GENE DUPLICATION AND EXPRESSION OF SPECIALIZED METABOLIC PATHWAY GENES IN CANNABIS SATIVA

by

Christian Cizek

B.Sc., Northwestern University, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Gene Duplication and Expression of Specialized Metabolic Pathway Genes in *Cannabis Sativa*

submitted by Christian Cizek in partial fulfillment of the requirements for

the degree of Master of Science

in Botany

Examining Committee:

Keith Adams, Professor, Botany, UBC
Supervisor

Joerg Bohlmann, Professor, Botany, UBC
Supervisory Committee Member

Quentin Cronk, Professor, Botany, UBC
Supervisory Committee Member

Simone Castellarin, Associate Professor, Land and Food Systems, UBC
Additional Examiner

Abstract

Cannabis sativa is now widely cultivated for recreational and medicinal markets (drug-type cannabis) in addition to fiber and grain for materials and foodstuffs (fiber-type cannabis). Cannabis often is grown for its female flowers, which are highly concentrated with the specialized metabolites cannabinoids and terpenes. These compounds are specifically found in glandular trichomes, which are abundant on the surfaces of maturing cannabis flowers. However, both the abundance and identity of specialized metabolites can vary considerably across different cannabis varieties. Drug-type cannabis commonly contains the cannabinoid, THC, and fiber-type cannabis contains CBD. In this research, 24 gene families from the biochemical pathways responsible for cannabinoid and terpenoid production were analyzed across five different genomes of *Cannabis sativa*, containing a diversity of chemical and physical phenotypes. Orthologous genes from hops and three other Rosales species also were included. Gene duplication patterns and copy number variation were investigated using phylogenetic trees to define the evolutionary patterns within these biochemical pathways. Additionally, the duplicated genes within these pathways were analyzed for gene expression in several organ types of the Purple Kush variety. When comparing the terpenoid and cannabinoid pathways, both the non-mevalonate (MEP) and mevalonate (MVA) pathways contained fewer duplicated genes than the genes involved in cannabinoid biosynthesis. The evolutionary origins of the olivetolic acid cyclase (OAC) and aromatic prenyltransferase (APT) genes were revealed when comparing with hops and other closely related species. The gene expression analysis of the Purple Kush cultivar indicated that genes involved in both terpenoid and cannabinoid pathways were expressed highest in flowers. However, the number of expressed copies and expression levels varied among genes, and different copies are expressed in different organ types. Overall, this thesis provides insights to the evolutionary histories and gene expression patterns of the biochemical pathways involved in cannabinoid and terpenoid biosynthesis of *Cannabis sativa*.

Lay Summary

Cannabis sativa is widely cultivated for its value for both medicinal and recreational use. The highly valued chemical compounds, cannabinoids (THC and CBD) and terpenes (aromas), are found in flowers. However, the evolutionary history of cannabis remains relatively unknown, particularly as it pertains to cannabinoid and terpene production. The aims of this research are to identify the number of genes responsible for producing these chemical compounds and to quantify the relative abundance of these genes in different plant tissues. This thesis explores the variance of gene copy number across several closely related species and the tissue-specific abundance of these cannabinoid and terpene genes.

Preface

The present study was conducted under the supervision of Dr. Keith Adams. Christian Cizek is responsible for the conception and design of the study with the guidance of Dr. Keith Adams. The preliminary research and genomic data consolidation were performed by Christian Cizek. Transcriptomic data used in the study was collected from: van Bakel, H et al. The draft genome and transcriptome of Cannabis sativa is from *Genome Biol* **12**, R102 (2011). Data analysis and interpretation of results was completed by Christian Cizek with guidance from Dr. Keith Adams. The manuscript was written by Christian Cizek and was reviewed and revised by Dr. Keith Adams.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAE – acyl-activating enzyme

ALDH – aldehyde dehydrogenase

APT – aromatic prenyltransferase

BBE – berberine bridge enzyme

CannS – cannabinoid synthase

CBD(A) – cannabidiol(ic acid)

CBDAS – CBDA synthase

CBG(A) – cannabigerol(ic acid)

CMK – 4-diphosphocytidyl-2-C-methyl-D-erythrito kinase

DMAPP – dimethylallyl diphosphate

DXR – 1-deoxy-D-xylulose 6-phosphate reductoisomerase

DXS – 1-deoxy-D-xylulose 6-phosphate synthase

FAD – fatty acid desaturase

FPP – farnesyl diphosphate

FPPS – farnesyl diphosphate synthase

GGPP – geranylgeranyl diphosphate

GGPPS – geranylgeranyl diphosphate synthase

GPP – geranyl diphosphate

GPPS – geranyl diphosphate synthase

HDR – 4-hydroxy-3-methyl-but-2-enyl diphosphate reductase

HDS – 4-hydroxy-3-methyl-but-2-enyl diphosphate synthase

HMGR – 3-hydroxy-3-methylglutaryl-CoA reductase

HMGS – 3-hydroxy-3-methylglutaryl-CoA synthase

HPL – hydroperoxide lyase

IDI – IPP isomerase

IPP – isopentenyl diphosphate

LOX - lipoxygenase

MEP – methylerythritol phosphate

MVA – mevalonate

MPDC – mevalonate-5-phosphate decarboxylase

OAC – olivetolic acid cyclase

OLS/PKS – olivetol synthase/polyketide synthase

PK – Purple Kush

PMK – phospho-mevalonate kinase

THC(A) – Δ 9 -tetrahydrocannabinol(ic acid)

THCAS – THCA synthase

Acknowledgements

I would like to acknowledge my lab, starting with Dr. Keith Adams, Ryan Bailey, Tonya Severson, Grant de Jong, Yihan Wu, and John Lee.

Also, many thanks to my committee, Dr. Keith Adams, Dr. Joerg Bohlmann and Dr. Quentin Cronk for their support.

Dedication

This thesis is dedicated to my family, friends, and those who have supported me.

Chapter 1: Introduction

The cultivation of *Cannabis sativa* can be traced back thousands of years (Li, 1973). Known for its medicinal and industrial properties, it is a plant of many uses: fiber, grain, and specialized metabolites. *Cannabis sativa* varieties have been characterized for commercial uses as either drug-type or fiber-type. Morphologically, drug-type cannabis is a short and dense plant, often producing several flowers from a single plant. These varieties are highly concentrated with specialized metabolites, predominantly cannabinoids and terpenoids. Fiber-type cannabis is considerably more elongated in structure and produces significantly fewer flowers and metabolites in favor of fibrous stalks and nutritional seeds (Kojoma et al., 2006; Piluzza et al., 2013).

Within all types and varieties of cannabis, there is a tremendous amount of metabolic variation. Currently, over 100 cannabinoids have been identified and over 90 terpenoids have been quantified in only a select few varieties (Berman et al., 2018; Booth et al., 2020; Shapira et al., 2019). However, tetrahydrocannabinol (THC), cannabidiol (CBD), cannabigerolic acid (CBGA), and cannabichromene (CBC) are the most abundant cannabinoids. Additionally, flavonoids and anthocyanins responsible for different colorations have yet to be chemically studied. Despite significant metabolic variation, a classification system has been developed to differentiate cannabis varieties by chemotype, which uses cannabinoid content as the sole factor (Small and Beckstead, 1973). This is still commonly used to date. Furthermore, statistical analysis on cannabis metabolite content has proposed new classification systems based on cannabinoid and terpenoid content (Fischedick J, 2015; Fischedick, 2017; Richins et al., 2018). However, classification systems based on genotypes could be considered as an alternative.

The advancement of sequencing technologies has coincided with the recent, widespread legalization of *Cannabis*, leading to a significant number of genomic sequencing efforts. The genome is diploid (2n=20) with nine autosomal chromosomes and two sex chromosomes. Fortuitously, the array of published cannabis genomes is diverse, which includes type-I (high THC/low CBD), type-II (mixed THC/CBD ratio), and type-III (high CBD/low THC) varieties of the cannabinoid classification system. The first two genomes that were publicly released are ‘Purple Kush’ and ‘Finola,’ which are type-I and type-III varieties, respectively. The former is a drug-type cannabis plant high in THC content. It is presumed to be high in terpene content, but

there is currently no available chemical profiling for this variety. The 'Finola' variety, a fiber-type plant, is significantly lower in cannabinoid content and produces predominantly cannabidiolic acid (CBDA) (Laverty et al., 2019; van Bakel et al., 2011). Another group published the genome of the variety, 'CBDRx.' This type-III plant, defined by CBDA production, is high in cannabinoids and therefore classified as drug-type (Grassa et al., 2018). Lastly, the type-II variety 'Jamaican Lion' has had two genomes published, one male and one female (McKernan et al., 2020). The cannabis genomes that have been published to date are not entirely complete. However, these five genomes are near completion, with BUSCO scores all above 95% (Grassa et al., 2018; Laverty et al., 2019; McKernan et al., 2020).

Gene duplication is a major feature of plant genomes. Several types of duplication events occur in plant genomes, including whole genome duplication, tandem duplication, segmental duplication, duplication due to transposable elements, and retroduplication (e.g., Panchy et al., 2016). As it relates to terpenoid and cannabinoid biosynthesis, duplicated genes often cluster together (Chae et al., 2014; Nützmann and Osbourn, 2014) in the form of tandem duplicates, which are gene copies that are direct neighbors, and proximal duplicates which are gene copies that are located on the same chromosome but are separated by only 10 or fewer genes (Qiao et al., 2019). In cannabis, gene clusters are observed in both cannabinoid and terpene synthases, which are both highly duplicated gene families (Booth et al., 2020; Laverty et al., 2019).

To date, a majority of studies investigating gene duplication in cannabis have primarily focused on cannabinoid and terpene synthases (TPS), the final enzymes in their respective biochemical pathways. Gene duplication in *Cannabis* of cannabinoid pathway genes is the likely cause of the production of its most abundant specialized metabolites: THCA, CBDA, CBCA (Weiblen et al., 2015). The neofunctionalization of cannabinoid synthase duplicates has led to the biosynthetic ability to produce both THCA and CBDA, and likely, the broad spectrum of cannabinoids (Vergara et al., 2019). Additionally, recent duplication events of these enzymes have caused substantial copy number variation between different cannabis varieties. Terpene synthases have undergone significant duplication as well. The high copy number of TPS enzymes within the cannabis genome (33 in Purple Kush), in addition to many plant species, is due to both recent and distant duplication events (Keeling et al., 2008; Allen et al., 2019). There are a number of repetitive clusters indicative of tandem duplication events as well (Booth et al., 2020).

Although widely known for the ability to produce cannabinoids, the cannabis plant produces a plethora of phytochemicals, including terpenoids, phenolics, and many more. Cannabinoids belong to a unique class of terpenoids called terpenophenolics. Bitter acids in hops are other members of this chemical class, displaying an interesting example of evolution manifested biochemically (Kovalchuk et al., 2020). Both of these compounds have a relatively simple, phenolic base structure, containing prenylated isoprenoid groups (Figure 5). In hops and cannabis, these are compounds synthesized from one phenolic compound and isoprenoid precursors. In cannabis, the phenolic compound is produced from a medium-chain fatty acid and three malonyl-CoA molecules via a polyketide synthase (PKS). For THCA and CBDA, the cannabinoid precursor is hexanoyl-CoA, which is synthesized through a series of enzymes from primary fatty acid metabolism (Marks et al., 2009). From the pool of hexanoyl-CoA, three enzymatic steps, catalyzed by four enzymes leads to the production of cannabinoids (Figure 4). A polyketide synthase (PKS) and olivetolic acid cyclase (OAC) produce the phenolic compound, then an aromatic prenyltransferase (APT) uses a GPP molecule to synthesize CBGA, and finally, a cannabinoid synthase produces the final cannabinoid products, CBDA or THCA.

Terpenoids are a common feature of plant metabolism, as they are fundamental to both primary and specialized metabolism. The phytol tail of chlorophyll is a derived from a terpenoid synthesized in the chloroplast, and volatile defense and attraction compounds can be produced in several compartments of the cell (Bohlmann et al., 1998). This is due to the two pathways responsible for terpene precursors, the non-mevalonate (MEP) and mevalonate (MVA) pathways, found in the chloroplast and cytosol, respectively.

The biosynthesis of terpenoids is divided between two major pathways in the cell. The MEP pathway is located within the chloroplast and converts both glyceraldehyde 3-phosphate (G3P) and pyruvate to isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) through a series of seven enzymatic reactions (Figure 1). In the cytosol, the MVA pathway produces IPP, which can then be isomerized to DMAPP through a series of six enzymes (Figure 2). These molecules are subsequently condensed by geranyl diphosphate (GPP) synthase (GPPS) and farnesyl diphosphate (FPP) synthase (FPPS) to form precursors to monoterpenoids and sesquiterpenoids, respectively (Figure 3). GPPS produces the 10-carbon precursor, GPP, to cannabinoids and monoterpenoids and can produce the 20-carbon GGPP, as exemplified in hops (Wang and Dixon, 2009).

In this thesis, I studied the duplication history and expression patterns of 24 gene families from the biochemical pathways responsible for cannabinoid and terpenoid production in *Cannabis sativa* along with orthologous genes from hops and three other Rosales species. I asked the following questions:

- How does gene duplication manifest across the landscape of cannabis specialized metabolism?
- Can we gain insights into when these duplicates originated?
- How does gene duplication vary across different cannabis varieties?
- How does gene expression differ across these duplicates?

Chapter 2: Materials and Methods

Data sources: Four different cannabis cultivars were used for this analysis, 'Purple Kush', 'Finola', 'CBDrx', and 'Jamaican Lion' (male and female). These genomes and reference annotations are available at https://genomevolution.org/coge/ with identifiers: 53042, 53059, 53165, 55360, and 55184, respectively.

Short read RNA-seq data from 'Purple Kush' and 'Finola' (van Bakel et al. 2011) are available under NCBI accession PRJNA73819 (Table 1). 'Purple Kush' data were available for roots, stems, shoots (shoot tips with young leaves and apical meristems), pre-flowers (shoot tips with flower primordia but no visible stigmas), early-stage flowers (flowers with visible stigmas), and mid-stage flowers (flowers with visible, non-withered stigmas and conspicuous trichomes). 'Finola' data were available from mid-flower.

For constructing phylogenies, several species within the Rosales order were used: *Humulus lupulus* (hops), *Malus domestica* (apple), *Fragaria vesca* (strawberry), and *Ficus carica* (fig). *Medicago trunculata* from the Fabales order was an outgroup species. For *Medicago*, *Humulus*, *Malus*, and *Fragaria*, the genomes and corresponding annotations are available at https://genomevolution.org/coge/ with identifiers: 42051, 53674, 54783, and 58093, respectively. The genome and mRNA coding sequences of *Ficus carica* are available at https://plantgarden.jp/.

Sequence Identification and Analysis: Genes implicated in the cannabinoid, MEP, and MVA pathways of cannabis were download from GenBank. These genes were used in tBLASTn searches to identify orthologous genes in the genome assemblies of the other studied species. The results of these searches were inputs for a species specific BLASTn search to identify additional homologues. The results of the second BLAST search were confirmed via BLASTx to verify the predicted gene function. Transit peptides were predicted based on translated sequences using a consensus from several prediction tools: TargetP 2.0 (Almagro Armenteros et al., 2019), MultiLoc2 (Blum et al., 2009), YLoc (Briesemeister et al., 2010), SherLoc2 (Briesemeister et al., 2009), DeepLoc (Almagro Armenteros et al., 2017), and LocTree3 (Goldberg et al., 2014).

Several gene models proposed by *Cannabis sativa* reference gene annotations were incomplete, with two separate proposed genes comprising two halves of a published sequence. In order to fill gaps, RNA-seq data was used to construct novel gene models. The reference annotation was manually edited by replacing the incomplete gene models with one refined by RNA-seq analysis.

Phylogeny Construction: Sequences were examined and translated to construct coding sequences. Amino acid sequences from each species were aligned using the MUSCLE algorithm (Edgar, 2004). The alignments were visually inspected, and trees were constructed using RAxML (Stamatakis, 2014) using a GTR substitution model with 250 bootstrapping replicates. Bootstrap values are displayed on tree branches as a percentage of the replicates.

RNA-Seq Analysis: Raw FASTQ files from six different organ types and developmental stages from 'Purple Kush', and mid-flower tissue from 'Finola', were trimmed and quality filtered using Trimmomatic v0.36 with default settings for Illumina single-end reads. The resulting reads of 'Purple Kush' and 'Finola' were mapped to their corresponding genomes using HISAT2 v2.2.0 with default settings (Kim et al., 2015), but the multi-mapping function was eliminated, and reads were only mapped once to the genome. A reference guided genome was calculated using StringTie v2.1.4 (Kovaka et al., 2019).

Due to the presence of recently duplicated genes, in some cases the sequence similarity was high enough that RNA-seq reads were unable to be differentiated between two duplicated genes. Therefore, in these cases, the aggregate expression was used and applied for these similar duplicates. Thus, we acknowledge RNA-seq could not differentiate expression levels for these highly similar genes.

Chapter 3: Results

3.1: MEP Pathway

DXS: The first step in the MEP pathway is catalyzed by the DXS gene. DXS genes analyzed from Rosales and *Medicago truncatula* were separated into three clades, which is indicative of at least two ancient duplication events (Figure 6). Each clade contained at least one DXS gene from each species except for clade 3, which was missing genes from both hops and cannabis. Additionally, within clade 2, DXS genes of fig (*Ficus carica*), hops (*Humulus lupulus*), and cannabis were all duplicated in an ancient duplication event. Of all species analyzed, apple (*Malus domestica*) had the most DXS copies with six. Both strawberry (*Fragaria vesca*) and *Medicago* genomes contains three DXS copies each, and both fig and hops genomes contain four DXS copies each.

Within cannabis, there is DXS copy number variation between varieties, with CBDRx and both Jamaican Lion genomes having three copies, whereas purple Kush and Finola contain four copies (Table 2). Gene expression analysis in Purple Kush organs showed that only one specific copy of DXS is actively expressed, *Cs_PK_DXS_1* (Figure 7). Furthermore, DXS is most highly expressed in mid flower. The expression values in early flower, pre flower, shoot, and stem were all very similar and slightly less than mid flower, but expression in root was substantially lower.

DXR: DXR catalyzes the subsequent reaction in the MEP pathway. In contrast to the phylogenetic tree of DXS, there is one major clade among the species analyzed (Figure 8). There are multiple copies of DXR in many species, but these are likely due to more recent duplication events. In Rosales, the strawberry genome contained the most DXR copies with four. *Medicago*, apple, and hops each have two specific copies, and fig only has one DXR copy. Of each cannabis variety analyzed, Purple Kush contained the most DXR copies with four. Both Finola and the Jamaican Lion male genomes contain three copies, and CBDRx and the Jamaican Lion female genomes each have only two DXR copies. Within cannabis, the DXR gene is present in a tandem duplicate array. In each variety except Finola, one DXR gene belongs to the Duplication 1 group and the other belongs to the Duplication 2 group. In Finola, there are 3 DXR tandem duplicated genes and two belong to the Duplication 1 group.

Gene expression patterns of DXR in cannabis are clearly different from those of DXS. Multiple copies are expressed from both duplication groups (Figure 7). In Purple Kush, the DXR copy with consistently the highest expression, *Cs_PK_DXR_2b*, is not located within the tandem duplicate array. Gene expression is clearly the highest in the flowers. Of the two tandem duplicates, there is a small amount of variance in organ-specific gene expression, with the copies expressed at either a similar level or Duplication 1 is expressed at a higher level.

MCT: The MCT gene catalyzes the third overall step in the MEP pathway. The phylogenetic tree structure of MCT is similar to that of DXR in that there are no ancient duplication events identified (Figure 9). Similarly, there are several recent duplications in various species. In the Rosales species, strawberry and apple have three and two MCT copies, respectively. Each of the other analyzes species has only a single MCT copy. In cannabis, both Purple Kush and Finola genomes have MCT tandem duplicates, and each additional analyzed variety only has a singular MCT copy. In Purple Kush, the two MCT copies were both expressed in several organs (Figure 7). One of the tandem duplicates (*Cs_PK_MCT_2*) has higher gene expression in each organ type. Both MCT copies showed the highest level of expression in shoot, pre flower, and early flower, whereas roots had the lowest levels of expression.

CMK: At the center point of the MEP pathway, the CMK gene encodes the enzyme for the fourth step. The CMK gene in Rosales provides a phylogenetic tree structure that follows similar trends seen in the previous two genes (Figure 10). There are the fewer copies among the analyzed species than for other genes in the pathway. In most species, and in some cannabis varieties, there is only a singular copy of the CMK gene. The varieties of cannabis that have two copies of CMK are Purple Kush and Finola. The two copies of CMK in Purple Kush were too similar in sequence to properly differentiate gene expression values, but it is clear that CMK is expressed the highest in early flower (Figure 2). Furthermore, CMK appears to be the lowest expressed member of the MEP pathway, potentially indicative of a limiting step of the metabolic flux.

MDS: The next gene in the MEP pathway is MDS, which adds a second phosphate group to create the first diphosphate intermediate. In the phylogenetic tree for MDS, no novel ancient

duplication events were identified, but there are several recent duplications across the Rosales species (Figure 11). The hops genome has the most MDS copies with four, and apple has three. The remainder of the analyzed species each have a single copy. Also, each cannabis variety, with the exception of the Jamaican Lion male, has a single MDS gene. Gene expression of the single copy MDS gene in Purple Kush showed that expression is highest in flowers and shoots (Figure 7). Expression was relatively equal among the three flower developmental stages and slightly higher than shoots.

HDS: The penultimate gene in the MEP pathway is HDS. The phylogenetic tree is similar to many of the trees throughout the MEP pathway, containing no ancient duplication events (Figure 12). Species outside of cannabis have one copy, except for apple which has two. Purple Kush is the only cannabis variety with multiple copies, also having two. Both copies of HDS in Purple Kush are expressed in all organs surveyed, with higher levels of HDS2 in all organ types (Figure 7). Both copies were most highly expressed in mid flower and had high expression levels in other flower stages and in shoots.

HDR: The final gene involved in the MEP pathway is HDR. This phylogeny of this gene is consistent with the majority of MEP pathway genes in that it does not have an ancient duplication event (Figure 13). Furthermore, the only Rosales species found to have multiple copies is apple, containing two HDR genes. Within cannabis, all varieties except Finola have a single HDR gene in their respective genomes. Finola was found to have two HDR genes, and they are highly similar in sequence. Gene expression analysis of the HDR gene in Purple Kush showed that it was the highest expressed gene throughout the entire MEP pathway, with the highest level of expression occurring in mid flower (Figure 7). Gene expression was found to be very high in the two additional flower stages as well as in shoots. Gene expression in stem and in root were also substantially higher relative to the expression of other MEP pathway genes in these two organ types.

3.2: MVA Pathway

HMGS: The first step of the MVA pathway is catalyzed by the HMGS gene. Among the analyzed species, several recent duplication events were identified, but there were no ancient duplication events. The phylogenetic tree structure is similar to many previously described from the MEP pathway (Figure 14). Species that were found to have multiple duplicates of the HMGS gene were *Medicago*, apple, and hops (Table 3). Each of these species has two separate copies, and these appear to be recent duplications unique to each species. Within cannabis, Purple Kush with two copies is the only variety with multiple copies of HMGS. The sequence similarity between the 2 HMGS copies in Purple Kush is nearly identical, making it impossible to differentiate gene expression between the two. The collective gene expression of both HMGS genes in Purple Kush is equal across stem, shoot, and flower organs, but substantially higher in root (Figure 15). This is clearly different than the genes in the MEP pathway where flowers have the highest expression levels.

HMGR: The second step of the MVA pathway is catalyzed by the HMGR enzyme. The phylogenetic tree structure is the most complex of all the genes in the pathway because it contains the most copies across all species (Figure 16). The phylogenetic tree shows two Rosales clades of HMGR genes and all genes in the *Medicago* outgroup branching separately, suggesting an ancient duplication event specific to Rosales.

In addition to multiple copies resulting from ancient duplications, recent duplication events have led to increased copy number in these species (Table 3). The outgroup species, *Medicago*, has seven total HMGR copies across the phylogeny. The apple genome has the most copies with nine total HMGR genes, with four located in the first clade, zero in the second clade, and five in the third clade. The strawberry genome has five HMGR copies in the first clade, one in the second clade and one in the third clade, totaling seven genes. In fig, there are only two copies of HMGR present, with each belonging to clade one and two. The hops genome contains a total of six genes, with two located in the first clade and four located in the second clade. In cannabis, the varieties have variable copy numbers. The Finola genome had the most HMGR copies with two copies in the first clade and one copy in the second, totaling three. Purple Kush, CBDRx, and the Jamaican Lion female each have two HMGR copies belonging to the first two

clades. The Jamaican Lion male only has a single copy of HMGR located in the first clade. The gene expression analysis of both HMGR copies in Purple Kush revealed interesting patterns across the two duplicates (Figure 15). In mid flower, both copies were expressed at a similar, high level. However, in early flower, pre flower, shoot and stem, the second clade copy (*Cs_PK_HMGR_1*) was expressed substantially higher. In roots, *Cs_PK_HMGR_1* was highly expressed, near the level of flowers, but the first clade copy (*Cs_PK_HMGR_2*) was expressed approximately four-fold.

MVA Kinase: The third enzyme in the MVA pathway is catalyzed by the MVA kinase. The phylogenetic tree is simple because there are no ancient duplication events, and there are very few recent duplications (Figure 17). The two species with multiple copies are apple and hops, with two genes encoding MVA kinases each. In cannabis, there is no variety specific variation, as each variety has only a single copy. In Purple Kush, MVA kinase has the lowest levels of gene expression throughout the MVA pathway (Figure 15). There is expression throughout all organ types, with the highest being in roots. Gene expression in roots was approximately four-fold higher than each of the other organ types.

PMK: PMK catalyzes the next step of the MVA pathway, forming a diphosphate intermediate. The phylogenetic tree of the PMK gene is nearly identical to that of the MVA kinase due to the similar duplication patterns and lack of ancient duplications (Figure 18). Recent duplications resulted in two PMK copies in both apple and hops. Each variety of cannabis only has a single PMK gene.

Expression of PMK in Purple Kush is highest in roots (Figure 15). This is a continuation of the pattern seen throughout the MVA pathway. PMK expression in flowers and shoots was approximately three times less than in roots, whereas expression in stems was approximately half.

MPDC: The final step of the MVA pathway is carried out by the MPDC enzyme, which synthesizes IPP. This is subsequently isomerized to DMAPP by additional enzymes. The phylogenetic tree for MPDC follows a similar pattern consistent with many of the findings previously stated (Figure 19). There are no ancient duplication events, but there are several

recent duplication events. In both apple and hops, recent duplications resulted in two copies of MPDC in each species. In cannabis, there is variety specific variation, with both Finola and Purple Kush genomes containing two MPDC copies as tandem duplicates. However, CBDRx and Jamaican Lion only have one copy of MPDC. Expression of the two MPDC genes in Purple Kush could not be fully differentiated because of the sequences were nearly identical (Figure 15). The combined expression of both MPDC copies was substantially higher in roots compared to other organ types, whereas the lowest expression was in flowers.

3.3: Geranyl/Farnesyl Diphosphate Enzymes

IDI: The precursors to geranyl-PP and products of the MEP and MVA pathways, IPP and DMAPP, can isomerize to each other via the IDI enzyme. The tree structure is simple, as there is no indication of an ancient duplication (Figure 20). There are a few recent duplication events across Rosales species (Table 4). These events occurred in both apple and hops which led to two and three IDI copies, respectively. Within cannabis, each variety only has a single IDI copy, with the exception of CBDRx, where the genome was missing the IDI gene. In Purple Kush, the single copy of the IDI gene was highly expressed in all organ types (Figure 21). The highest expression levels were in roots and in mid flowers, which both had similar levels of expression. Expression in early flower was lower but was substantially higher than in stems, shoots, and pre flowers.

GPPS: The first step after the MEP and MVA pathways is the formation of the direct monoterpenoid and cannabinoid precursor compound, GPP. The enzyme, GPPS, combines one DMAPP and one IPP together forming geranyl diphosphate. This enzyme is interesting because it is comprised of both a small subunit and a large subunit. The products of some GPP synthases can vary across species, producing either GPP or GGPP (Wang and Dixon, 2009). The phylogenetic tree confirms three different clades of GPPS genes (Figure 22). The first corresponds to a small subunit GPPS, which generally is responsible for only GPP production. The second clade represents the heterodimeric small subunit group, which is also involved in the production of GPP. Lastly, the third clade is comprised of large subunit group GPPS. This

subclade encodes GPPS enzymes that are able to catalyze the formation of either GPP or GGPP molecules, depending on the species-specific enzyme.

Within each clade, *Medicago* GPPS genes are identified, which indicates two ancient duplications occurred before the divergence of Fabales and Rosales from a common ancestor. The GPPS small subunit clade indicates a duplication event that is specific to both hops and cannabis, which led to two GPPS.ssu copies in both species. In addition, there are cannabis variety specific differences within this clade. There is only one copy in CBDRx, located in the 1a subclade. The Jamaican Lion male genome has two copies, with one in each subclade. Purple Kush, Finola, and the Jamaican Lion male genomes each have two copies, with each unique to the subclades. Interestingly, the duplication in hops resulted in tandem duplicates, but in cannabis, these duplicates are no longer in tandem arrays. Additionally, GPPS.ssu genes seem to have been lost in both apple and strawberry, as they were not found in either species.

In the GPPS heterodimeric small subunit clade, each species has representative genes, but there is no indication of more ancient duplication events. Apple, fig, and hops each have two copies, whereas *Medicago* and strawberry both have a single copy. Among cannabis varieties, CBDRx, Finola, and the Jamaican Lion female genomes each have a single copy within the heterodimeric clade. Purple Kush and the Jamaican Lion male both have two copies, but the duplicates are in a tandem array in the Purple Kush genome only.

The final clade of the GPPS genes includes enzymes responsible for the large subunit and GGPPS activity. Similar to the other clades, this does not have an indication of ancient duplication events, but there are many recent duplication events. Two unique *Medicago* genes were identified, and apple and strawberry were found to have four and three GPPS.lsu copies, respectively. The fig genome has four copies as a result of recent tandem duplication. In Cannabaceae species, hops has two GPPS.lsu copies, but cannabis only has a single copy. This is consistent across all cannabis varieties that were analyzed.

Gene expression analysis of the GPPS genes showed variance across both clades and organ types (Figure 23). Within the small subunit clade, *Cs_PK_GPPS_1a* was expressed in all organs, with highest expression in early flower and mid flower, and much lower expression in roots and stems. In contrast, *Cs_PK_GPPS_1b* was not expressed at all. The GPPS heterodimeric subunit genes were tandem duplicates in Purple Kush, and both were actively expressed. The second gene in the array, *Cs_PK_GPPS_2b*, was expressed at higher levels in all

organ types. This copy was expressed the highest in roots and at moderate levels in the other organ types. The first gene in the tandem array, *Cs_PK_GPPS_2a*, was expressed highest in early flower, but it was still expressed at a lower level than *Cs_PK_GPPS_2b* in all organs. The GPPS large subunit in Purple Kush, *Cs_PK_GPPS_3*, was expressed in all organs, but peaked in shoot, pre flower, and early flower. Gene expression levels dropped by over half in mid flower and the gene was also expressed lower in roots and stems.

FPPS: The direct precursor to sesquiterpenoids, farnesyl diphosphate, is produced from one DMAPP and two IPP molecules by FPPS. The phylogenic tree of FPPS reveals an ancient duplication event that is specific to Rosales, resulting in major two clades within Rosales (Figure 24). In the first clade, only apple has undergone recent duplication within its genome, containing two FPPS copies. Each of the other species included in the analysis is single copy within this clade. Additionally, there is no variety-specific copy number variation in cannabis. In the second clade, apple, strawberry, and hops have undergone recent duplications with each possessing two FPPS copies. Also, there is no copy of FPPS in the second clade. In cannabis, each FPPS gene in the second clade is a single copy, meaning there is no variety specific variation in copy number. Gene expression analysis of FPPS copies in Purple Kush showed that both are actively expressed in all organs (Figure 23). In stems, shoot, pre flower, and early lower, the gene expression values are very similar between the two copies. However, in roots and mid flower, the FPPS copy from clade one, *Cs_PK_FPPS_1*, was expressed substantially higher than *Cs_PK_FPPS_2*. Additionally, these two organs showed similar levels of expression.

3.4: Pre-Cannabinoid Pathway

Desaturase: The addition of a double bond to an oleic acid molecule is catalyzed by a desaturase enzyme, which is critical for the bridge between primary metabolism and cannabinoid synthesis. The phylogenetic tree is comprised of two separate clades, indicative of ancient duplication events occurring prior to the divergence of the Rosales and *Medicago* from a common ancestor (Figure 25). Several recent duplication events within the analyzed species have occurred, but in the case of fig, there is no desaturase gene in the lower clade. Also, within the

lower clade, there have been several duplication events specific to Cannabaceae leading to three subclades in hops and cannabis.

Within the first clade of desaturase genes, *Medicago*, strawberry, fig, and some cannabis varieties each have a single desaturase copy. Two copies are present in apple, which is a result of recent duplication. Within cannabis, Finola has two desaturase copies that are a result of tandem duplication. The second clade contains single copies of desaturase from both apple and strawberry. Within this clade, hops and cannabis desaturase genes are organized into three subclades. In both species, proximal and tandem duplicates are located in all three subclades, indicating that a tandem duplication of desaturase genes in a common ancestor has been conserved in Cannabaceae. However, further expansion of desaturase genes has occurred within cannabis. The duplication patterns substantially vary across cannabis varieties. In Finola, there is a single duplicate cluster, containing six consecutive tandem duplicates (2, 3a, 3b, 4a, 6a, 6b), whereas in CBDRx, there are two proximal clusters, each containing two tandem duplicates (2,3 and 4,5). In Purple Kush, there is a single tandem duplicate array containing four duplicates (2b,3a, 3b,3c), which is in close proximity to three additional, proximal duplicated genes (2a, 4, 5). In both male and female Jamaican Lion, there are several tandem duplicate clusters. The first array contains three genes (4a, 4b, 5a) and the second has only two duplicates (3,6). However, there is one additional pair of tandem duplicates in the male Jamaican Lion (4c, 5b).

In Purple Kush, gene expression varies among the different genes of the tandem array (Figure 26). The first three tandem duplicates, Cs_PK_FAD_2a, Cs_PK_FAD_2b, Cs_PK_FAD_3a, are all expressed in each tissue type, but Cs_PK_FAD_3a has the highest expression level of the three. The highest expression level is in shoots and in flower stages, whereas expression is lower in roots and stems. The fourth gene in the tandem array, Cs_PK_FAD_3b, is expressed even less and only in shoots and flowers. The fifth gene in the tandem array, Cs_PK_FAD_3c, was not expressed in any of the cannabis organs studied here. The next two desaturase genes, Cs_PK_FAD_5 and Cs_PK_FAD_4, are both expressed in only mid flowers with the former being expressed approximately four times higher than the latter. The only desaturase gene not located in a tandem array, Cs_PK_FAD_1, was highly expressed in all organ types. This gene was expressed highest in roots, shoots, and pre flowers.

LOX: The next step in the pre-cannabinoid pathway is the lipoxygenase (LOX) enzyme, which begins the cleavage process of the double bond on the linoleic acid molecule. There are many members of the LOX gene family because of the varying lengths of fatty acid chains. Within the LOX phylogenetic tree, there is evidence of several ancient duplication events (Figure 27). However, because bootstrap values are low, the precise timing of a few of these events is not discernable. Within the LOX gene family, there are two distinct groups of genes in the Rosales that show evidence of an ancient duplication event. The first group is comprised of the LOX_1 genes, and the second groups contains many more LOX copies, LOX_2-6. Subsequent duplications, particularly in Cannabaceae, have led to a considerably larger Group 2 clade. The first clade shows several different tandem duplication events in both apple and strawberry that are specific to each species (Figure 27). Md_FAD_1a-c and Md_FAD_1d-e are two different tandem duplicates in apple, and Fv_FAD_1a-d and Fv_FAD_2a-b are two tandem duplications in strawberry. In contrast, desaturase genes in hops, fig, and some cannabis varieties have not been duplicated within this group. In the second clade, there have been additional duplications specific to fig, hops, and cannabis, resulting in four subclades. The fig duplicates are located in only two of the subclades and they are all in a single tandem duplicate array. Additional duplications are present in hops and cannabis, which has resulted in four total subclades. Several of these duplicated genes are tandem duplicates, found in clusters with genes of different LOX subclades.

The hops genome contains a total of 16 LOX genes, and many of these are located within a few tandem duplicate arrays. Of the cannabis varieties, Purple Kush has the most LOX genes with 14 total copies. Both CBDRx and Finola have six LOX genes. Within Jamaican Lion, the female and male genomes have six and ten LOX gene copies, respectively. The proximity of these duplicates varies across cannabis varieties. In both CBDRx and Finola, all LOX duplicates are located in a single tandem duplicate array. In Purple Kush, 10 of 14 LOX duplicates are located in three different tandem duplicate clusters. The first array contains six consecutive LOX genes (1a, 2a, 2c, 4a, 5a, 6a) and the other array contains four genes (1b, 4b, 5b, 6b). Two additional genes, Cs_PK_LOX_2b and Cs_PK_LOX_2d, are located in another tandem array, which is proximal to the six gene cluster. The other two genes, Cs_PK_LOX_3a and Cs_PK_LOX_3b, are located proximal to one of the arrays. In the Jamaican Lion female, the duplicates are arranged in two tandem duplicate arrays, containing two and four duplicates.

Duplicated genes in the Jamaican Lion male are organized in three unique clusters, comprised of two, three and four tandem duplicates.

Gene expression analysis of the LOX genes in Purple Kush indicated that organ-specific expression patterns vary across genes in the tandem duplicate arrays (Figure 28). The first gene in one of the arrays, Cs_PK_LOX_2b, was primarily expressed in early flowers but maintains lower levels of expression in shoot and mid flowers. The next gene in this array, Cs_PK_LOX_2d, was expressed at a very low level. In a separate tandem array, Cs_PK_LOX_6a, was expressed at very low levels and Cs_PK_LOX_5a was expressed at a very high level only in mid flowers. The next two genes in the array, Cs_PK_LOX_1a and Cs_PK_LOX_4a, were expressed in all organs except roots. Both were expressed highest in mid flowers and were expressed in other flower stages and shoots at lower levels. The fifth gene in the array, Cs_PK_LOX_2a, was expressed in all organs, but expression was greatest in mid flowers, early flowers, and shoots. The final gene of this array, Cs_PK_LOX_2c, showed little to no expression in all organs. In the third tandem array, the first two genes, Cs_PK_LOX_1b and Cs_PK_LOX_4b, were most highly expressed in mid flowers, whereas each were expressed at lower levels in every other organ except roots. The final two genes of the array, Cs_PK_LOX_5b and Cs_PK_LOX_6b, showed low levels of expression in each organ type except mid flowers, where there was a considerably higher level of expression. Both Cs_PK_LOX_3a and Cs_PK_LOX_3b, which are proximal duplicates, were not expressed in any of the studied organ types of Purple Kush.

HPL: The hydroperoxide lyase (HPL) catalyzes the break of the fatty acid chain. In pre-cannabinoid synthesis, one of the products is hexanal, the six-carbon precursor to THC and CBD. The phylogenetic tree is relatively simple compared to the other pre-cannabinoid pathway genes (Figure 29). There is a single *Medicago* HPL gene at the base of the tree. Within the Rosales species, there is no evidence of ancient duplication, and there is only evidence of more recent duplication in some species. Both strawberry and apple have two HPL copies, whereas fig and hops both have a single copy (Table 5). Within cannabis, CBDRx and both Jamaican Lion genomes only contain a single HPL gene. In contrast, both Purple Kush and Finola contain two copies of the HPL gene, and in both varieties the genes are tandem duplicates. In Purple Kush, gene expression between the two HPL copies varies, as the first duplicate, Cs_PK_HPL_2,

shows substantially lower expression than the second duplicate, Cs_PK_HPL_1, in all organs and flower stages (Figure 30). The first duplicate was solely expressed in flowers, and the second duplicate was expressed in all tissue types. Expression was highest in shoots and pre flowers, but gene expression in the two other flower stages, especially mid flower, was still high.

ALDH: One hypothesis for the synthesis of cannabinoid precursors is that hexanal is oxidized to hexanoic acid by an aldehyde dehydrogenase (ALDH). The genes that encode ALDH enzymes are part of a large superfamily that is conserved across nearly all living organisms. In this analysis, a smaller subset of ALDH genes were selected which are potentially linked to the biosynthesis of cannabinoid precursors (Marks et al., 2009). The ALDH superfamily contains many ancient duplication events that have been covered outside of the scope of this analysis. Subsequently, within the clade hypothesized to be involved in cannabinoid biosynthesis, there is no indication of additional ancient duplication events, but instead there are species-specific, recent duplications (Figure 31). Furthermore, within the specific ALDH clade of interest, both apple and hops have two copies each. *Medicago*, strawberry, fig are all species with a single copy of ALDH within this subclade. In cannabis, each variety has a single copy with the exception of Purple Kush, which has two ALDH genes. The gene expression of ALDH in Purple Kush shows that the two duplicates exhibit a very similar expression pattern across all tissue types (Figure 32). The highest expression is in mid flower but was consistently high across all flower stages. The lowest levels of expression were found in stem and shoot.

AAE: In order to undergo further cannabinoid biosynthesis, hexanoic acid must be carried by a coenzyme-A molecule. An acyl-activating enzyme (AAE) is responsible for creating a bond between the carboxylic acid and the thiol carrier molecule. Plants are comprised of many AAE genes which allow for binding with many different types of acyl molecules. Due to the large size of the AAE gene family, a phylogeny was constructed with the characterized AAE enzyme for cannabinoid biosynthesis (Figure 33). There is no evidence of ancient duplication within this specific subclade. There has been tandem duplication within cannabis to create two separate clades. Purple Kush has a third AAE gene which is an additional tandem duplicate. In Purple Kush, gene expression of the first two tandem duplicates, Cs_PK_AAE_2a and Cs_PK_AAE_2b, was substantially higher than the final gene in the array (Figure 34). Both of

these AAE genes were expressed highest in mid flower and had high levels of expression in shoots and other flower stages. The final gene in the array, Cs_PK_AAE_1, was expressed substantially lower and in similar amounts in all tissue types.

3.5: Cannabinoid Pathway

PKS: The first step of the cannabinoid pathway is catalyzed by a polyketide synthase (PKS), which has been characterized as an olivetolic acid synthase (OLS). This class of enzyme, which is common to many plant species, is substantially duplicated throughout the species that were analyzed. The PKS gene family is complex and contains multiple ancient duplication events that likely occurred prior to the split between Fabales and Rosales from a common ancestor (Figure 35). Several different clades of PKS genes occur, and in some cases, species are missing from certain clades. In others, there is a large number of duplicates, often resulting from recent duplication. In Purple Kush and Finola, there is a total of 10 PKS gene copies, which is the largest number of all the cannabis varieties. Both the male and female Jamaican Lion genomes have nine PKS copies, and CBDRx has the fewest with seven copies. Hops has more PKS genes than cannabis with a total of 16 copies, many of which belong to several tandem duplicate arrays. In cannabis varieties, there is a variance in the number of duplicates in different clades of the phylogenetic tree. In the first clade, there are three PKS duplicates in Finola, two in the Jamaican Lion female, but only single copies in the remaining varieties. Similarly, in the lowest clade, there are differing numbers of tandem duplications across varieties. CBDRx (5,6a,6b), Finola(5,6a,6b), and Purple Kush(5,6a,6b) each have three duplicates in a tandem array, but Jamaican Lion male (5,6b,6c,7) and female (5,6a,6b,7) have four duplicated genes in a tandem array. Also, this specific set of tandem duplications contains the functionally characterized PKS, which emphasizes its importance to cannabinoid biosynthesis (Raharjo et al., 2004).

Gene expression analysis in Purple Kush indicated that four of ten PKS genes were actively expressed in cannabis tissues (Figure 36). Within one of the tandem duplication arrays, the first gene, Cs_PK_PKS_5, was not highly expressed in any tissue type. The second gene in the array, Cs_PK_PKS_6a, is expressed considerably higher with maximum expression in mid flower. However, there was little to no expression in roots and stems. The final PKS gene in the

array, Cs_PK_PKS_6b, is expressed considerably higher than the second gene in all organs and flower stages except for roots and stems, where there was no expression. A set of proximal duplicates in the PKS family, Cs_PK_PKS_4a and Cs_PK_PKS_4b, have lower expression levels. The PKS gene Cs_PK_PKS_6c is expressed at a very high level in mid flowers followed by early flowers but has no expression in roots and stems. The gene Cs_PK_PKS_2 was expressed highest in shoots and less in pre-flowers. This gene is also expressed at a high level in early flowers and mid flowers but is lowly expressed in roots.

OAC: The olivetolic acid cyclase (OAC) works in tandem with the OLS to synthesize the cannabinoid precursor, olivetolic acid. This enzyme is derived from a member of the DABB protein family, which was used to construct this phylogeny. The structure of the tree is comprised of two main clades (Figure 37). The first clade is relatively simple with no indication of any ancient duplication events, and there are only duplications that have resulted from recent events. The second clade is more complex because of the large number of cannabis duplicates, from which the OAC gene responsible for cannabinoid biosynthesis arose (Figure 37). Additionally, the OAC gene has undergone two recent tandem duplications in the Cs_OAC_3 and Cs_OAC_4 groups, which are responsible for the higher copy number within the subclade. Hops has two copies in the first clade which resulted from recent duplication and one gene in the second clade. In the first clade, each cannabis variety has a single copy except the Jamaican Lion male, which has two copies. In the second clade, the most basal cluster of genes is comprised of a single copy in both CBDRx and two copies in Finola and Purple Kush. This specific cluster does not contain any genes from Jamaican Lion. The two unique clusters at the distal end of the tree contain genes from each of the cannabis varieties that were analyzed. In the first cluster, there are three copies in Purple Kush, and two copies in CBDRx and Jamaican Lion. There is also a short branch in this subclade, belonging to Finola (Cs_Finola_OAC_3). In the lowest cluster, there are three copies in Jamaican Lion male and Purple Kush, two copies in Jamaican Lion female, and single copies in Finola and CBDRx.

Gene expression analysis within the OAC/DABB family of Purple Kush revealed that a single member of a four gene tandem array (Cs_PK_OAC_3c) was primarily responsible for cannabinoid biosynthesis (Figure 38). This OAC gene is the second in this tandem array. This gene was expressed at high levels in all organs except roots and stems and was highest in the

later flower developmental stages. Furthermore, Cs_PK_OAC_3c which has been functionally characterized, directly participates in cannabinoid biosynthesis (Gagne et al., 2012). The first gene of this OAC tandem duplicate array is Cs_PK_OAC_3a, which is the second highest expressed gene of these duplicates. The expression occurred only in shoots and flower stages. The next gene in the tandem array, Cs_PK_OAC_3b, exhibited a very different expression pattern, with low levels of expression in all examined organs/stages. The final gene in this tandem array, Cs_PK_OAC_2a, was ubiquitously expressed across all tissues, albeit considerably lower than Cs_PK_OAC_3c. The OAC/DABB duplicate that is located in the same phylogenetic cluster, Cs_PK_OAC_2b, has similar expression patterns and levels across all tissues. Outside of the OAC specific cluster are tandem duplicates, Cs_PK_OAC_4a and Cs_PK_OAC_4b, which are proximal to the duplicate, Cs_PK_OAC_4c. These three genes share a similar pattern in which there is little to no expression in roots, stems, and mid flowers, but expression is highest in shoots, followed by pre flowers and early flowers. Furthermore, the second tandem duplicate, Cs_PK_OAC_4a, is expressed significantly lower than the first duplicate despite having a similar expression pattern. The OAC/DABB gene from the first clade, Cs_PK_OAC_1, is expressed in all tissues, but considerably higher in stems.

APT: The formation of CBGA from olivetolic acid and GPP is catalyzed by an aromatic prenyltransferase (APT). The phylogenetic tree of the APT gene family is primarily composed of cannabis genes due to the large number of recent duplications within the species (Figure 39). In the tree, there are two clades that include genes from all species, indicating a shared ancient duplication before the Rosales and Fabales diverged from a common ancestor. There are two *Medicago* APT genes, which are at the base of both clades. Both strawberry and fig have two copies each and are located in both clades. Apple has a total of four APT genes, which have resulted from recent duplication in both clades. Hops has eight total APT genes, with two in the first clade and six in the second clade. In cannabis, Finola has seven total copies, and CBDRx has eleven. Both of these varieties have two genes in the first clade but vary in the number of genes in the second clade. Purple Kush has three genes in the first clade and six genes in the second clade. Both male and female Jamaican Lion genomes have two APT genes located in the first clade, but in the second clade, the male and female genomes have ten and nine APT genes,

respectively. Thus, there is quite a bit of variation in copy number of APT genes in different cannabis varieties (Table 5).

The first clade contains two copies of APT in cannabis derived from a duplication during the evolution of the Cannabaceae. The second clade is much more complex due to the increased number of duplications particularly within Cannabaceae. The larger subclade, which contains only hops and cannabis APT genes is the location of the functionally characterized APT gene (Cs_APT_6) for cannabinoid biosynthesis (Rea et al., 2019). In cannabis, several of these duplicates are organized into tandem duplicate clusters. In CBDRx and the Jamaican Lion female, six APT genes comprise tandem duplicate arrays (3b, 3c, 4b, 4c, 5, 6 and 3c, 4a, 4b, 4c, 5, 6). However, in Purple Kush (3b, 4, 5, 6) and Finola (3, 4, 5, 6), the tandem arrays containing the putative APT gene for cannabinoid biosynthesis only contain four genes. In the Jamaican Lion male, there are two tandem arrays each containing four duplicated genes (3b, 4a, 5a, 6a and 3c, 4b, 5b, 6b). Hops has 3 genes (4a, 4b, 4c) that branch with the preceding tandem duplicate arrays in cannabis, where Hl_APT_4a and Hl_APT_4c are tandem duplicates. Another pair of APT genes in hops (2a, 2b) branch with single copy cannabis genes.

In Purple Kush, the characterized APT gene for cannabinoid biosynthesis, Cs_PK_APT_6, is located in a tandem array comprising four different genes. The first gene in this tandem array, Cs_PK_APT_5, is expressed in flowers and shoots (Figure 40). The second gene in the array, Cs_PK_APT_6, is the highest expressed APT, which is most highly expressed in mid flowers. However, this gene is not expressed in roots or stems and expression is substantially lower in shoots and pre flowers. The third APT gene in the tandem array, Cs_PK_APT_4, is highly expressed in shoots and flowers. The final gene in this tandem array, Cs_PK_APT_3b, is also the least expressed overall, with expression in only shoots and flowers. Outside of the tandem array but in the expanded APT clade, Cs_PK_APT_3a, is not highly expressed in any tissue type and not present in roots. Cs_PK_APT_2 is expressed in all tissue types, with maximum expression in shoots, pre flowers, and early flowers. Cs_PK_APT_1c is expressed in a similar pattern but is located in the first clade of the phylogenetic tree. The proximal duplicates in the first clade, Cs_PK_APT_1a and Cs_PK_APT_1b, are both lowly expressed.

CannS/Berberine: The final step of cannabinoid synthesis is catalyzed by a cannabinoid synthase enzyme, which belongs to the berberine bridge enzyme (BBE) gene family. Within Rosales, species often contain many BBE genes, including the species analyzed in this study. Due to the large number of genes and previous, extensive study of cannabinoid synthases, a phylogenetic tree was not constructed (van Velzen and Schranz, 2020; Vergara et al., 2019; Weiblen et al., 2015).

There was a large variance of BBE copy number between species. The *Medicago* genome contains six unique copies of BBE genes, and strawberry contains 18 genes, which is the fewest amount of the Rosales species (Table 5). Apple, fig, and hops each have 20, 28, and 34 copies of BBE genes, respectively. In cannabis, the Jamaican Lion male genome contains 30 unique BBE copies, and the Jamaican Lion female has two additional copies, with 32. Both CBDRx and Finola have 31 unique BBE genes, and Purple Kush has 36 copies. In Purple Kush, many of the genes in the BBE gene family are not expressed, with 14 copies having no expression (Figure 41). Various members of this gene family are expressed in differing tissues, but only one gene is predominantly expressed in flower stages, primarily mid flowers. This specific gene, *Cs_PK_BBE/CannS_13*, is the likely THCA synthase due to the greater than 99% sequence similarity to the functionally characterized gene.

3.6 Tables and Figures

Table 1: Purple Kush RNA, Illumina 1 x 100bp single-end reads (from van Bakel et al. 2011).

	Raw Reads	Raw nt (Gb)	Filtered Reads	Filtered nt (Gb)
PK-Mid-flower	37,835,287	3.8	25,687,331	2.3
PK-Early-flower	37,472,665	3.7	25,434,724	2.3
PK-Pre-flower	54,026,640	5.4	35,522,980	3.2
PK-Shoot	55,653,984	5.6	36,204,828	3.3
PK-Stem	60,353,149	6.0	39,274,463	3.5
PK-Root	37,374,640	3.7	24,904,927	2.2

Table 2: MEP Pathway Gene Copy Numbers

	Purple Kush	Finola	CBDRx	Jamaican Lion (♀)	Jamaican Lion (♂)	Hops	Fig	Strawberry	Apple	Medicago
DXS	4	4	3	3	3	4	4	3	6	3
DXR	4	3	2	2	3	2	1	4	2	2
MCT	2	2	1	1	1	1	1	3	2	1
CMK	2	2	1	1	1	1	1	1	1	1
MDS	1	1	1	1	2	4	1	1	3	1
HDS	2	1	1	1	1	1	1	1	2	1
HDR	1	2	1	1	1	1	1	1	2	1

Table 3: MVA Pathway Gene Copy Numbers

	Purple Kush	Finola	CBDRx	Jamaican Lion (♀)	Jamaican Lion (♂)	Hops	Fig	Strawberry	Apple	Medicago
HMGS	2	1	1	1	1	2	1	1	2	2
HMGR	2	3	2	2	1	8	2	7	9	7
MVA Kinase	1	1	1	1	1	2	1	1	2	1
PMK	1	1	1	1	1	21*	1	1	2	1
MPDC	2	2	1	1	1	2	1	1	2	1

Table 4: Diphosphate Synthesis Copy Numbers

	Purple Kush	Finola	CBDRx	Jamaican Lion (♀)	Jamaican Lion (♂)	Hops	Fig	Strawberry	Apple	Medicago
IDI	1	1	0	1	1	3	1	1	2	1
GPPS.ssu	4	3	2	3	4	4	3	1	2	2
GPPS.lsu	1	1	1	1	1	2	3	3	5	2
FPPS	2	2	2	2	2	3	2	3	4	1

Table 5: Cannabinoid Pathway Copy Numbers

	Purple Kush	Finola	CBDRx	Jamaican Lion (♀)	Jamaican Lion (♂)	Hops	Fig	Strawberry	Apple	Medicago
Desaturase	4	10	5	6	7	4	1	2	3	3
LOX	14	8	8	10	8	16	3	8	7	4
HPL	2	2	2	2	2	2	1	2	2	1
AAE	18	19	17	18	19	21	16	18	16	15
PKS	10	10	7	9	9	16	3	4	12	24
OAC/DABB	9	5	5	7	5	3	2	2	4	3
PT	9	7	11	11	12	7	2	2	4	2
CannS/BBE	36	31	31	32	30	34	28	18	20	6

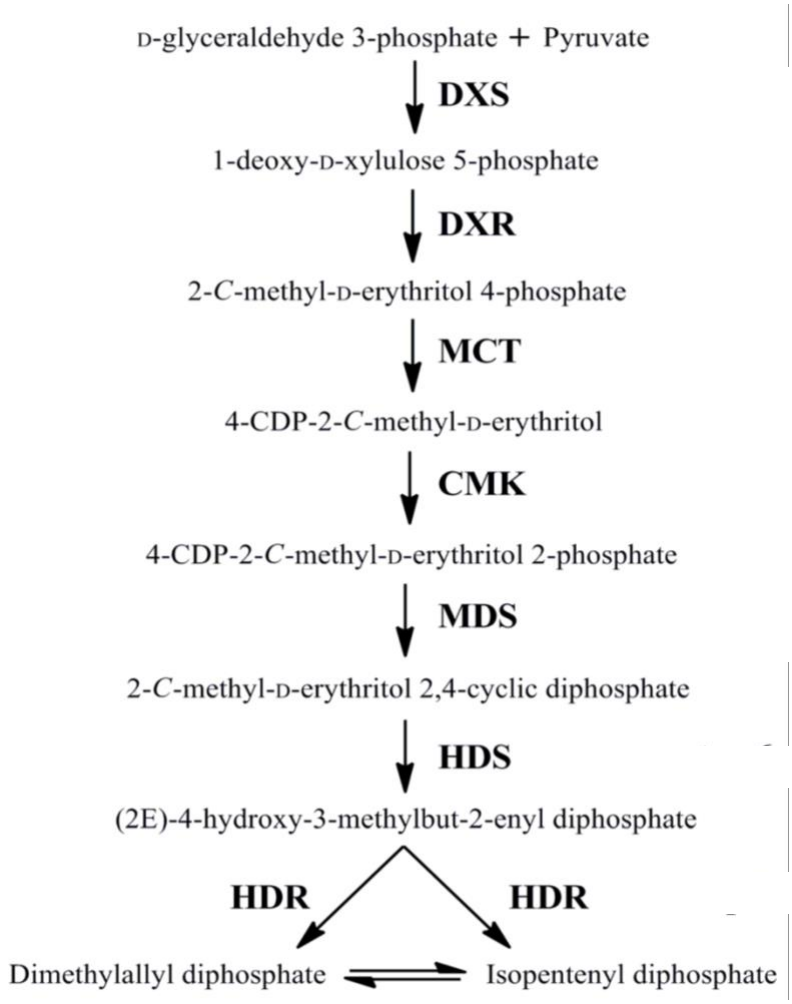

Figure 1: MEP Pathway

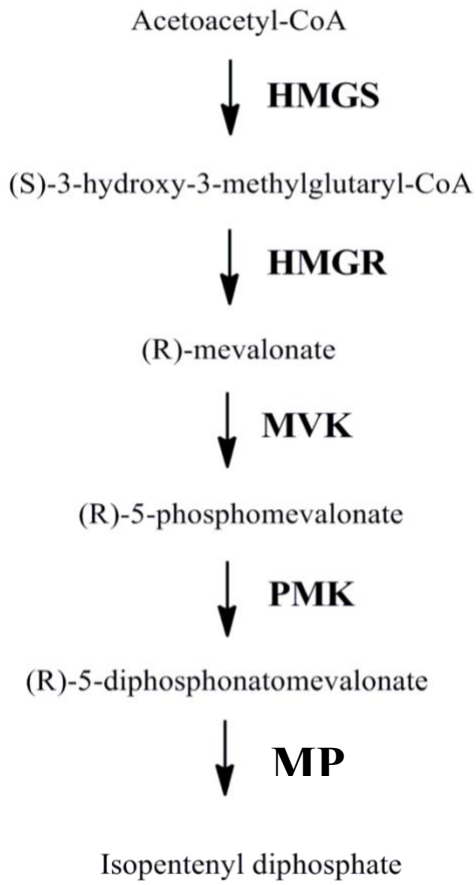

Figure 2: MVA Pathway

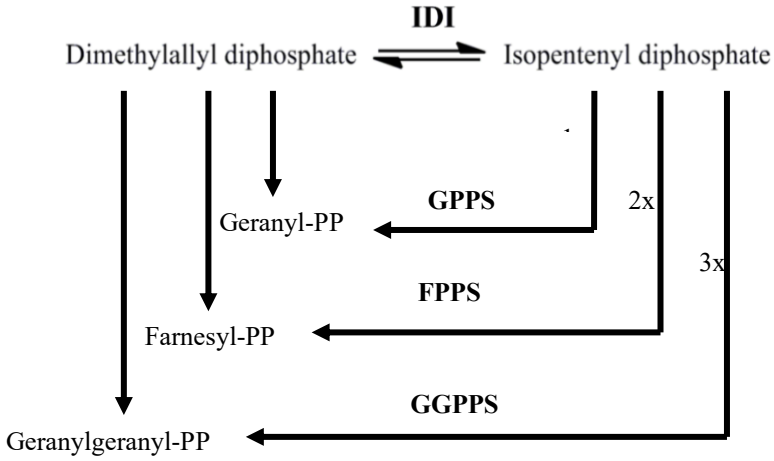

Figure 3: Diphosphate Synthase Reactions

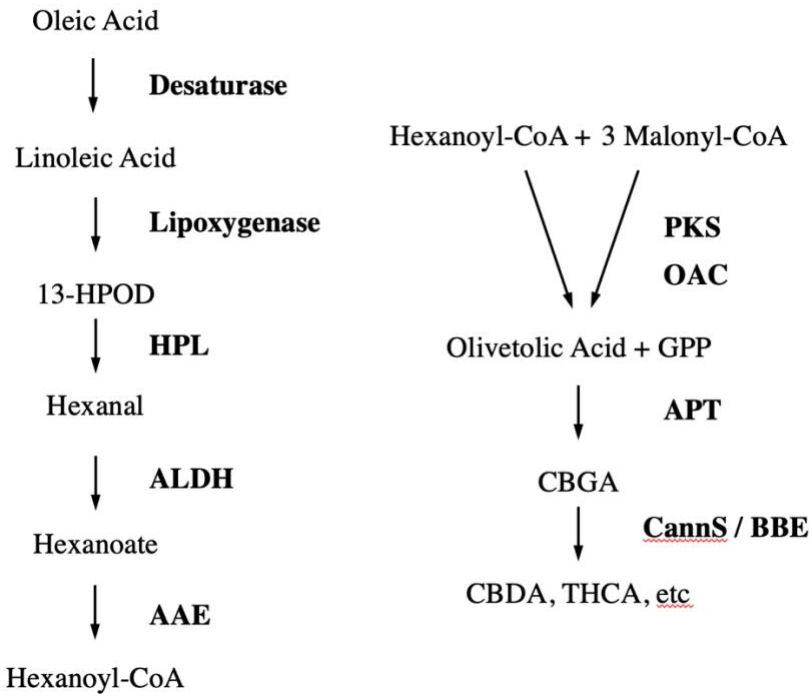

Figure 4: Biosynthesis Pathway of Cannabinoids

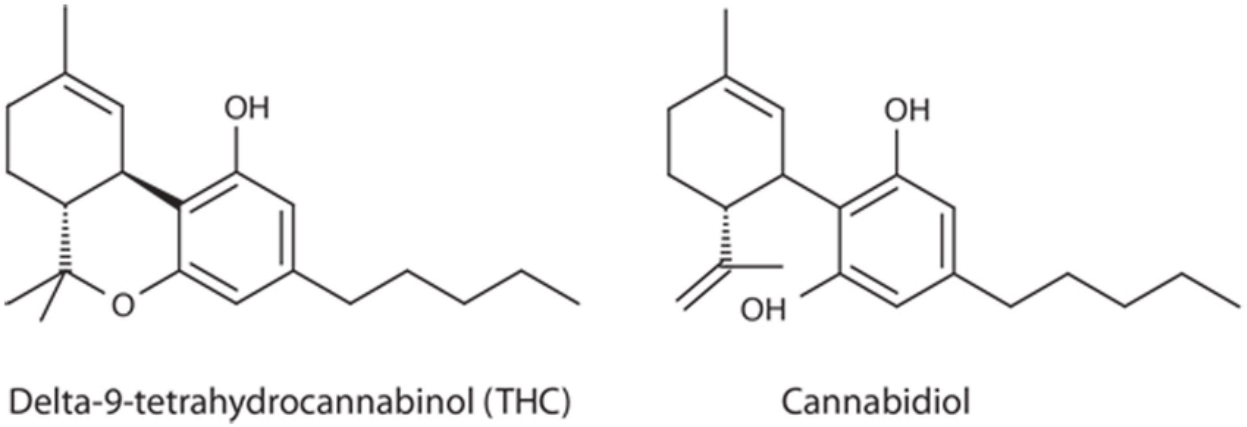

Figure 5: Molecular structures of THC and CBD

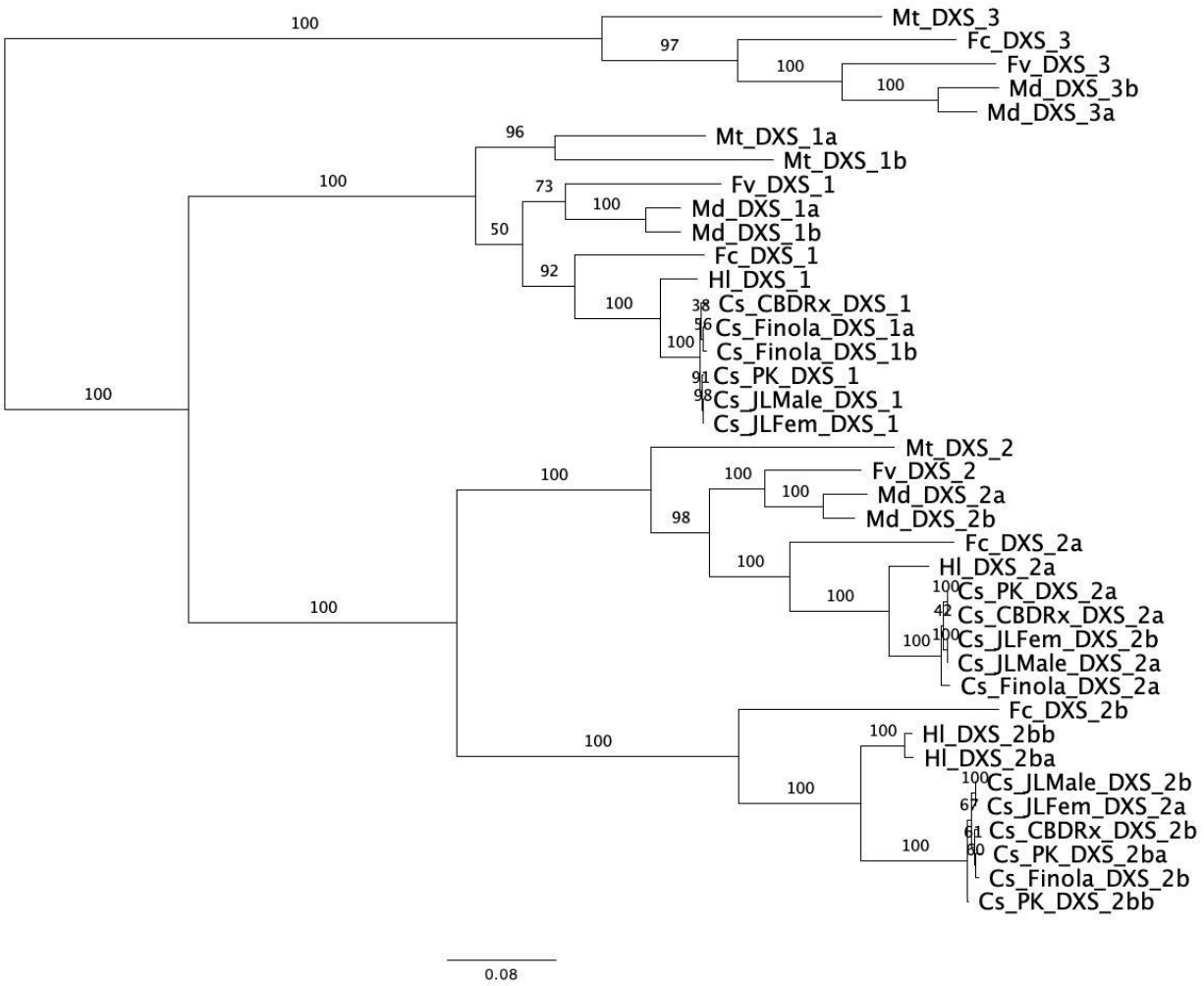

Figure 6:Phylogenetic tree of DXS genes. Coding sequences for DXS genes were aligned, and a tree was constructed using RAxML.

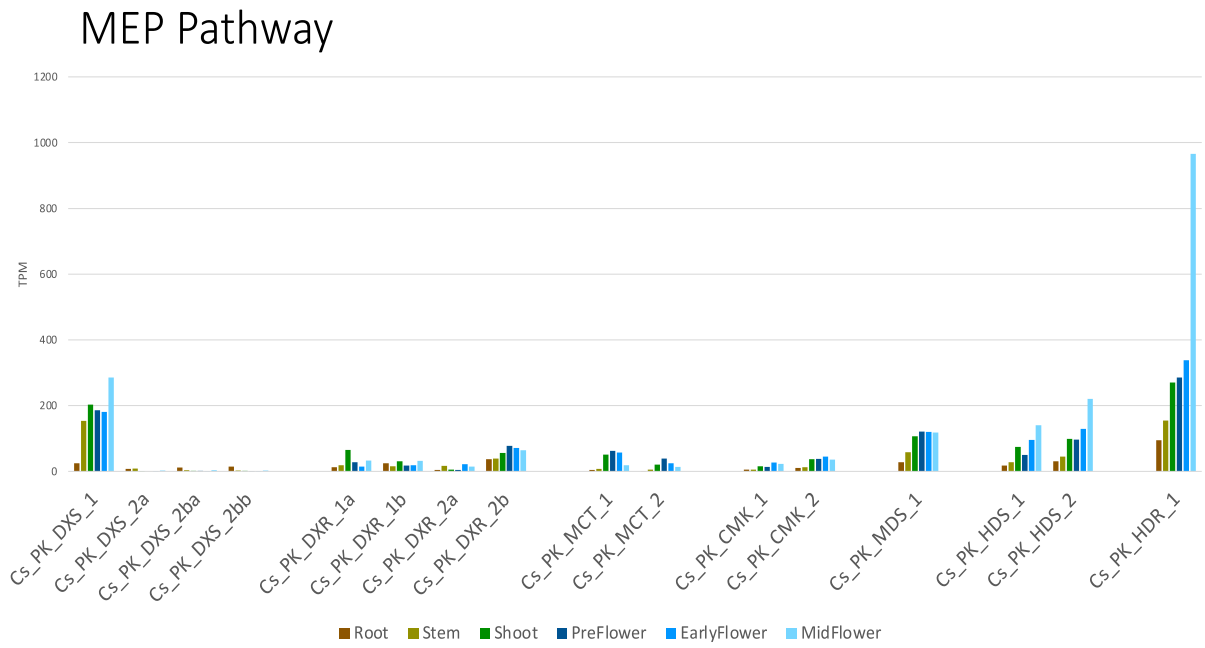

Figure 7: Gene expression of MEP pathway genes. RNA reads were aligned and TPM values were counted for each gene copy of the pathway.

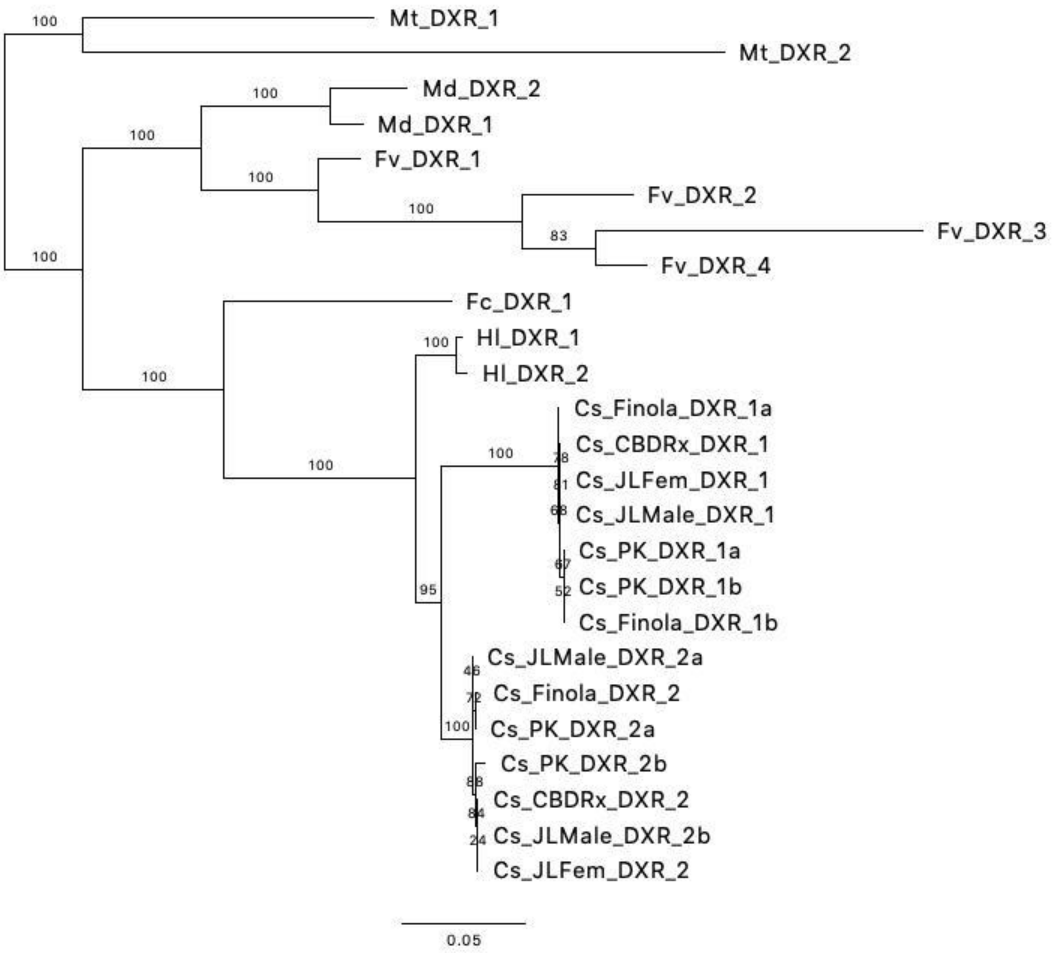

Figure 8: Phylogenetic tree of DXR genes

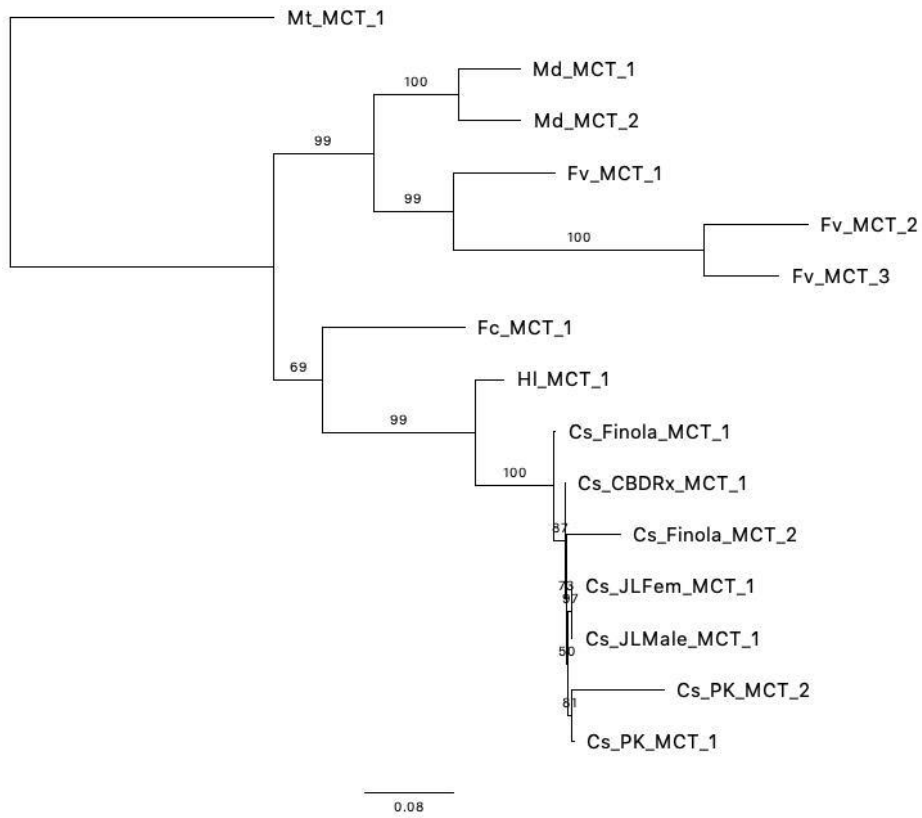

Figure 9: Phylogenetic tree of MCT genes

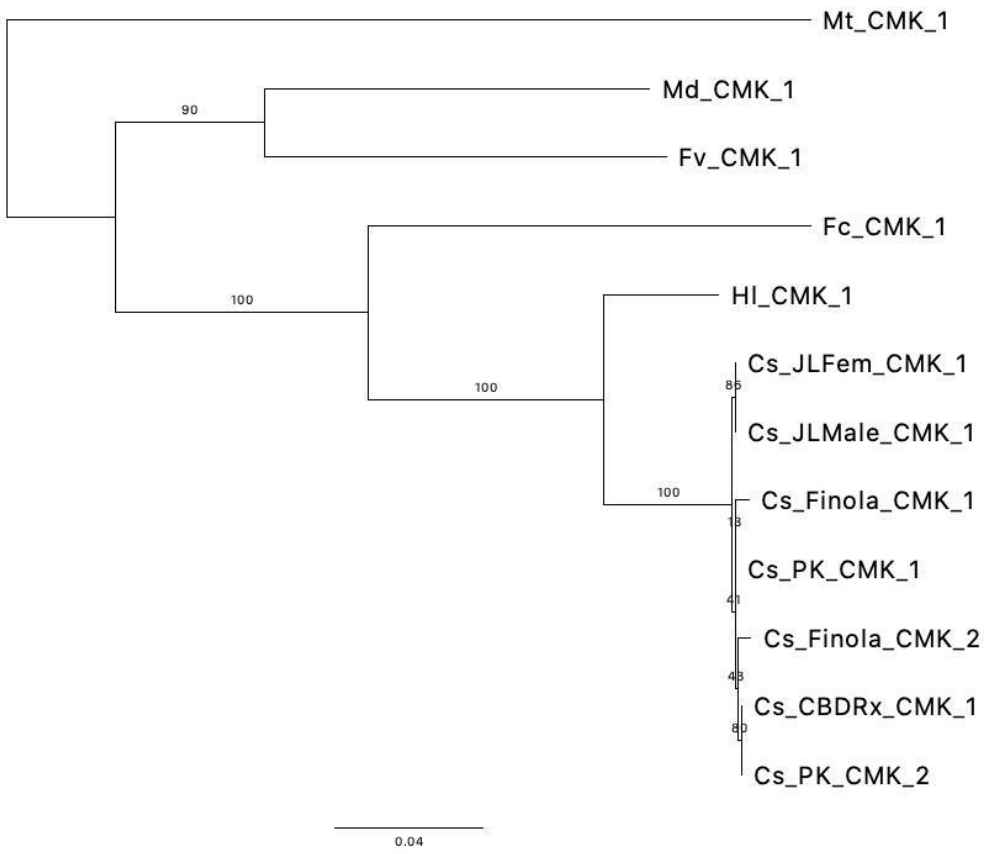

Figure 10: Phylogenetic tree of CMK genes

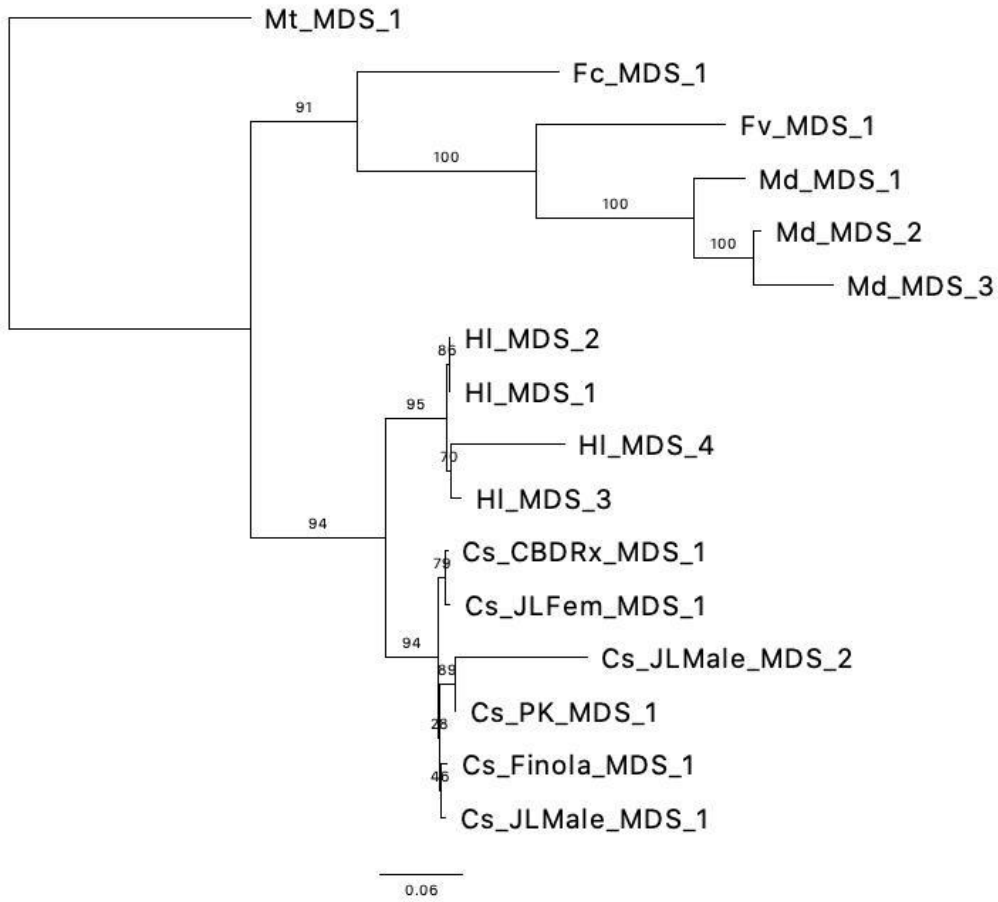

Figure 11: Phylogenetic tree of MDS genes

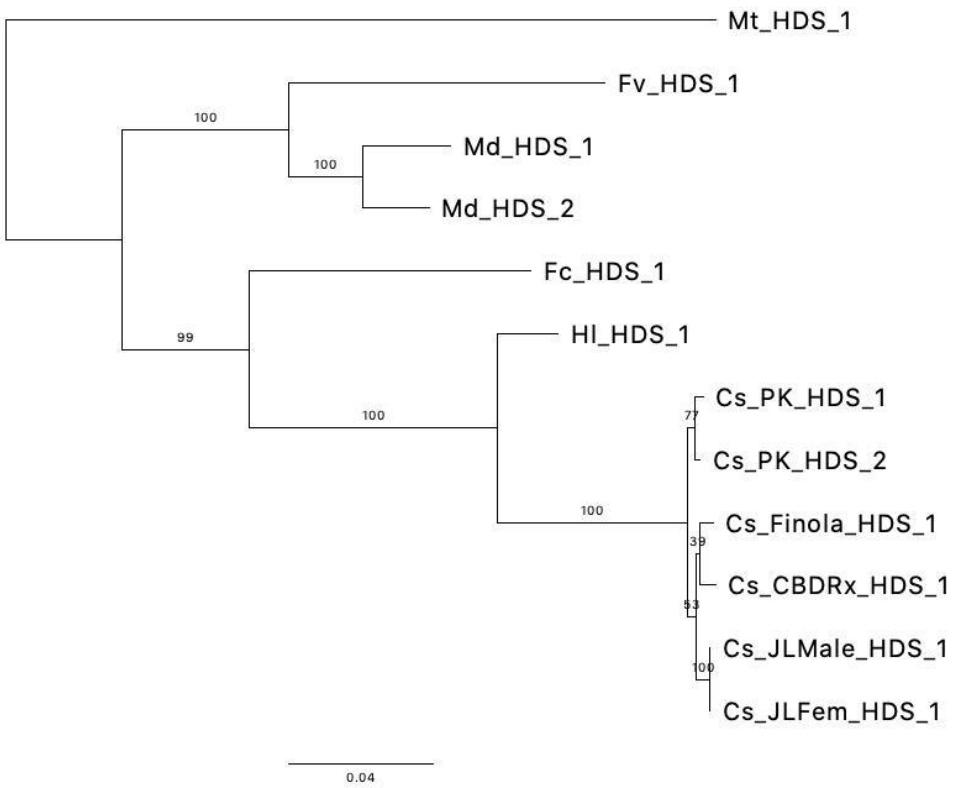

Figure 12: Phylogenetic tree of HDS genes

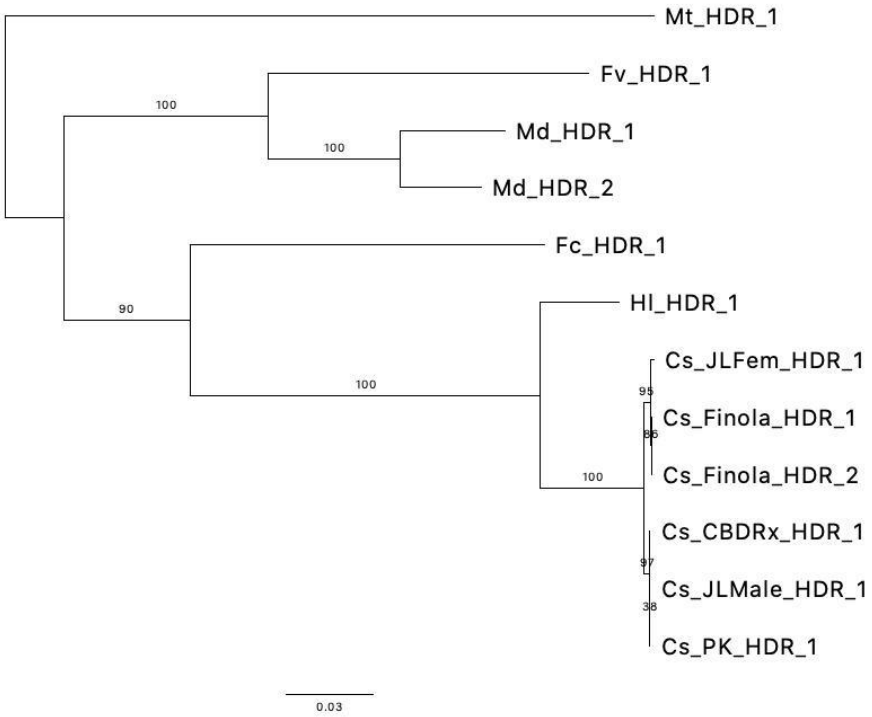

Figure 13: Phylogenetic tree of HDR genes

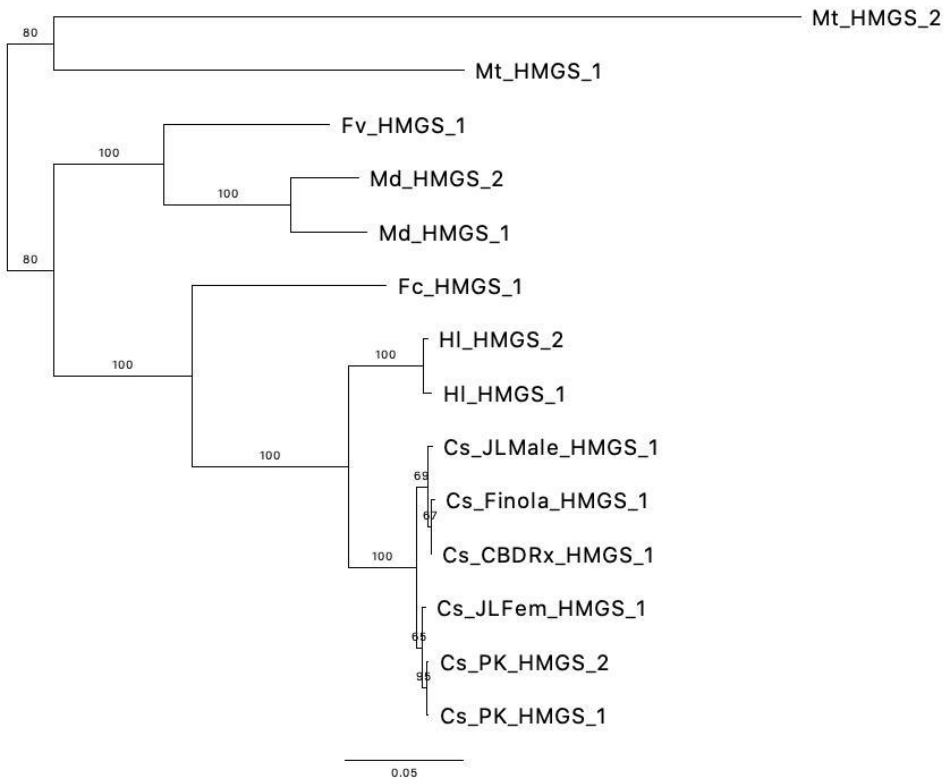

Figure 14: Phylogenetic tree of HMGR genes

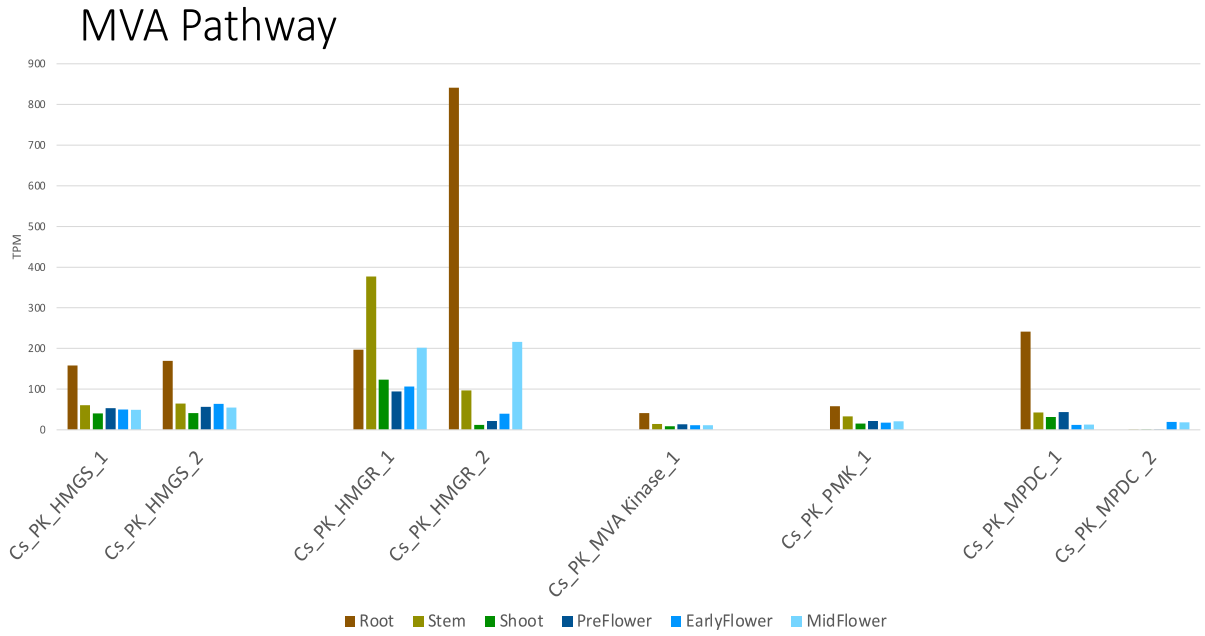

Figure 15: Gene expression of MVA pathway genes

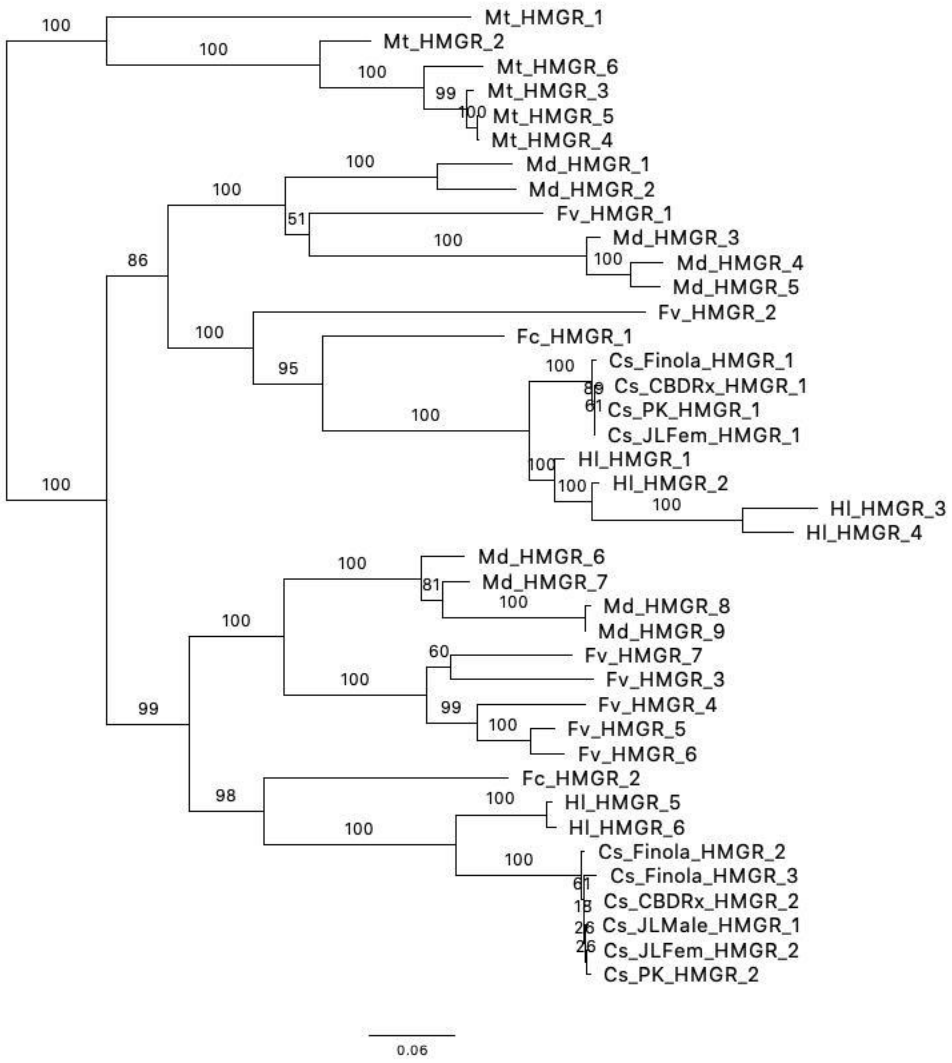

Figure 16: Phylogenetic tree of HMGR genes

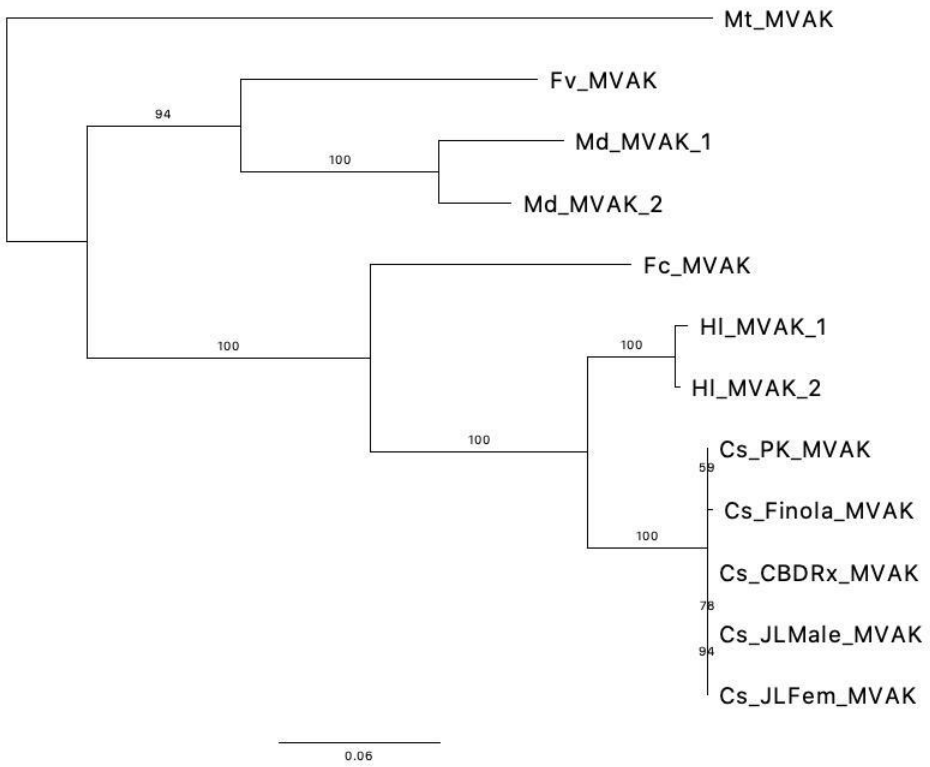

Figure 17: Phylogenetic tree of MVA Kinase genes

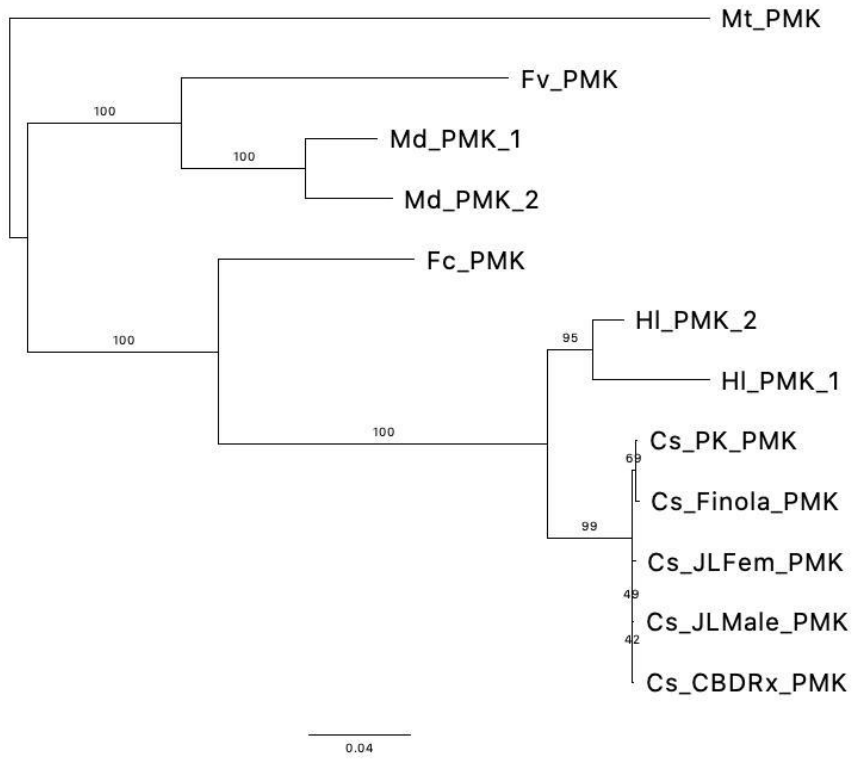

Figure 18: Phylogenetic tree of PMK genes

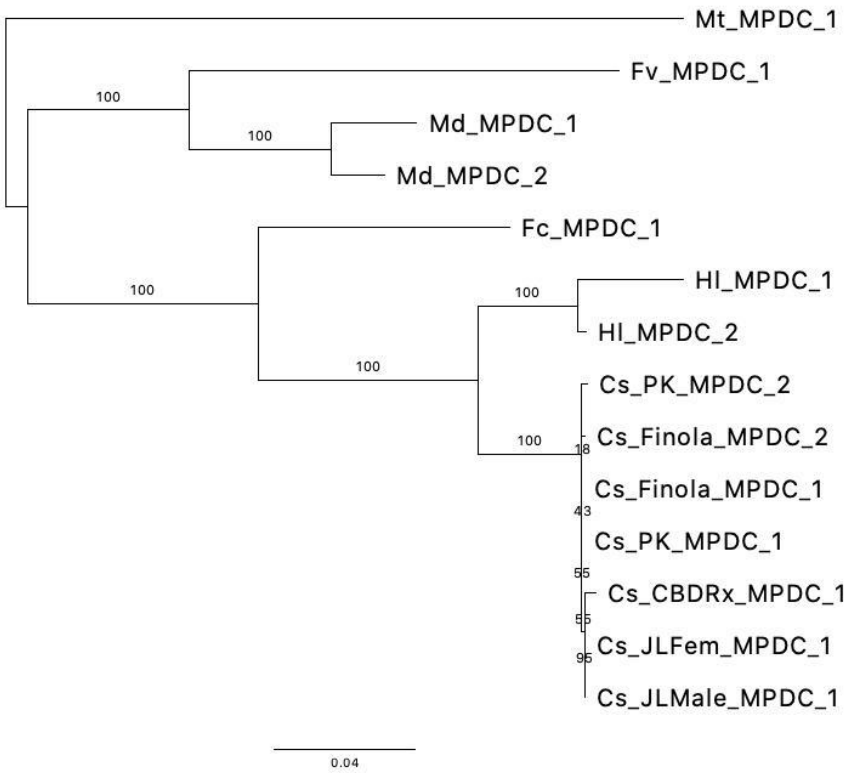

Figure 19: Phylogenetic tree of MPDC genes

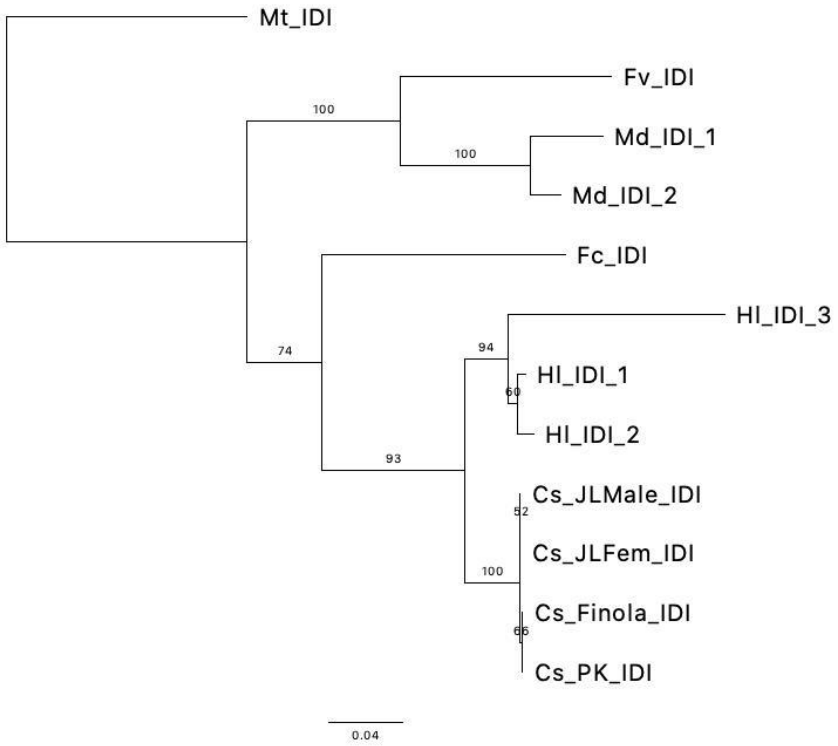

Figure 20: Phylogenetic tree of IDI genes

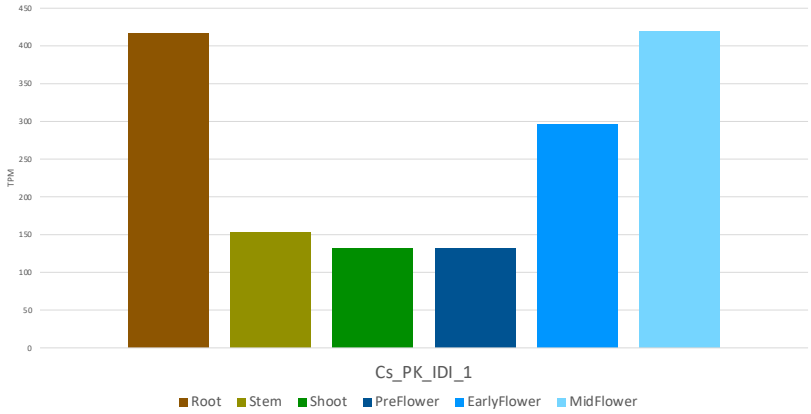

Figure 21: IDI gene expression

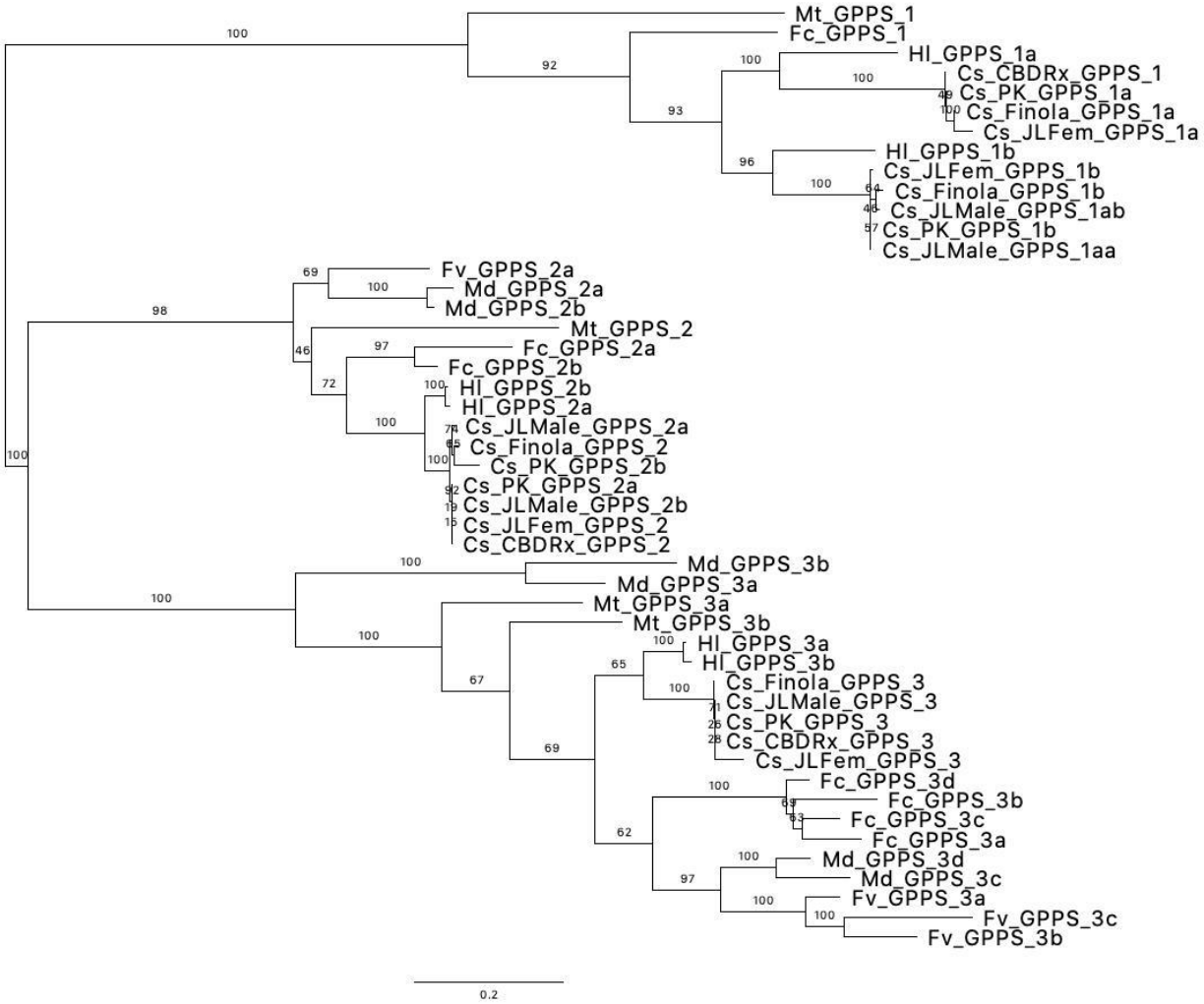

Figure 22: Phylogenetic tree of GPPS genes

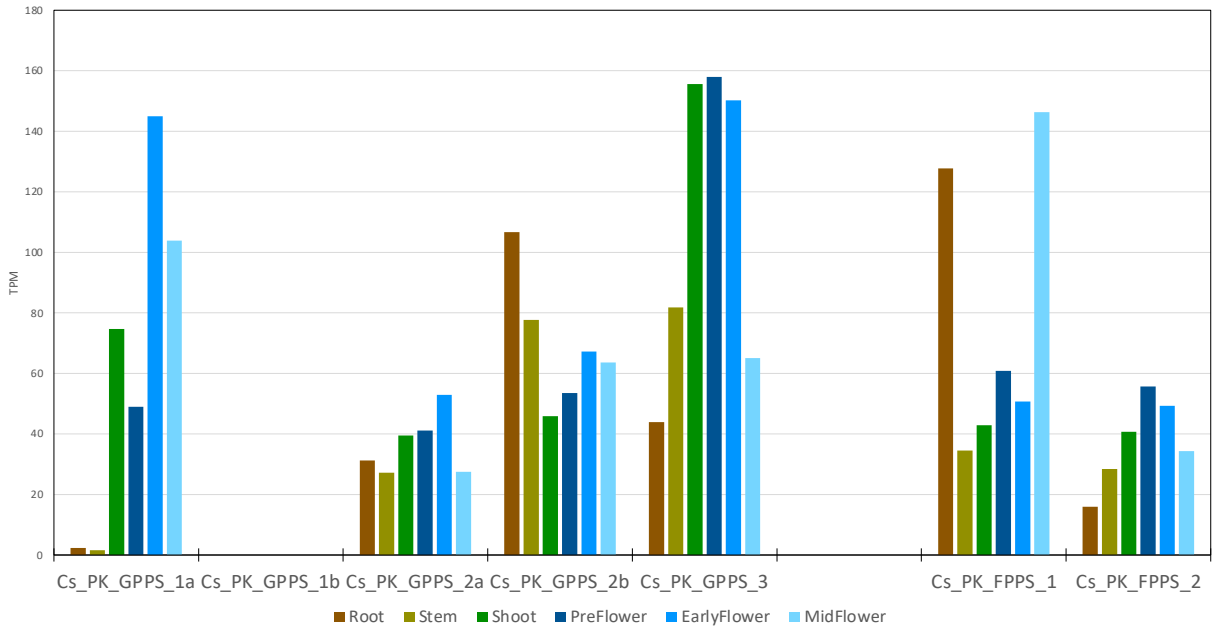

Figure 23: Gene expression of diphosphate synthase genes

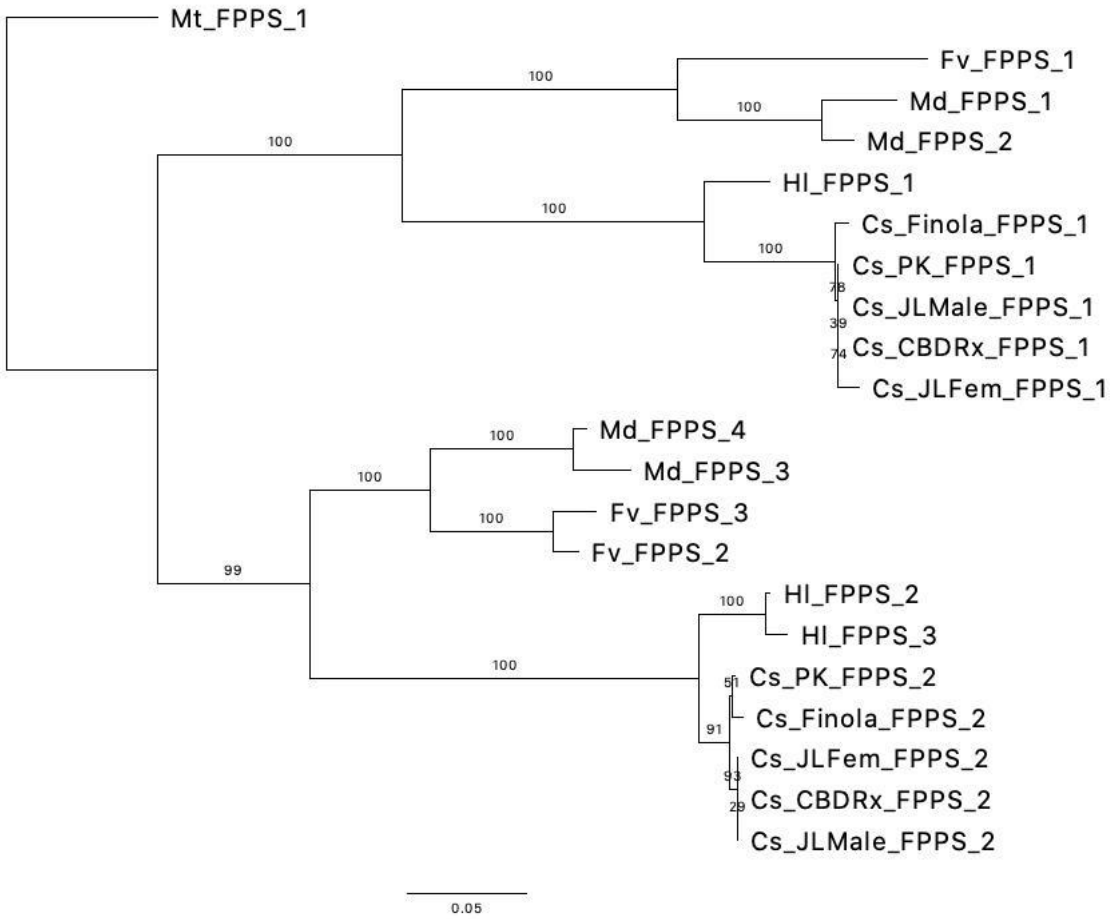

Figure 24: Phylogenetic tree of FPPS genes

Figure 25: Phylogenetic tree of Desaturase (FAD) genes

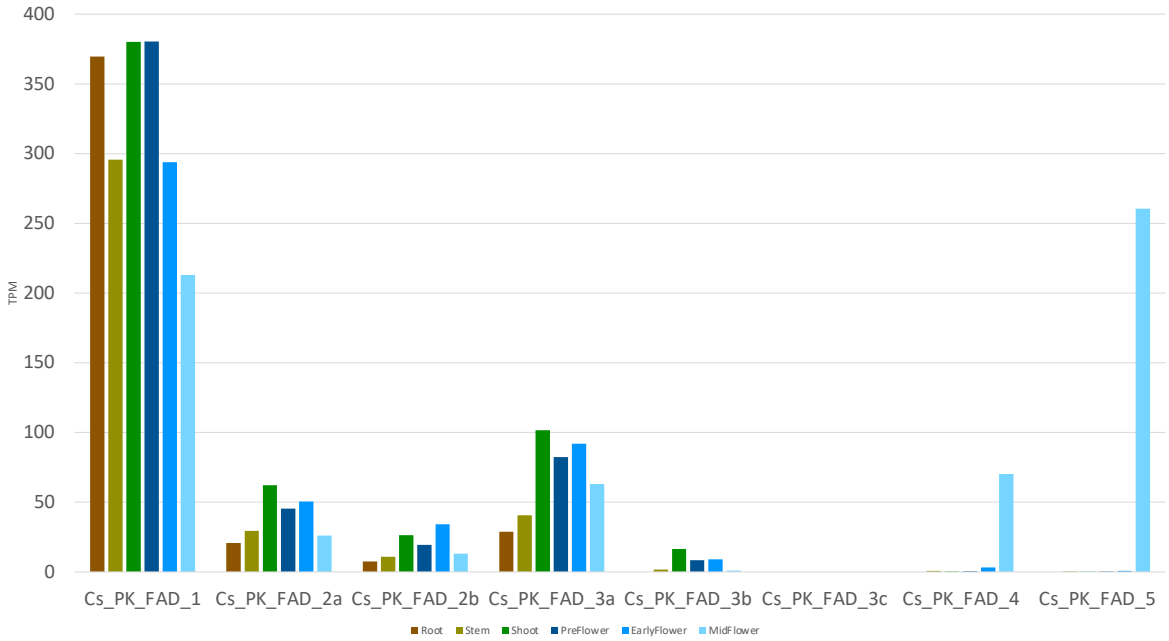

Figure 26: Gene expression of fatty acid desaturase (FAD) genes

Figure 27: Phylogenetic tree of lipoxygenase (LOX) genes

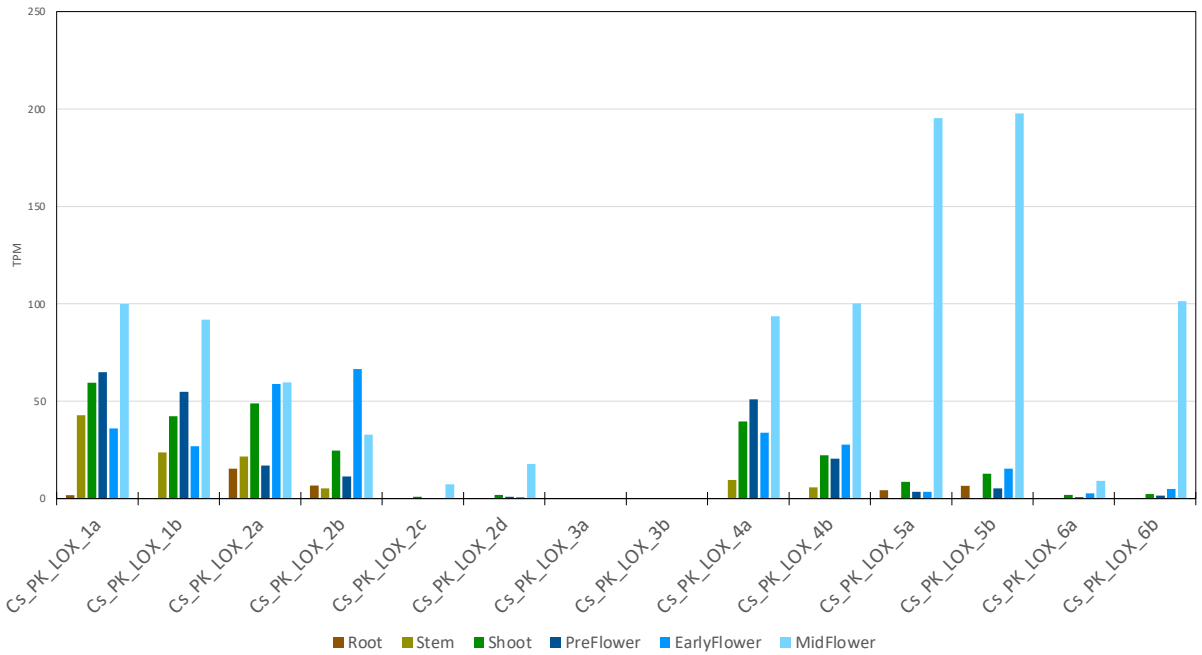

Figure 28: Gene expression of lipoxygenase (LOX) genes

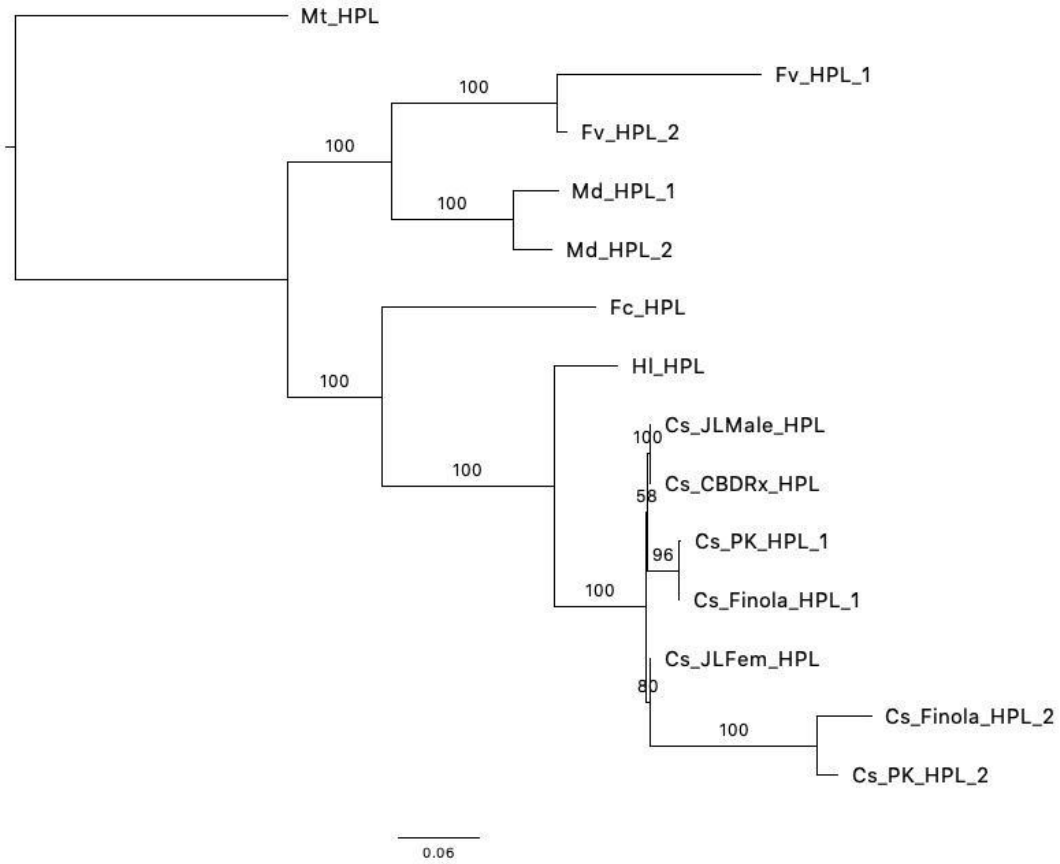

Figure 29: Phylogenetic tree of HPL genes

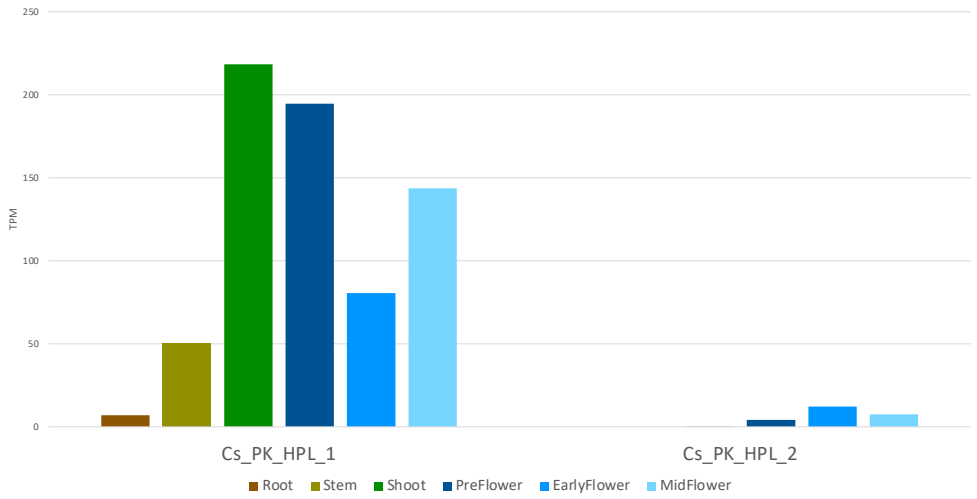

Figure 30: HPL gene expression

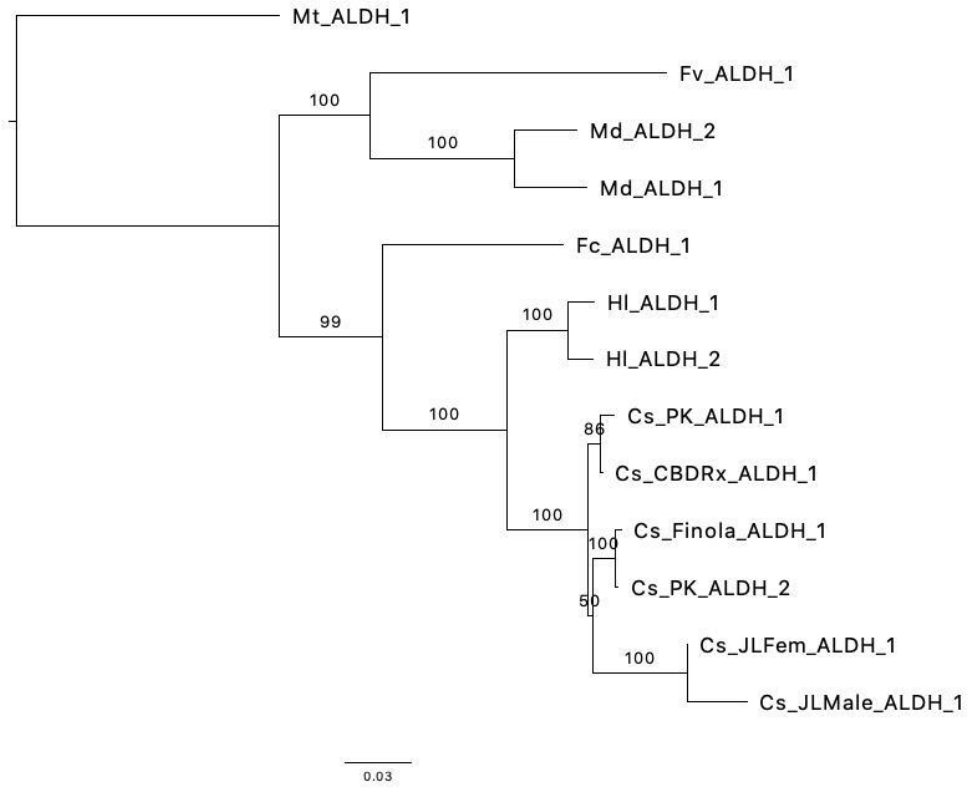

Figure 31: Phylogenetic tree of ALDH genes

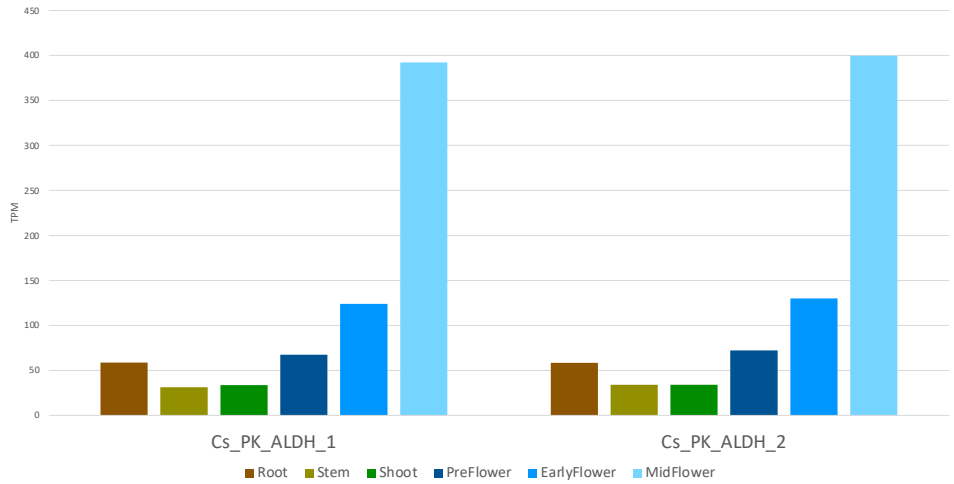

Figure 32: Gene expression of ALDH genes

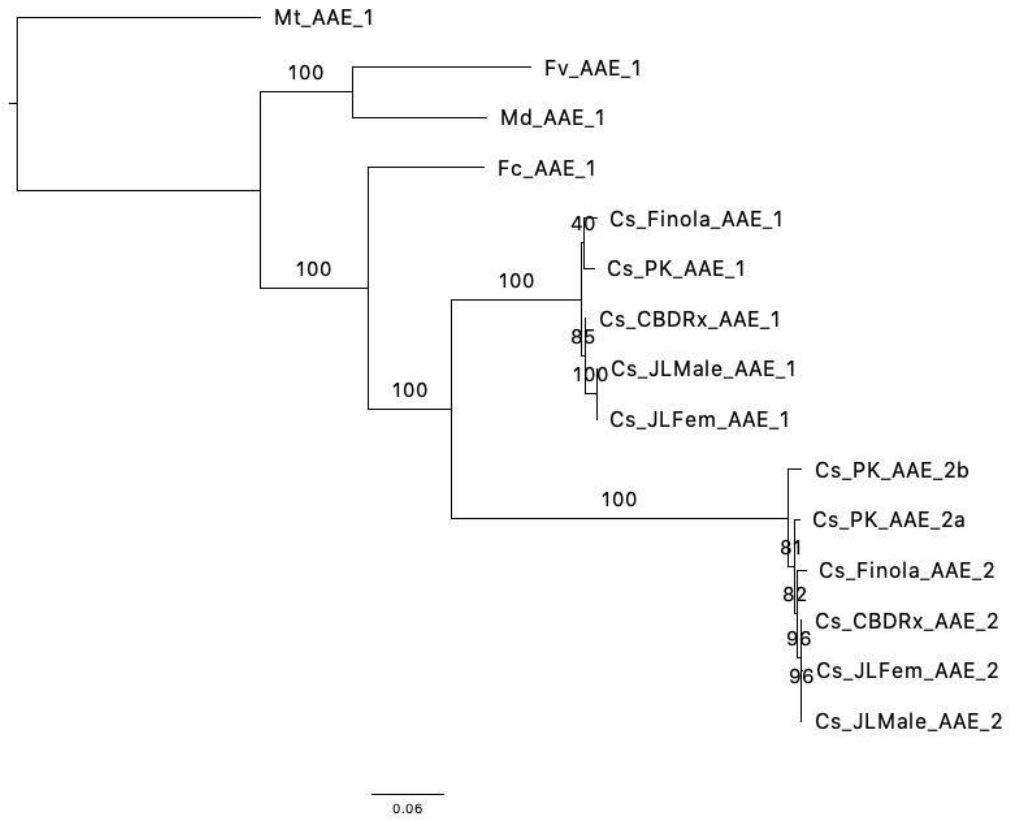

Figure 33: Phylogenetic tree of AAE genes

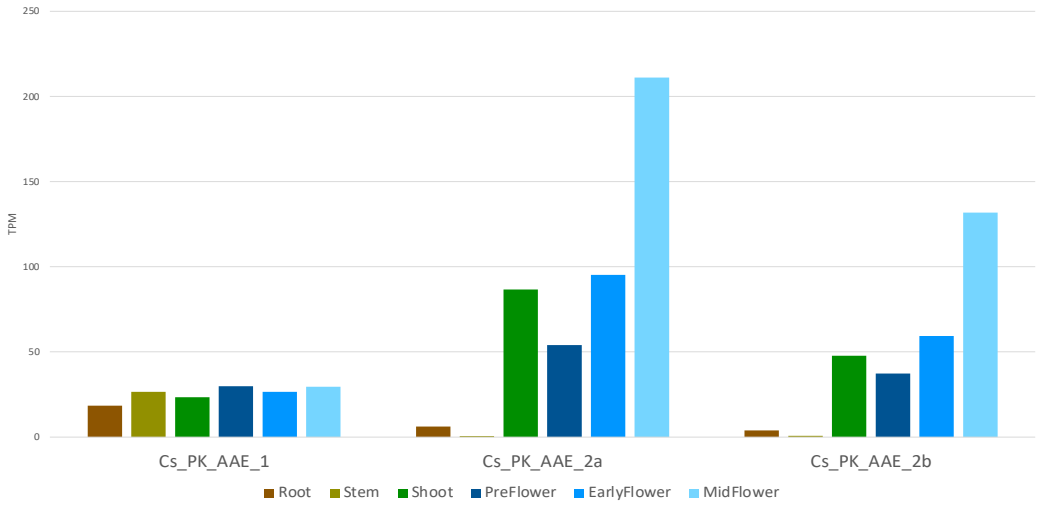

Figure 34: Gene expression of AAE genes

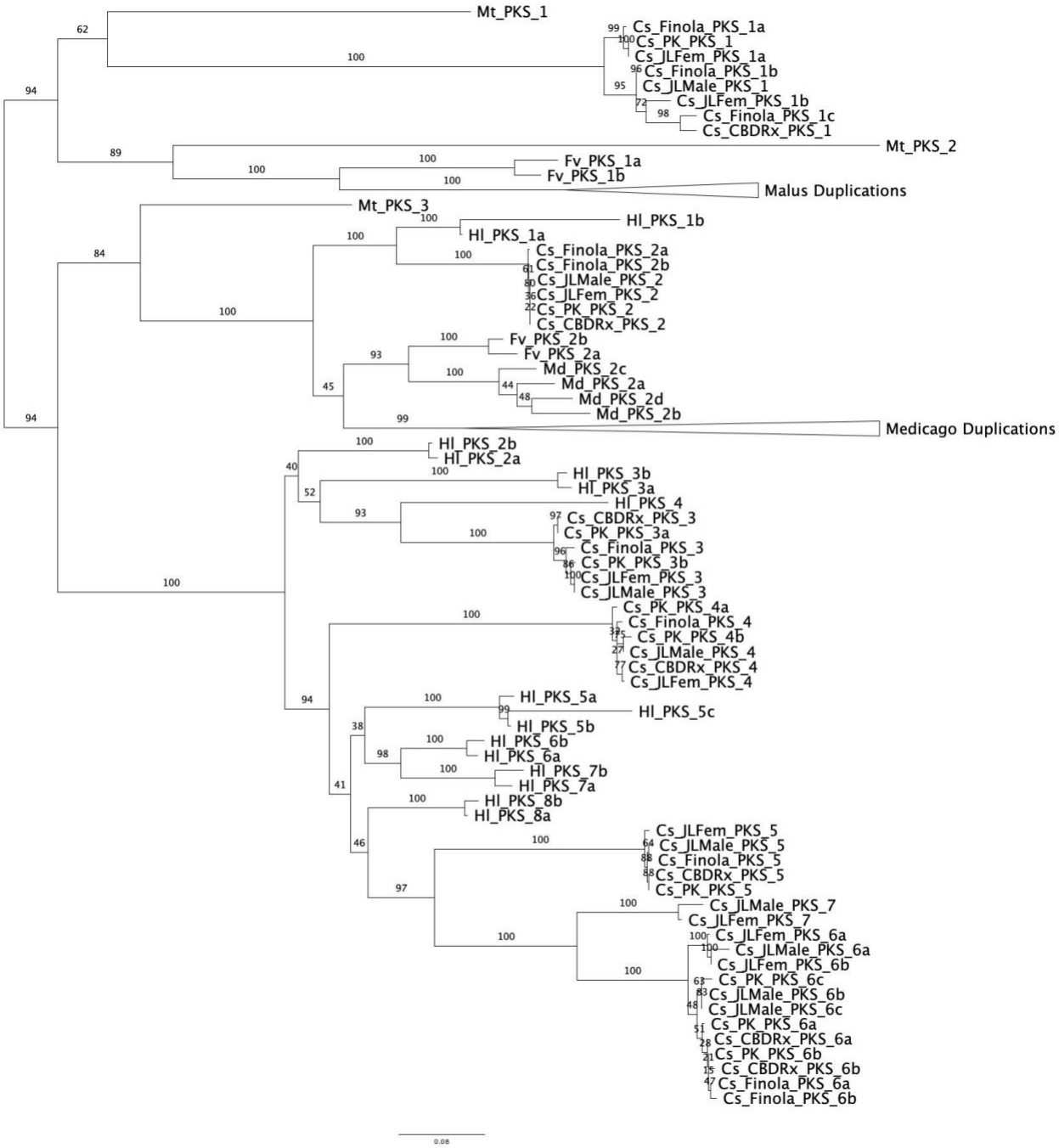

Figure 35: Phylogenetic tree of PKS genes

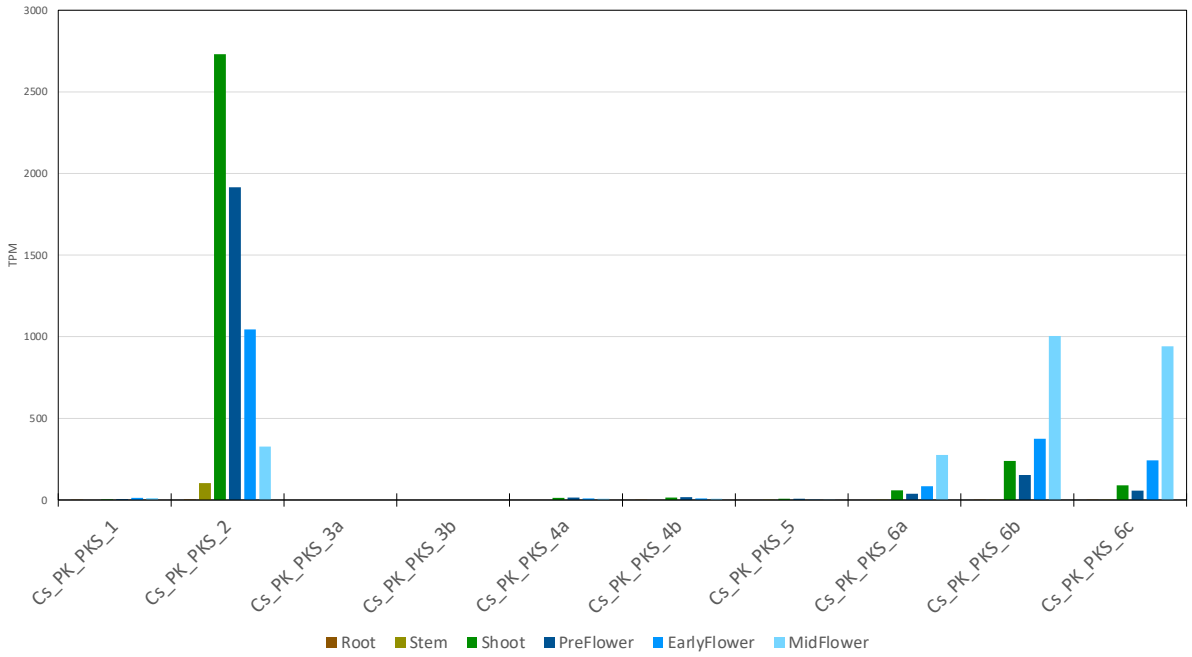

Figure 36: Gene expression of PKS genes

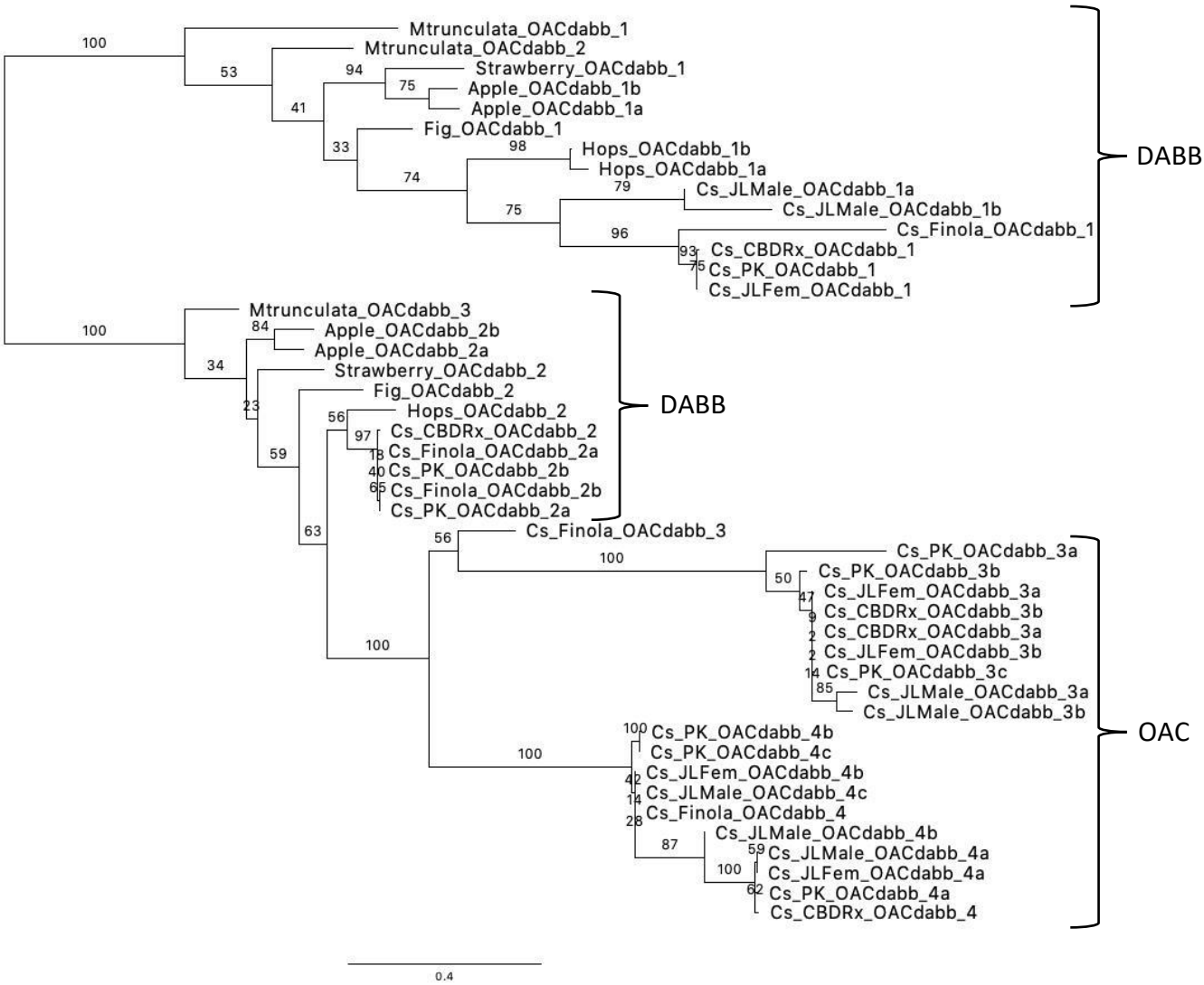

Figure 37: Phylogenetic tree of OAC/DABB genes

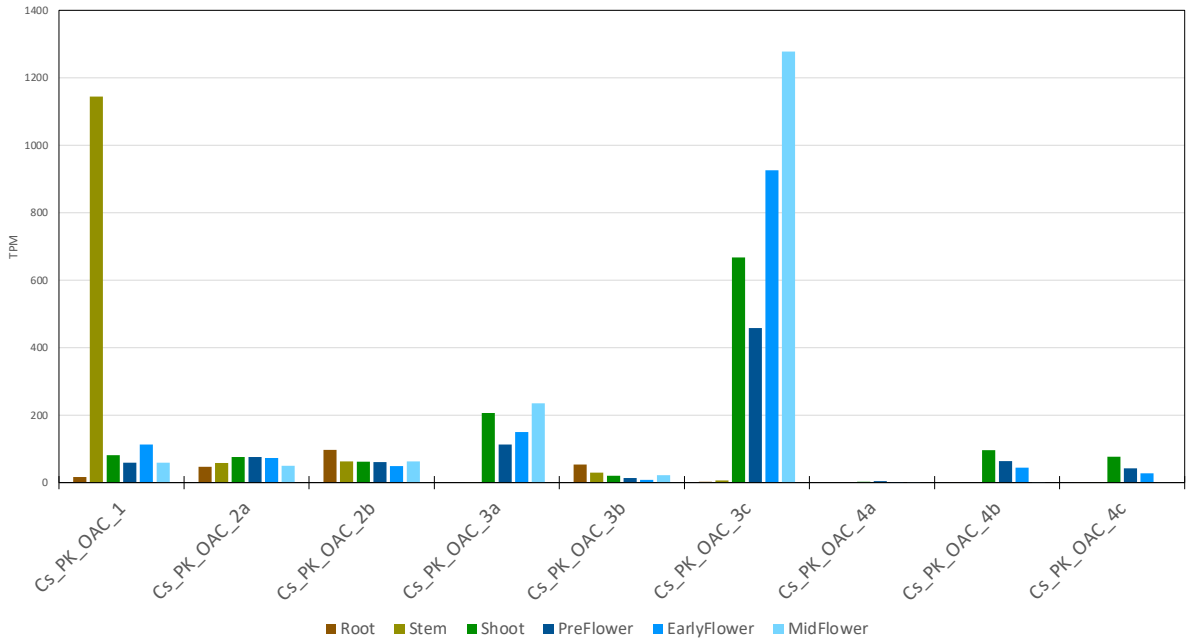

Figure 38: Gene expression Olivetolic Acid Cyclase/DABB genes

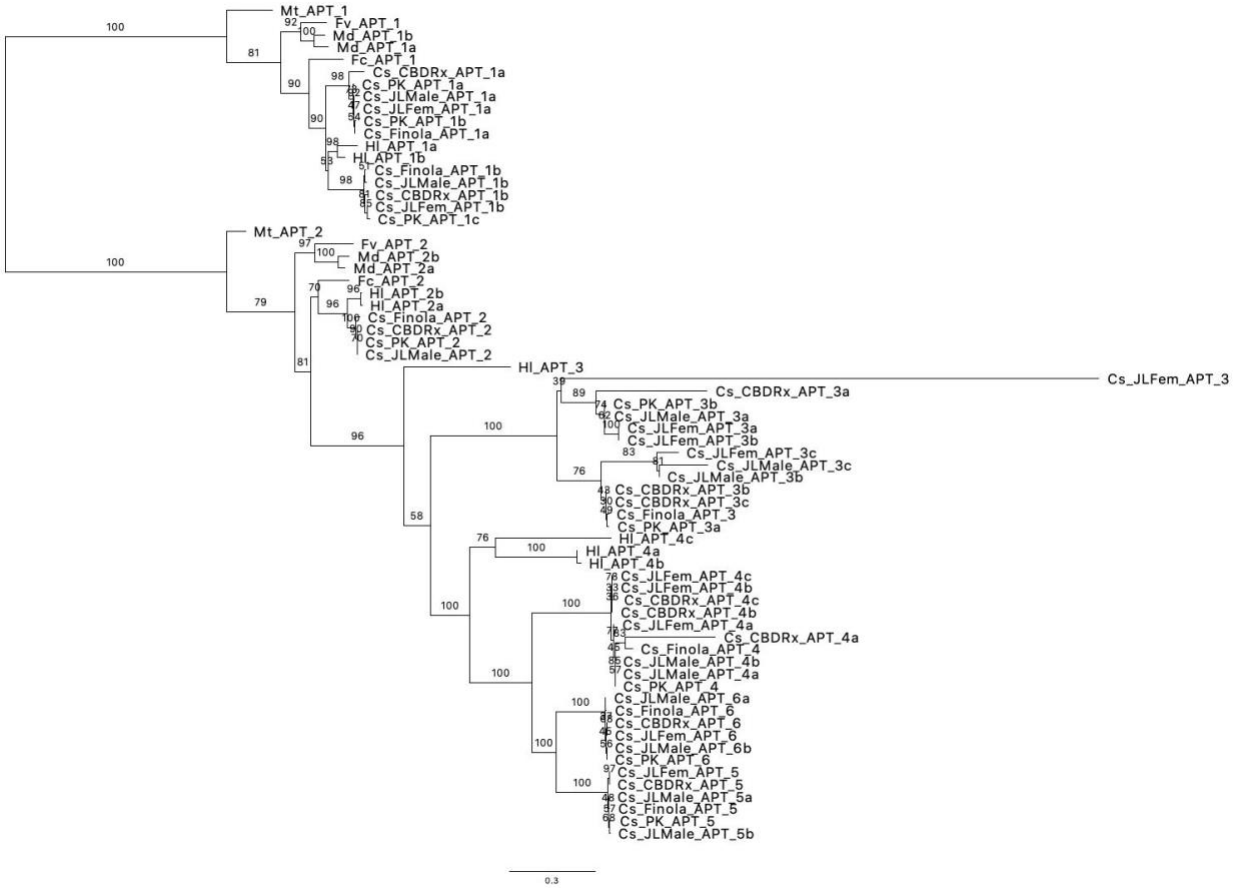

Figure 39: Phylogenetic tree of aromatic prenyltransferase (APT) genes

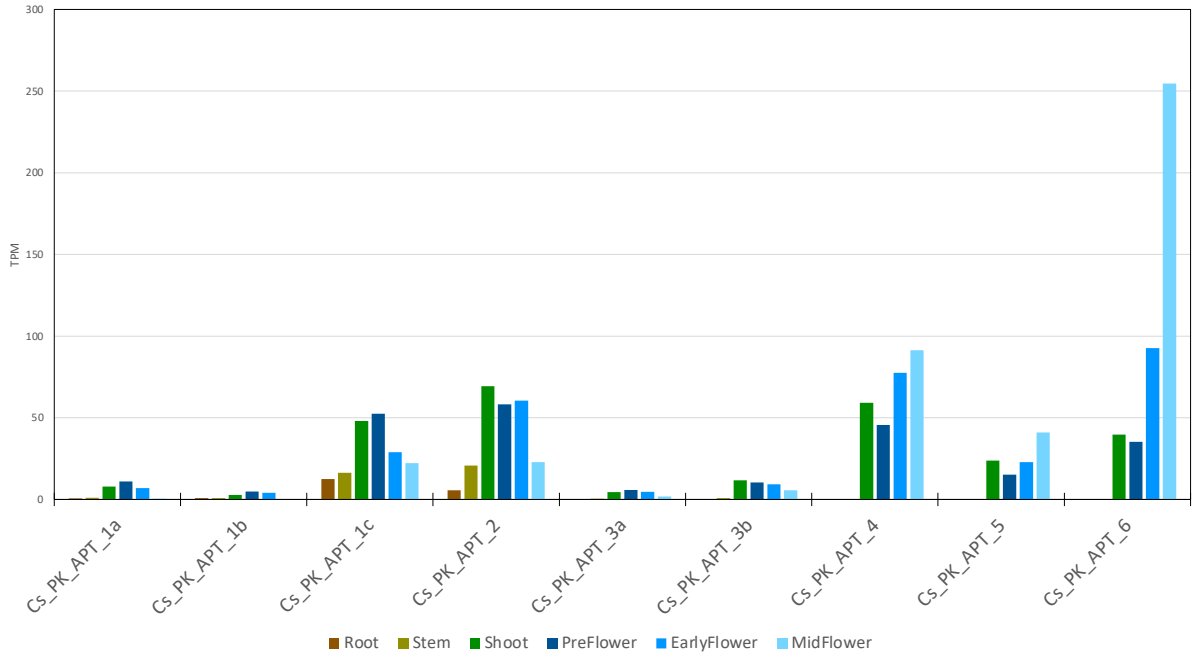

Figure 40: Gene expression of APT genes

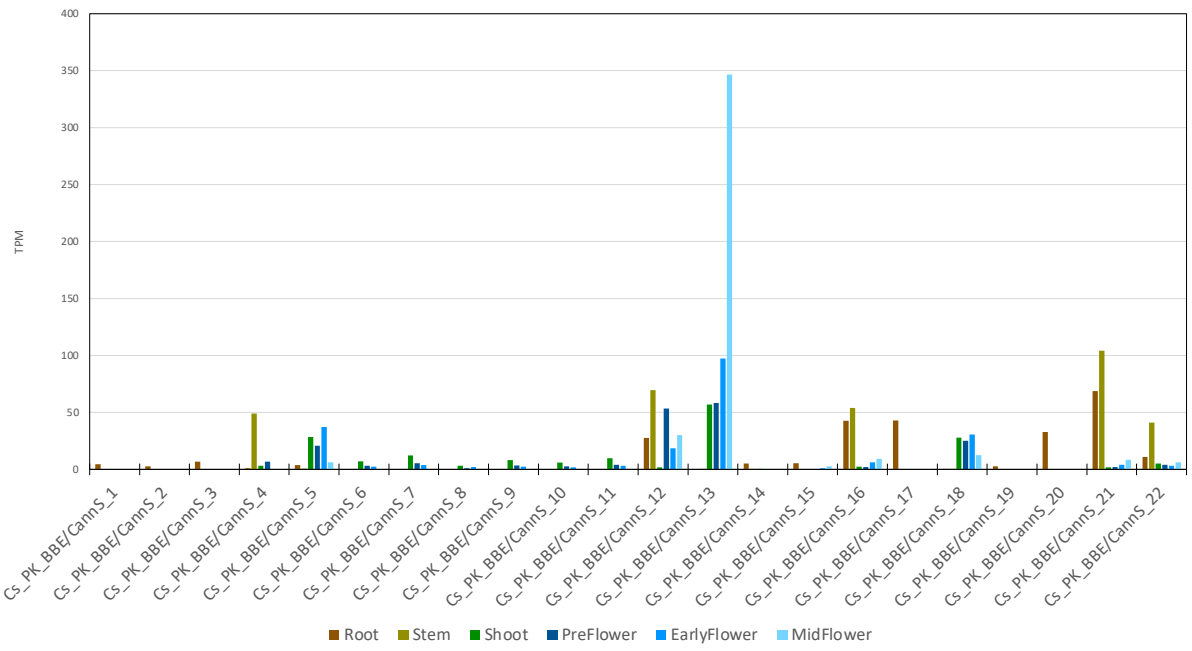

Figure 41: Berberine Bridge Enzyme/Cannabinoid Synthase Gene Expression

Chapter 4: Discussion

4.1 Overview

In *Cannabis sativa*, the highly valued metabolites terpenoids and cannabinoids are products of the MEP, MVA, and cannabinoid metabolic pathways which show many gene duplications and diverse expression patterns. Both the MEP and MVA pathways are highly conserved pathways which are responsible for many terpenoid compounds vital for plant growth and defense. The cannabinoid pathway responsible for the production of CBD and THC has a more recent evolutionary history and is specific to *Cannabis sativa*. This pathway is derived from the more ancient flavonoid pathway through the evolution of polyketide synthases. Additionally, prenyltransferase enzymes have also evolved in Cannabaceae, gaining the function of adding prenyl groups to terpenophenolic compounds. The gene expression profiles across these specialized metabolite pathways highlights the specific copies responsible for biosynthesis and where they are most highly expressed.

4.2 Terpenoid Pathways

The biosynthesis of terpenoid precursors is carried out by the MEP and MVA pathways which are comprised of seven and five genes, respectively. These pathways each vary in copy number across Rosales species. The majority of the variance in copy number is due to recent duplications across all species. Additionally, more genes in the MEP pathway are duplicated than in the MVA pathway. Furthermore, the number of duplicates per gene in the MEP pathway is also higher. The findings within cannabis contrast with previous findings in *A. thaliana* and *O. sativa* which showed that the MEP pathway is more resistant to gene duplication (Zeng and Dehesh, 2021). Furthermore, ancient duplication events before the evolution of the Rosales are only responsible for increased copy number in one gene in each of the MEP and MVA pathways, *DXS* and *HMGR*.

The MEP pathway is the first biosynthetic process for many essential plant molecules, including chlorophylls, quinones, and plant hormones such as gibberellins and abscisic acid. This pathway is also involved in the production of monoterpenoids, which are produced at high levels in flowers of *Cannabis sativa*. The expression of genes within the MEP pathway reveals that the

pathway is active in all tissues, emphasizing its importance for producing essential primary metabolites (Pulido et al., 2012; Zeng and Dehesh, 2021). The low expression in roots in stems indicates these organs require or are producing compounds of the MEP pathway. The increased expression in shoots is likely due to the increased role of the MEP pathway for the production of chlorophylls and not terpenoid production. This is because monoterpenoids in cannabis are found only at minor levels in shoots and leaves (Booth et al., 2020). However, the common trend in the cannabis MEP pathway is that all genes are highly expressed in flowers. This corresponds with the increased monoterpenoid production in those organs.

The expression of genes in the MVA pathway contrasts with the MEP pathway in terms of organ-specific expression patterns. Despite producing high levels of sesquiterpenoids, cannabis flowers did not show the highest level of expression of genes in the MVA pathway. The highest expression levels occurred in roots in all MVA pathway genes. This indicates that triterpenoids and sterol-like compounds may play a more critical role in the roots of cannabis, as they are more highly expressed in those organs (Jin et al., 2020). Further chemical analysis will identify these specific compounds and the functional importance to cannabis. However, there is still considerable expression in flower tissue, which is the likely source of flower sesquiterpenes, such as β-caryophyllene.

4.3 Cannabinoid Pathways

The biosynthesis pathways of cannabinoids in cannabis and bitter acids in hops share many similarities. In both species, a PKS and prenyltransferase are involved in the production of these specific metabolites. Throughout the species analyzed, the PKS gene family showed high copy numbers resulting from many duplications. However, duplications specific to Cannabaceae are responsible for the increased PKS copy number in cannabis and hops, which are found in the lower subclade of *Figure 31.* These Cannabaceae-specific subclades contain the functionally characterized PKS enzymes, which correspond to the Cs_PKS_6 genes in the phylogeny (Raharjo et al., 2004)(Castro et al., 2008). Additionally, Cannabaceae-specific duplications within the APT family have contributed to the rise of the biosynthesis of cannabinoids and bitter acids. Both hops and cannabis have increased copies of APT genes compared to other species, similar to PKS genes. Furthermore, the functionally characterized APT genes (Cs_APT_6) for

cannabinoid biosynthesis and genes for bitter acid biosynthesis (Hl_APT_4) are located in the Cannabaceae specific clade of Figure 23 (Rea et al., 2019) (Li et al., 2015). Due to the presence of critical APT genes in a clade with solely cannabis and hops, the Cannabaceae-specific duplication of APT genes is another contributing factor to a similar rise of cannabinoids and bitter acids.

Another critical event for the biosynthesis of cannabinoids in cannabis was the evolution of the OAC enzyme. Arising from the DABB gene family, the OAC gene (Cs_OAC_3) in cannabis is an example of a neofunctionalized gene duplicate gaining a metabolic role, assisting in the cyclization of olivetolic acid (Gagne et al., 2012). In Finola, this specific OAC gene has a significantly shorter branch, indicating potential rapid sequence evolution in certain cannabis varieties. This is interesting because Finola does not produce cannabinoids at high levels and has evolved separately than the other varieties in this study, being more similar to the ancestral Cannabis variety (Lynch et al., 2016). Cannabis has also retained multiple DABB genes, presumably for a stress response role more similar to the ancestral function as opposed to OAC function (Gagne et al., 2012). The duplication event that led to the rise of the OAC gene and its tandem duplicates is unique to cannabis. This is evident in the second clade of the OAC phylogenetic tree which has numerous, cannabis-specific gene duplicates. This tandem duplication is recent and specific to cannabis, indicating another crucial evolutionary step for the biosynthesis of cannabinoids.

The high level of similarity of many of these copies in cannabis highlights a limitation of the study. The gene expression analysis is somewhat limited because of the high similarity between genes, which in turn makes it difficult to impossible to properly differentiate expression values between these highly similar genes, with greater than 98% sequence identity.

Both cannabis and hops are members of the family Cannabaceae and are phylogenetically very close. In addition to their genetic similarities, hops and cannabis produce many similar specialized metabolites. For example, caryophyllene, humulene, and myrcene are all abundant terpenes produced by both species (Nuutinen, 2018). Also, cannabinoids and bitter acids are both similar terpenophenolic compounds produced by cannabis and hops, respectively (Kovalchuk et al., 2020). The observed gene duplication patterns across these biosynthesis pathways indicate that their produced terpenophenolics can be traced back to duplication events in a common ancestor. Throughout the specialized metabolism pathways of terpenoids, both cannabis and

hops share duplication events, particularly in *DXS* and *GPPS*. In addition, the phylogenetic trees of *desaturase*, *LOX*, *PKS*, and *APT* reveal recent duplication events specific to Cannabaceae. Furthermore, tandem duplicate arrays in *desaturase*, *LOX*, and *PKS* are conserved across both species without considerable gene dispersion and gene loss. These duplication events were likely present in a common ancestor of cannabis and hops that eventually led to the rise of both cannabinoids and bitter acids in each species.

Throughout the specialized metabolic pathways in cannabis, there are a large number of duplicated genes. The cannabinoid pathway has the highest number of duplicates across all pathway genes, compared to the MEP and MVA pathways. Despite the large number of duplicated genes, only a small number of them have been functionally characterized (Gagne et al., 2012; Luo et al., 2019; Raharjo et al., 2004; Rea et al., 2019; Stout et al., 2012; Taura et al., 2007). Additionally, there have been no functionally characterized enzymes in the upstream cannabinoid pathway, including *desaturase*, *LOX*, *HPL*, and *ALDH*, which directly affect the pool of hexanoate acid for THCA or CBDA production. Potentially, duplicates of these genes might be responsible for producing butanoic acid, the precursor for the less common C3-cannabinoids, THCVA and CBDVA (Welling et al., 2019). In addition to the cannabinoid precursor genes, the downstream cannabinoid biosynthesis pathway genes, *PKS*, *OAC*, *APT*, *CannS/BBE*, have many duplicates that are yet to be characterized. Elucidating the enzymatic function of some of these unknown duplicates might contribute to the discovery of potentially novel or rare cannabinoids.

Of the pathways considered in this study, the cannabinoid pathway is clearly the most duplicated specialized metabolism pathway in cannabis. Nearly each gene in the pathway has been duplicated, and the number of duplicates is high (Table 5). This has been caused by gene duplication events at different evolutionary time scales, from ancient duplications before the divergence of Rosales and Fabales from a common ancestor to recent cultivar-specific events in cannabis. These gene duplications are likely a contributing factor to the diversity of cannabinoid compounds produced in cannabis. Cannabis has been shown to produce over 100 different types of cannabinoids, despite being dominated primarily by THCA and CBDA (Berman et al., 2018). Although gene expression analysis indicated that only one primary cannabinoid synthase is active, cannabis varieties can have over 30 cannabinoid synthase-like genes. This high copy number, along with high copy numbers throughout the pre-cannabinoid and cannabinoid

pathways, could provide avenues for a variety of potentially unknown, minor cannabinoid compounds.

Many cannabis varieties, including Purple Kush, have varying organs that change color from green to purple during their life cycles. This phenomenon is commonly attributed to the production of anthocyanins, which belong to the parent class, flavonoids (Khan et al., 2012; Sun et al., 2016). Although the flavonoid/anthocyanin pathway was not investigated, *PKS* genes, *Cs_PK_PKS_1* and *Cs_PK_PKS_2*, are responsible for the first steps of this pathway in cannabis (Rea et al., 2019). The expression of these genes is highest in shoots, indicating that they are not involved in cannabinoid biosynthesis, but rather in flavonoid/anthocyanin biosynthesis.

Drug-type cannabis is universally cultivated for the production of high value cannabinoids, THC and CBD, which are concentrated in glandular trichomes found on flowers (Booth et al., 2017; Livingston et al., 2020). The high levels of gene expression of functionally characterized cannabinoid pathway genes confirmed that the production of these compounds is located in flowers. Additionally, there was hardly any gene expression in roots or stems, which reveals that cannabinoids are unlikely to be produced in those organs, which is supported by a study, albeit limited by a small number of varieties (Jin et al., 2020). Low levels of *PKS*, *OAC*, and *APT* expression in roots make it unlikely that cannabinoids are produced in 'Purple Kush,' but *CannS/BBE* genes are expressed in roots, indicating that there could be additional functions for these genes yet to be characterized.

The high number of duplicates in all cannabinoid synthesis pathway genes is a major contributor to the rise of cannabinoids. This large number of duplicates is not unique to cannabis but also hops because of the large number of shared duplications across several subclades of those genes. Thus, a common ancestor in Cannabaceae is the likely origin of these duplications and therefore the biosynthesis of cannabinoids in cannabis and bitter acids in hops. However, in cannabis there are unique duplication events that are the likely contributors to the specific production of cannabinoids. Specifically, OAC and CannS/BBE both have cannabis specific duplications that have been functionally characterized to participate in cannabinoid biosynthesis.

4.4 Conclusion

This study investigated duplication and expression patterns of genes involved in cannabis specialized metabolism. Through incorporating multiple varieties of cannabis as well as other

Rosales species, the evolutionary histories of 24 gene families in both the terpenoid and cannabinoid metabolic pathways were elucidated. The comparison between the terpenoid (MEP and MVA) and cannabinoid pathways revealed that there are more duplications within the latter. Furthermore, the duplications within the terpenoid pathways are more ancient than those in the cannabinoid pathway. Additionally, the evolution of both cannabinoids and bitter acids found in hops can be traced back to a common ancestor in this family. The overall gene expression data from the variety Purple Kush confirmed that the biosynthesis genes in these two pathways are expressed predominantly in flowers. With respect to genes present in multiple copies, the specific number of copies that is expressed is variable throughout the studied pathways. In some cases, only one copy is actively expressed in a specific organ or set of organs, and in other cases, several different gene duplicates are ubiquitously expressed across organ types.

The identification duplicated genes in specialized metabolic pathways of cannabis will aid with the future breeding efforts to control the metabolic profiles of newly developed varieties. Additionally, many of these duplicated genes have not yet been functionally characterized, providing future research opportunities to investigate the function of these enzymes. Increased genome and transcriptome sequencing will expand upon this thesis by providing further knowledge of copy numbers in other varieties in addition to expression patterns in different cannabis organs and varieties.

References

Allen, K.D., McKernan, K., Pauli, C., Roe, J., Torres, A., Gaudino, R., 2019. Genomic characterization of the complete terpene synthase gene family from Cannabis sativa. PLOS ONE 14, e0222363. https://doi.org/10.1371/journal.pone.0222363

Almagro Armenteros, J.J., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., Nielsen, H., 2019. Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance 2, e201900429. https://doi.org/10.26508/lsa.201900429

Almagro Armenteros, J.J., Sønderby, C.K., Sønderby, S.K., Nielsen, H., Winther, O., 2017. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33, 3387–3395. https://doi.org/10.1093/bioinformatics/btx431

Berman, P., Futoran, K., Lewitus, G.M., Mukha, D., Benami, M., Shlomi, T., Meiri, D., 2018. A new ESI-LC/MS approach for comprehensive metabolic profiling of phytocannabinoids in Cannabis. Sci. Rep. 8, 14280. https://doi.org/10.1038/s41598-018-32651-4

Blum, T., Briesemeister, S., Kohlbacher, O., 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinformatics 10, 274. https://doi.org/10.1186/1471-2105-10-274

Bohlmann, J., Meyer-Gauen, G., Croteau, R., 1998. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. Proc. Natl. Acad. Sci. 95, 4126–4133. https://doi.org/10.1073/pnas.95.8.4126

Booth, J.K., Page, J.E., Bohlmann, J., 2017. Terpene synthases from Cannabis sativa. PLOS ONE 12, e0173911. https://doi.org/10.1371/journal.pone.0173911

Booth, J.K., Yuen, M.M.S., Jancsik, S., Madilao, L.L., Page, J.E., Bohlmann, J., 2020. Terpene Synthases and Terpene Variation in *Cannabis sativa*. Plant Physiol. 184, 130–147. https://doi.org/10.1104/pp.20.00593

Briesemeister, S., Blum, T., Brady, S., Lam, Y., Kohlbacher, O., Shatkay, H., 2009. SherLoc2: A High-Accuracy Hybrid Method for Predicting Subcellular Localization of Proteins. J. Proteome Res. 8, 5363–5366. https://doi.org/10.1021/pr900665y

Briesemeister, S., Rahnenfï¿½hrer, J., Kohlbacher, O., 2010. YLoc—an interpretable web server for predicting subcellular localization. Nucleic Acids Res. 38, W497–W502. https://doi.org/10.1093/nar/gkq477

Castro, C.B., Whittock, L.D., Whittock, S.P., Leggett, G., Koutoulis, A., 2008. DNA Sequence and Expression Variation of Hop (Humulus lupulus) Valerophenone Synthase (VPS), a Key Gene in Bitter Acid Biosynthesis. Ann. Bot. 102, 265–273. https://doi.org/10.1093/aob/mcn089

Chae, L., Kim, T., Nilo-Poyanco, R., Rhee, S.Y., 2014. Genomic Signatures of Specialized Metabolism in Plants. Science 344, 510–513. https://doi.org/10.1126/science.1252076

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Fischedick J, E.S., 2015. Cannabinoids and Terpenes as Chemotaxonomic Markers in Cannabis. Nat. Prod. Chem. Res. 03. https://doi.org/10.4172/2329-6836.1000181

Fischedick, J.T., 2017. Identification of Terpenoid Chemotypes Among High (−)- *trans* -Δ^9 -Tetrahydrocannabinol-Producing *Cannabis sativa* L. Cultivars. Cannabis Cannabinoid Res. 2, 34–47. https://doi.org/10.1089/can.2016.0040

Gagne, S.J., Stout, J.M., Liu, E., Boubakir, Z., Clark, S.M., Page, J.E., 2012. Identification of olivetolic acid cyclase from Cannabis sativa reveals a unique catalytic route to plant

polyketides. Proc. Natl. Acad. Sci. 109, 12811–12816. https://doi.org/10.1073/pnas.1200330109
Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Altermann, U., Angerer, P., Ansorge, S., Balasz, K., Bernhofer, M., Betz, A., Cizmadija, L., Do, K.T., Gerke, J., Greil, R., Joerdens, V., Hastreiter, M., Hembach, K., Herzog, M., Kalemanov, M., Kluge, M., Meier, A., Nasir, H., Neumaier, U., Prade, V., Reeb, J., Sorokoumov, A., Troshani, I., Vorberg, S., Waldraff, S., Zierer, J., Nielsen, H., Rost, B., 2014. LocTree3 prediction of localization. Nucleic Acids Res. 42, W350–W355. https://doi.org/10.1093/nar/gku396
Grassa, C.J., Wenger, J.P., Dabney, C., Poplawski, S.G., Motley, S.T., Michael, T.P., Schwartz, C.J., Weiblen, G.D., 2018. A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content (preprint). Genomics. https://doi.org/10.1101/458083
Jin, D., Dai, K., Xie, Z., Chen, J., 2020. Secondary Metabolites Profiled in Cannabis Inflorescences, Leaves, Stem Barks, and Roots for Medicinal Purposes. Sci. Rep. 10, 3309. https://doi.org/10.1038/s41598-020-60172-6
Keeling, C.I., Weisshaar, S., Lin, R.P.C., Bohlmann, J., 2008. Functional plasticity of paralogous diterpene synthases involved in conifer defense. Proc. Natl. Acad. Sci. 105, 1085–1090. https://doi.org/10.1073/pnas.0709466105
Khan, J.I., Kennedy, T.J., Christian, D.R., 2012. Cannabis, in: Basic Principles of Forensic Chemistry. Humana Press, Totowa, NJ, pp. 145–156. https://doi.org/10.1007/978-1-59745-437-7_12
Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360. https://doi.org/10.1038/nmeth.3317
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, 132–140. https://doi.org/10.1016/j.forsciint.2005.07.005
Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., Pertea, M., 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20, 278. https://doi.org/10.1186/s13059-019-1910-1
Kovalchuk, I., Pellino, M., Rigault, P., van Velzen, R., Ebersbach, J., Ashnest, J.R., Mau, M., Schranz, M.E., Alcorn, J., Laprairie, R.B., McKay, J.K., Burbridge, C., Schneider, D., Vergara, D., Kane, N.C., Sharbel, T.F., 2020. The Genomics of *Cannabis* and Its Close Relatives. Annu. Rev. Plant Biol. 71, 713–739. https://doi.org/10.1146/annurev-arplant-081519-040203
Laverty, K.U., Stout, J.M., Sullivan, M.J., Shah, H., Gill, N., Holbrook, L., Deikus, G., Sebra, R., Hughes, T.R., Page, J.E., van Bakel, H., 2019. A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the *THC/CBD acid synthase* loci. Genome Res. 29, 146–156. https://doi.org/10.1101/gr.242594.118
Li, H., Ban, Z., Qin, H., Ma, L., King, A.J., Wang, G., 2015. A Heteromeric Membrane-Bound Prenyltransferase Complex from Hop Catalyzes Three Sequential Aromatic Prenylations in the Bitter Acid Pathway. Plant Physiol. 167, 650–659. https://doi.org/10.1104/pp.114.253682
Li, H.-L., 1973. An archaeological and historical account of cannabis in China. Econ. Bot. 28, 437–448. https://doi.org/10.1007/BF02862859

Livingston, S.J., Quilichini, T.D., Booth, J.K., Wong, D.C.J., Rensing, K.H., Laflamme-Yonkman, J., Castellarin, S.D., Bohlmann, J., Page, J.E., Samuels, A.L., 2020. Cannabis glandular trichomes alter morphology and metabolite content during flower maturation. Plant J. 101, 37–56. https://doi.org/10.1111/tpj.14516

Luo, X., Reiter, M.A., d'Espaux, L., Wong, J., Denby, C.M., Lechner, A., Zhang, Y., Grzybowski, A.T., Harth, S., Lin, W., Lee, H., Yu, C., Shin, J., Deng, K., Benites, V.T., Wang, G., Baidoo, E.E.K., Chen, Y., Dev, I., Petzold, C.J., Keasling, J.D., 2019. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. Nature 567, 123–126. https://doi.org/10.1038/s41586-019-0978-9

Lynch, R.C., Vergara, D., Tittes, S., White, K., Schwartz, C.J., Gibbs, M.J., Ruthenburg, T.C., deCesare, K., Land, D.P., Kane, N.C., 2016. Genomic and Chemical Diversity in *Cannabis*. Crit. Rev. Plant Sci. 35, 349–363. https://doi.org/10.1080/07352689.2016.1265363

Marks, M.D., Tian, L., Wenger, J.P., Omburo, S.N., Soto-Fuentes, W., He, J., Gang, D.R., Weiblen, G.D., Dixon, R.A., 2009. Identification of candidate genes affecting Δ9-tetrahydrocannabinol biosynthesis in Cannabis sativa. J. Exp. Bot. 60, 3715–3726. https://doi.org/10.1093/jxb/erp210

McKernan, K.J., Helbert, Y., Kane, L.T., Ebling, H., Zhang, L., Liu, B., Eaton, Z., McLaughlin, S., Kingan, S., Baybayan, P., Concepcion, G., Jordan, M., Riva, A., Barbazuk, W., Harkins, T., 2020. Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes (preprint). Genomics. https://doi.org/10.1101/2020.01.03.894428

Nützmann, H.-W., Osbourn, A., 2014. Gene clustering in plant specialized metabolism. Curr. Opin. Biotechnol. 26, 91–99. https://doi.org/10.1016/j.copbio.2013.10.009

Nuutinen, T., 2018. Medicinal properties of terpenes found in Cannabis sativa and Humulus lupulus. Eur. J. Med. Chem. 157, 198–228. https://doi.org/10.1016/j.ejmech.2018.07.076

Panchy, N., Lehti-Shiu, M., Shiu, S.-H., 2016. Evolution of Gene Duplication in Plants. Plant Physiol. 171, 2294–2316. https://doi.org/10.1104/pp.16.00523

Piluzza, G., Delogu, G., Cabras, A., Marceddu, S., Bullitta, S., 2013. Differentiation between fiber and drug types of hemp (Cannabis sativa L.) from a collection of wild and domesticated accessions. Genet. Resour. Crop Evol. 60, 2331–2342. https://doi.org/10.1007/s10722-013-0001-5

Pulido, P., Perello, C., Rodriguez-Concepcion, M., 2012. New Insights into Plant Isoprenoid Metabolism. Mol. Plant 5, 964–967. https://doi.org/10.1093/mp/sss088

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., Paterson, A.H., 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol. 20, 38. https://doi.org/10.1186/s13059-019-1650-2

Raharjo, T.J., Chang, W.-T., Verberne, M.C., Peltenburg-Looman, A.M.G., Linthorst, H.J.M., Verpoorte, R., 2004. Cloning and over-expression of a cDNA encoding a polyketide synthase from Cannabis sativa. Plant Physiol. Biochem. 42, 291–297. https://doi.org/10.1016/j.plaphy.2004.02.011

Rea, K.A., Casaretto, J.A., Al-Abdul-Wahid, M.S., Sukumaran, A., Geddes-McAlister, J., Rothstein, S.J., Akhtar, T.A., 2019. Biosynthesis of cannflavins A and B from Cannabis sativa L. Phytochemistry 164, 162–171. https://doi.org/10.1016/j.phytochem.2019.05.009

Richins, R.D., Rodriguez-Uribe, L., Lowe, K., Ferral, R., O'Connell, M.A., 2018. Accumulation of bioactive metabolites in cultivated medical Cannabis. PLOS ONE 13, e0201119. https://doi.org/10.1371/journal.pone.0201119

Shapira, A., Berman, P., Futoran, K., Guberman, O., Meiri, D., 2019. Tandem Mass Spectrometric Quantification of 93 Terpenoids in *Cannabis* Using Static Headspace Injections. Anal. Chem. 91, 11425–11432. https://doi.org/10.1021/acs.analchem.9b02844

Small, E., Beckstead, H.D., 1973. Cannabinoid Phenotypes in Cannabis sativa. Nature 245, 147–148. https://doi.org/10.1038/245147a0

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stout, J.M., Boubakir, Z., Ambrose, S.J., Purves, R.W., Page, J.E., 2012. The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in Cannabis sativa trichomes: A cytoplasmic acyl-activating enzyme involved in cannabinoid biosynthesis. Plant J. no-no. https://doi.org/10.1111/j.1365-313X.2012.04949.x

Sun, B., Zhu, Z., Cao, P., Chen, H., Chen, C., Zhou, X., Mao, Y., Lei, J., Jiang, Y., Meng, W., Wang, Y., Liu, S., 2016. Purple foliage coloration in tea (Camellia sinensis L.) arises from activation of the R2R3-MYB transcription factor CsAN1. Sci. Rep. 6, 32534. https://doi.org/10.1038/srep32534

Taura, F., Sirikantaramas, S., Shoyama, Yoshinari, Yoshikai, K., Shoyama, Yukihiro, Morimoto, S., 2007. Cannabidiolic-acid synthase, the chemotype-determining enzyme in the fiber-type *Cannabis sativa*. FEBS Lett. 581, 2929–2934. https://doi.org/10.1016/j.febslet.2007.05.043

van Bakel, H., Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R., Page, J.E., 2011. The draft genome and transcriptome of Cannabis sativa. Genome Biol. 12, R102. https://doi.org/10.1186/gb-2011-12-10-r102

van Velzen, R., Schranz, M.E., 2020. Origin and evolution of the cannabinoid oxidocyclase gene family (preprint). Evolutionary Biology. https://doi.org/10.1101/2020.12.18.423406

Vergara, D., Huscher, E.L., Keepers, K.G., Givens, R.M., Cizek, C.G., Torres, A., Gaudino, R., Kane, N.C., 2019. Gene copy number is associated with phytochemistry in Cannabis sativa. AoB PLANTS 11, plz074. https://doi.org/10.1093/aobpla/plz074

Wang, G., Dixon, R.A., 2009. Heterodimeric geranyl(geranyl)diphosphate synthase from hop (Humulus lupulus) and the evolution of monoterpene biosynthesis. Proc. Natl. Acad. Sci. 106, 9914–9919. https://doi.org/10.1073/pnas.0904069106

Weiblen, G.D., Wenger, J.P., Craft, K.J., ElSohly, M.A., Mehmedic, Z., Treiber, E.L., Marks, M.D., 2015. Gene duplication and divergence affecting drug content in *Cannabis sativa*. New Phytol. 208, 1241–1250. https://doi.org/10.1111/nph.13562

Welling, M.T., Liu, L., Raymond, C.A., Kretzschmar, T., Ansari, O., King, G.J., 2019. Complex Patterns of Cannabinoid Alkyl Side-Chain Inheritance in Cannabis. Sci. Rep. 9, 11421. https://doi.org/10.1038/s41598-019-47812-2

Zager, J.J., Lange, I., Srividya, N., Smith, A., Lange, B.M., 2019. Gene Networks Underlying Cannabinoid and Terpenoid Accumulation in Cannabis. Plant Physiol. 180, 1877–1897. https://doi.org/10.1104/pp.18.01506

Zeng, L., Dehesh, K., 2021. The eukaryotic MEP-pathway genes are evolutionarily conserved and originated from Chlaymidia and cyanobacteria. BMC Genomics 22, 137. https://doi.org/10.1186/s12864-021-07448-x

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., Gaut, B.S., 2019. The population genetics of structural variants in grapevine domestication. Nat. Plants 5, 965–979. https://doi.org/10.1038/s41477-019-0507-8