiation. *Talanta*, 230, 122313. <https://doi.org/10.1016/j.talanta.2021.122313>
- Chen, G., Wang, Y., Liu, X., Duan, S., Jiang, S., Zhu, J., Zhang, Y., & Hou, H. (2023). The *DmIR156n* regulates drought tolerance and flavonoid synthesis in apple calli and *Arabidopsis*. *International Journal of Molecular Sciences*, 24(7), 6049. <https://doi.org/10.3390/ijms24076049>
- Chen, Y., Lu, H., Zhang, N., Zhu, Z., Wang, S., & Li, M. (2020). PremPS: Predicting the impact of missense mutations on protein stability. *PLoS Computational Biology*, 16(12), e1008543. <https://doi.org/10.1371/journal.pcbi.1008543>
- Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: A complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*, 89(4), 789–804. <https://doi.org/10.1111/tpj.13415>
- Chu, J., Sadeghi, S., Raymond, A., Jackman, S. D., Nip, K. M., Mar, R., Mohamadi, H., Butterfield, Y. S., Gordon Robertson, A., & Birol, I. (2014). BioBloom tools: Fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics*, 30(23), 3402–3404. <https://doi.org/10.1093/bioinformatics/btu558>
- Chua, C. S., Biermann, D., Goo, K. S., & Sim, T.-S. (2008). Elucidation of active site residues of *Arabidopsis thaliana* flavonol synthase provides a molecular platform for engineering flavonols. *Phytochemistry*, 69(1), 66–75. <https://doi.org/10.1016/j.phytochem.2007.07.006>
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain¹¹¹⁸; iso-2; iso-3. *Fly*, 6(2), 80–92.
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., & McKenna, A. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–498. <https://doi.org/10.1038/ng.806>
- Dias, M. C., Pinto, D. C. G. A., & Silva, A. M. S. (2021). Plant flavonoids: Chemical characteristics and biological activity. *Molecules*, 26(17), 5377. <https://doi.org/10.3390/molecules26175377>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- EiSohly, M. A., & Gul, W. (2014). Constituents of *Cannabis sativa*. In R. Pertwee (Ed.), *Handbook of cannabis* (pp. 3–22). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199662685.003.0001>
- EiSohly, M. A., & Slade, D. (2005). Chemical constituents of marijuana: The complex mixture of natural cannabinoids. *Life Sciences*, 78(5), 539–548. <https://doi.org/10.1016/j.lfs.2005.09.011>
- Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3), 276–278. <https://doi.org/10.1038/s41587-020-0439-x>
- Ferber, S. G., Namdar, D., Hen-Shoval, D., Eger, G., Koltai, H., Shoval, G., Shbiro, L., & Weller, A. (2020). The “entourage effect”: Terpenes coupled with cannabinoids for the treatment of mood disorders and anxiety disorders. *Current Neuropharmacology*, 18(2), 87–96. <https://doi.org/10.2174/1570159X17666190903103923>
- Fraser, C. M., & Chapple, C. (2011). The phenylpropanoid pathway in *Arabidopsis*. *The Arabidopsis Book*, 9(January), e0152. <https://doi.org/10.1199/tab.0152>
- Fujita, A., Goto-Yamamoto, N., Aramaki, I., & Hashizume, K. (2006). Organ-specific transcription of putative flavonol synthase genes of grapevine and effects of plant hormones and shading on flavonol biosynthesis in grape berry skins. *Bioscience, Biotechnology, and Biochemistry*, 70(3), 632–638. <https://doi.org/10.1271/bbb.70.632>
- Grassa, C. J., Weiblen, G. D., Wenger, J. P., Dabney, C., Poplawski, S. G., Timothy Motley, S., Michael, T. P., & Schwartz, C. J. (2021). A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytologist*, 230(4), 1665–1679. <https://doi.org/10.1111/nph.17243>
- Hamberger, B., Ellis, M., Friedmann, M., De Azevedo, C., Souza, B. B., & Douglas, C. J. (2007). Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: The *Populus* lignin toolbox and conservation and diversification of angiosperm gene families. *Canadian Journal of Botany*, 85(12), 1182–1201. <https://doi.org/10.1139/B07-098>
- Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3), 499–508. <https://doi.org/10.1002/prot.22458>
- Izzo, L., Castaldo, L., Narváez, A., Graziani, G., Gaspari, A., Rodríguez-Carrasco, Y., & Ritieni, A. (2020). Analysis of phenolic compounds in commercial *Cannabis sativa* L. inflorescences using UHPLC-Q-Orbitrap HRMS. *Molecules*, 25(3), 631. <https://doi.org/10.3390/molecules25030631>
- Jones, P., Messner, B., Nakajima, J.-I., Schäffner, A. R., & Saito, K. (2003). UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in *Arabidopsis thaliana*. *Journal of Biological Chemistry*, 278(45), 43910–43918. <https://doi.org/10.1074/jbc.M303523200>
- Khoo, H. E., Azlan, A., Tang, S. T., & Lim, S. M. (2017). Anthocyanidins and anthocyanins: Colored pigments as food, pharmaceutical ingredients, and the potential health benefits. *Food & Nutrition Research*, 61(1), 1361779. <https://doi.org/10.1080/16546628.2017.1361779>
- Knoch, E., Sugawara, S., Mori, T., Nakabayashi, R., Saito, K., & Yonekura-Sakakibara, K. (2018). UGT79B31 is responsible for the final modification step of pollen-specific flavonoid biosynthesis in *Petunia hybrida*. *Planta*, 247(4), 779–790. <https://doi.org/10.1007/s00425-017-2822-5>
- Koltai, H., & Namdar, D. (2020). Cannabis phytomolecule ‘entourage’: From domestication to medical use. *Trends in Plant Science*, 25(10), 976–984. <https://doi.org/10.1016/j.tplants.2020.04.007>

- Kumar, S., & Pandey, A. K. (2013). Chemistry and biological activities of flavonoids: An overview. *The Scientific World Journal*, 2013, 1–16. <https://doi.org/10.1155/2013/162750>
- Kundan, M., Gani, U., Fayaz, M., Angmo, T., Kesari, R., Rahul, V. P., Gairola, S., & Misra, P. (2022). Two R2R3-MYB transcription factors, CsMYB33 and CsMYB78 are involved in the regulation of anthocyanin biosynthesis in *Cannabis sativa* L. *Industrial Crops and Products*, 188, 115546. <https://doi.org/10.1016/j.indcrop.2022.115546>
- Laskowski, R. A., & Swindells, M. B. (2011). LigPlot+: Multiple ligand–protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling*, 51(10), 2778–2786. <https://doi.org/10.1021/ci200227u>
- Lee, J., Durst, R. W., & Wrolstad, R. E. (2005). Determination of total monomeric anthocyanin pigment content of fruit juices, beverages, natural colorants, and wines by the pH differential method: Collaborative study. *Journal of AOAC International*, 88(5), 1269–1278. <https://doi.org/10.1093/jaoac/88.5.1269>
- Li, P., Li, Y.-J., Zhang, F.-J., Zhang, G.-Z., Jiang, X.-Y., Hui-Min, Y., & Hou, B.-K. (2017). The Arabidopsis UDP-glycosyltransferases UGT79B2 and UGT79B3, contribute to cold, salt and drought stress tolerance via modulating anthocyanin accumulation. *The Plant Journal*, 89(1), 85–103. <https://doi.org/10.1111/tpj.13324>
- Liu, Y., Yang, X., Gan, J., Chen, S., Xiao, Z.-X., & Cao, Y. (2022). CB-Dock2: Improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting. *Nucleic Acids Research*, 50(W1), W159–W164. <https://doi.org/10.1093/nar/gkac394>
- Louveau, T., & Osbourn, A. (2019). The sweet side of plant-specialized metabolism. *Cold Spring Harbor Perspectives in Biology*, 11(12), a034744.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). IDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21), 2722–2728. <https://doi.org/10.1093/bioinformatics/btt473>
- Martens, S., Preuß, A., & Matern, U. (2010). Multifunctional flavonoid dioxygenases: Flavonol and anthocyanin biosynthesis in *Arabidopsis thaliana* L. *Phytochemistry*, 71(10), 1040–1049. <https://doi.org/10.1016/j.phytochem.2010.04.016>
- Martens, S., & Mithöfer, A. (2005). Flavones and flavone synthases. *Phytochemistry*, 66(20), 2399–2407. <https://doi.org/10.1016/j.phytochem.2005.07.013>
- McPartland, J. M., Hegman, W., & Long, T. (2019). *Cannabis* in Asia: Its center of origin and early cultivation, based on a synthesis of subfossil pollen and archaeobotanical studies. *Vegetation History and Archaeobotany*, 28, 691–702. <https://doi.org/10.1007/s00334-019-00731-8>
- Mierziak, J., Kostyn, K., & Kulma, A. (2014). Flavonoids as important molecules of plant interactions with the environment. *Molecules*, 19(10), 16240–16265. <https://doi.org/10.3390/molecules191016240>
- Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nature Methods*, 19(6), 679–682. <https://doi.org/10.1038/s41592-022-01488-1>
- Nakamura, T., Yamada, K. D., Tomii, K., & Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*, 34(14), 2490–2492. <https://doi.org/10.1093/bioinformatics/bty121>
- Osmani, S. A., Bak, S., & Möller, B. L. (2009). Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. *Phytochemistry*, 70(3), 325–347. <https://doi.org/10.1016/j.phytochem.2008.12.009>
- Owens, D. K., Alerding, A. B., Crosby, K. C., Bandara, A. B., Westwood, J. H., & Winkel, B. S. J. (2008). Functional analysis of a predicted flavonol synthase gene family in Arabidopsis. *Plant Physiology*, 147(3), 1046–1061. <https://doi.org/10.1104/pp.108.117457>
- Paquette, S. M., Jensen, K., & Bak, S. (2009). A web-based resource for the Arabidopsis P450, cytochromes *b*₅, NADPH-cytochrome P450 reductases, and family 1 glycosyltransferases (<http://www.P450.kvl.dk>). *Phytochemistry*, 70(17–18), 1940–1947. <https://doi.org/10.1016/j.phytochem.2009.08.024>
- Passeri, V., Koes, R., & Quattrocchio, F. M. (2016). New challenges for the design of high value plant products: Stabilization of anthocyanins in plant vacuoles. *Frontiers in Plant Science*, 7(February), 153. <https://doi.org/10.3389/fpls.2016.00153>
- Pelletier, M. K., Murrell, J. R., & Shirley, B. W. (1997). Characterization of flavonol synthase and leucoanthocyanidin dioxygenase genes in Arabidopsis (further evidence for differential regulation of “early” and “late” genes). *Plant Physiology*, 113(4), 1437–1445. <https://doi.org/10.1104/pp.113.4.1437>
- Preuß, A., Stracke, R., Weisshaar, B., Hillebrecht, A., Matern, U., & Martens, S. (2009). *Arabidopsis thaliana* expresses a second functional flavonol synthase. *FEBS Letters*, 583(12), 1981–1986. <https://doi.org/10.1016/j.febslet.2009.05.006>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3), e9490. <https://doi.org/10.1371/journal.pone.0009490>
- Pucker, B., Reiher, F., & Schilbert, H. M. (2020). Automatic identification of players in the flavonoid biosynthesis with application on the medicinal plant *Croton tiglium*. *Plants*, 9(9), 1103. <https://doi.org/10.3390/plants9091103>
- Radwan, M. M., Chandra, S., Gul, S., & ElSohly, M. A. (2021). Cannabinoids, phenolics, terpenes and alkaloids of *Cannabis*. *Molecules*, 26(9), 2774. <https://doi.org/10.3390/molecules26092774>
- Rull, V. (2022). Origin, early expansion, domestication and anthropogenic diffusion of *Cannabis*, with emphasis on Europe and the Iberian Peninsula. *Perspectives in Plant Ecology, Evolution and Systematics*, 55, 125670. <https://doi.org/10.1016/j.ppees.2022.125670>
- Russo, E. B. (2011). Taming THC: Potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. *British Journal of Pharmacology*, 163(7), 1344–1364. <https://doi.org/10.1111/j.1476-5381.2011.01238.x>
- Saito, K., Yonekura-Sakakibara, K., Nakabayashi, R., Higashi, Y., Yamazaki, M., Tohge, T., & Fernie, A. R. (2013). The flavonoid biosynthetic pathway in Arabidopsis: Structural and genetic diversity. *Plant Physiology and Biochemistry*, 72, 21–34. <https://doi.org/10.1016/j.plaphy.2013.02.001>
- Sawler, J., Stout, J. M., Gardner, K. M., Hudson, D., Vidmar, J., Butler, L., Page, J. E., & Myles, S. (2015). The genetic structure of marijuana and hemp. *PLoS ONE*, 10(8), e0133292. <https://doi.org/10.1371/journal.pone.0133292>
- Schilbert, H. M., Schöne, M., Baier, T., Busche, M., Viehöver, P., Weisshaar, B., & Holtgräwe, D. (2021). Characterization of the *Brassica napus* flavonol synthase gene family reveals bifunctional flavonol synthases. *Frontiers in Plant Science*, 12(October), 733762. <https://doi.org/10.3389/fpls.2021.733762>
- Setonix Supercomputer, Pawsey Supercomputing Research Centre. (2023). Setonix supercomputer. <https://doi.org/10.48569/18sb-8s43>
- Sievers, F., & Higgins, D. G. (2014). Clustal omega. *Current Protocols in Bioinformatics*, 48(1), 3 - 13. <https://doi.org/10.1002/0471250953.bi0313s48>
- Small, E. (2015). Evolution and classification of *Cannabis sativa* (marijuana, hemp) in relation to human utilization. *The Botanical Review*, 81(3), 189–294. <https://doi.org/10.1007/s12229-015-9157-3>



- Sommano, S. R., Chittasupho, C., Ruksiriwanich, W., & Jantrawut, P. (2020). The cannabis terpenes. *Molecules*, 25(24), 5792. <https://doi.org/10.3390/molecules25245792>
- The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Tulio, A. Z. Jr., Neil Reese, R., Wyzgoski, F. J., Rinaldi, P. L., Ruiling, F., Scheerens, J. C., & Miller, A. R. (2008). Cyanidin 3-rutinoside and cyanidin 3-xylosylrutinoside as primary phenolic antioxidants in black raspberry. *Journal of Agricultural and Food Chemistry*, 56(6), 1880–1888. <https://doi.org/10.1021/jf072313k>
- Turnbull, J. J., Nakajima, J.-i., Welford, R. W. D., Yamazaki, M., Saito, K., & Schofield, C. J. (2004). Mechanistic studies on three 2-oxoglutarate-dependent oxygenases of flavonoid biosynthesis. *Journal of Biological Chemistry*, 279(2), 1206–1216. <https://doi.org/10.1074/jbc.M309228200>
- Turner, C. E., Elshohly, M. A., & Boeren, E. G. (1980). Constituents of *Cannabis sativa* L. XVII. A review of the natural constituents. *Journal of Natural Products*, 43(2), 169–234. <https://doi.org/10.1021/np50008a001>
- Verdan, A. M., Wang, H. C., García, C. R., Henry, W. P., & Brumaghim, J. L. (2011). Iron binding of 3-hydroxychromone, 5-hydroxychromone, and sulfonated morin: Implications for the antioxidant activity of flavonols with competing metal binding sites. *Journal of Inorganic Biochemistry*, 105(10), 1314–1322. <https://doi.org/10.1016/j.jinorgbio.2011.07.006>
- Vogt, T. (2010). Phenylpropanoid biosynthesis. *Molecular Plant*, 3(1), 2–20. <https://doi.org/10.1093/mp/ssp106>
- Wang, Y., Shi, Y., Li, K., Yang, D., Liu, N., Zhang, L., Zhao, L., Zhang, X., Liu, Y., Gao, L., Xia, T., & Wang, P. (2021). Roles of the 2-oxoglutarate-dependent dioxygenase superfamily in the flavonoid pathway: A review of the functional diversity of F3H, FNS I, FLS, and LDOX/ANS. *Molecules*, 26(21), 6745. <https://doi.org/10.3390/molecules26216745>
- Welford, R. W. D., Clifton, I. J., Turnbull, J. J., Wilson, S. C., & Schofield, C. J. (2005). Structural and mechanistic studies on anthocyanidin synthase catalysed oxidation of flavanone substrates: The effect of C-2 stereochemistry on product selectivity and mechanism. *Organic & Biomolecular Chemistry*, 3(17), 3117–3126. <https://doi.org/10.1039/b507153d>
- Wilmouth, R. C., Turnbull, J. J., Welford, R. W. D., Clifton, I. J., Prescott, A. G., & Schofield, C. J. (2002). Structure and mechanism of anthocyanidin synthase from *Arabidopsis thaliana*. *Structure*, 10(1), 93–103. [https://doi.org/10.1016/S0969-2126\(01\)00695-5](https://doi.org/10.1016/S0969-2126(01)00695-5)
- Wisman, E., Hartmann, U., Sagasser, M., Baumann, E., Palme, K., Hahlbrock, K., Saedler, H., & Weisshaar, B. (1998). Knock-out mutants from an En-1 mutagenized *Arabidopsis thaliana* population generate phenylpropanoid biosynthesis phenotypes. *Proceedings of the National Academy of Sciences*, 95(21), 12432–12437. <https://doi.org/10.1073/pnas.95.21.12432>
- Wu, Q., Peng, Z., Zhang, Y., & Yang, J. (2018). COACH-D: Improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Research*, 46(W1), W438–W442. <https://doi.org/10.1093/nar/gky439>
- Wu, X., Gu, L., Prior, R. L., & McKay, S. (2004). Characterization of anthocyanins and proanthocyanidins in some cultivars of *Ribes*, *Aronia*, and *Sambucus* and their antioxidant capacity. *Journal of Agricultural and Food Chemistry*, 52(26), 7846–7856. <https://doi.org/10.1021/jf0486850>
- Xu, F., Li, L., Zhang, W., Cheng, H., Sun, N., Cheng, S., & Wang, Y. (2012). Isolation, characterization, and function analysis of a flavonol synthase gene from *Ginkgo biloba*. *Molecular Biology Reports*, 39(3), 2285–2296. <https://doi.org/10.1007/s11033-011-0978-9>
- Yan, Y., Song, C., Falginella, L., & Castellarin, S. D. (2020). Day temperature has a stronger effect than night temperature on anthocyanin and flavonol accumulation in ‘Merlot’ (*Vitis vinifera* L.) grapes during ripening. *Frontiers in Plant Science*, 11(July), 1095. <https://doi.org/10.3389/fpls.2020.01095>
- Yoshihara, N., Imayama, T., Fukuchi-Mizutani, M., Okuhara, H., Tanaka, Y., Ino, I., & Yabuya, T. (2005). cDNA cloning and characterization of UDP-glucose: Anthocyanidin 3-O-glucosyltransferase in *Iris hollandica*. *Plant Science*, 169(3), 496–501. <https://doi.org/10.1016/j.plantsci.2005.04.007>
- Zhang, X., Xu, G., Cheng, C., Lei, L., Sun, J., Xu, Y., Deng, C., Dai, Z., Yang, Z., Chen, X., & Liu, C. (2021). Establishment of an *Agrobacterium*-mediated genetic transformation and CRISPR/Cas9-mediated targeted mutagenesis in hemp (*Cannabis sativa* L.). *Plant Biotechnology Journal*, 19(10), 1979–1987. <https://doi.org/10.1111/pbi.13611>
- Zhang, Y. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7), 2302–2309. <https://doi.org/10.1093/nar/gki524>
- Zhao, Z. C., Gui Bing, H., Fu Chu, H., Wang, H. C., Yang, Z. Y., & Lai, B. (2012). The UDP glucose: Flavonoid-3-O-glucosyltransferase (UFGT) gene regulates anthocyanin biosynthesis in litchi (*Litchi chinensis* Sonn.) during fruit coloration. *Molecular Biology Reports*, 39(6), 6409–6415. <https://doi.org/10.1007/s11033-011-1303-3>
- Zhu, X., Mi, Y., Meng, X., Zhang, Y., Chen, W., Cao, X., Wan, H., Yang, W., Li, J., Wang, S., Xu, Z., Wahab, A. T., Chen, S., & Sun, W. (2022). Genome-wide identification of key enzyme-encoding genes and the catalytic roles of two 2-oxoglutarate-dependent dioxygenase involved in flavonoid biosynthesis in *Cannabis sativa* L. *Microbial Cell Factories*, 21(1), 215. <https://doi.org/10.1186/s12934-022-01933-y>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gagalova, K. K., Yan, Y., Wang, S., Matzat, T., Castellarin, S. D., Birol, I., Edwards, D., & Schuetz, M. (2024). Leaf pigmentation in *Cannabis sativa*: Characterization of anthocyanin biosynthesis in colorful *Cannabis* varieties. *Plant Direct*, 8(11), e70016. <https://doi.org/10.1002/pld3.70016>