**Short title:** Coregulation of Cannabinoid and Terpenoid Pathways

# Gene Networks Underlying Cannabinoid and Terpenoid Accumulation in Cannabis[1]

Jordan J. Zager[a], Iris Lange[a], Narayanan Srividya[a], Anthony Smith[b], B. Markus Lange[a]

[a] Institute of Biological Chemistry and M.J. Murdock Metabolomics Laboratory, Washington State University, Pullman, WA 99164-6340, USA

[b] Evio Labs, 540 E Vilas Road Suite F, Central Point, OR 97502, USA

**Contact details for corresponding author**

Bernd Markus Lange, Ph.D.
Professor, Institute of Biological Chemistry
Director, M.J. Murdock Metabolomics Laboratory
Washington State University
Pullman, WA 99164-6340, USA
Tel:     xx1-509-335-3794
Fax:     xx1-509-335-7643
e-Mail: lange-m@wsu.edu

1

**ONE-SENTENCE SUMMARY**

Metabolite and transcriptome profiling of cannabis glandular trichomes differentiates strains and provides evidence for co-regulation of cannabinoid and terpenoid volatile biosynthesis.

**AUTHOR CONTRIBUTIONS**

Author contributions: J.J.Z., A.S. and B.M.L. designed the experiments; A.S. harvested and extracted plant materials; A.S. performed metabolite analyses; J.J.Z. and I.L. cloned terpene synthase genes and performed functional assays; J.J.Z., A.S. and B.M.L. analyzed the data; J.J.Z and B.M.L. wrote the manuscript, with input from all authors.

**Key words:** cannabinoid, cannabis, coexpression network, functional characterization, glandular trichome, metabolic regulation, terpene synthase, terpenoid.

**Abbreviations and Acronyms:**

CBD, cannabidiol; CBDA, cannabidiolic acid; BN, cannabinol; HCA, hierarchical clustering analysis; MEP, 2C-methyl-D-erythritol 4-phosphate; MVA, mevalonic acid; OPLS-DA, orthogonal projections to latent structures discriminant analysis; PCA, principal component analysis; RNA-seq, ribonucleic acid sequencing; SCC, Spearman correlation coefficient; THC, tetrahydrocannabinol; THCA, tetrahydrocannabinolic acid; TPM, transcripts per kilobase million; TPS, terpene synthase; WGCNA, weighted gene correlation network analysis.

**DISCLOSURE DECLARATION**

JJZ and BML are members of Dewey Scientific LLC, a biotechnology company based in Pullman, WA, USA.  A.S. serves as Chief Scientific Officer for Evio Labs, a cannabis analytical testing company based on Central Point, OR, USA.

# Abstract

Glandular trichomes are specialized anatomical structures that accumulate secretions with important biological roles in plant–environment interactions. These secretions also have commercial uses in the flavor, fragrance, and pharmaceutical industries. The capitate-stalked glandular trichomes of *Cannabis sativa* (cannabis), situated on the surfaces of the bracts of the female flowers, are the primary site for the biosynthesis and storage of resins rich in cannabinoids and terpenoids. In this study, we profiled nine commercial cannabis strains with purportedly different attributes, such as taste, color, smell and genetic origin. Glandular trichomes were isolated from each of these strains and cell type-specific transcriptome data sets were acquired. Cannabinoids and terpenoids were quantified in flower buds. Statistical analyses indicated that these data sets enable the high-resolution differentiation of strains by providing complementary information. Integrative analyses revealed a coexpression network of genes involved in the biosynthesis of both cannabinoids and terpenoids from imported precursors. Terpene synthase genes involved in the biosynthesis of the major mono- and sesquiterpenes routinely assayed by cannabis testing laboratories were identified and functionally evaluated. In addition to cloning variants of previously characterized genes, specifically Cs*TPS14CT* ((-)-limonene synthase) and Cs*TPS15CT* (β-myrcene synthase) we functionally evaluated genes that encode enzymes with activities not previously described in cannabis, namely CsTPS18VF and CsTPS19BL (nerolidol/linalool synthases); CsTPS16CC (germacrene B synthase); and CsTPS20CT (hedycaryol synthase). This study lays the groundwork for developing a better understanding of the complex chemistry and biochemistry underlying resin accumulation across commercial cannabis strains.

3

## INTRODUCTION
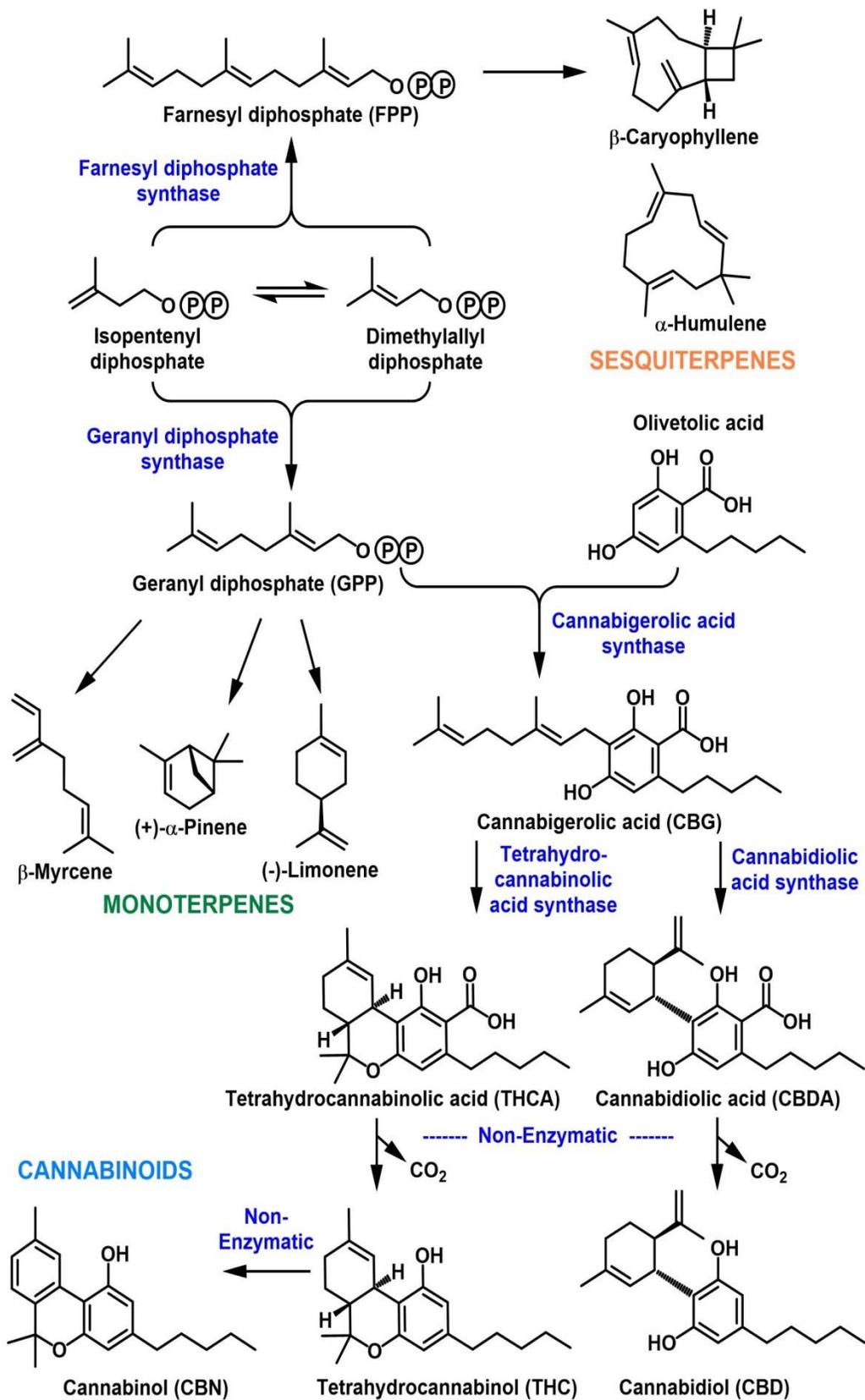
*Cannabis sativa* (cannabis) was originally discovered in Central Asia and has likely been cultivated for tens of thousands of years by human civilizations, with the first mention about 5,000 years ago in Chinese texts (Unschuld, 1986). Whereas the initial utility was primarily as a source of grain and fiber, strains with medicinal properties were already in use in northwest China some 2,700 years ago, as evidenced by the detection of the psychoactive cannabinoid, (−)-trans-$\Delta^9$-tetrahydrocannabinol (THC), in plant residues recovered from an ancient grave (Russo et al., 2008). Cannabis strains containing less THC but more of the non-psychoactive cannabidiol (CBD), commonly referred to as hemp, were grown in Roman Britain for grain and fiber, but later found additional uses as a medicine during the Anglo-Saxon period (Grattan and Singer, 1952). The 1925 Geneva International Opium Convention required signatories to control the trade of certain drugs (including cannabis), which was followed by increasingly restrictive resolutions by the League of Nations and later United Nations (United Nations, 1966). Until very recently, cannabis was considered an illicit substance of abuse by many governments, and could only be researched by selected, authorized scientists in tightly supervised laboratories. Despite these restrictions, evidence for the medicinal potential was sufficiently convincing that, by the mid-1980s, the synthetic cannabinoids nabilone and dronabinol had been granted approval by the U.S. Food and Drug Administration to suppress nausea during chemotherapy (Abuhasira et al., 2018). The discovery of the existence of a high-affinity cannabinoid receptor in the rat brain during the late 1980s (Devane et al., 1988) prompted further research to identify the endogenous ligands. This resulted in the characterization, beginning in the early 1990s, of several lipid-based retrograde neurotransmitters (endocannabinoids) and multiple enzymes involved in their biosynthesis, trafficking and perception (the endocannabinoid system), which were subsequently demonstrated to regulate a multitude of physiological and cognitive processes in humans and other animals (Devane et al., 1992). With receptor targets in hand, follow-up research and clinical trials brought several additional cannabis-related products to the pharmaceutical marketplace, including nabiximols (marketed as Sativex® in Canada since 2005), a cannabis extract used to treat symptoms of multiple sclerosis, and a formulation of

56    highly-purified, plant-sourced CBD (marketed as Epidiolex® in the USA since early 2018) to treat

57    certain forms of epilepsy. In the meanwhile, several jurisdictions and even entire countries

58    changed their policies on cannabis, endorsing laws that allow its therapeutic use, and

59    decriminalizing or even legalizing it for recreational purposes (Abuhasira et al., 2018).

60    Legislation has not been able to keep up with these recent developments, and specific labeling

61    regulations with regard to the composition of active ingredients, serving sizes and

62    recommended doses are woefully lacking (Subritzky et al., 2016). This situation is exacerbated

63    by an inadequate understanding of how the chemistry (cannabinoids and other specialized

64    metabolites) of cannabis extracts and formulations relates to their biological effects.

65

66    Since the original structural elucidation, during the early 1960s, of THC as a psychoactive

67    principle in cannabis (Gaoni and Mechoulam, 1964), the structures of more than 90 biogenic

68    cannabinoids have been reported to occur in members of the genus *Cannabis* (Andre et al.,

69    2016), with a handful of constituents being the most prominent across strains (Fig. 1). These

70    cannabinoids accumulate primarily in capitate-stalked glandular trichomes of female plants at

71    the flowering stage (Mahlberg and Kim, 2003). A second class of metabolites with high

72    abundance and even greater chemical diversity in cannabis glandular trichomes are

73    monoterpenes and sesquiterpenes (Brenneisen, 2007) (Fig. 1). These volatile terpenoids are

74    responsible for the distinctive aromas of different cannabis strains.  The popular press and

75    trade magazines liberally use the term "entourage effect" to suggest that synergism among

76    cannabinoids or between cannabinoids and other constituents (in particular terpenoids) may

77    contribute to different psychological perceptions of cannabis varieties by users. In support of

78    this view, β-caryophyllene, a sesquiterpene with almost ubiquitous occurrence in plant oils and

79    resins, was demonstrated to bind with high affinity to the CB2 cannabinoid receptor, and has

80    therefore been referred to as a dietary cannabinoid (Gertsch et al., 2008). However, there is

81    only limited clinical evidence for entourage effects of terpenoids in cannabis formulations

82    (Gertsch et al., 2010; Russo, 2011). Irrespective of these considerations, the chemical

83    composition of each cannabis strain is unique and acquiring a "metabolic fingerprint" is an

84    excellent first step in building a more robust scientific foundation for assessing the correlation

5

**Farnesyl diphosphate (FPP)**

**Farnesyl diphosphate synthase**

**β-Caryophyllene**

**α-Humulene**

**SESQUITERPENES**

Isopentenyl diphosphate ⇌ Dimethylallyl diphosphate

**Geranyl diphosphate synthase**

Olivetolic acid

**Geranyl diphosphate (GPP)**

**Cannabigerolic acid synthase**

β-Myrcene    (+)-α-Pinene    (-)-Limonene

**MONOTERPENES**

Cannabigerolic acid (CBG)

**Tetrahydro-cannabinolic acid synthase**

**Cannabidiolic acid synthase**

Tetrahydrocannabinolic acid (THCA)    Cannabidiolic acid (CBDA)

------- **Non-Enzymatic** -------

**CANNABINOIDS**

$CO_2$    $CO_2$

**Non-Enzymatic**

Cannabinol (CBN)    Tetrahydrocannabinol (THC)    Cannabidiol (CBD)

85    between the composition of plant material and the perception by users (Fischedick et al.,

6

86    2010).

87

88        Most of the cannabis products traded licitly or illicitly today are sourced from strains for

89    which minimal documentation is available in the public domain, and for which the primary goal

90    was clearly to breed high-THC strains (Fidelia et al., 2012). In other words, the genetics

91    underlying chemical diversity in commercial cannabis strains is currently poorly understood

92    (Welling et al., 2016). In this context, it is interesting to note that cannabinoids and terpenoids

93    share a common biosynthetic origin. The biosynthesis of the prominent cannabinoids involves

94    two direct precursor pathways. The polyketide pathway gives rise to olivetolic acid from a

95    short-chain fatty acid intermediate (hexanoyl-CoA), whereas the methylerythritol 4-phosphate

96    (MEP) pathway provides geranyl diphosphate (GPP) (Fellermeier et al., 2001; Taura et al., 2009;

97    Gagne et al., 2012; Stout et al., 2012; Page and Stout, 2013) (Fig. 1). An aromatic

98    prenyltransferase catalyzes the formation of cannabigerolic acid from oilvetolic acid and GPP

99    (Fellermeier and Zenk, 1998; Page and Boubakir, 2012). The pathway then branches again

100    toward different cyclized products, such as tetrahydrocannabinolic acid (THCA), cannabidiolic

101    acid (CBDA), and cannabichromanic acid (Marimoto et al., 1998; Sirikantaramas et al., 2005;

102    Taura et al., 2007) (Fig. 1). Reduced metabolic products of these acids are formed non-

103    enzymatically by exposure to heat (Degenhardt et al., 2017). Plant monoterpenes are mostly

104    derived from the plastid-localized MEP pathway, whereas the cytosolic/peroxisomal

105    mevalonate (MVA) pathway is a common source of precursors for sesquiterpenes, although

106    crosstalk between both pathways has also been reported (Hemmerlin et al., 2012) (Fig. 1).

107    Terpene synthases catalyze the first committed step in the biosynthesis of a specific terpenoid

108    from a prenyl diphosphate precursor of the appropriate chain length. To date, monoterpene

109    synthases (accepting a C10 precursor) and sesquiterpene synthases (acting on a C15 precursor)

110    that are responsible for the production of about half a dozen terpenoids in cannabis have been

111    reported (Günnewich et al., 2008; Booth et al., 2017) (Fig. 1), with many more awaiting

112    functional characterization. In the current manuscript, we report the chemical profiles and

113    corresponding gene networks across several cannabis strains, thereby building the foundation

114    for a better understanding of their chemical and biochemical diversity.

115

116

## RESULTS

### Strategic Considerations for Logistics, Strain Selection, and Experimental Design

One of the goals of this pilot study was to test the utility of combining metabolic and transcriptomic data to differentiate cannabis strains with regard to the most relevant traits. To ensure the consistency of data sets, all plant materials were sourced from the same facility, where they had been maintained under comparable growth conditions (Shadowbox Farms in Williams, OR, USA). Plant harvest was performed when the appearance of glandular trichome content had changed from a turbid white to clear, and before another change to an amber-like color occurred. For most strains, the pistils had changed color from white to a yellow or orange. These are the visual cues used by experienced growers to indicate optimal harvest time. All further processing was performed with fresh (uncured) material to avoid the previously reported loss of terpenoid volatiles during drying (Ross and El Sohly, 1996). Cannabinoids and terpenoids were extracted and quantified at a testing facility licensed according to the National Environmental Laboratory Accreditation Program's TNI 2009 Standard (Evio Labs, Central Point, OR, USA). At this facility, fractions highly enriched in glandular trichomes were obtained and RNA was isolated, with minor modifications, using previously established protocols (Lange et al., 2000). Glandular trichome-specific RNA sequencing (RNA-seq) data were then acquired by a commercial service provider (Quick Biology Inc., Pasadena, CA). Metabolite and transcriptome data were acquired for three biological replicates per strain.

This study involved a selection of strains with *Cannabis sativa* L. ancestry, whereas *Cannabis indica* Lam. (formally classified as *Cannabis sativa* forma *indica*) was dominant in others (Fig. 2). Strains of *C. sativa* provenance are generally characterized by fairly thin and narrow leaves, comparatively longer flowering cycles, and a relatively tall stature. A typical example in the current study is 'Mama Thai', which is generally considered a landrace of *C. sativa*. In contrast, *C. indica* strains ordinarily have large and thick leaves, a rather short flowering cycle (6-8 weeks), and a proportionately short habitus (Fig. 2A). Our pilot study featured 'Blackberry Kush'

9

## A



**Black Berry Kush**

**Black Lime**

**Canna Tsu**

**Mama Thai**

**Valley Fire**

**Cherry Chem**

**Terple**

**Sour Diesel**

**White Cookies**

## B

| Strain | Origin | Aromatic Description | Lineage |
|---|---|---|---|
| Blackberry Kush' | Indica Dominant Hybrid | Hashy, jet fuel aroma with a sweet berry taste | Afghani x Blackberry |
| Black Lime' | Sativa Dominant Hybrid | Pine, lemon, and black pepper | Northern Lights x Purple Kush |
| Canna Tsu' | Sativa Dominant Hybrid | Sweet, earthy, citrus | Cannatonic x Sour Tsunami |
| Mama Thai' | Sativa | Fruity, citrus | Likely a landrace |
| Valley Fire' | Sativa Dominant Hybrid | Earthy, pine, citrus, flowery | The White x Fire OG |
| Cherry Chem' | Indica Dominant Hybrid | Sweet cherry, grassy, spicy | Cherry Pie x Chemdawg |
| Terple' | Unclear | Earthy, woody, tobacco, spicy, slightly fruity | Unknown |
| Sour Diesel' | Sativa Dominant Hybrid | Pungent diesel like aroma, earthy | Chemdawg 91 x Super Skunk |
| White Cookies' | Sativa Dominant Hybrid | Sweet, earthy | White Widow x Girl Scout Cookies |

146    as a *C. indica* dominant strain. The remaining strains were hybrids of mixed *C. sativa* and *C.*

10

147     *indica* lineage, plus one strain ('Terple') with poorly documented origin (Fig. 2B).

148

149     To address our goal of assessing the utility of our data for classifying strains, RNA-seq and

150     chemical data (cannabinoid and terpenoid profiles) were subjected to multivariate statistical

151     analyses. We then tested the hypothesis that cannabinoid and terpenoid pathways are co-

152     regulated by performing gene coexpression network analyses. A combination of gene network

153     and phylogenetic analyses was subsequently used to identify candidate genes for hitherto

154     uncharacterized terpene synthases that contribute significantly to the cannabis volatile

155     bouquet.

156

157 **Strain Differentiation Based on RNA-seq Data**

158

159     High-quality libraries reflecting transcripts expressed in isolated glandular trichomes were

160     subjected to RNA-seq analysis (nine strains, three biological replicates each, 27 samples total)

161     on the Illumina HiSeq 4000 platform. A *de novo* consensus transcriptome assembly was

162     generated using the Trinity suite (Haas et al., 2013) (assembly statistics in Supplemental Table

163     S1). The reads were assembled into contigs covering a total of 305 Mbp of sequence with a GC

164     content of 40.4%. The resulting assembly produced an N50 value of 833 bp, containing 514,208

165     contigs of at least 201 bp in length. The assembled transcriptome data set was searched against

166     the NCBI non-redundant protein database, which resulted in the annotation of 82,523

167     sequences at e-values < 1e-5. Read counts for each transcript in each sample were then

168     processed with the RSEM software package (Li and Dewey, 2011) to calculate normalized

169     expression levels as <u>T</u>ranscripts <u>P</u>er <u>K</u>ilobase <u>M</u>illion (TPM). Transcripts with TPM values lower

170     than 5.0 across all varieties were removed from subsequent analysis, resulting in 46,559

171     predicted genes with significant expression (Supplemental Table S2).

172

173     As a first step to investigate the utility of RNA-seq for strain categorization, transcriptome

174     data sets were subjected to <u>P</u>rincipal <u>C</u>omponent <u>A</u>nalysis (PCA), a statistical procedure that

175     reduces attribute space from a larger number of variables to a smaller number of so-called

11
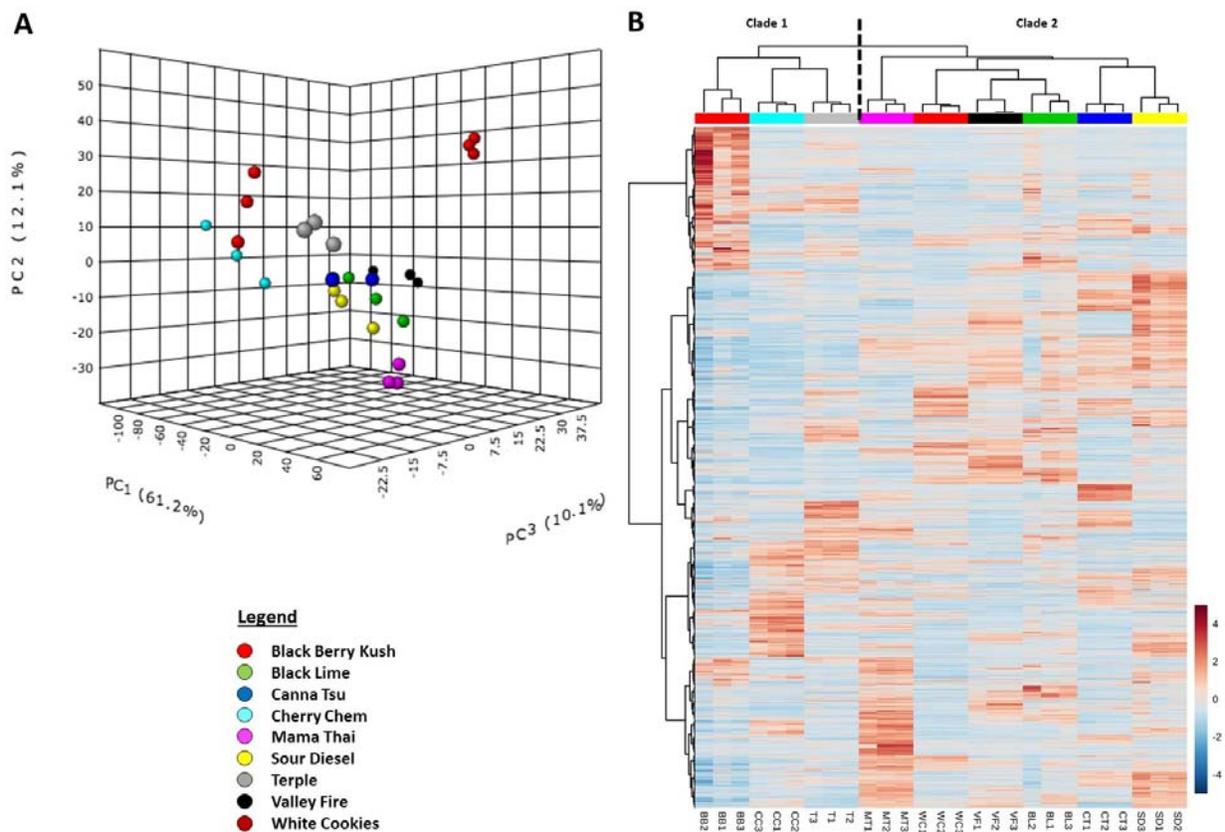
principal components, thereby decreasing the dimensionality of the original data. The first

three principal components accounted for 83% of the variability in the data set (Fig. 3A). The

replicates for each strain clustered together in a three-dimensional PCA plot, whereas the

component scores for each strain were separated from those of all other strains, indicating that

the overall transcriptome of each strain is unique (Fig. 3A). Processing of RNA-seq data by

Hierarchical Clustering Analysis (HCA), which builds a cluster hierarchy that is commonly

displayed as a dendrogram, grouped strains into two major clades (Fig. 3B). The first clade

contained 'Blackberry Kush', 'Cherry Chem' and 'Terple', whereas the second consisted of

'Mama Thai', 'White Cookies', 'Valley Fire', 'Black Lime', 'Canna Tsu' and 'Sour Diesel', indicating

a clear separation of strains by heritage (*C. indica* for clade 1 and *C. sativa* for clade 2).


**Strain Differentiation Based on Metabolite Profiling Data**

189    The highly robust analytical platforms that served as the basis for the analysis of six
190    cannabinoids and 24 terpenoids were described in a previous report (Fischedick et al., 2010)
191    and used here with minor modifications. Cannabinoid concentration was highest in 'White
192    Cookies' (28.4% of flower bud dry weight), with relatively high contents also occurring in
193    'Cherry Chem' (17.7%), 'Black Lime' (17.5%), 'Backberry Kush' (15.8%), 'Valley Fire' (15.7%),
194    'Terple' (15.6%), 'Sour Diesel' (12.4%), and 'Canna Tsu' (12.2%) (Table 1).  Significantly lower
195    concentrations were detected in 'Mama Thai' (6.4%). In eight of the nine strains investigated,
196    THCA was the major cannabinoid, ranging from 26.3% of the flower bud dry weight in 'White
197    Cookies' to 5.9% in 'Mama Thai' (Table 1). The only exception was the 'Canna Tsu' strain, in
198    which CBDA (7.8% of flower bud dry weight) dominated over THCA (3.2%), whereas CBDA in all
199    other strains remained at 1% or less. Two additional cannabinoids of fairly high abundance
200    were cannabinol (CBN), which accumulated to 0.2–1.7% of flower bud dry weight, and
201    tetrahydrocannabinol, which amounted to 0.2–1.6% (Table 1) (for structures see Fig. 1).
202    Cannabichromene was not detected in any of the sampled varieties.

203

204    Terpenoid content was highest in 'Black Lime' (8.8% of flower bud dry weight), with fairly
205    high contents also occurring in 'White Cookies' (4.8%), 'Terple' (4.8%), 'Valley Fire' (3.8%),
206    'Cherry Chem' (3.5%), 'Blackberry Kush' (3.4%), and 'Canna Tsu' (3.3%) (Table 1).  Significantly
207    lower concentrations were detected in 'Sour Diesel' (1.8%) and 'Mama Thai' (0.7%). The
208    monoterpene (C10) to sesquiterpene (C15) ratio was generally very high (> 10), with only three
209    strains in which the ratio was below 3 ('Cherry Chem', 'Mama Thai', and 'Sour Diesel') (Table 1).
210    It should be noted that this ratio only applies to the terpenoids we were able to quantify based
211    on the availability of authentic standards. β-Myrcene was the most abundant monoterpene in
212    most strains (up to 4.3% of flower bud dry weight in 'Black Lime'). The only exceptions were
213    'Mama Thai' (generally low terpenoid contents, with terpinolene as most abundant
214    monoterpene at 0.1%) and 'White Cookies' (with limonene at 1.5%) (Table 1). Limonene
215    content was also high in 'Black Lime' (0.9%) and 'Valley Fire' (0.7%). α-Pinene and β-pinene
216    amounts were quite high in 'Black Lime' (2.0% and 0.5%, respectively). 1,8-Cineole was
217    particularly abundant in 'Canna Tsu' and 'Cherry Chem' (0.5% in both) (Table 1). All other

13

**Table 1.** Constituents of Cannabis Female Flower Buds (Metabolite Content in Nine Strains Expressed as Percent of Dry Weight).
Abbreviations: n.d., not detectable.

| Metabolite | Black Berry Kush | Black Lime | Canna Tsu | Cherry Chem | Valley Fire | Mamma Thai | Sour Diesel | Terple | White Cookies |
|---|---|---|---|---|---|---|---|---|---|
| **Cannabinoids** | | | | | | | | | |
| Tetrahydrocannabinolic acid | 13.56 ± 0.90 | 15.02 ± 1.10 | 3.19 ± 0.20 | 16.55 ± 0.81 | 13.89 ± 1.33 | 5.91 ± 0.60 | 11.31 ± 1.04 | 13.72 ± 1.36 | 26.33 ± 0.54 |
| Tetrahydrocannabinol | 0.31 ± 0.02 | 1.62 ± 0.19 | 0.55 ± 0.055 | 0.15 ± 0.008 | 0.41 ± 0.049 | 0.14 ± 0.02 | 0.22 ± 0.027 | 1.15 ± 0.12 | 0.86 ± 0.091 |
| Cannabidiolic acid | 0.45 ± 0.02 | 0.12 ± 0.012 | 7.76 ± 0.63 | 0.079 ± 0.007 | 0.037 ± 0.001 | 0.016 ± 0.003 | 0.032 ± 0.002 | 0.067 ± 0.002 | 0.088 ± 0.004 |
| Cannabidiol | 0.95 ± 0.07 | 0.139 ± 0.016 | 0.085 ± 0.013 | 0.079 ± 0.008 | 0.12 ± 0.004 | 0.047 ± 0.005 | 0.086 ± 0.006 | 0.11 ± 0.008 | 0.098 ± 0.013 |
| Cannabigerol | 0.12 ± 0.015 | 0.086 ± 0.008 | 0.093 ± 0.008 | 0.051 ± 0.005 | 0.15 ± 0.027 | 0.016 ± 0.001 | 0.052 ± 0.004 | 0.093 ± 0.002 | 0.25 ± 0.005 |
| Cannabinol | 1.74 ± 0.20 | 0.55 ± 0.019 | 0.53 ± 0.051 | 0.83 ± 0.019 | 1.12 ± 0.14 | 0.29 ± 0.028 | 0.68 ± 0.033 | 0.502 ± 0.007 | 0.78 ± 0.025 |
| Cannabichromene | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| Total Cannabinoids | 15.87 ± 1.13 | 17.53 ± 1.25 | 12.20 ± 0.85 | 17.74 ± 0.82 | 15.70 ± 1.53 | 6.41 ± 0.65 | 12.387 ± 1.05 | 15.64 ± 1.47 | 28.40 ± 0.54 |
| **Monoterpenes** | | | | | | | | | |
| β-Myrcene | 2.35 ± 0.2 | 4.34 ± 0.36 | 1.70 ± 0.15 | 1.61 ± 0.049 | 2.24 ± 0.28 | 0.11 ± 0.009 | 0.70 ± 0.046 | 2.96 ± 0.25 | 1.14 ± 0.17 |
| (-)-Limonene | 0.29 ± 0.02 | 0.89 ± 0.08 | 0.16 ± 0.021 | 0.23 ± 0.015 | 0.65 ± 0.098 | 0.03 ± 0.003 | 0.17 ± 0.011 | 0.23 ± 0.019 | 1.53 ± 0.24 |
| α-Pinene | 0.015 ± 0.001 | 1.99 ± 0.12 | 0.38 ± 0.039 | 0.016 ± 0.001 | 0.044 ± 0.008 | 0.007 ± 0.001 | 0.004 ± 0 | 0.82 ± 0.051 | 0.20 ± 0.032 |
| β-Pinene | 0.086 ± 0.005 | 0.50 ± 0.034 | 0.18 ± 0.025 | 0.056 ± 0.003 | 0.11 ± 0.013 | 0.026 ± 0.002 | 0.039 ± 0.002 | 0.31 ± 0.022 | 0.04 ± 0.007 |
| 1,8-Cineole | 0.26 ± 0.02 | 0.38 ± 0.038 | 0.52 ± 0.075 | 0.464 ± 0.012 | 0.22 ± 0.028 | 0.057 ± 0.007 | 0.11 ± 0.011 | 0.00 ± 0 | 0.31 ± 0.037 |
| Linalool | 0.082 ± 0.005 | 0.079 ± 0.004 | 0.052 ± 0.005 | 0.13 ± 0.003 | 0.16 ± 0.027 | 0.023 ± 0.002 | 0.074 ± 0.005 | 0.067 ± 0.006 | 0.57 ± 0.072 |
| Terpinolene | 0.019 ± 0.001 | 0.034 ± 0.003 | 0.019 ± 0.002 | 0.019 ± 0.001 | 0.02 ± 0.003 | 0.13 ± 0.016 | 0.017 ± 0.001 | 0.02 ± 0.002 | 0.041 ± 0.006 |
| Borneol | 0.039 ± 0.002 | 0.041 ± 0.003 | n.d. | 0.032 ± 0.002 | 0.033 ± 0.005 | 0.021 ± 0.002 | 0.026 ± 0.002 | 0.036 ± 0.002 | 0.048 ± 0.008 |
| β-Ocimene | n.d. | 0.039 ± 0.003 | n.d. | n.d. | 0.006 ± 0.001 | 0.13 ± 0.014 | n.d. | 0.086 ± 0.007 | 0.015 ± 0.002 |
| Camphene | n.d. | 0.089 ± 0.008 | 0.055 ± 0.007 | n.d. | 0.004 ± 0.001 | n.d. | n.d. | 0.019 ± 0.002 | 0.07 ± 0.012 |
| δ-3-Carene | 0.029 ± 0.002 | 0.052 ± 0.006 | 0.003 ± 0 | 0.008 ± 0.001 | 0.022 ± 0.003 | 0.003 ± 0.001 | n.d. | 0.027 ± 0.002 | 0.016 ± 0.002 |
| Camphor | 0.044 ± 0.003 | 0.006 ± 0.001 | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | 0.101 ± 0.013 |
| (+)-Terpinene | 0.001 ± 0.001 | n.d. | n.d. | n.d. | n.d. | 0.005 ± 0.001 | n.d. | n.d. | 0.002 ± 0 |
| Total monoterpenes | 3.23 ± 0.26 | 8.43 ± 0.66 | 3.07 ± 0.32 | 2.56 ± 0.085 | 3.52 ± 0.47 | 0.54 ± 0.057 | 1.14 ± 0.078 | 4.57 ± 0.36 | 4.09 ± 0.60 |
| **Sesquiterpenes** | | | | | | | | | |
| β-Caryophyllene | 0.13 ± 0.01 | 0.24 ± 0.023 | 0.21 ± 0.022 | 0.74 ± 0.012 | 0.23 ± 0.034 | 0.12 ± 0.013 | 0.45 ± 0.026 | 0.15 ± 0.009 | 0.60 ± 0.068 |
| α-Humulene | 0.03 ± 0.002 | 0.06 ± 0.005 | 0.051 ± 0.005 | 0.20 ± 0.011 | 0.087 ± 0.014 | 0.068 ± 0.008 | 0.19 ± 0.009 | 0.058 ± 0.003 | 0.15 ± 0.018 |
| Nerolidol | n.d. | 0.06 ± 0.004 | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| Total sesquiterpenes | 0.16 ± 0.015 | 0.361 ± 0.032 | 0.26 ± 0.027 | 0.93 ± 0.019 | 0.32 ± 0.048 | 0.19 ± 0.021 | 0.64 ± 0.035 | 0.21 ± 0.012 | 0.75 ± 0.086 |
| **Total terpenoids** | 3.39 ± 0.27 | 8.79 ± 0.69 | 3.33 ± 0.35 | 3.49 ± 0.10 | 3.84 ± 0.51 | 0.73 ± 0.078 | 1.78 ± 0.11 | 4.78 ± 0.38 | 4.83 ± 0.69 |

1

218 monoterpenes had concentrations below 0.2%. All strains contained sesquiterpenes, of which

219 β-caryophyllene was consistently the most abundant (0.1–0.7% of flower bud dry weight). α-

220 Humulene was also detectable in all strains (< 0.2%), whereas 'Black Lime' was the only strain in

221 which the nerolidol concentration rose above the limit of quantitation (< 0.1%) (Table 1).

222

223 Processing of the metabolite data (cannabinoids and terpenoid profiles) by PCA resulted in

224 a clear separation of the strains, with individual biological replicates clustering closely together

225 (Fig. 4A). Remarkably, 99% of the data variation across genotypes was captured by the first

226 three principal components. Application of Orthogonal Projections to Latent Structures

227 Discriminant Analysis (OPLS-DA), a statistical modeling tool used commonly in metabolomics

228 research (Worley, 2013), indicated a separation of strains into two groups based on our

229 metabolite profiling data, one representing the *C. indica*-dominant strains, whereas another

230 constituted the *C. sativa*-dominant strains (Fig. 4B). Biological replicates for each strain once

231 again clustered together, whereas significant separation was observed across strains. In

232 summary, glandular trichome-specific gene expression and metabolite data were consistent in

233 differentiating cannabis strains.

234

14
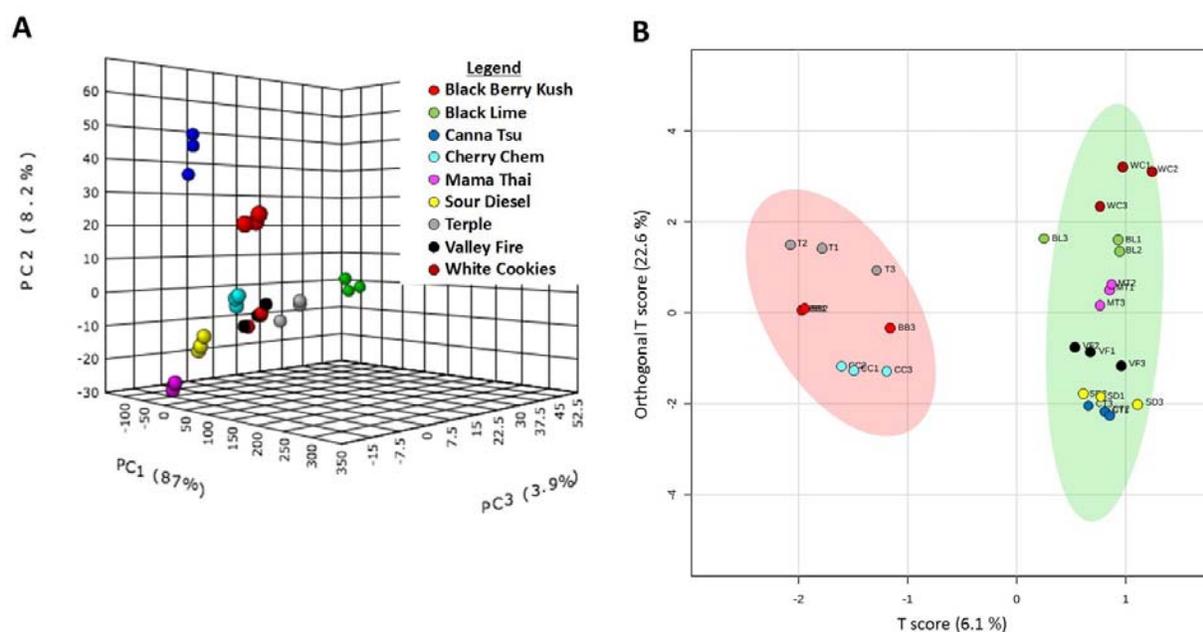
**Evidence for Co-Expression of Cannabinoid and Terpenoid Pathways**

Our glandular trichome RNA-seq data sets were filtered to eliminate genes with consistently low expression levels (below 50 TPM), thereby retaining roughly 16,000 expressed genes with significant expression levels in at least one strain. Gene abundance across strains was then evaluated using the <u>W</u>eighted <u>G</u>ene <u>C</u>orrelation <u>N</u>etwork <u>A</u>nalysis (WGCNA) package in *R* (Langfelder and Horvath, 2008), which resulted in the binning of genes (only those with Spearman Correlation Coefficients (SCCs) of ≥ 0.8 were considered) into seven co-expression modules (Supplemental Table S3). Further analysis using the 'moduleEigengenes' function indicated that the accumulation of CBDA, the signature cannabinoid of the 'Canna Tsu' strain, was highly correlated (SCC of 0.97, *P*-value of 2e-17) with one of the co-expression modules (indicated by brown color in Fig. 5A). Interestingly, this module contained the gene coding for CBDA synthase, the enzyme responsible for the conversion of cannabigerolic acid to CBDA (Table 2). An analogous analysis for THCA or THC (which correlated with a module indicated by yellow color in Fig. 5A) and THCA synthase was not possible, because single nucleotide polymorphisms in this gene (and not lack of expression) result in an inactive enzyme in strains that accumulate primarily CBDA (Kojoma et al., 2006; Laverty et al., 2018) (Table 2).

15
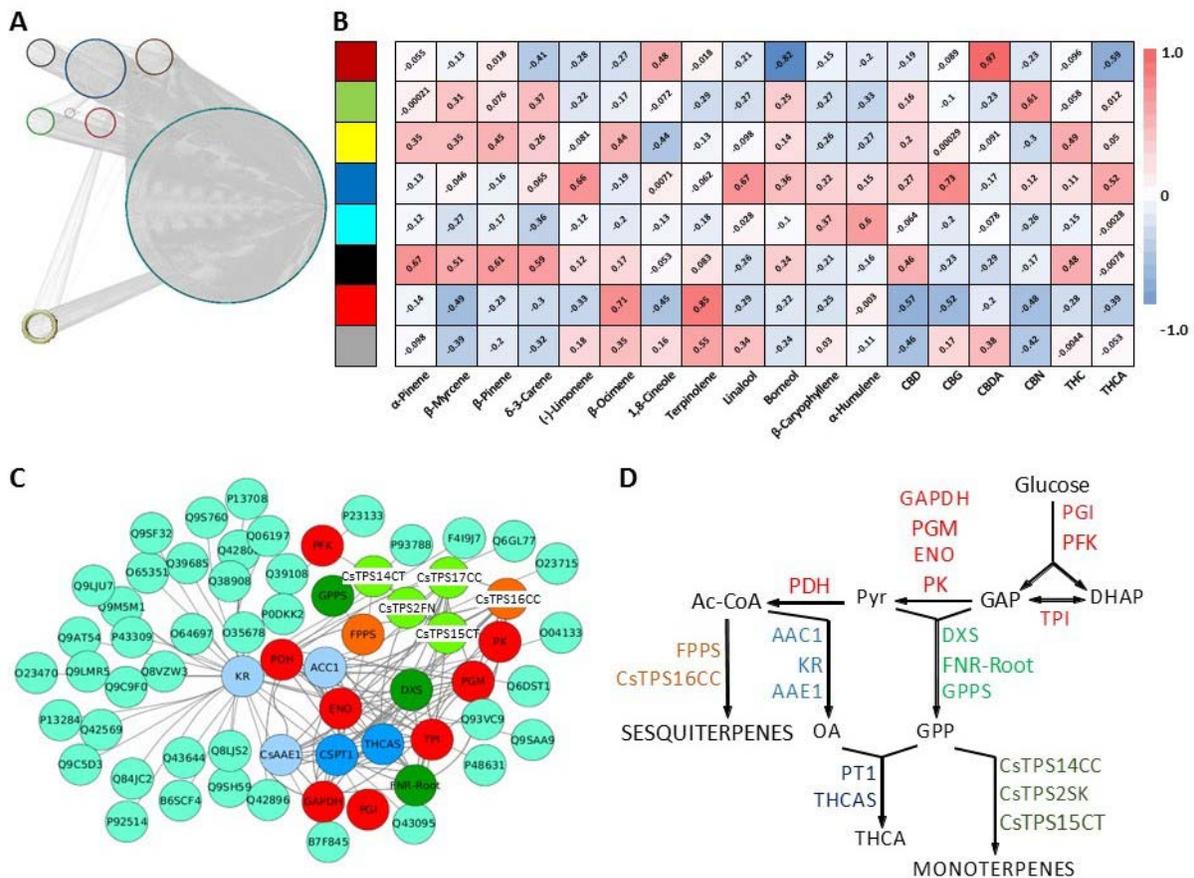
Interestingly, the THCA synthase sequences were essentially identical, with the exception of that of the 'Canna Tsu' strain, the only CBDA accumulator in our pilot study (Supplemental Fig. S1). Consequently, a full-length CBDA synthase gene was expressed only in the 'Canna Tsu' strain (Supplemental Fig S2), which is novel information that furthers our understanding of the mechanisms underlying CBDA accumulation. Finally, the yellow-colored module (which as mentioned above contained THCA synthase) also comprised cannabigerolic acid synthase (Table 2), the gene preceding tetrahydrocannabinolic acid synthase in the cannabinoid pathway (Fig. 1), thereby providing additional evidence for gene-to-metabolite correlation in the cannabinoid pathway.

We then asked if similar gene-to-metabolite correlations occurred in the terpenoid pathway. Interestingly, two co-expression modules (indicated by black and yellow color in Fig.

16

**Table 2.** Transcript Abundance for Genes Involved in the Biosynthesis of Cannabinoids and Terpenoids in Cannabis Strains.
Abbreviations: n.d., not detectable; MEP, 2C-methyl-D- erythritol 4-phosphate; MVA, mevalonate.

| Gene Annotation | UniProt Identifier | Transcript Abundance [Transcripts Per Kilobase Million] | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Black Berry Kush | Black Lime | Canna Tsu | Cherry Chem | Mama Thai | Sour Diesel | Terple | Valley Fire | White Cookies |
| **Cannabinoid pathway** | | | | | | | | | | |
| Acyl activating enzyme 1 | H9A1V3_CANSA | 80.63 | 160.44 | 316.06 | 840.92 | 377.29 | 397.99 | 93.59 | 188.84 | 229.65 |
| Olivetol synthase | OLIS_CANSA | 3946.85 | 9454.00 | 10400.03 | 14619.66 | 17955.05 | 4984.60 | 9706.06 | 11374.75 | 12373.11 |
| Geranyl diphosphate: olivetolate geranyltransferase | CsPT1 | 422.42 | 222.42 | 189.43 | 407.76 | 649.37 | 263.62 | 246.13 | 175.87 | 115.21 |
| CBDA synthase | CBDAS_CANSA | n.d. | n.d. | 1282.46 | n.d. | n.d. | 18.39 | n.d. | n.d. | n.d. |
| THCA synthase | THCAS_CANSA | 885.17 | 423.29 | 1203.31 | 2321.64 | 2317.68 | 1557.54 | 619.22 | 309.23 | 524.08 |
| **MEP pathway** | | | | | | | | | | |
| 1-Deoxy-D-xylulose-5-phosphate synthase | A0A1V0QSH6_CANSA | 221.85 | 284.41 | 412.74 | 1627.02 | 319.76 | 1751.70 | 533.57 | 288.69 | 16.32 |
| 1-Deoxy-D-xylulose 5-phosphate reductoisomerase | A0A1V0QSG8_CANSA | 172.63 | 228.15 | 185.07 | 667.96 | 304.62 | 117.92 | 176.79 | 256.25 | 16.01 |
| 2-C-Methyl-D-erythritol 4-phosphate cytidylyltransferase | A0A1V0QSI6_CANSA | 36.77 | 95.99 | 96.25 | 168.24 | 160.40 | 146.38 | 46.96 | 75.40 | 64.73 |
| 4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase | A0A1V0QSI2_CANSA | 35.20 | 3.70 | 67.94 | 211.85 | 212.43 | 109.88 | 57.60 | 104.05 | 80.23 |
| 2-C-Methyl-D-erythritol 2,4,-cyclodiphosphate synthase | G9C075_HUMLU | 67.75 | 118.23 | 315.86 | 338.21 | 184.98 | 419.84 | 69.75 | 171.17 | 207.15 |
| (E)-4-Hydroxy-3-methylbut-2-enyl-diphosphate synthase | A0A1V0QSG3_CANSA | 107.65 | 287.57 | 794.25 | 744.09 | 444.09 | 596.36 | 349.56 | 297.07 | 317.55 |
| (E)-4-Hydroxy-3-methylbut-2-enyl-diphosphate reductase | A0A1V0QSH9_CANSA | 1485.98 | 561.96 | 3447.50 | 3468.57 | 3090.49 | 3024.22 | 1889.37 | 1031.90 | 4544.35 |
| Isopentenyldiphosphate isomerase | A0A1V0QSG5_CANSA | 165.10 | 272.72 | 433.46 | 1836.07 | 306.03 | 347.85 | 476.86 | 509.70 | 9.96 |
| **MVA pathway** | | | | | | | | | | |
| Acetoacetyl-CoA thiolase | A0A1V0QSH3_CANSA | 38.35 | 11.90 | 253.38 | 302.58 | 313.99 | 134.71 | 252.40 | 54.35 | 248.13 |
| 3-Hydroxy-3-methylglutaryl-coenzyme A synthase | A0A1V0QSH3_CANSA | 13.44 | 22.98 | 20.81 | 21.60 | 27.81 | 34.33 | 9.24 | 19.32 | 91.24 |
| 3-Hydroxy-3-methylglutaryl-coenzyme A reductase | A0A1V0QSF5_CANSA | 26.69 | 56.93 | 21.92 | 43.41 | 29.05 | 107.71 | 19.75 | 69.30 | 48.26 |
| Mevalonate kinase | A0A1V0QSI0_CANSA | 1.63 | 1449.32 | 3.63 | 3.41 | 5.81 | 4.75 | 2.45 | 5.93 | 5.05 |
| Phosphomevalonate kinase | A0A1V0QSH8_CANSA | 3.68 | 7.58 | 7.99 | 6.63 | 8.09 | 6.03 | 3.81 | 7.40 | 305.27 |
| Mevalonate diphosphate decarboxylase | A0A1V0QSG4_CANSA | 5.00 | 11.89 | 10.21 | 14.89 | 21.24 | 19.39 | 9.67 | 9.64 | 9.96 |

1

264 5A) correlated with β-myrcene accumulation (Fig. 5B). This metabolite is formed by a

265 monoterpene synthase encoded by the Cs*TPS3FN* gene (Booth et al., 2017), which was

266 contained in one of these modules (yellow color in Fig. 5A) (Table 3). Analogous gene-to-

267 metabolite correlations were observed for limonene and Cs*TPS1FN*, α-pinene and Cs*TPS2FN*, β-

268 ocimene and Cs*TPS6FN*, and β-caryophyllene/α-humulene and Cs*TPS9FN* (color of modules in

269 Fig. 5A: black, yellow, yellow, turquoise, respectively; terpene synthase annotation based on

270 Günnewich et al., 2007 and Booth et al., 2017) (Fig. 5B). Transcripts corresponding to Cs*TPS5FN*

271 (β-myrcene/α-pinene synthase), Cs*TPS4FN* (alloaromadendrene synthase), Cs*TPS8FN* (γ-

272 eudesmol/valencene synthase) and Cs*TPS13PK* (a second β-ocimene synthase) (Booth et al.,

273 2017) remained below the threshold expression level in our data sets. The corresponding

274 terpenoids were not detected in the strains investigated, indicating that the expressed gene

275 complement was generally sufficient to account for the presence of the major terpenoids

17

**Table 3.** Transcript Abundance for Terpene Synthases Across Cannabis Strains.
Symbols: *, functionally characterized as part of the present study; #, from Booth et al., 2017

| Gene | GenBank Accession | CsTPS Identifier | Transcript Abundance (TPM) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Black Berry Kush | Black Lime | Canna Tsu | Cherry Chem | Valley Fire | Mamma Thai | Sour Diesel | Terple | White Cookies |
| **Monoterpene synthases (TPS-b clade)** | | | | | | | | | | | |
| (-)-Limonene synthase * | MK801766 | CsTPS14CT | 646.24 | 898.94 | 612.37 | 651.86 | 2272.48 | 751.48 | 201.94 | 2.46 | 895.86 |
| (+)-α-Pinene synthase # | KY014565 | CsTPS2FN | 217.36 | 2041.33 | 1554.77 | 101.32 | 96.90 | n.d. | n.d. | 1298.95 | 49.52 |
| β-Myrcene synthase* | MK801765 | CsTPS15CT | 183.29 | 597.88 | 325.85 | 272.65 | 157.78 | 254.10 | 183.29 | 436.63 | n.d. |
| β-Myrcene/(-)-α-pinene synthase # | KY014560 | CsTPS5FN | 217.59 | 640.97 | 483.09 | 547.24 | 157.78 | 445.85 | 125.94 | 472.33 | 50.51 |
| (E)-β-Ocimene synthase # | KY014563 | CsTPS6FN | n.d. | n.d. | n.d. | n.d. | n.d. | 103.41 | n.d. | 191.65 | n.d. |
| (Z)-β-Ocimene synthase # | KY014558 | CsTPS13PK | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| **Acyclic terpene synthases (TPS-g clade)** | | | | | | | | | | | |
| (E)-Nerolidol/(+)-linalool synthase * | MK801764 | CsTPS18VF | 2.82 | 9.41 | 2.62 | 16.21 | 16.39 | 2.51 | 4.80 | 16.77 | 8.76 |
| (E)-Nerolidol/linalool synthase * | MK801763 | CSTPS19BL | 56.78 | 81.13 | 27.22 | 80.23 | 249.23 | 62.53 | 47.73 | 90.86 | 66.47 |
| **Sesquiterpene synthases (TPS-a clade)** | | | | | | | | | | | |
| Alloaromadendrene synthase # | KY014564 | CsTPS4FN | n.d. | 108.92 | n.d. | 639.56 | n.d. | 329.87 | 148.17 | n.d. | 323.36 |
| γ-Eudesmol/valencene synthase # (putative) | KY014556 | CsTPS8FN | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| δ-Selinene synthase # (putative) | KY014554 | CsTPS7FN | 356.34 | n.d. | 367.47 | n.d. | 316.74 | 210.58 | n.d. | n.d. | 268.50 |
| β-Caryophyllene/α-humulene synthase # | KY014555 | CsTPS9FN | 764.18 | 794.46 | 435.11 | 3241.85 | 1090.94 | 738.74 | 555.25 | 495.72 | 591.86 |
| Germacrene B synthase * | MK131289 | CsTPS16CC | 16.14 | 19.44 | 9.13 | 156.08 | 20.60 | 40.36 | 20.22 | 7.19 | 22.72 |
| Hedycaryol synthase * | MK801762 | CSTPS20CT | 310.43 | 27.00 | 498.70 | 98.21 | 19.35 | 11.98 | 17.67 | 0.00 | 17.02 |

1

276 (Table 3). Linalool and nerolidol were exceptions for which the corresponding terpene

277 synthases had hitherto not been identified from cannabis. Notably, genes involved in the

278 formation of these terpenoids (and others) were cloned and functionally characterized as part

279 of the current study (details below), which contributes significantly to a better understanding of

280 the genetic underpinnings of terpenoid diversity.

281

282 The yellow module featured prominently in our gene-to-metabolite correlation analysis for

283 the cannabinoid and terpenoid pathways. Interestingly, a Gene Ontology (GO) analysis implied

284 a substantial enrichment of genes involved in terpenoid biosynthesis in the yellow module (*P*-

285 value of 1.4e-05) (Supplemental Table S3) (note that GO terms for cannabinoid biosynthesis as

286 a biological process have not yet been released). Interestingly, a total of 22 genes involved in

287 the conversion of precursor metabolites into cannabinoid and terpenoid end products were co-

18

288 expressed with THCA synthase (Fig. 5C). Specifically, these genes code for enzymes involved in

289 glycolysis (conversion of an imported carbon source into triose phosphates and pyruvic acid),

290 the MEP pathway toward GPP and ultimately monoterpenes, the production of sesquiterpenes,

291 the formation of olivetolic acid from fatty acid precursors, and the incorporation of olivetolic

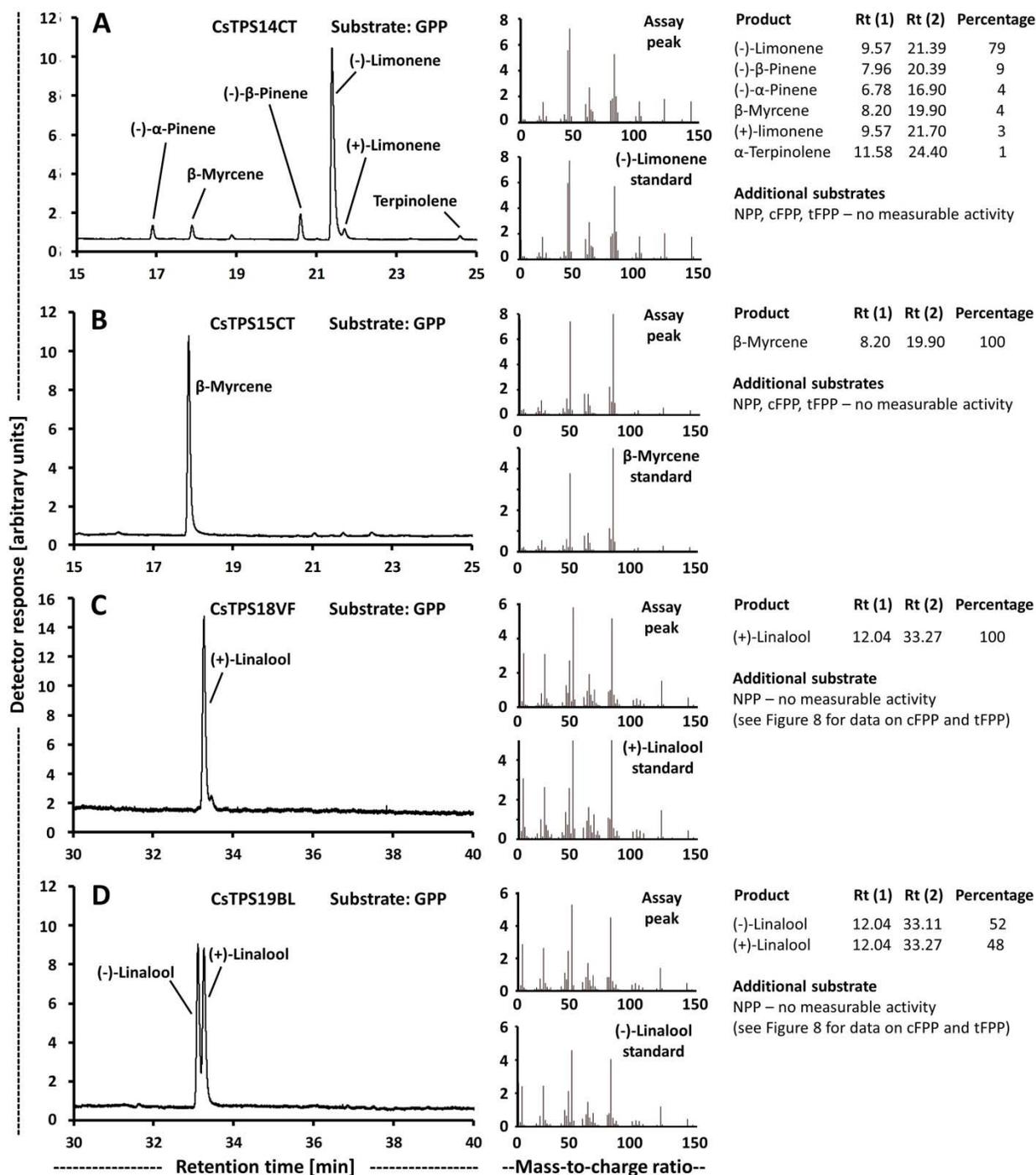292 acid and GPP into cannabinoids (Fig. 5D).

293

294 **Target Gene Identification and Characterization**

295

296 Building on our terpenoid profiling and glandular trichome-specific transcriptome data sets, we

297 embarked on gene discovery efforts aimed at characterizing terpene synthases associated with

298 the biosynthesis of major mono- and sesquiterpenes routinely quantified in commercial

299 cannabis testing, as well as other terpenoids that are not assayed routinely. The analytical

300 chemistry data were employed to assess which genes would be expected to be expressed to

301 support the observed terpenoid profiles. We then performed BlastX searches with previously

302 characterized terpene synthases to identify contigs with high sequence identity in our

303 transcriptome data sets.  We then asked which of the putative cannabis terpene synthases

304 were expressed at appreciable levels in particular cannabis strains. Sequences of selected

305 contigs were then chosen to perform a sequence relatedness analysis with previously

306 characterized terpene synthases, thereby enabling their categorization by class. cDNAs of

307 putative terpene synthases were cloned into appropriate vectors, expressed heterologously in

308 *Escherichia coli*, the corresponding recombinant proteins purified, and assays performed with

309 appropriate prenyl diphosphate substrates. Expression for genes putatively encoding geranyl

310 diphosphate synthase and *trans,trans*-farnesyl diphosphate synthase was readily detectable in

311 transcriptome data sets of all strains; in contrast, no putative orthologs of neryl diphosphate

312 synthase and *cis,cis*-farnesyl diphosphate synthase were recognizable based on sequence

313 identity (Supplemental Table S1 and S2). Nevertheless, terpene synthase assays were

314 performed with GPP, neryl diphosphate (NPP), 2-*trans,6-trans* farnesyl diphosphate (tFPP) and

315 2-*cis,6-cis*-farnesyl diphosphate (cFPP).

316

19

317      β-Myrcene and (-)-limonene were principal monoterpenes in all strains (Table 1) and,

318 expectedly, contigs with high sequence identity to the previously characterized β-myrcene and

319 (-)-limonene synthases of cannabis (Günnewich et al., 2007; Booth et al., 2017), which belong to

320 the TPS-b clade of terpene synthases (Fig. 6 and Supplemental Table S4), were expressed at

321 high levels across most strains investigated in the present study (Table 2). Cloning was

322 successful for the corresponding cDNAs from the 'Canna Tsu' strain (Cs*TPS14CT* and

323 Cs*TPS15CT*), and a functional evaluation confirmed the annotation ((-)-limonene synthase and

324 β-myrcene, respectively) (Fig. 7A, B). The translated peptide sequences of β-myrcene synthases

325 (CsTPS3FN and CsTPS15CT; excluding plastidial targeting sequence) had thirteen mismatches

326 (Supplemental Fig. S3) but identical specificity (100% β-myrcene as product with GPP as

327 substrate). The sequence of the (-)-limonene synthase characterized as part of the present

328 study (CsTPS14CT; excluding plastidial targeting sequence) had two mismatches when

329 compared to CsTPS1SK and nine mismatches when compared to CsTPS1FN (Supplemental Fig.

330 S3). As described for CsTPS1SK, CsTPS14CT generated several other products, and we report the

331 stereochemistry of those (Fig. 7A).

332

333      The monoterpene linalool was accumulated to fairly high amounts in the 'Valley Fire' and

334 'White Cookies' strains, whereas the sesquiterpene nerolidol was quantifiable only in the 'Black

335 Lime' strain (Table 1). Contigs with moderate sequence identity (slightly above 50%) to

336 bifunctional nerolidol/linalool synthases (strawberry: Aharoni et al., 2004; snapdragon:

337 Nagegowda et al., 2008) and considerable expression in glandular trichomes were identified in

338 our transcriptome data sets (Table 3), and corresponding cDNAs were cloned from the 'Valley

339 Fire' (Cs*TPS18VF*) and 'Black Lime' (Cs*TPS19BL*) strains. These sequences belong to the TPS-g

340 clade of terpene synthases (Fig. 6 and Supplemental Table S4). Heterologous expression and

341 functional characterization confirmed that the corresponding recombinant proteins were able

342 to catalyze the formation of (*E*)-nerolidol from tFPP and linalool from GPP, but no activity was

343 detected with NPP or cFPP (Fig. 8A, B). Interestingly, follow-up chiral separation of products

344 from assays performed with GPP as substrate indicated that CsTPS18VF generated almost

345 exclusively (+)-linalool, whereas CsTPS19BL produced a mixture of (-)-linalool and (+)-linalool

20

Figure showing GC-MS chromatograms and mass spectra.

Panel A: CsTPS14CT, Substrate: GPP. Peaks labeled (-)-α-Pinene, β-Myrcene, (-)-β-Pinene, (-)-Limonene, (+)-Limonene, Terpinolene. Assay peak and (-)-Limonene standard mass spectra.

| Product | Rt (1) | Rt (2) | Percentage |
|---|---|---|---|
| (-)-Limonene | 9.57 | 21.39 | 79 |
| (-)-β-Pinene | 7.96 | 20.39 | 9 |
| (-)-α-Pinene | 6.78 | 16.90 | 4 |
| β-Myrcene | 8.20 | 19.90 | 4 |
| (+)-limonene | 9.57 | 21.70 | 3 |
| α-Terpinolene | 11.58 | 24.40 | 1 |

**Additional substrates**
NPP, cFPP, tFPP – no measurable activity

Panel B: CsTPS15CT, Substrate: GPP. Peak labeled β-Myrcene. Assay peak and β-Myrcene standard mass spectra.

| Product | Rt (1) | Rt (2) | Percentage |
|---|---|---|---|
| β-Myrcene | 8.20 | 19.90 | 100 |

**Additional substrates**
NPP, cFPP, tFPP – no measurable activity

Panel C: CsTPS18VF, Substrate: GPP. Peak labeled (+)-Linalool. Assay peak and (+)-Linalool standard mass spectra.

| Product | Rt (1) | Rt (2) | Percentage |
|---|---|---|---|
| (+)-Linalool | 12.04 | 33.27 | 100 |

**Additional substrate**
NPP – no measurable activity
(see Figure 8 for data on cFPP and tFPP)

Panel D: CsTPS19BL, Substrate: GPP. Peaks labeled (-)-Linalool, (+)-Linalool. Assay peak and (-)-Linalool standard mass spectra.

| Product | Rt (1) | Rt (2) | Percentage |
|---|---|---|---|
| (-)-Linalool | 12.04 | 33.11 | 52 |
| (+)-Linalool | 12.04 | 33.27 | 48 |

**Additional substrate**
NPP – no measurable activity
(see Figure 8 for data on cFPP and tFPP)

Detector response [arbitrary units] -------- Retention time [min] -------- --Mass-to-charge ratio--

346  (Fig. 7C, D). Sequence differences across sesquiterpene synthases with different product

347  profiles included residues with potential roles in catalysis (Fig. 9), and the implications are

348  evaluated in the Discussion section.

349

21
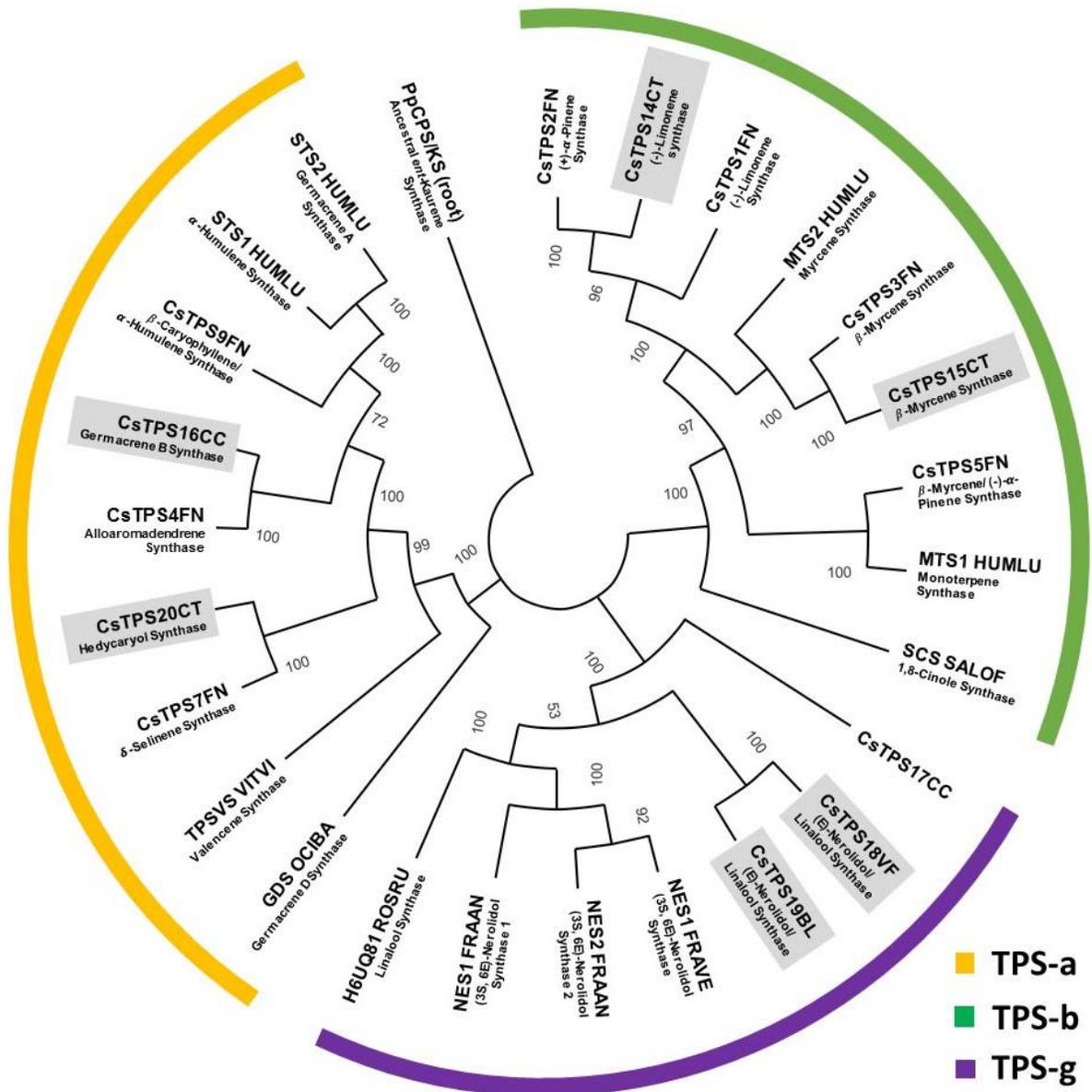
To further investigate the genetic potential for generating terpenoid chemical diversity, two representatives of the TPS-b clade of terpene synthases (CsTPS16CC and CsTPS20CT) were selected for functional characterization. Cs*TPS16CC* had very high expression levels in the
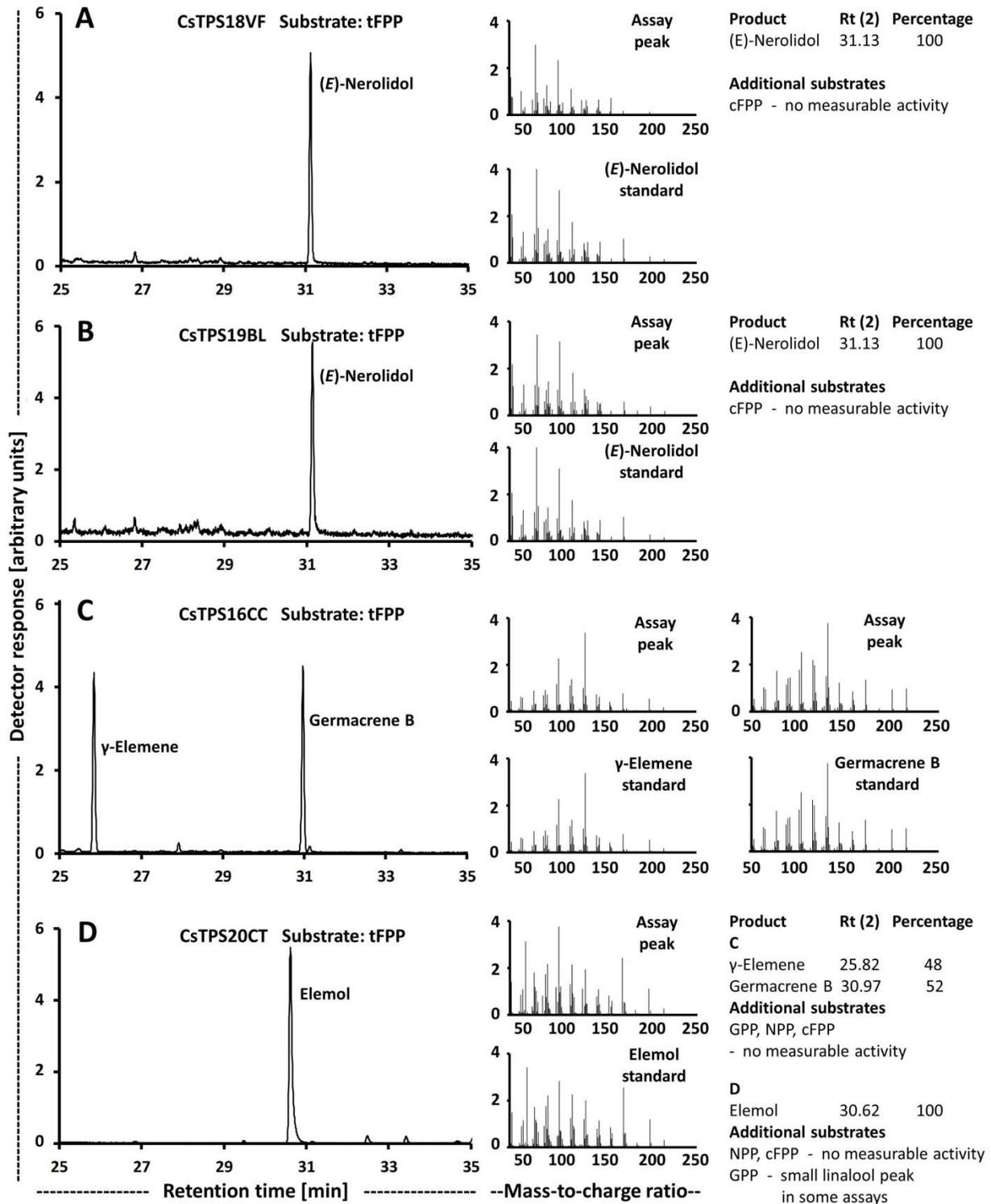
22

## A

```
CsTPS18VF    ISDAWKCLNKECILRNPAFPPPFLKASLNLARLVPLMYNYDH-NQRLPHLE--EHIKSLL 554
CsTPS19BL    ISDAWKCLNKECILRNPAFPPPFLKASLNLARLVPLMYNYDH-NQRLPHLE--EHIKSLL 554
AmNES/LIS12  ISSEWKLLNKECFNLNHVSTSSIKKAALNTARMVPLMYSYDE-NQGLPILEEYVKIMLFD 563
AmNES/LIS2   ISSEWKLLNKECFSLNHVSTSSLKKAALNTAKIVPLMYSYDE-NQRLPILEEYVKIMLFD 563
FvNES1       ISDEWKKLNRELLSPN-PFPATFTSASLNLARMIPLMYSYDG-NQSLPSLKEYMKLMLYE 573
FaNES2       ISDEWKKLNRELLSPN-PFPATITLASLNLARMIPLMYSYDG-NQCLPSLKEYMKLMLYE 571
FaNES1       ISDEWKKLNRELLSPN-PFPASFTLASLNLARMIPLMYSYDG-NQCLPSLKEYMKLMLYE 513


CsTPS16CC    VVNLWKEINQEFLR-PTSMPSSILVRILNFTKVLDIIYKEGD-GYTHVGKLVKDSVAALL 565
CsTPS20CT    VDTHWKEINEDFIR-PAVVPFPILVRVLNFTKIVDLLYKEGDDQYTNVGKVLKESIAALL 545
```

## B
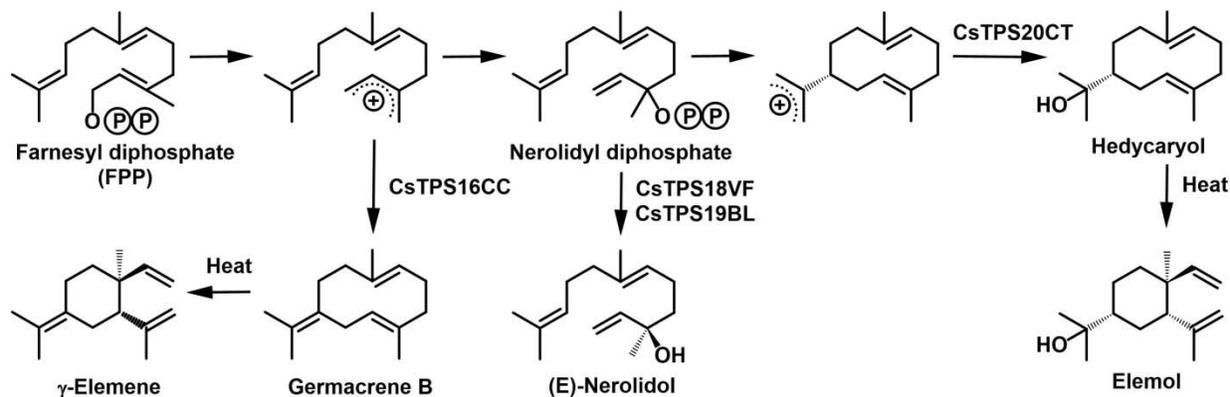


353 'Cherry Chem' strain (Table 3). The sequence was most similar to that of the previously

354 characterized alloaromadendrene synthase (Booth et al., 2017) (Fig. 6 and Supplemental Table

355 S4). In our assays, the recombinant protein generated germacrene B from tFPP (Fig. 8C), with γ-

**A** CsTPS18VF Substrate: tFPP

(*E*)-Nerolidol

Assay peak

(*E*)-Nerolidol standard

| Product | Rt (2) | Percentage |
|---|---|---|
| (E)-Nerolidol | 31.13 | 100 |

**Additional substrates**
cFPP - no measurable activity

**B** CsTPS19BL Substrate: tFPP

(*E*)-Nerolidol

Assay peak

(*E*)-Nerolidol standard

| Product | Rt (2) | Percentage |
|---|---|---|
| (E)-Nerolidol | 31.13 | 100 |

**Additional substrates**
cFPP - no measurable activity

**C** CsTPS16CC Substrate: tFPP

γ-Elemene

Germacrene B

Assay peak

γ-Elemene standard

Assay peak

Germacrene B standard

**D** CsTPS20CT Substrate: tFPP

Elemol

Assay peak

Elemol standard

| Product | Rt (2) | Percentage |
|---|---|---|
| **C** | | |
| γ-Elemene | 25.82 | 48 |
| Germacrene B | 30.97 | 52 |

**Additional substrates**
GPP, NPP, cFPP
- no measurable activity

| | | |
|---|---|---|
| **D** | | |
| Elemol | 30.62 | 100 |

**Additional substrates**
NPP, cFPP - no measurable activity
GPP - small linalool peak
in some assays

---------------- Retention time [min] ---------------- --Mass-to-charge ratio--

356    elemene being detected as a thermal breakdown product (de Kraker et al., 1998). Other prenyl

357    diphosphate substrates were not accepted as substrates with appreciable conversion rates (Fig.

358    8). The 'Canna Tsu' strain had a particularly high expression level of Cs*TPS20CT* (Table 3). Its

24

359  closest neighbor in the sequence relatedness tree was a putative δ-selinene synthase from

360  cannabis (Booth et al., 2017) (Fig. 6 and Supplemental Table S4). Functional assays with the

361  purified, recombinant protein indicated a conversion of tFPP to elemol, a thermal breakdown

362  product of the sesquiterpene hedycaryol (Koo and Gang, 2012; Hattan et al., 2016), but there

363  was little or no activity with other prenyl diphosphate substrates (Fig. 8D). In summary, we

364  demonstrate that the resources and approaches described here can be employed to identify

365  candidates and subsequently characterize functions of terpene synthase genes that belong to

366  three different clades, thereby contributing to a better understanding of the genetic

367  determinants of terpenoid chemical diversity in cannabis.

368

369

**DISCUSSION**

**Utility of Transcript Profiling for Strain Differentiation**

Competition in decriminalized retail markets for cannabis has put pressure on breeders to differentiate their product from that of their competitors. This has led to branding with a plethora of distinct and memorable names, which has caused both confusion and controversy (Small, 2015). Chemical profiling can be employed as a powerful tool in strain differentiation but adding genotyping information further increases the resolution of the analysis. The differentiation of drug-type and fiber-type cannabis strains can be achieved with standard genotyping analyses (Piluzza et al., 2013). However, a differentiation of genetically related strains has been much more challenging (Sawler et al., 2015; Punja et al., 2017). Traditional genotyping approaches benefit significantly from high-quality reference genome sequences (Scheben et al., 2017) but, unfortunately, only fairly low-quality genome sequences have been published for two cannabis strains (van Bakel et al., 2011). We employed RNA-seq as an alternative approach for genotyping (Haseneyer et al., 2011), which does not depend on prior sequence data (Wang et al., 2009). We used RNA-seq to obtain the transcriptome of glandular trichome cells of nine selected cannabis strains (with three biological replicates each). Importantly, statistical analyses of these data sets allowed the differentiation of strains into broader clades (descendants of landraces of *C. sativa* or *C. indica*), but also resulted in the full separation of all individual strains (with biological replicates clustering closely together) (Fig. 3). We fully recognize that RNA-seq is not a viable option for routine genotyping but it can be used to develop Single Nucleotide Polymorphism (SNP)-based genotyping platforms. This approach has been employed successfully for a number of crops, including alfalfa (*Medicago sativa*; Yang et al., 2011), maize (*Zea mays*; Hansey et al., 2012), and wheat (*Triticum aestivum*; Ramirez-Gonzalez et al., 2015). Our data sets are therefore highly valuable for building resources for follow-up research with cannabis. As an added benefit, RNA-seq data can be used for gene expression analysis, thereby providing a functional context, which is discussed in more detail below.

26

399

**Utility of Metabolite Profiling for Strain Differentiation**

401

402 We assessed the utility of cannabinoid and terpenoid profiling, in addition to strain

403 differentiation by genotyping as discussed above, to demarcate nine commercial cannabis

404 strains. Two independent statistical approaches, PCA and OPLS-DA, grouped biological

405 replicates closely together, while still separating individual strains and classes of strains (those

406 of *C. sativa* or *C. indica* heritage) (Fig. 4). Several authors have advocated the profiling of both

407 cannabinoids and terpenoids in recent publications (Fischedick et al., 2010; Elzinga et al., 2015;

408 Aizpurua-Olaizola et al., 2016; Hazekamp et al., 2016; Fischedick, 2017; Lewis et al., 2018; Orser

409 et al., 2018; Richins et al., 2018; Sexton et al., 2018). The key advantage of this approach over

410 merely profiling cannabinoids lies in the enormous diversity of terpenoids accumulated in

411 cannabis (and in other plants as well), which significantly increases the power of statistical

412 analyses. It also reflects the fact that many users select cannabis strains based on both the

413 reported THC content and aroma (which is largely imparted by terpenoids) (Gilbert and DiVerdi,

414 2018). A comprehensive analysis of cannabis strains recently indicated the presence of close to

415 200 detectable volatiles, which were tentatively identified based on searches against various

416 spectral databases (Rice and Koziel, 2015). A notable challenge with terpenoid profiling pertains

417 to the limitation that authentic standards are often very costly or unavailable from commercial

418 sources, which is particularly true for sesquiterpenes (dozens detected by Rice and Koziel,

419 2015). Commercial cannabis testing laboratories therefore rarely offer services that comprise

420 more than 20 terpenoids. While such analyses may detect the most abundant terpenoids for

421 popular strains, it is not unlikely that important aroma volatiles with a low odor detection

422 threshold could be missed (Chin and Marriott, 2015). Another reason why a comprehensive

423 profiling of terpenoids would be desirable relates to testing the validity of the "entourage

424 effect", the proposed synergism between cannabinoids and other constituents (in particular

425 terpenoids) that might affect the experience of the user (Gertsch et al., 2010; Russo, 2011).

426 Should such effects be substantiated by empirical evidence, it would be advisable to reconsider

427 the current laws and rules for formulations containing cannabis extracts, which are based solely

27

428    on THC. An improved understanding of terpenoid phytochemistry in cannabis would be an

429    important first step in this direction (Booth and Bohlmann, 2019).

430

431    **Co-Regulation of Metabolic Pathways in Cannabis is Consistent with Gene Expression**

432    **Patterns Commonly Observed in Glandular Trichomes**

433

434    Our statistical analyses using the WGCNA package indicated a tight correlation of biosynthetic

435    genes with cannabinoid and terpenoid end products (Fig. 5). We recently performed a meta-

436    analysis of gene expression patterns in glandular trichomes across various species (Zager and

437    Lange, 2018). One of the conclusions, consistent with the data presented here, was that gene

438    expression patterns correlate well with the metabolic specialization in these anatomical

439    structures. Co-regulation has been observed for genes across multiple pathways of specialized

440    metabolism, such as cannabinoids and terpenoids (this study), monoterpenes and diterpenes

441    (*Salvia pomifera*; Trikka et al., 2015), flavonoids and acyl sugars (*Salpiglossis sinuata* and

442    *Solanum quitoense*; Moghe et al., 2017), and bitter acids and prenylflavonoids (*Humulus

443    lupulus*; Kavalier et al., 2011; Clark et al. 2013). These tight gene-to-metabolite correlations

444    were also reflective of predicted fluxes through the relevant pathways (Zager and Lange, 2018).

445    In contrast, gene expression patterns appear to be less predictive of fluxes through central

446    carbon metabolism, where regulation at the protein level plays a more significant role (Paul and

447    Pellny, 2003; Koch, 2004; Gibon et al., 2006; Rocca et al., 2014; Schwender et al., 2014). This

448    does not mean that feedback regulation of specialized metabolism is negligible in glandular

449    trichomes; there is just a particularly strong overall gene-to-metabolite correlation, and

450    unraveling the details will be an exciting topic for future research.

451

452    **Functional Characterization of Terpene Synthases Contributes to an Improved Understanding**

453    **the Genetic Determinants of Terpenoid Diversity**

454

455    The observed gene-to-metabolite correlations in cannabis glandular trichomes provided

456    opportunities for gene discovery efforts.  Booth et al. (2017) analyzed transcriptome data sets

457    obtained with the 'Finola' and 'Purple Kush' strains to obtain candidate genes for terpene

28

458   synthases that were subsequently characterized to encode enzymes for the production of 14

459   mono- and sesquiterpenes. Those that contribute to the formation of some of the common

460   mono- and sesquiterpenes (e.g., β-myrcene, (-)-limonene, α-pinene, β-caryophyllene, and α-

461   humulene) were found to be expressed at fairly high levels across the strains included in the

462   present analysis, whereas those that generate less common products (e.g., (Z)-β-ocimene, γ-

463   eudesmol, alloaromadendrene, δ-selinene, and valencene) were found to be expressed only in

464   a limited number of strains or not at all (Table 3). To assess sequence variation among these

465   genes, we cloned genes with high sequence identity to the previously characterized β-myrcene

466   and (-)-limonene synthases.

467

468   Prior to the present study, a notable gap existed with regard to the terpene synthases

469   underlying the formation of the monoterpene linalool and the sesquiterpene nerolidol, which

470   are both common constituents in cannabis resin. We now identified a gene coding for an

471   enzyme (CsTPS19BL) that generates a mixture of (+)-linalool and (-)-linalool from GPP and (E)-

472   nerolidol from tFPP in the 'Black Lime' strain. We also cloned a putative ortholog from the

473   'Valley Fire' strain to evaluate the effects of sequence variation. Interestingly, the encoded

474   enzyme (CsTPS18VF) had the same specificity as CsTPS19BL with regard to the tFPP substrate

475   ((E)-nerolidol as product); however, with GPP as substrate, (+)-linalool was detected as the

476   essentially exclusive product. This difference in specificity is surprising given that the peptide

477   sequences have only three mismatches (Supplemental Fig. S3).

478

479   Finally, we cloned genes that, based on sequence relatedness, were expected to code for

480   enzymes that generate sesquiterpene products not previously detected in assays with cannabis

481   terpene synthases. Indeed, CsTPS16CC was demonstrated to produce germacrene B and

482   CsTPS20CT formed hedycaryol as primary product. In assays with CsTPS16CC, γ-elemene was

483   also detected, but this is a well-known product of thermal degradation in the GC inlet (de

484   Kraker et al., 1998). Elemol was the sole product of assays with CsTPS20 CT, which is also a

485   thermal degradation product, in this case of hedycaryol (Koo and Gang, 2012; Hattan et al.,

486   2016). Consequently, the enzyme activities are referred to as germacrene B synthase and

29

487    hedycaryol synthase, respectively. To the best of our knowledge, the sesquiterpenes generated

488    by these terpene synthases (germacrene B and hedycaryol) have not been identified in

489    cannabis samples yet, indicating the need for a more comprehensive coverage of terpenoids to

490    better understand strain-specific aroma profiles. It should also be noted that several recent

491    studies reporting on comprehensive chemical and sensory analyses of volatiles emitted from

492    cannabis found that non-terpenoid alcohols and aldehydes have potent odor impacts (Rice and

493    Koziel, 2015; Wiebelhaus et al., 2016; Calvi et al., 2018). These considerations indicate that

494    more emphasis needs to be placed on comprehensive metabolite profiling, including

495    cannabinoids and terpenoids but also extending to other volatiles, for future efforts focused on

496    strain characterization.

497

498    With a larger number of functionally characterized genes in cannabis, sequence

499    comparisons are now allowing us to ask questions about some of the determinants of

500    specificity. The overall sequence identity of the sesquiterpene synthases characterized here is

501    fairly low (< 70% at the amino acid level) but there are striking differences in the nature of a

502    conserved aromatic residue (Y527) that had previously been hypothesized to stabilize the

503    positive charge of the carbocation occurring during the formation of a germacrene

504    intermediate in the *epi*-aristolochene synthase catalytic sequence (Starks et al., 1997). The

505    equivalent residues in sesquiterpene synthases that catalyze the formation of cyclic products

506    (CsTPS16CC and CsTPS20CT) are also tyrosines (Fig. 9). In contrast, glutamine residues occupy

507    this position in CsTPS18VF, CsTPS19BL and other characterized enzymes of the TPS-g clade

508    (Aharoni et al., 2004; Nagegowda et al., 2008) (Fig. 9A) which, possibly because of insufficient

509    carbocation stabilization, generate (E)-nerolidol as a non-cyclic product (Fig. 9). Testing this

510    hypothesis will be an important future goal for follow-up research.

511

512    **MATERIALS AND METHODS**

513

514    **Plant Materials and Chemicals**

515

30

516    Clonal plant cuttings of nine strains ('Sour Diesel', 'Canna Tsu', 'Black Lime', 'Valley Fire', 'White

517    Cookies', 'Mama Thai', 'Terple', 'Cherry Chem', and 'Blackberry Kush') were placed in 250 l pots

518    and grown in hoop-style, light-deprivation greenhouses (Shadowbox Farms, Williams, OR, USA)

519    under a 18-h light/6-h dark regime (natural light) to stimulate vegetative growth, before shifting

520    to a 12-h light/12-h dark cycle to induce flowering. The length of these time periods varied from

521    strain to strain and was adjusted based on phenotypic evaluations. All aspects of plant growth,

522    harvest and transport were performed in accordance with the laws and rules under Chapter

523    475B, as released by the Oregon Liquor Control Commission. Plant harvest was performed

524    when the consistency of glandular trichome content had changed from a turbid white to clear,

525    and before another change to an amber-like color occurred. For most strains the pistils had

526    changed color from white to a yellow or orange. Buds were harvested, parts with low glandular

527    trichome content removed using scissors, and the remainder placed on ice until further

528    processing (always within 3 h). Monoterpene and sesquiterpene reference standards were

529    purchased from Restek (Bellefort, PA, USA). Cannabinoid reference standards were obtained

530    from Sigma-Aldrich (St. Louis, MO, USA). Solvents for extraction were procured from Sigma-

531    Aldrich (St. Louis, MO, USA). Solvents and chemicals for chromatography were sourced from

532    Burdick & Jackson (Morris, Plains, NJ, USA).  Substrates for enzyme assays (GPP, NPP and $E,E$-

533    FPP) were prepared synthetically (Davisson et al., 1986) or obtained from a commercial source

534    ($Z,Z$-FPP, Echelon Biosciences, Salt Lake City, UT, USA). The sources of standards for enzyme

535    assays were as follows: germacrene B, isolated as a side product from assays with germacrene C

536    synthase (Colby et al., 1998)); γ-elemene, obtained by heating germacrene B under argon (de

537    Kraker et al., 1998); elemol, institutional chemical repository (originally purchased from

538    Parchem, New Rochelle, NY, USA); hedycaryol, institutional chemical repository (source

539    unknown); (S)-(+)-linalool, isolated from coriander oil; (-)-limonene, (+)-limonene, ($R$)-(-)-

540    linalool, β-myrcene, (E)-nerolidol, (-)-α-pinene, (-)-β-pinene, α-terpinolene, all purchased from

541    Sigma-Aldrich, St. Louis, MO, USA).

542

543    **Metabolite Extraction and Analysis**

544

31

545    Cannabinoids and terpenoids were extracted and quantified according to Fischedick et al.

546    (2010), with modifications, at a testing facility with accreditation by ISO/IEC 17025 and licensed

547    through the National Environmental Laboratory Accreditation Program (Evio Labs, Central

548    Point, OR, USA). Briefly, roughly 2.0 g of fresh bud tissue was crushed in a falcon tube,

549    suspended in 10 ml methyl tert-butyl ether (containing 1-octanol as internal standard) with

550    gentle shaking for 15 min, followed by centrifugation at 2,000 x $g$ for 5 min. The supernatant

551    was transferred to a new vial and the plant material extracted two more times as above (no

552    addition of internal standard to solvent). The combined supernatants were filtered through a

553    polytetrafluoroethylene syringe filter (0.45 µm pore size, 25 mm diameter) and an aliquot

554    transferred to a screw-cap glass vial, which was stored at -20°C until further analysis. Following

555    extraction, the remaining plant material was dried in an oven (50°C) and weighed to determine

556    dry weights for each sample.

557

558    Cannabinoids were separated via high performance liquid chromatography (model LC-2030C,

559    Shimadzu, Columbia, MD, USA) using a Kinetex C18 reversed phase column (50 x 4.6 mm, 2.6

560    µm particle size; Phenomenex, Torrance, CA) and a binary gradient of solvent A (water

561    containing 0.1% (v/v) formic acid and 10 mM ammonium formate) and solvent B (methanol

562    containing 0.05% (v/v) formic acid) with the following settings: 0–9 min, 68–78% B; 9–11.9 min,

563    78–100% B; 11.9–13.5 min, hold at 100% B. Analytes were monitored at 228 nm in a diode

564    array detector. Peak identification was achieved based on comparisons of retention times and

565    spectral characteristics with those of authentic cannabinoid reference standards. Analytes were

566    quantified based on calibration curves acquired with authentic standards. The validation of the

567    analytical method was performed according to Fischedick et al. (2010).

568

569    Terpenoids were separated via gas chromatography (model 6890, Agilent Technologies, Santa

570    Clara, CA, USA) using a DB5 column (30 m x 25 mm, 25 µm film thickness; Agilent Technologies,

571    Santa Clara, CA, USA) and detected with a flame ionization detector. The conditions for

572    separation were as follows: injector at 250°C, 20 : 1 split injection mode (1 µL injected);

573    detector at 250°C ($H_2$ flow at 30 ml/min, airflow at 400 ml/min, makeup flow (He) at 25

574 ml/min); oven heating from 40°C to 120°C at 2°C/min, then ramped to 200°C at 50°C/min, with

575 a final hold at 200°C for 2 min. GC peaks were identified based on comparisons of retention

576 times of authentic standards (purchased from Sigma-Aldrich, St. Louis, MO, USA). Analytes were

577 quantified based on calibration curves acquired with authentic standards. The validation of the

578 analytical method was performed according to Fischedick et al. (2010).

579

580 **RNA Isolation from Glandular Trichomes and cDNA Library Preparation**

581

582 Secretory cells of glandular trichomes were removed from 10–15 g of bud tissue by surface

583 abrasion and then collected by filtering through a series of nylon meshes (Lange et al., 2000).

584 Total RNA was isolated from secretory cells using the RNeasy Plant kit (Qiagen, Germantown,

585 MD, USA) according to the manufacturer's instructions. RNA integrity was determined using a

586 BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). cDNA libraries from 1–2 µg of

587 total RNA were generated using the SuperScript III Reverse Transcriptase kit (Invitrogen,

588 Carlsbad, CA, USA) according to the manufacturer's instructions.

589

590 **RNA Sequencing and Transcriptome Assembly**

591

592 RNA sequencing libraries were prepared from 250 ng total glandular trichome RNA with the

593 Stranded mRNA-Seq Poly(A) Selection kit (KAPA Biosystems, Wilmington, MA, USA). The quality

594 and quantity of the sequencing library was assessed using a Bioanalyzer 2100 and a Qubit 3.0

595 Fluorometer (Agilent Technologies; Life Technologies, Carlsbad, CA, USA). Sequencing of 150-bp

596 paired end reads was performed on a HiSeq 4000 instrument (Illumina, San Diego, CA, USA).

597 Sequenced reads were trimmed of adapter sequences with Trimmomatic (Bolger et al., 2014)

598 and sequence quality was checked with FastQC (Andrews, 2010). Trimmed sequences were

599 merged and assembled using the Trinity *de novo* assembler and downstream functional

600 annotation of the assembly was performed with Trinotate (Haas et al., 2013). The resulting

601 transcriptome assembly contained 514,208 contigs, with a mean contig length of 875 bp and an

602 N50 value of 1529 bp.  Transcript abundance in each RNA-seq data set (3 biological replicates

603 per strain) was determined with RSEM (Li and Dewey, 2011).

33

604

**Analysis of Global Gene Expression Patterns and Gene Ontology Enrichment**

606

607 Testing for differential gene expression across strains was performed using the Bioconductor

608 package DESeq2 (version 1.18.1) (Love et al., 2014). *P*-values were adjusted using the

609 Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). An adjusted *P*-value (false

610 discovery rate) ≤ 1.0e-10 and $\log_2$ ratio ≥ 3.0 were set as thresholds. A Cluster analysis of gene

611 expression patterns between strains was performed within the Trinity suite (Haas et al., 2013)

612 by partitioning genes into clusters by cutting the hierarchically clustered gene tree at 60%

613 height of the tree.  A Gene Ontology (GO) enrichment analysis of differentially expressed genes

614 was performed using the GOseq package in *R* (Young et al., 2010). GO terms with an adjusted *P*-

615 value < 0.01 were considered significantly enriched.

616

**Gene Co-Expression Network Analysis**

618

619 A gene co-expression network was built using the WGCNA package in *R* (Langfelder and

620 Horvath, 2008). Transcriptome datasets were filtered to remove genes with an average

621 expression value of 50 TPM or smaller. Co-expression modules were identified using the

622 function blockwiseModules with the following settings: power at 7, mergeCutHeight at 0.55,

623 and minModuleSize at 30. Eigengene values were determined for each co-expression module to

624 test for association significance. Modules with similar eigengene values were merged to obtain

625 the final co-expression modules.

626

**Phylogenetic Analysis of TPS Candidates**

628

629 The identification of TPS candidate genes was accomplished by searching the translated

630 transcriptome consensus assembly against a manually curated protein database specific to

631 characterized plant TPSs using the Blastx algorithm. A reciprocal search (tBlastn) was performed

632 with sequences of 114 characterized angiosperm TPSs against the assembly for each individual

633 strain. Predicted TPS sequences were then analyzed for gene expression values across strains.

634 Translated amino acid sequences of these and reference TPSs (from *Cannabis sativa* and

635 *Humulus lupulus*) were aligned using the MUSCLE algorithm. Alignments were analyzed with

636 maximum likelihood analysis using a Jones-Taylor-Thorton model with Gamma distribution for

637 rates among amino acid sites. One thousand bootstrap replicates were then used to construct a

638 phylogeny using MEGA7 (Jones et al., 1992; Kumar et al., 2015).

639

640 **Cloning of TPS cDNAs**

641

642 First-strand cDNA was prepared from RNA with the SuperScript III First Strand Synthesis kit

643 (Invitrogen) with random hexamer oligonucleotides. Open reading frames for TPSs were

644 amplified using gene-specific primers (Supplemental Table S5) (amplicons for full-length cDNAs

645 were generated for putative sesquiterpene synthases, whereas cDNAs devoid of the plastidial

646 targeting sequence were amplified for putative monoterpene synthases). Amplicons were

647 ligated into the pGEM-T Easy vector (Promega, Fitchburg, WI, USA) and sequence-verified. For

648 expression in *E. coli*, full-length or truncated genes were subcloned into the pSBET expression

649 vector (predigested with *Nde*I and *Bam*HI) (Steinbiss et al., 1995). Several terpene synthase

650 cDNAs (Cs*TPS18VF*, Cs*TPS19B*L and Cs*TPS20CT*) were purchased as synthetic products (in the

651 pET28B expression vector) from GenScript (Piscataway, NJ, USA).

652

653 *In Vitro* **Functional Assays for Recombinant TPSs**

654

655 Plasmids were transformed into chemically competent cells of several *E. coli* strains (BL21

656 (DE3), C41 (DE3), C43 (DE3), C43 (DE3) pLysS and ArcticExpress (DE3)), which were then grown

657 in 25 ml of liquid LB medium at 37°C with shaking to an $OD_{600}$ of 0.8. Expression of TPS genes

658 was induced with 0.1 or 0.5 mM Isopropyl β-D-1-thiogalactopyranoside (Goldbio, St. Louis, MO,

659 USA) and cells grown for another 24 h at three different temperatures (16°C, 10 °C and 4 °C).

660 Bacterial cells were harvested by centrifugation at 5,000 x *g* and resuspended in 300 ul MOPSO

661 buffer, pH 7.0, supplemented with 1 mM dithiothreitol (DTT; Goldbio, St. Louis, MO, USA). Cells

35

662  were lysed using a model 475 sonicator (VirTis, Gardiner, NY, USA), with three 15 s bursts and

663  cooling on ice between bursts. The resulting homogenate was centrifuged at 15,000 x $g$ for 30

664  min at 4°C, and the clear supernatant mixed with ceramic hydroxyapatite (Biorad, Hercules, CA,

665  USA). The purification of recombinant protein was performed as described in Srividya et al.

666  (2016) for constructs in the pSBET expression vector, whereas those in the pET28B expression

667  vector were purified over $Ni^{2+}$ affinity columns according to the manufacturer's instructions

668  (Novagen-EMD Millipore, Burlington, MA, USA). *In vitro* assays were performed in 2-ml glass

669  vials containing 200 µg purified enzyme in MOPSO buffer containing DTT and $MgCl_2$ (total

670  volume 100 µl). A prenyl diphosphate substrate (GPP, NPP, tFPP or cFPP) was added to a final

671  concentration of 0.5 mM. The assay mixtures were overlaid with 100 µl n-hexane (Avantor,

672  Center Valley, PA, USA) and incubated at 30°C for 16 h on a multi-tube rotator (Labquake,

673  Barnstead Thermolyne, Ramsey, MN, USA). The enzymatic reaction was stopped by vigorous

674  mixing of the contents of the tubes, followed by 30 min at -80°C for phase separation. The

675  organic phase was removed and transferred to glass vial inserts and stored in GC vials at -20°C

676  until further analysis.

677

678      Enzymatically formed products were analyzed on a 6890N gas chromatograph coupled to a

679  5973 mass selective detector (Agilent, Santa Clara, CA, USA). Analyte separation was achieved

680  under the conditions developed by Adams (2007), which includes a comprehensive resource for

681  spectral comparisons of volatiles. The chiral separation of monoterpenes was achieved as

682  described in Turner et al. (2019). Enzymatically generated products were identified based on

683  retention times and mass spectral properties when compared to those of authentic standards.

684

685  **Statistical Analyses**

686

687  For metabolite analyses, statistical analyses were performed in *R* using the MetaboAnalystR package

688  (Chong and Xia, 2018). Quantitative terpenoid and cannabinoid data were scaled by dividing mean

689  centered values by the standard deviation of each variable to generate principal component (PC)

690  loadings. Principal components were then plotted in three dimensions within the *R* environment. OPLS-

691  DA analysis was also performed in the same way using the MetaboAnalystR package. Differential gene

36

692  expression patterns were assessed using the Bioconductor package DESeq2 (version 1.18.1) (Love et al.,

693  2014), with the P-value for the Benjamini-Hochberg false discovery threshold being adjusted to ≤ 1.0e-

694  10 and the $\log_2$ fold-change ratio to ≥ 3.0. Cluster analysis of differential gene expression was performed

695  within the Trinity suite (Haas et al., 2013) by cutting the clustered gene tree at 60% tree height and

696  differentially expressed genes subjected to further analysis within GOSeq as described above (Young et

697  al., 2010). TPS candidates were identified based on sequence identity with functionally characterized

698  TPSs in tBLASTn searches. Candidates with e-values > 0.001 and bitscores < 250 were removed from

699  further consideration.

700

701  **Accession Numbers**

702

703  The raw transcriptome sequence data for cannabis strains are available at the NCBI Sequence

704  Read Archive, project number PRJNA498707. Nucleotide sequences for genes characterized as

705  part of this study were deposited in GenBank and received the accession numbers MK131289

706  (Cs*TPS16CC*), MK801762 (Cs*TPS20CT*), MK801763 (Cs*TPS19BL*), MK801764 (Cs*TPS18VF*),

707  MK801765 (Cs*TPS15CT*), and MK801766 (Cs*TPS14CT*).

708

## Supplemental Data

**Supplemental Figure S1.** Alignment of translated peptide sequences, based on RNA-seq data, of tetrahydrocannabinolic acid synthase across cannabis strains.

**Supplemental Figure S2.** Nucleotide and translated peptide sequence, based on RNA-seq data, of cannabidiolic acid synthase from the cannabis strain 'Canna Tsu'.

**Supplemental Figure S3.** Alignment of terpene synthase sequences.

**Supplemental Table S1.** Statistics of *de novo* assemblies performed based on cannabis glandular trichome-specific RNA-seq data sets.

37

**Supplemental Table S2.** Annotation of transcripts represented in cannabis glandular trichome-specific RNA-seq data sets.

**Supplemental Table S3.** Clustering of genes into coexpression modules obtained by Weighted Gene Correlation Network Analysis of cannabis glandular trichome-specific RNA-seq data sets.

**Supplemental Table S4.** Accession numbers and sequences of terpene synthases considered for phylogenetic analysis.

**Supplemental Table S5.** Primers used to clone cannabis cDNAs for functional characterization.

709

710 **ACKNOWLEDGEMENTS**

711 This study was supported by gifts from private individuals and we are grateful for their
712 generosity. We would also like to thank Shadowbox Farms for allowing A.S. to harvest plant
713 materials.

714

715 **TABLES**

716

38

**FIGURE LEGENDS**

718

719 **Figure 1**. Shared origin of the cannabinoid and terpenoid biosynthetic pathways. A circled "P"
720 denotes phosphate moieties.

721

722 **Figure 2.** Characteristics of cannabis strains. A, Floral phenotype. B, Origin and aroma
723 description (according to https://www.leafly.com).

724

725 **Figure 3.** Cannabis strain differentiation based on glandular trichome-specific RNA-seq data. A,
726 Three-dimensional plot representing outcomes of a Principal Component Analysis. B, Heatmap
727 of a two-way Hierarchical Clustering Analysis. The numerical values and red-white-blue color
728 code indicate the $\log_2$ fold-change compared to the average gene expression value across all
729 strains. Abbreviations at the bottom of panel B: BB, 'Black Berry Kush'; BL; 'Black Lime'; CC,
730 'Cherry Chem'; CT, 'Canna Tsu'; MT, 'Mama Thai'; SD, 'Sour Diesel'; T, 'Terple'; VF, 'Valley Fire';
731 WC, 'White Cookies'.

732

733 **Figure 4.** Cannabis strain differentiation based on cannabinoid and terpenoid profiles. A, Three-
734 dimensional plot representing outcomes of a Principal Component Analysis. B, Two-dimensional
735 plot of the outcomes of an Orthogonal Projections to Latent Structures Discriminant Analysis.

736

737 **Figure 5.** Co-expression of genes involved in cannabinoid and terpenoid biosynthesis. A,
738 Weighted Gene Correlation Network Analysis (WGCNA) of glandular trichome-specific RNA-seq
739 data categorizes transcripts into eight color-coded modules (for gene lists see Supplemental
740 Table S3). B, Correlation of WGCAN modules with metabolites. A color code is used to visualize
741 the Spearman Correlation Coefficients (SCCs) for each module-metabolite pair, with red color
742 representing positive and blue color indicating negative SCCs. C, Genes involved in cannabinoid
743 and terpenoid biosynthesis are enriched in the yellow co-expression module obtained by
744 WGCNA. Color code for pathways: light blue, hexanoate formation; dark green, precursors for
745 monoterpenes; light green, monoterpene synthases; orange, sesquiterpenes; dark blue,

746 cannabinoids; cyan, remaining genes. D, Functional context of genes highlighted in (C) in
747 simplified metabolic pathway scheme. Abbreviations: *ACC1*, acetyl-CoA carboxylase; Ac-CoA,
748 acetyl coenzyme A; *AAE1*, acyl activating enzyme for short-chain fatty acids;
749 Cs*TPS1FN*/Cs*TPS14C*T, (-)-limonene synthase; Cs*TPS2SK,* (+)-α-pinene synthase;
750 Cs*TPS3FN*/Cs*TPS15CT*, β-myrcene synthase; Cs*TPS16CC*, germacrene B synthase; *PT1*,
751 cannabigerolic acid synthase; DHAP, dihydroxyacetone phosphate; *DXS*, 1-deoxy-D-xylulose-5-
752 phosphate synthase; *ENO*, enolase; *FNR-Root*, ferredoxin-NADP+ reductase (isoform of roots
753 and glandular trichomes); *FPPS*, farnesyl diphosphate synthase; GAP, glyceraldehyde-3-
754 phosphate; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; GPP, geranyl diphosphate;
755 GPPS, geranyl diphosphate synthase; *KR*, β-ketoacyl reductase (fatty acid synthase complex);
756 OA, olivetolic acid; *PDH*, pyruvate dehydrogenase; *PFK*, phosphofructokinase; *PGI*,
757 phosphoglucoisomerase; *PGM*, phosphoglucomutase; *PK*, pyruvate kinase;  Pyr, pyruvate;
758 THCA, tetrahydrocannabinolic acid; *THCAS*, tetrahydrocannabinolic acid synthase; and *TPI*,
759 triose phosphate isomerase.

760

761 **Figure 6.** Maximum likelihood phylogenetic tree of selected, functionally characterized terpene
762 synthases. The tree is rooted with the ancestral *ent*-kaurene synthase of *Physcomitrella patens*
763 (Pp*CPS/KS*). A color code is used to indicate different clades (yellow, TPS-a; green, TPS-b; and
764 purple, TPS-g). Abbreviations: BL, 'Black Lime'; CC, 'Cherry Chem'; CT, 'Canna Tsu'; Cs, *Cannabis*
765 *sativa*; FN, 'Finola'; FRAAN, *Fragaria* x *ananassa*; FRAVE, *Fragaria vesca*; HUMLU, *Humulus*
766 *lupulus*; OCIBA, *Ocimum basilicum*; ROSRU, *Rosa rugosa*; SALOF, *Salvia officinalis*; and VF,
767 'Valley Fire'; and VITVI, *Vitis vinifera.* The accession numbers and sequences of the terpene
768 synthases are provided in Supplemental Table S4.

769

770 **Figure 7.** Functional characterization of cannabis terpene synthases that act on GPP as
771 substrate. A to D, Left panel: chiral GC chromatogram. Center panel: mass spectra of primary
772 products. Right panel: product distribution. A, (-)-limonene synthase (CsTPS14CT). B, β-myrcene
773 synthase (CsTPS15CT). C, (E)-nerolidol/(+)-linalool synthase (CsTPS18VF). D, (E)-nerolidol/(+)-
774 linalool synthase (CsTPS19BL).

40

775

**Figure 8.** Functional characterization of cannabis terpene synthases that act on tFPP as substrate. A to D, Left panel, GC-MS chromatogram. Center panel, mass spectra of primary products. Right panel, product distribution. A, (E)-nerolidol/(+)-linalool synthase (CsTPS18VF). B, (E)-nerolidol/(+)-linalool synthase (CsTPS19BL). C, germacrene B synthase (CsTPS16CC). D, hedycaryol synthase (CsTPS20CT).

781

**Figure 9.** Variation of the residue putatively stabilizing carbocation intermediates correlates with outcome of catalysis in cannabis sesquiterpene synthases. A, Sequence alignment of sesquiterpene synthases (with carbocation-stabilizing residue highlighted). B, Proposed cyclization reactions catalyzed by sesquiterpene synthases. Identifiers for sequences from the literature (Aharoni et al., 2004; Nagegowda et al., 2008): AmNES/LIS1, EF433761; AmNES/LIS2, EF433762; FvNES1, AX529002; FaNES2, AX529067; FaNES1, KX450224 (species abbreviations: Am, *Antirrhinum majus*; Fa, *Fragaria* x *ananassa*; Fv, *Fragaria vesca*).

41

# Parsed Citations

Abuhasira R, Shbiro L, Landschaft Y (2018) Medical use of cannabis and cannabinoids containing products–Regulations in Europe and North America. Eur J Intern Med 49: 2-6
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ahkami A, Johnson SR, Srividya N, Lange BM (2015) Multiple levels of regulation determine monoterpenoid essential oil compositional variation in the mint family. Mol Plant 8: 188–191
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ahmed SA, Ross SA, Slade D, Radwan MM, Khan IA, ElSohly MA (2008a) Structure determination and absolute configuration of cannabichromanone derivatives from high potency Cannabis sativa. Tetrahedron Lett 49: 6050–6053
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ahmed SA, Ross SA, Slade D, Radwan MM, Zulfiqar F, ElSohly MA (2008b) Cannabinoid ester constituents from high-potency Cannabis sativa. J Nat Prod 71: 536–542
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Andre CM, Hausman J-F, Guerriero G (2016) Cannabis sativa: the plant of the thousand and one molecules. Front Plant Sci 7: 19
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc Series B 57: 289–300
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Booth JK, Page JE, Bohlmann J (2017) Terpene synthases from Cannabis sativa. PLoS One 12: e0173911
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Booth JK, Bohlmann J (2019) Terpenes in Cannabis sativa – from plant genome to humans. Plant Sci 284: 67-72.
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Calvi L, Pentimalli D, Panseri S, Giupponi L, Gelmini F, Beretta G, Vitali D, Bruno M, Zilio E, Pavlovic R (2018) Comprehensive quality evaluation of medical Cannabis sativa L. inflorescence and macerated oils based on HS-SPME coupled to GC–MS and LC-HRMS (q-Exactive Orbitrap®) approach. J Pharm Biomed Anal 150: 208–219
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Chin ST, Marriott PJ (2015) Review of the role and methodology of high resolution approaches in aroma analysis. Anal Chim Acta 854: 1–12
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Chong J, Xia J (2018) MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics 34: 4313–4314
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Clark SM, Vaitheeswaran V, Ambrose SJ, Purves RW, Page JE (2013) Transcriptome analysis of bitter acid biosynthesis and precursor pathways in hop (Humulus lupulus). BMC Plant Biology 13: 12
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Davisson VJ, Woodside AB, Neal TR, Stremler KE, Muehlbacher M, Poulter CD (1986) Phosphorylation of isoprenoid alcohols. J Org Chem 51: 4768–4779
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Devane W, Hanus L, Breuer A, Pertwee R, Stevenson L, Griffin G, Gibson D, Mandelbaum A, Etinger A, Mechoulam R (1992) Isolation and structure of a brain constituent that binds to the cannabinoid receptor. Science 258: 1946–1949

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Devane WA, Dysarz FA, Johnson MR, Melvin LS, Howlett AC (1988) Determination and characterization of a cannabinoid receptor in rat brain. Mol Pharmacol 34: 605–613

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

de Kraker JW, Franssen MC, de Groot A, Konig WA, Bouwmeester HJ (1998) (+)-Germacrene A biosynthesis. The committed step in the biosynthesis of bitter sesquiterpene lactones in chicory. Plant Physiol 117: 1381–1392.

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

ElSohly MA, Slade D (2005) Chemical constituents of marijuana: The complex mixture of natural cannabinoids. Life Sci 78: 539–548

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Elzinga S, Fischedick J, Podkolinski R, Raber JC (2015) Cannabinoids and terpenes as chemotaxonomic markers in cannabis. Nat Prod Chem Res 3: 2

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Fellermeier M, Zenk MH (1998) Prenylation of olivetolate by a hemp transferase yields cannabigerolic acid, the precursor of tetrahydrocannabinol. FEBS Lett 427: 283–285

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Fischedick JT (2017) Identification of terpenoid chemotypes among high (−)-trans-Δ9-tetrahydrocannabinol-producing Cannabis sativa L. cultivars. Cannabis Cannabinoid Res 2: 34–47

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Fischedick JT, Hazekamp A, Erkelens T, Choi YH, Verpoorte R (2010) Metabolic fingerprinting of Cannabis sativa L. cannabinoids and terpenoids for chemotaxonomic and drug standardization purposes. Phytochemistry 71: 2058–2073

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gagne SJ, Stout JM, Liu E, Boubakir Z, Clark SM, Page JE (2012) Identification of olivetolic acid cyclase from Cannabis sativa reveals a unique catalytic route to plant polyketides. Proc Natl Acad Sci USA 109: 12811–12816.

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gertsch J, Leonti M, Raduner S, Racz I, Chen J-Z, Xie X-Q, Altmann K-H, Karsak M, Zimmer A (2008) Beta-caryophyllene is a dietary cannabinoid. Proc Natl Acad Sci USA 105: 9099–9104

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gibon Y, Usadel B, Blaesing OE, Kamlage B, Hoehne M, Trethewey R, Stitt M (2006) Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. Genome Biol 7: R76

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gilbert AN, DiVerdi JA (2018) Consumer perceptions of strain differences in Cannabis aroma. PloS One 13: e0192247

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Grattan JHG, Singer CJ (1952) Anglo-Saxon Magic and Medicine. Oxford University Press, London, UK, 234 pages

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Guennewich N, Page JE, Köllner TG, Degenhardt J, Kutchan TM (2007) Functional expression and characterization of trichome-specific (-)-limonene synthase and (+)-α-pinene synthase from Cannabis sativa. Nat Prod Commun 2: 223–232

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Prot 8: 1494–1512

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hansey CN, Vaillancourt B, Sekhon RS, De Leon N, Kaeppler SM, Buell CR (2012) Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. PloS One 7: e33071

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Haseneyer G, Schmutzer T, Seidel M, Zhou R, Mascher M, Schön C-C, Taudien S, Scholz U, Stein N, Mayer KF (2011) From RNA-seq to large-scale genotyping-genomics resources for rye (Secale cereale L.). BMC Plant Biol 11: 131**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Hattan J, Shindo K, Ito T, Shibuya Y, Watanabe A, Tagaki C, Ohno F, Sasaki T, Ishii J, Kondo A, Misawa N (2016) Identification of a novel hedycaryol synthase gene isolated from Camellia brevistyla flowers and floral scent of Camellia cultivars. Planta 243: 959–972**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Hazekamp A, Fischedick JT (2012) Cannabis - from cultivar to chemovar: Towards a better definition of Cannabis potency. Drug Test Anal 4: 660–667**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Hazekamp A, Tejkalová K, Papadimitriou S (2016) Cannabis: from cultivar to chemovar II-a metabolomics approach to Cannabis classification. Cannabis Cannabinoid Res 1: 202–215**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

<span style="color:red">**Hemmerlin A, Harwood JL, Bach TJ (2012) A raison d'être for two distinct pathways in the early steps of plant isoprenoid biosynthesis? Prog Lipid Res 51: 95–148**</span>

**Hillig KW (2004) A chemotaxonomic analysis of terpenoid variation in Cannabis. Biochem Syst Ecol 32: 875–891**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Hillig KW, Mahlberg PG (2004) A chemotaxonomic analysis of cannabinoid variation in Cannabis (Cannabaceae). Am J Bot 91: 966–975**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8: 275–282**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Kassirer JP (1997) Federal foolishness and marijuana. N Engl J Med 336: 366–367**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Kavalier AR, Litt A, Ma C, Pitra NJ, Coles MC, Kennelly EJ, Matthews PD (2011) Phytochemical and morphological characterization of hop (Humulus lupulus L.) cones over five developmental stages using high performance liquid chromatography coupled to time-of-flight mass spectrometry, ultrahigh performance liquid chromatography photodiode array detection, and light microscopy techniques. J Agric Food Chem 59: 4783–4793**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Koch K (2004) Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. Curr Opin Plant Biol 7: 235–246**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**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**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Koo HJ, Gang DR (2012) Suites of terpene synthases explain differential terpenoid production in ginger and turmeric tissues. PlosOne 7: e51481**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33: 1870–1874**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Lange BM, Wildung MR, Stauber EJ, Sanchez C, Pouchnik D, Croteau R (2000) Probing essential oil biosynthesis and secretion by functional evaluation of expressed sequence tags from mint glandular trichomes. Proc Natl Acad Sci USA 97: 2934–2939**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Laverty KU, Stout JM, Sullivan MJ, Shah H, Gill N, Bolbrook L, Deikus G, Sebra R, Hughes TR, Page JE, van Bakel H (2018) A physical and genetic map of Cannabis sativa identifies extensive rearrangement at the THC/CBD acid synthase locus. Genome Res, in press, doi: 10.1101/gr.242594.118
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Lewis MA, Russo EB, Smith KM (2018) Pharmacological foundations of Cannabis chemovars. Planta Med 84: 225–233
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Malingre T, Hendriks H, Batterman S, Bos R, Visser J (1975) The essential oil of Cannabis sativa. Planta Med 28: 56–61
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Marks MD, Tian L, Wenger JP, Omburo SN, Soto-Fuentes W, He J, Gang DR, Weiblen GD, Dixon RA (2009) Identification of candidate genes affecting Δ9-tetrahydrocannabinol biosynthesis in Cannabis sativa. J Exp Bot 60: 3715–3726
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

de Meijer EPM, Bagatta M, Carboni A, Crucitti P, Moliterni VMC, Ranalli P, Mandolino G (2003) The inheritance of chemical phenotype in Cannabis sativa L. Genetics 163: 335–346
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Orser C, Johnson S, Speck M, Hilyard A, Afia I (2018) Terpenoid chemoprofiles distinguish drug-type Cannabis sativa L. cultivars in Nevada. Nat Prod Chem Res 6: 304
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Paul MJ, Pellny TK (2003) Carbon metabolite feedback regulation of leaf photosynthesis and development. J Exp Bot 54: 539–547
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

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 Res Crop Evol 60: 2331–2342
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Punja ZK, Rodriguez G, Chen S (2017) Assessing Genetic Diversity in Cannabis sativa Using Molecular Approaches. In S Chandra, H Lata, MA ElSohly, eds, Cannabis sativa L. - Botany and Biotechnology. Springer International Publishing, Cham, pp 395–418
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Ramirez-Gonzalez RH, Segovia V, Bird N, Fenwick P, Holdgate S, Berry S, Jack P, Caccamo M, Uauy C (2015) RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. Plant Biotechnol J 13: 613–624
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Rice S, Koziel JA (2015) Characterizing the smell of marijuana by odor impact of volatile compounds: an application of simultaneous chemical and sensory analysis. PloS One 10: e0144160
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Richins RD, Rodriguez-Uribe L, Lowe K, Ferral R, O'Connell MA (2018) Accumulation of bioactive metabolites in cultivated medical Cannabis. PloS One 13: e0201119
  Pubmed: Author and Title
  Google Scholar: Author Only Title Only Author and Title

Rocca JD, Hall EK, Lennon JT, Evans SE, Waldrop MP, Cotner JB, Nemergut DR, Graham EB, Wallenstein MD (2015) Relationships between protein-encoding gene abundance and corresponding process are commonly assumed yet rarely observed. ISME J 9: 1693

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Russo EB (2011) Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. Br J Pharmacol 163: 1344–1364**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Russo EB, Jiang H-E, Li X, Sutton A, Carboni A, del Bianco F, Mandolino G, Potter DJ, Zhao Y-X, Bera S, et al (2008) Phytochemical and genetic analyses of ancient cannabis from Central Asia. J Exp Bot 59: 4171–4182**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Sawler J, Stout JM, Gardner KM, Hudson D, Vidmar J, Butler L, Page JE, Myles S (2015) The genetic structure of marijuana and hemp. PloS One 10: e0133292**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Scheben A, Batley J, Edwards D (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. J Biotechnol 15: 149–161**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Schwender J, König C, Klapperstück M, Heinzel N, Munz E, Hebbelmann I, Hay JO, Denolf P, De Bodt S, Redestig H (2014) Transcript abundance on its own cannot be used to infer fluxes in central metabolism. Front Plant Sci 5: 668**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Sexton M, Shelton K, Haley P, West M (2018) Evaluation of cannabinoid and terpenoid content: Cannabis flower compared to supercritical CO2 concentrate. Planta Med 84: 234–241**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

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

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Small E (2015) Evolution and classification of Cannabis sativa (marijuana, hemp) in relation to human utilization. Bot Rev 81: 189–294**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Srividya N, Lange I, Lange BM (2016) Generation and functional evaluation of designer monoterpene synthases. Methods Enzymol 576: 147–165**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Starks CM, Back K, Chappell J, Noel JP (1997) Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. Science 277: 1815–1820**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Stout JM, Boubakir Z, Ambrose SJ, Purves RW, Page JE (2012) The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in Cannabis sativa trichomes. Plant J 71: 353–365**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Sun SY, Jiang WG, Zhao YP (2010) Characterization of the aroma-active compounds in five sweet cherry cultivars grown in Yantai (China). Flav Fragr J 25: 206–213**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Taura F, Sirikantaramas S, Shoyama Y, Shoyama Y, Morimoto S (2007) Phytocannabinoids in Cannabis sativa: Recent studies on biosynthetic enzymes. Chem Biodiv 4: 1649–1663**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Trikka FA, Nikolaidis A, Ignea C, Tsaballa A, Tziveleka L-A, Ioannou E, Roussis V, Stea EA, Božić D, Argiriou A (2015) Combined metabolome and transcriptome profiling provides new insights into diterpene biosynthesis in S. pomifera glandular trichomes. BMC Genomics 16: 935**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Turner GW, Parrish AN, Zager JJ, Fischedick JT, Lange BM (2019) Assessment of flux through oleoresin biosynthesis in epithelial cells of loblolly pine resin ducts. J Exp Bot 70: 217–230**

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**United Nations (1966) Commission on Narcotic Drugs. Document E/4294; Economic and Social Council: Official Records**
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Unschuld PU (1986) Medicine in China: A History of Pharmaceutics. University of California Press, Berkeley, CA, USA, 384 pages**
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Wiebelhaus N, Kreitals NM, Almirall JR (2016) Differentiation of marijuana headspace volatiles from other plants and hemp products using capillary microextraction of volatiles (CMV) coupled to gas-chromatography–mass spectrometry (GC–MS). Forensic Chem 2: 1–8**
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A (2017) A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. Plant Cell 29: 944–959**
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung H-JG, Vance CP, Gronwald JW (2011) Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. BMC Genomics 12: 199**
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11: R14**
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

**Zager JJ, Lange BM (2018) Assessing flux distribution associated with metabolic specialization of glandular trichomes. Trends Plant Sci 23: 638–647**
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title