frontiers in Plant Science

ORIGINAL RESEARCH
published: 09 May 2018
doi: 10.3389/fpls.2018.00537

Identification of Putative Precursor Genes for the Biosynthesis of Cannabinoid-Like Compound in *Radula marginata*

Tajammul Hussain[1]*, *Blue Plunkett*[2], *Mahwish Ejaz*[3], *Richard V. Espley*[2] and *Oliver Kayser*[1]*

[1] Department of Technical Biochemistry, TU Dortmund University, Dortmund, Germany, [2] The New Zealand Institute for Plant & Food Research Limited (PFR), Auckland, New Zealand, [3] Max Planck Institute for Plant Breeding Research, Cologne, Germany

OPEN ACCESS

Edited by:
Marco Fondi,
Università degli Studi di Firenze, Italy

Reviewed by:
Hamed Bostan,
North Carolina State University, United States
Xiyin Wang,
North China University of Science and Technology, China

***Correspondence:**
Tajammul Hussain
tajammul.hussain@tu-dortmund.de
Oliver Kayser
oliver.kayser@tu-dortmund.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Plant Science

Received: 29 January 2018
Accepted: 06 April 2018
Published: 09 May 2018

Citation:
Hussain T, Plunkett B, Ejaz M, Espley RV and Kayser O (2018) Identification of Putative Precursor Genes for the Biosynthesis of Cannabinoid-Like Compound in Radula marginata. Front. Plant Sci. 9:537. doi: 10.3389/fpls.2018.00537

The liverwort *Radula marginata* belongs to the bryophyte division of land plants and is a prospective alternate source of cannabinoid-like compounds. However, mechanistic insights into the molecular pathways directing the synthesis of these cannabinoid-like compounds have been hindered due to the lack of genetic information. This prompted us to do deep sequencing, *de novo* assembly and annotation of *R. marginata* transcriptome, which resulted in the identification and validation of the genes for cannabinoid biosynthetic pathway. In total, we have identified 11,421 putative genes encoding 1,554 enzymes from 145 biosynthetic pathways. Interestingly, we have identified all the upstream genes of the central precursor of cannabinoid biosynthesis, cannabigerolic acid (CBGA), including its two first intermediates, stilbene acid (SA) and geranyl diphosphate (GPP). Expression of all these genes was validated using quantitative real-time PCR. We have characterized the protein structure of stilbene synthase (STS), which is considered as a homolog of olivetolic acid in *R. marginata*. Moreover, the metabolomics approach enabled us to identify CBGA-analogous compounds using electrospray ionization mass spectrometry (ESI-MS/MS) and gas chromatography mass spectrometry (GC-MS). Transcriptomic analysis revealed 1085 transcription factors (TF) from 39 families. Comparative analysis showed that six TF families have been uniquely predicted in *R. marginata*. In addition, the bioinformatics analysis predicted a large number of simple sequence repeats (SSRs) and non-coding RNAs (ncRNAs). Our results collectively provide mechanistic insights into the putative precursor genes for the biosynthesis of cannabinoid-like compounds and a novel transcriptomic resource for *R. marginata*. The large-scale transcriptomic resource generated in this study would further serve as a reference transcriptome to explore the *Radulaceae* family.

Keywords: *Radula marginata*, *de novo* transcriptomic assembly, biosynthetic metabolomics pathways, plant cannabinoids, *Radula*, a prospective alternate to cannabinoids

INTRODUCTION

The liverwort *Radula marginata* belongs to the bryophyte division of land plants. Bryophytes are a distinctive group of an early-diverged lineage of non-vascular land plants comprising around 20,000 species. This division includes mosses, hornworts, and liverworts. Liverworts, or Marchantiophyta, are the most abundant phylum, comprising almost 6,000–8,000 species with high diversity in their ecology, morphology and genetic variation (Qiu et al., 2007; Rubinstein et al., 2010; Ruhfel et al., 2014). These non-vascular plants are very simple, with small flattened bodies with overlapping scales (Buck, 1986). Liverworts evolved (from algae) as the earliest bryophyte group and thus are considered ancestors of the mosses, hornworts, and land plants (embryophytes) (Willis and McElwain, 2002). Because of their distinct evolutionary position as a basal group for the colonization of land plants, bryophytes are ideal for exploring the evolution of plants, genetic networks, and developmental variation (Bowman et al., 2007; Sharma et al., 2014).

Among three classes of liverworts, the Jungermanniopsida encompasses 85% of known liverwort species. This is further subdivided into three subclasses and each subclass has three orders. The order Jungermanniales of the subclass Jungermanniidae consists of 13 suborders and 47 families (Stotler and Crandall-Stotler, 1977; Ruggiero et al., 2015). The *Radulaceae* family comprises the *Radula* genus, with 283 species. The most important *Radula* species are *R. complanta, R. demissa* (Renner et al., 2013), *R. jonesii* (Bouman et al., 1998), *R. kojana, R. laxiramea, R. visianica, R. perrotteii,* and *R. marginata* (Losada-Lima et al., 2007). *Radula* species have been reported to be rich in habitual diversity and are found in almost all ecosystems such as trees, rocks and soils throughout the world, from Antarctica's coastal area to the northern hemisphere and from Australian semi-arid regions to the Amazon rainforest (Hallingbäck and Hodgetts, 2001).

Two decades ago, *Radula* species were shown to be rich in secondary metabolites, such as terpenoids and phenolics. For example, prenyl bibenzyls were identified in *R. perrottetti*, *R. complanta*, and *R. kojana* species (Asakawa et al., 1991; Toyota et al., 1994). These compounds have a distinct carbon backbone structures which acts as a marker to differentiate *Radula* species (Ludwiczuk and Asakawa, 2008). These compounds have suggested biological as well as pharmaceutical significance, having antifungal, antioxidant, antimicrobial and cytotoxic activities. Perrottetinene and its acid (perrottetinenic acid) have been identified in *R. marginata* (Toyota et al., 2002; Park and Lee, 2010). Interestingly, these compounds are structural analogs of tetrahydrocannabinol (Δ9-THC), a psychopharmacological compound in *Cannabis sativa L.*

Cannabinoids are plant secondary metabolites and belong to the class of terpenophenolics which are predominantly found in *C. sativa*. These compounds have proven to have a wide-ranging role in numerous clinical applications. These compounds accumulate in specialized glandular structures known as trichomes (Flemming et al., 2009; Happyana et al., 2013). Since cannabinoids have been recognized for their clinical value, research has also been carried out on plants other than *C. sativa* that contain cannabinoid-like compounds. This has led to the identification of cannabidiol-like (CBD-like) compounds in *Linum usitatissimum* (Styrczewska et al., 2012). Likewise, cannabigerol-like (CBG-like) compounds were found in the South African flowering plant *Helichrysum umbraculigerum* (Bohlmann and Hoffmann, 1979). In addition, sesquiterpenoid-like β-caryophyllene was discovered in *Piper nigrum* (Tisserand and Young, 2014). Despite some reports on the identification of cannabinoid-like compounds in plants, their complex plant architecture makes it challenging as an alternative source of cannabinoids. Although cannabinoids have been identified in different plant species, there has been no report on psychopharmacologically essential compounds except the THC-like metabolites found in *R. marginata*. Recently, the reported agonistic activity of THC-like natural products from *R. marginata* with the cannabinoid receptor 1 (CB1) further confirmed its significance (Russo, 2016; Gachet et al., 2017; Soethoudt et al., 2017). Therefore, *R. marginata* could be a suitable alternate source because of its relatively simple architecture and a diversity of natural habitats. Thus a genetic-level understanding of secondary metabolic pathways that lead to the synthesis of cannabinoid-like natural compounds would be desirable.

Since whole genome sequencing is resource-intensive, we performed a *de novo* approach to assemble the *R. marginata* transcriptome to establish a reference dataset. In addition to the transcriptome, micro-transcriptomic traits like transcription factors (TFs), simple sequence repeats (SSRs), and non-coding RNAs (ncRNAs) were also identified to understand the regulation of these genes. Interestingly, candidate genes for almost all the enzymes required for the conversion of primary metabolites into geranyl diphosphate (GPP) were identified. Expression of all the genes was confirmed by quantitative real time PCR (RT-qPCR). In addition, stilbene synthase (STS) was identified, structurally analyzed, functionally validated and considered as the first intermediate rather than olivetolic acid for the production of CBGA-like compound. This study was designed firstly to discover genes, identify biosynthetic pathways and the regulation of these genes to understand the biological processes involved. Secondly, being the first such gene expression profiling for this family, it would also serve as the reference transcriptome in future exploration of the *Radulaceae*. Furthermore, it would provide the link to study the transition from liverworts to higher plants during evolution.

Abbreviations: BLAST, Basic local alignment search tool; GO, Gene ontology; BP, Biological Process; MF, Molecular Function; CC, Cellular Components; KEGG, Kyoto Encyclopedia of Genes and Genomes; NGS, Next Generation Sequencing; RT-qPCR, Real-time quantitative polymerase chain reaction; RPKM, Reads per kilobases per million reads; ncRNA, non-coding RNA; RSEM, RNA-seq by Expectation-Maximization; FPKM, Fragment per kilobase of transcript per million mapped reads; TPM, Transcripts per million; TFs, transcription factors; SSRs, simple sequence repeats; *Rm-Tct*, *Radula marginata* tentative consensus transcripts; SA, Stilbene acid; OA, Olivetolic acid; STS, Stilbene synthase; CHS, Chalcone synthase; PDB, Protein databank.

RESULTS

Illumina NGS Sequencing and *de Novo* Transcriptome Assembly

In the absence of a reference genome, *de novo* assembly of *R. marginata* was performed to obtain a comprehensive reference transcriptome and to identify novel genes for the biosynthesis of secondary metabolites, in particular, cannabinoid biosynthesis. For this purpose, we performed paired-end RNA sequencing to determine the distant connections between the transcripts. RNA-sequencing yielded ∼30 million raw reads for each pair with an average length of 250 bp. After filtering the raw reads for low base quality, trimming and adapter removal (see section Materials and Methods for details), we selected 91% (28,231,052) reads for the assembly (**Table 1**, **Supplementary Table 1**). Raw reads were *in-silico* normalized and assembled using Trinity (version 2.2.0). The *de novo* assembly resulted a total of 1,580,612 transcripts, with a median contig size of 303 bp, and a maximum of 24,899 bp. Since *de novo* assembly also generates many isoforms of a gene, particularly at high coverage areas, we only selected the longest isoform per gene, resulting in 1,482,641 transcripts (**Supplementary Table 2**). In addition, clustering analysis resulted in 501,849 non-redundant transcripts of an average length of 484 bp and N50 of 459 bp. To perform the assembly assessment, raw reads were mapped back to the *de novo* assembled transcriptome using Bowtei2 (version 2.2.6) (Langmead and Salzberg, 2012). In total, 98% of raw reads successfully mapped back to the transcripts, of which 72% of the reads were mapped to both forward and reverse reads (**Supplementary Table 3**).

To determine the potential of all the putative genes to translate into full-length protein, open reading frames (ORFs) of these putative genes were predicted using the TransDecoder software package (Haas et al., 2013). The homology of these genes was searched from the protein database UniProtKB/Swiss-Prot (http://www.uniprot.org/) followed by the identification of the protein domains by the Hidden Markov Model based approach using PFAM (http://pfam.xfam.org/). Results from both searches were incorporated with the longest ORFs. The ORFs showing homology to known proteins were selected. Based on the gene structure, we have predicted four types of ORFs: complete, three prime partial (3′p), five prime partial (5′p) and the internal ORFs. The complete ORFs were defined as the target genes which were transcribed with a start codon followed by five prime UTR, exon(s) and stop codon with three prime UTR (5′UTR+exon+3′UTR). The 3′p ORFs lacked a stop codon consisting of only 5′UTR and exon, while the start codon was absent in 5′p ORFs, only having exon and 3′UTR. In contrast, the internal ORFs were transcribed from the start to the end of the transcripts and did not contain start and stop codons. We have identified 2,880, 2,977, and 5,104 ORFs that belonged to complete, 3′p and 5′p ORF types, respectively. The highest number of the ORFs (76,499) belonged to the internal type. Moreover, for each type of ORF, the number of ORFs identified on the positive and negative strands was similar as shown in the percentages (**Supplementary Table 4**).

Finally, after the selection of the longest isoform, clustering of assembled transcripts, mapping of raw reads, quantification and retention of the longest ORF peptide candidates resulted in a total of 87,460 transcripts. This high quality unique transcriptomic dataset were named as unigenes. The length of unigenes ranged from 278 to 23,874 bp, with most of the unigenes (93%) having a length of up to 1 kb (**Table 2**). These unigenes were further used for the functional annotation and other downstream analysis. Furthermore, these unigenes were considered as the tentative consensus transcripts (Tct) and were assigned unique identifiers with the prefix *Rm-Tct* (*R. marginata* tentative consensus transcripts). The raw sequencing data have been submitted to NCBI under the BioProject No. PRJNA430820, BioSample No. 08378951 and Sequence Read Archive No SRR6489347.

TABLE 1 | Sequencing summary.

No of raw reads	30,981,418
No of filtered reads	28,231,052 (91%)
No of assembled bases	608,201,850
No of assembled transcripts	1,580,612
Read mapped	20,315,784 (72%)
Maximum length (bp)	23,874
Median contig length	303
Minimum length (bp)	278

Functional Annotation

BLAST Functional Analysis

Annotation is one of the most critical steps for interpreting and precisely evaluating the transcriptomic assembly. For the preliminary functional analysis, homology of all the unigenes was searched using the blastx algorithm (Altschul et al., 1990) against the National Centre for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov/ non-redundant (nr) protein database. High quality identifiers for 68,715 (78%) unigenes were found with known protein coding potential with a cut-off *E*-value of 1e-05. The remaining 18,746 (21.5%) unigenes could not be matched with known protein entries in

TABLE 2 | Assembly statistics.

	Transcripts	Contigs	Unigenes
Number	1,580,612	217,702	87,460
Length	881,230,701	77,717,062	36,765,277
N50 (bp)	604	314	324
Percent GC	54.25	55.7	55.75
Median length (bp)	311	308	316
Averager length (bp)	557	359	425
200–500 nt	1,262,140	209,643	81,695
500–1,000 nt	205,061	3,788	1,984
1,000–2,000 nt	49,077	1,699	1,356
2,000–5,000 nt	43,227	2,205	2,064
5,000–10,000 nt	17,345	331	326
>10,000 nt	49	35	35

the database, because of lack of an existing genome sequence and EST information for *R. marginata*. Currently there are only 34 nucleotide and 13 protein entries available at NCBI. Percentage identity and *E*-values are two important parameters determining the quality of BLAST results. In our analysis, we found 98% of the blasted unigenes had sequence similarity between 50 and 90%, except 1.76% of the transcripts, which had a minimum sequence similarity of 31% (**Figure 1B**). Similarly, 23, 31, and 24% of unigenes had an *E*-value of up to 1e-15, 1e-30, and 1e-45, respectively. Only 2% of the sequences had an *E*-value equal to zero (**Figure 1C**). Higher percentage identity as well as a lower *E*-value of the majority of the blasted unigenes reflects the strength of the blast analysis. Furthermore, blast species distribution (**Supplementary Data 1**) showed that *Marchantia polymorphia* and *Physcomitrella patens* were the most abundant homologous species. Indeed, these two species are the closest relatives to *R. marginata*. The most represented species in blast hits are shown in **Figure 1A**.

InterProScan Functional Classification

To enable a functional classification of unigenes, InterProScan (InterPro) https://www.ebi.ac.uk/interpro/ (release 60) was used. From this analysis, 65,415 (74%) unigenes matched with at least one protein signature. In total 3,487 protein family, 3,529 domains, 515 sites and 108 repeats had significant similarity, with 19,760, 60,021, 1,922, and 4,378 unigenes, respectively. "Heat shock protein 70 family" was the most abundant, with 399 unigenes, followed by the Short-chain dehydrogenase/reductase SDR and "Chaperonin Cpn60/TCP-1 family" that had 357 and 284 unigenes for each category (**Figure 2A**). Within the domain signatures, 2,935 unigenes had the "P-loop containing nucleoside triphosphate hydrolase" domain (**Figure 2B**). The Tetra-tri-co-peptide repeat, the Leucine-rich repeat and the WD40 repeat were the most recurrent that were found in 232, 200 and 186 transcripts, respectively (**Figure 2C**). Furthermore, out of the predicted site signatures, 68% (349) of the signatures were from the conserved category, followed by 19% (96) that belonged to the active sites. The serine/threonine-protein kinase active site was the most prevalent site present in 205 unigenes. The top most representative protein signatures for the four protein classification databases are shown in **Figures 2A–D**, **Supplementary Figure 1A** and **Supplementary Data 2**.

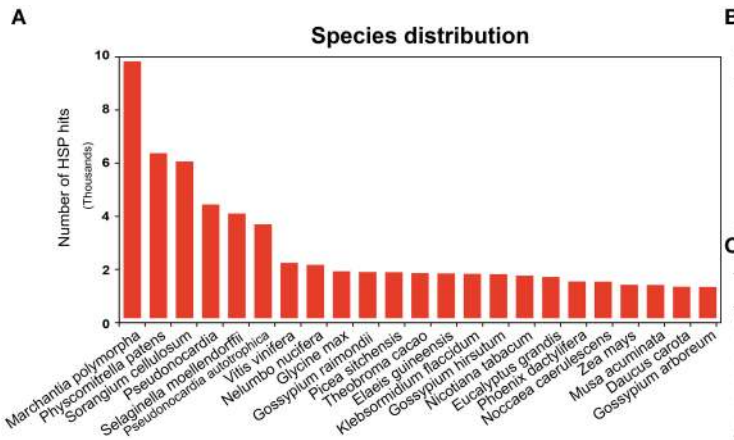
Gene Ontology Classification

To determine the shared dynamic proteins among organisms and to understand the putative biological functions, gene ontology (GO) was analyzed (Ashburner et al., 2000). Based on the homology results from the NCBI nr database, GO terms were mapped to each unigene using a comprehensive annotation suite BLAST2GO (Conesa et al., 2005). Predicted InterPro scan protein signatures were also mapped for GO terms and finally an integrated annotation of unigenes was obtained after merging with blast-derived GO annotation. From the analysis, a total of 116,543 GOs were predicted from 4,820 functions. These GO terms were divided into three major categories: molecular function, biological process and cellular component. Fifty-one percent of the GOs (59,279) were predicted for the molecular function category (MF) followed by 34% (39,950) for biological processes (BP) and 15% (17,314) for cellular components (CC). Within the MF category, ATP binding, DNA binding, and oxidoreductase activity were significantly represented, with 5,359, 2,957, and 2,606 GOs for each type of function. Metabolic process (1,884) and regulation of transcription, DNA templated (1,744) were the most active BPs. CC were enriched with the membrane, membrane and cytoplasm integral components, with 4,080, 3,445, and 1,386 GOs. In total, 4,820 GO categories from 2,353 molecular functions, 1,849 biological processes and 618 cellular components were significantly present in the liverwort transcriptome. The total numbers of unigenes mapped with each GO category were 39,864, 31,268, and 14,369 for MF, BP, and CC, respectively. The ratio of GOs/unigenes was 1.49, with a minimum of one GO term per unigene to 27 GOs per unigenes (**Figure 3A**, **Supplementary Figures 1B,C** and **Supplementary Data 3**). From the complete annotation process we were able to generate 49,167 unigenes with functional annotations. In summary, 56% of the *de novo* assembled transcriptomic pool was annotated (**Figure 3B**).

The remaining 44% of the non-annotated were considered as unknown or orphan genes. Despite the lack of significant annotation, these unknown genes may have a possible functional potential as they were generated from the expressed transcripts (mRNA). They might either have potential binding sites for transcription factors (TFs), SSRs, or be ncRNAs.

KEGG Pathway Network Analysis

One of the objectives in establishing a reference liverwort (*R. marginata*) transcriptome was to provide a resource for further NGS and proteomics studies. We further estimated the coverage of our newly assembled annotated reference transcriptome by determining the minimum coverage of general pathways involved in the MF and BP in *R. marginata*. To determine the rate of coverage, we analyzed more than 518 pathways listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. To identify active biological pathways, unigenes were scanned for the pathways based on analysis using KEGG (http://www.genome.jp/kegg/). As a result, 11,421 candidate genes were identified that encoded 1,554 enzymes in 145 biosynthesis pathways. These pathways belonged to 15 functional categories from "network of metabolism," "environmental information processing," "organismal processes" and "genetic information processing" (**Figure 4A** and **Supplementary Data 4**). Among the identified pathways, the carbohydrate metabolism pathway was mapped with the highest number of 323 enzymes. We identified 53 enzymes for secondary metabolite biosynthesis and 29 from the terpenoid metabolism and polyketide pathways (**Supplementary Table 5**). Moreover, the identified enzymes were broadly distributed into six major enzyme classifications (**Figure 4B**, **Supplementary Table 6**). Transferases were the most abundant enzymes, followed by the oxidoreductases and hydrolases. Distribution of the identified enzymes per pathways is shown in **Figure 4C**.



B

Sequence similarity

Similarity %	Transcripts %
30-50	1.76
50-59	13.93
60-69	23.37
70-79	24.88
80-89	21.59
90-100	14.47

C

E-value distribution

E-value	Transcripts %
0 ~ 1e-100	1.92
1e-100 ~ 1e-60	6.09
1e-60 ~ 1e-45	13.83
1e-45 ~ 1e-30	23.31
1e-30 ~ 1e-15	29.77
1e-15 ~ 1e-5	21.70

FIGURE 1 | BLAST analysis. **(A)** Blast species distribution of closely related bryophyte species against NCBI non redundant (nr) database with an *E*-value of 1e-03. **(B)** Sequence similarity and **(C)** *E*-value distribution of BLAST results.

Candidate Genes for the Biosynthesis of Cannabinoid-Like Compounds in *R. marginata*

In *C. sativa* the biosynthesis of cannabinoids, such as tetrahydrocannabinolic acid (THCA), is the result of oxidative cyclization of cannabigerolic acid (CBGA) (Sirikantaramas et al., 2004; Taura et al., 2007; Stout et al., 2012; Andre et al., 2016; Zirpel et al., 2017). CBGA is the central precursor for cannabinoid biosynthesis, which is catalyzed by the alkylation of olivetolic acid (OA) with geranyl diphosphate (GPP). OA and GPP are derived from the polyketide and 2-C-Methyl-D-erthritol 4-phosphate (MEP) pathways, respectively. GPP is derived from a condensation reaction between two isoprene units, dimethylallyl pyrophosphate and isopentenyl pyrophosphate (Croteau and Purkett, 1989). Remarkably, we have identified all the potential homologous genes in *R. marginata* encoding the enzymes required for the conversions of pyruvate and D glyceraldehyde 3 phosphate into GPP, and also confirmed the expression through qRT-PCR (**Figures 5A–C**). The genes identified for GPP synthase had homology of 79.7% at the protein level with the GPP synthase in *Capsicum annuum,* and 75% with the large subunit of GPP synthase in *C. sativa*. In addition, we found that all the 15 potential binding sites of both isoprene units were conserved.

In contrast to OA in *C. sativa*, which is synthesized from the condensation of aliphatic CoA-tethered hexanoyl (hexanoyl-CoA), we found an analogous compound, stilbene acid (SA), in *R. marginata*. Compared with OA, SA catalyzes the condensation reaction from the CoA-tethered phenyl propanoid i.e., coumaroyl (P-coumaroyl-CoA). However, both require three molecules of malonyl-CoA by polyketide synthase (Fellermeier et al., 2001; Austin et al., 2004). Acetyl CoA carboxylase is the enzyme required for the carboxylation of acetyl-CoA to form malonyl-CoA, and has been identified in this study. Also, unlike the hexanoyl-acyl carrier protein, P-coumaroyl-CoA provides the "shuttle service" for the intermediate polyketide as well as substrates in the formation of SA by the enzymatic reaction of STS. The STS identified from *R. marginata* had 85% homology with chalcone synthase (CHS) and 84% homology with stilbencarboxylate synthase from *M. polymorph.* STS diverged from CHS as a result of gene duplication, from an evolutionary perspective (Tropf et al., 1994). Both catalyze the same reaction with different cyclization mechanisms, which lead to different products (Austin and Noel, 2003).

The CBGA analog was identified with the HPLC-ESI-MS-MS approach (**Supplementary Figure 2A**) and the metabolite with the same m/z was also observed from GC-MS analysis (**Supplementary Figure 2B, Supplementary Data 5**). The positive ESI mass spectrum of the CBGA bibenzyl analog showed an [M+H]+ ion at m/z 395, a predominant [M+H-H2O]+ ion at m/z 377, and fragment ions at m/z 271 [M+H-C9H16] and m/z 267 [M+H-H2O-C8H4]. The ion m/z 253 is probably the result of degradation processes eliminating two protons, resulting in [M+H-H2O-C9H16]. Together these indicate the presence of the CBGA analogue's bibenzyl structure in the methanolic extracts of *R. marginata*.

To gain insights into the STS identified in *R. marginata*, we carried out structural analysis of the protein sequence. For this purpose, we used homology modeling to develop the structure of STS in comparison with the CHS-like PKS (PDB ID code 3awk) from *Huperzia serrata* and STS (PDB ID code 2p0u) from *Marchantia polymorpha*. Structural analysis revealed that the *R. marginata* STS shared 64 and 60% homology with these, respectively. In the model prediction, Lys-Ala was replaced with Thr-Ser, Thr-Glu, Arg-Lys, and Ala-Lys at positions 20–21, 46–47, 74–75, 312–313, respectively. Similarly, these variable sites except 20–21 were also found in STS from *M. polymorpha,* where they code Pro-Ala, Glu-Ser, and Ala-Lys amino acids. Likewise, OA in *C. sativa* also showed the same variable sites with the amino acid changes Thr-Ala, Thr-Gln, Ala-Gln, and Glu-Ala. Since STS diverged from CHS and OA was identified as

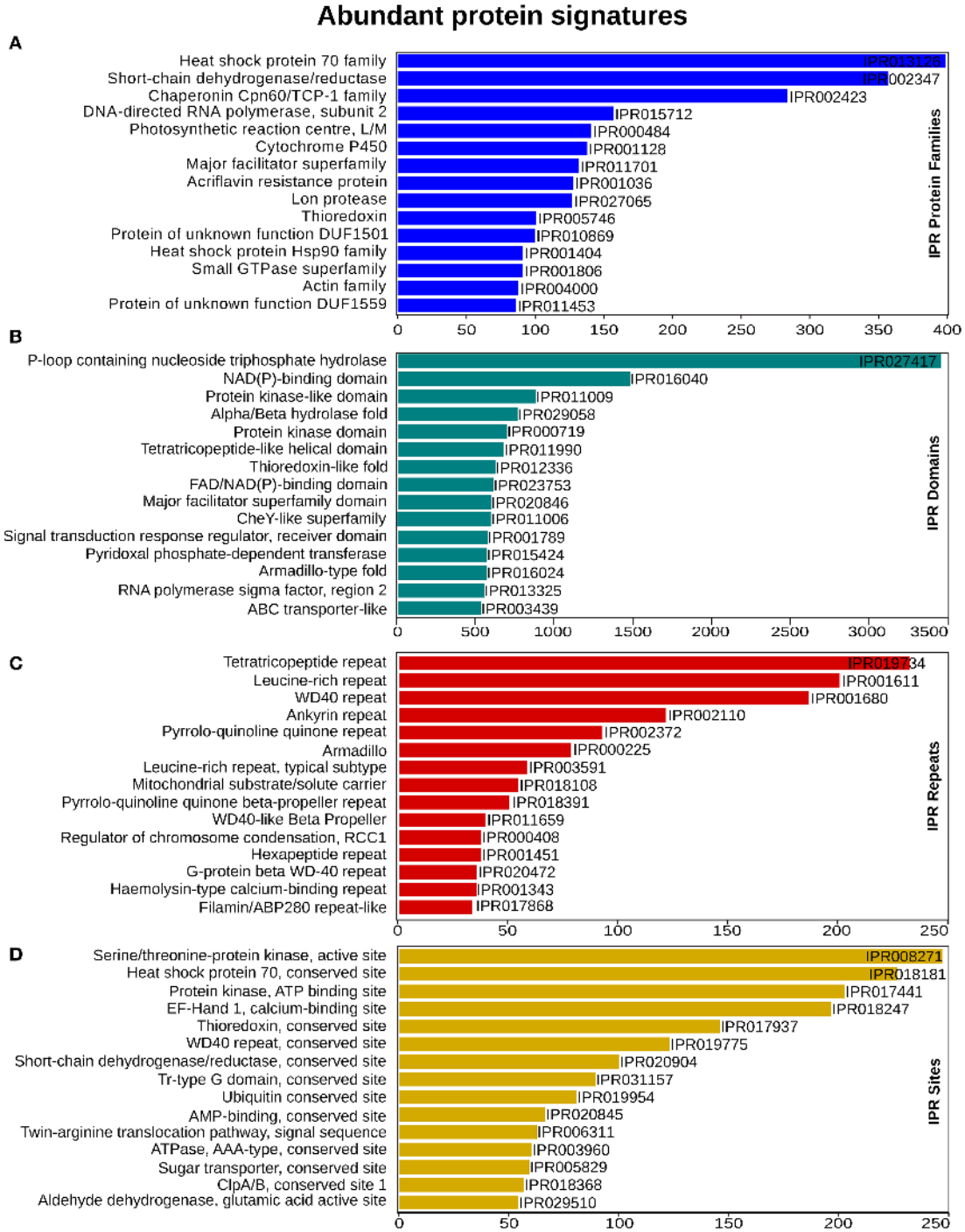


FIGURE 2 | InterProScan functional analysis. **(A)** Annotation distribution of most significant protein families **(B)** domain identified **(C)** prominent repeats observed and **(D)** important sites determined. X-axis represents the number of genes associated with the respective protein signatures on Y-axis.

the homolog of CHS, we speculate that these variable sites might have the potential for the diversification of the parental gene. However, the mechanism of the divergence and their catalytic activity of these potential sites need to be explored (**Figure 6**).

Identification of Transcription Factors

To understand gene regulation, unigenes were scanned using a Hidden Markov Models (HMM) (Finn et al., 2011)-based approach against the pfam (Finn et al., 2016) as well as the PlnTFDB (Pérez-Rodríguez et al., 2009) and PlantTFDB

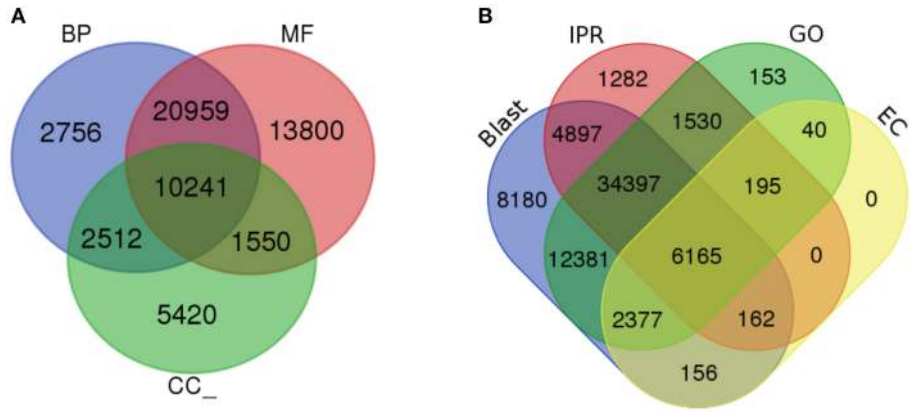
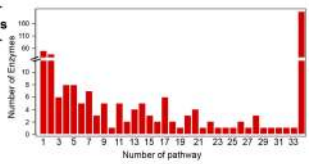


FIGURE 3 | Gene Ontology analysis. **(A)** Gene ontology functional annotation for biological process (BP), Molecular function (MF), and Cellular component (CC). **(B)** Overall annotation results from BLAST, InterProScan (IPR), Gene ontology (GO), and KEGG pathway databases.

A Identified enzymes for respective pathway category

Main category type	Pathway category	No. of pathways	No. of enzymes	No. of unigenes
Metabolism	Carbohydrate metabolism	15	323	2464
	Amino acid metabolism	14	275	2309
	Biosynthesis of antibiotics (Global and overview maps)	01	164	1573
	Metabolism of cofactors and vitamins	12	135	3491
	Lipid metabolism	17	114	879
	Energy metabolism	02	113	1456
	Xenobiotics biodegradation and metabolism	09	104	941
	Nucleotide metabolism	20	97	5917
	Metabolism of other amino acids	14	62	524
	Biosynthesis of other secondary metabolites	07	53	595
	Glycan biosynthesis and metabolism	17	48	227
	Metabolism of terpenoids and polyketides	11	29	248
Genetic Information Processing	Translation	01	22	241
Environmental Information Processing	Signal transduction	02	12	37
Organismal Systems	Immune system	02	03	256
		144	**1554**	**21158**

B Enzymes per pathway



C Enzymes code (EC) distribution

EC class	No of EC	No of unigenes
Oxidoreductases	392	2143
Transferases	493	2701
Hydrolases	264	4757
Lyases	201	562
Isomerases	97	527
Ligases	107	731
Total	**1554**	**11421**

FIGURE 4 | KEGG Pathway analysis. **(A)** Identification of genes encoding enzymes in 15 functional pathways among four main categories. **(B)** Distribution of enzymes per pathway, X-axis shows the number of pathways, while Y-axis represents the number of identified enzymes for each pathway. **(C)** Enzyme code distribution according to the enzyme class (EC) type.

(Jin et al., 2014) databases. TFs were predicted according to the family assessment rule as described in PlantTFDB. To increase the accuracy, TFs with an *E*-value less than 1e-05 as well as an identity percentage <50 were excluded from further analysis. After filtering we identified 1,085 TFs belonging to 39 TF families (**Figure 7**). We found that putative TF encoding regions were present in 3,449 unigenes. The total number of TFs identified in our study was higher than previously reported TFs in two desiccation-tolerant moss species, *Syntrichia caninervis* and *Bryum argenteum*, which contained 778 and 770 TFs, respectively. However, the number of TFs was less than in the model moss *P. patens*, which contains 1,156 TFs. Moreover, the 39 identified TF families in *R. marginata* were less than in the above-mentioned species, which contained 49, 50, and 53 TF families for each, respectively. Furthermore, the number of transcripts having a TF-encoding region was considerably higher than those in already-reported transcripts.

TF families are highly conserved in eukaryotic organisms, especially in plants. Despite the sequence conservation, the number of TFs for specific families varies among different species. This variation might be due to evolution or species specification. As for other plant species, evolutionary expansion/contraction was also observed in this study. For example, comparative analysis of three closely related moss species showed that the number of TFs within the ARF, C2H2, and CH3 families was higher in our data than in *B. argenteum*. In contrast, the B3 and bHLH families were relatively less abundant in the *R. marginata* transcriptome than in *S. caninervis* and *B. argenteum*. Moreover, six TF families—BSD, CSD, DbpA, FHA, LIM, tify, and TIG—were explicitly predicted in our data, confirming evolutionary expansion of TFs. On the other hand, we could not find any target sequences from B3 and trihelix TF families, which might have been due to evolutionary contraction. On this basis we might assume that the prediction of some TF families, as well as the apparent lack of other TF families, was due to evolutionary expansion and contraction, respectively.

The TF families: MYBs, Basic Leucine Zipper Domain (bZIP), AP2/ERF family proteins, NAC, Basic helix-loop-helix (bHLH),

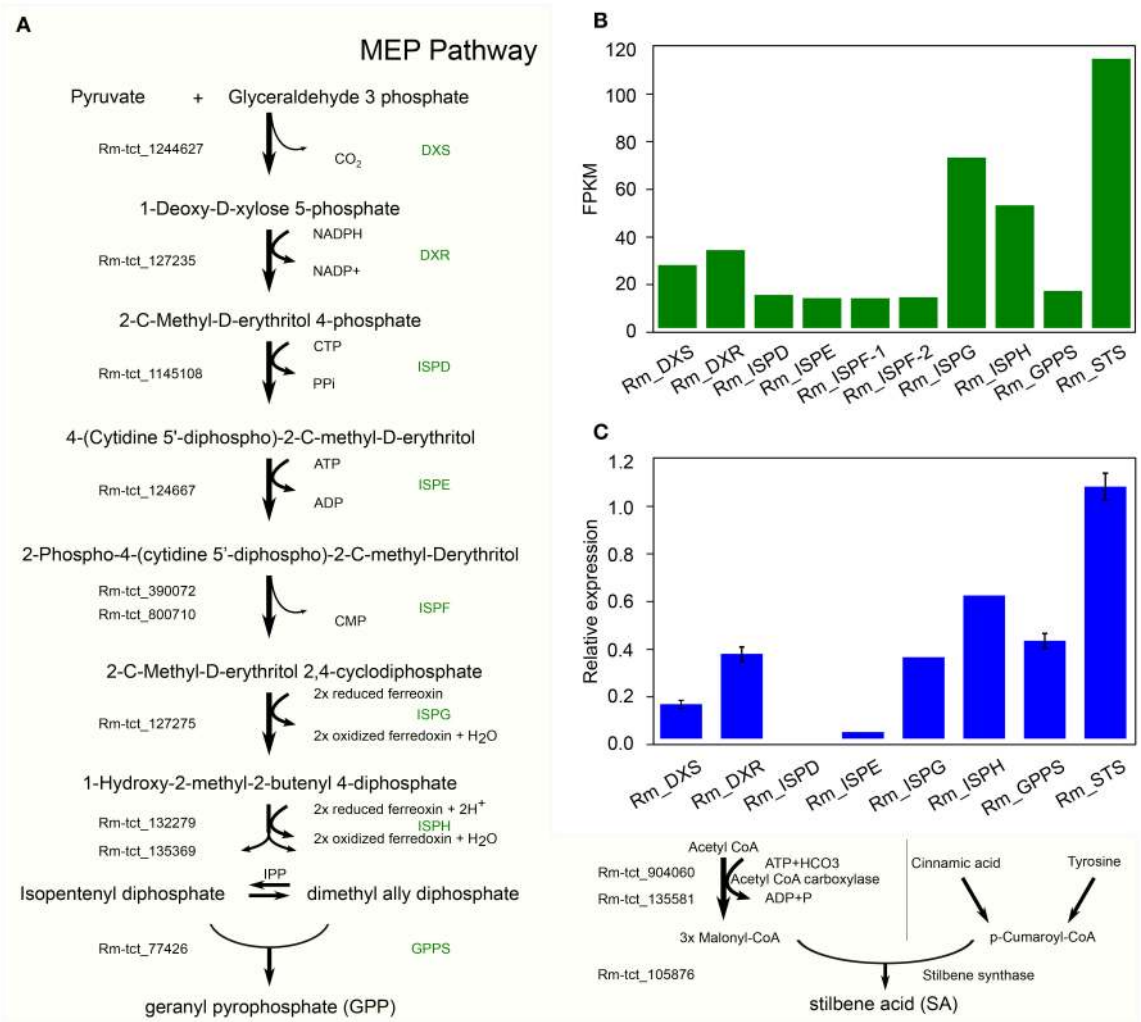


FIGURE 5 | Candidate precursor genes for cannabinoid biosynthesis in *Radula marginata*. **(A)** Identification of candidate genes for cannabinoid biosynthesis in *R. marginata*. GPP transcripts as well as all the upstream genes were identified. Genes with Rm-tct_xxx prefix were identified in this study. **(B)** Transcriptomic expression of identified genes determined from FPKM. **(C)** Relative expression of candidate gene by real time quantitative PCR (RT-qPCR).

and DNA-binding One Zinc Finger (DOF), are considered those most closely involved in the regulation of secondary metabolism, development and growth in many plant species (Vom Endt et al., 2002; Li et al., 2015; Zhu et al., 2015). Interestingly, we found 156 TFs belonging to the above-mentioned families, with 66, 39, 9/22, 10, 8, and 2 TFs, respectively. Of these, MYB TFs were the most abundant, contributing up to 42% of the total TFs, some of which are proposed to regulate secondary metabolism by forming regulatory complexes with other TFs, such as bHLH and WD40 repeats in the regulation of phenylpropanoids (Espley et al., 2007; Liu et al., 2015). Moreover, the regulation of terpenoids has been associated with the APETALA2 (AP2), WRKY, and MYC families (Broun et al., 2006; Spyropoulou et al., 2014). In total, nine TFs from AP2 and three from WRKY were identified from our data. These predictive TFs might regulate aspects of secondary metabolism and terpenoid biosynthesis in *R. marginata* (**Supplementary Table 7** and **Supplementary Data 6**).

Simple Sequence Repeats (SSRs) Identification

To identify SSRs, all the unigenes were screened using a microsatellite satellite identification tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/). A total of 2,041 SSRs (mono to hexa-) were identified from 87,460 assembled transcripts (**Figure 8A**). Mononucleotide repeats can encounter a problem with a higher rate of homo-polymer errors associated with genotyping as well as sequencing (Gilles et al., 2011; Zhang and Huang, 2016). Hence, these were excluded from further SSR analysis. Of the remaining 1,577 di-nucleotide to hexa-nucleotide repeats, 90% of the SSRs belonged were di-nucleotides and tri-nucleotides. Tetra-, penta-, and hexa-nucleotide SSRs were found only in trace amounts, with 4, 3, and 3%, respectively (**Figure 8B**). The

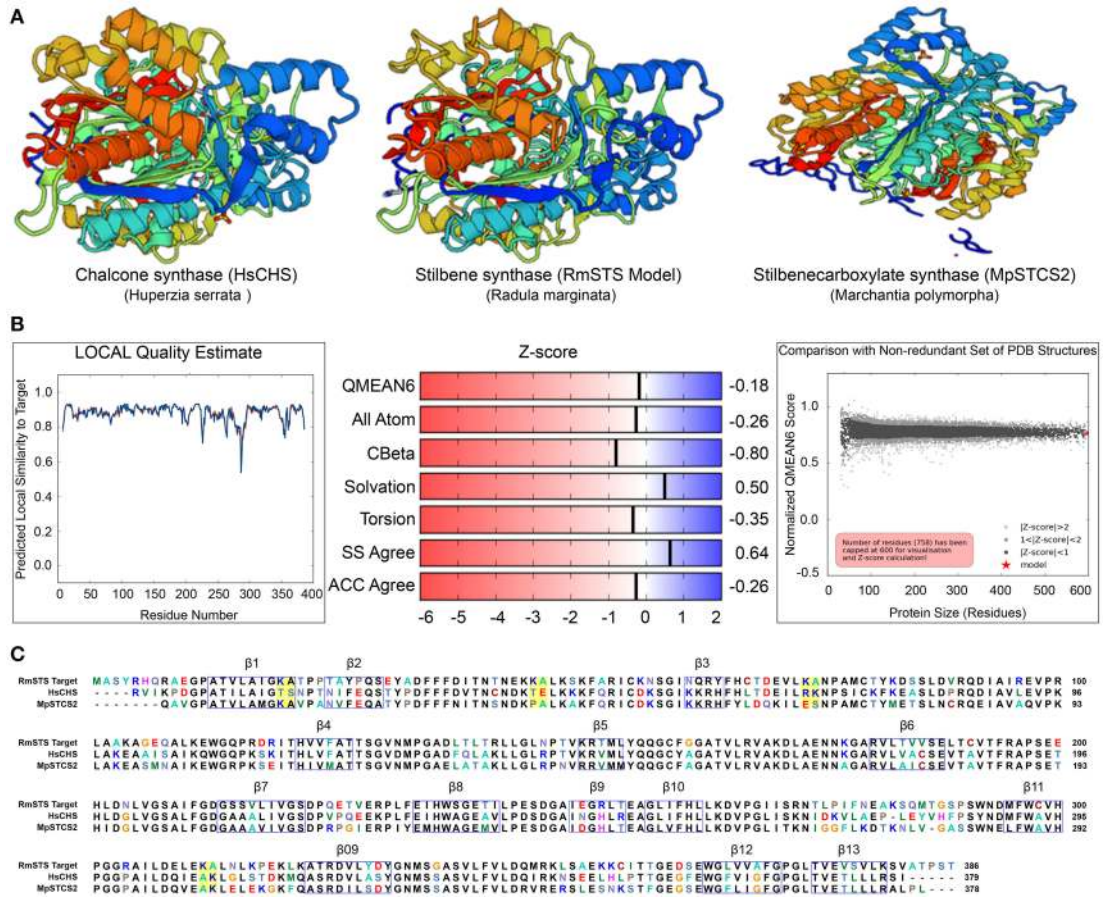


FIGURE 6 | (A) Crystal structure of stilbene synthase model compared with Chalcone synthase- like polyketide (Left, 3awk.1) and Stilbene carboxylate synthase from *Marchantia polymorpha* (right, 2p0u.1. A). **(B)** Quality estimate of local similarity, Z score of the predicted STS model in *Radula marginata*. **(C)** Alignment of the identified stilbene synthase in *Radula marginata* with the respective template model, blue boxes are the beta sheets, yellow boxes in the alignment are the variable sites for both models used as well as in olivetolic acid from *Cannabis sativa*. Taken with permission from "https://swissmodel.expasy.org" (Arnold et al., 2006; Biasini et al., 2014).

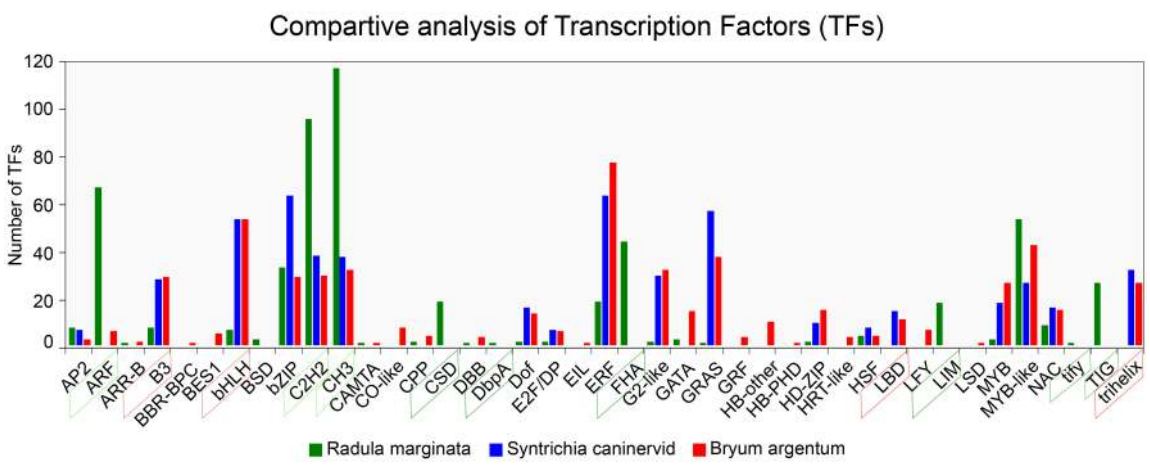


FIGURE 7 | Comparative analysis of Transcription Factor families from *Radula marginata* and related moss species. TF families are on X-axis while number of TFs for each family shown on Y-axis. Green dotted box represents abundant TF families; those in red dotted box are less abundant, green solid-line box is for TF families that are exclusively observed in *R. marginata* compared with related moss species, while red solid-line box represents TF families that were not observed in *R. marginata*.

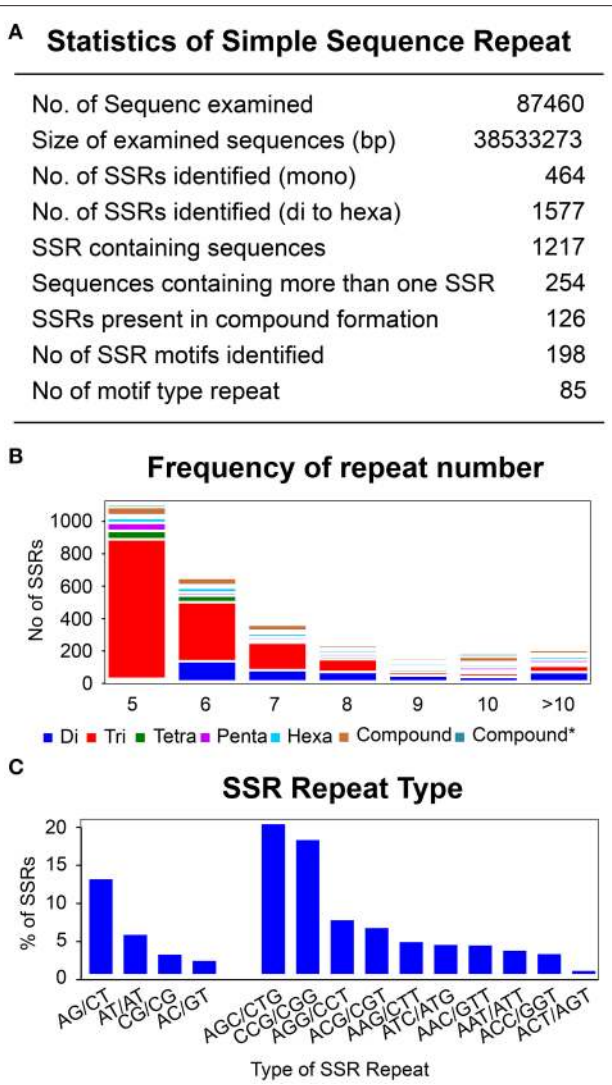


FIGURE 8 | Identification of simple sequence repeats. **(A)** Statistic of simple sequence repeats. **(B)** Frequency of repeat number from 5 to 86 for di-, tri-, tetra-, Penta-, and hexa- type of SSR repeats. Y-axis represents the number of genes for each category of repeats type according to the repeat number on X-axis. **(C)** Most abundant di- and tri- repeats type distribution.

linkage potential of SSRs to loci was much higher with those SSRs identified from the annotated unigenes than for non-annotated (Li et al., 2014). The 1,577 SSRs were annotated with 1,217 unigenes with a minimum length of 12 bp. Moreover, 20% of the annotated unigenes (254) were found to have more than one SSR and 7% of the annotated SSRs (126) were found in compound formation. Based on sequence complementation, a total of 198 identified motifs were classified into 85 repeat types. Analysis showed that for di-nucleotide repeats AG/CT (12%) and AT/AT (5%) were the most abundant, while for trinucleotide repeats AGC/CTG (19%) as well as CCG/CGG (17%) showed the highest frequency. These specific repeats accounted for 53% of the total repeats found (**Figure 8C**, **Supplementary Table 8** and **Supplementary Data 7**).

Repetition of the SSR motif varied from a minimum of five times up to 46 (86 for compound SSR). Although the repeat frequency was observed up to 86, the significance proportion of the motifs were below the repetition value of 10. Only 0.01% of the motifs were above 50, which was almost negligible as well as not being from the annotated SSRs. Similarly, a small fraction of motifs, 4.03%, were found to have a repetition of above 10. A significant proportion of the motifs lay within the repetition unit of 5 (48.54%), followed by the repetition of 6, 7, 8, and 9 with 26.46, 11.90, 5.65, and 1.46%, respectively. Furthermore, distribution of the repeat number within the repetition of 5 was decreased with the increase of motif length from tri to hexanucleotide. Trinucleotide motifs comprised the highest proportion, with 88%, followed by the tetra- (4%), penta- (3.4%), and hexa-nucleotides (1.4%) (La Rota et al., 2005; Hisano et al., 2008; Cloutier et al., 2009). This highest proportion of trinucleotide repeats was directly proportional to the accuracy. The possible reason for this is that expansion and contraction in tetra-, penta-, and hexa-nucleotide repeats may lead to a frame-shift mutation in the coding region (Metzgar et al., 2000; Morgante et al., 2002).

Prediction of Non-coding RNAs

To identify non-coding RNA genes (ncRNA), non-annotated unigenes were scanned for functional RNAs against Rfam database. Rfam http://rfam.xfam.org/ is a collection of RNA families and has three functional categories of non-coding RNA genes, structured *cis*-regulatory elements and self-splicing RNAs. ncRNA has been demonstrated to play roles in the regulation of gene expression at the post transcriptional level (Eddy and Hughes, 2001) as well as in maintaining genome stability by guiding RNA modification (Moazed, 2009). ncRNA genes produce a direct functional RNA molecule rather than being translated into proteins (Mattick and Mattick, 2010). From the analysis we predicted 2,043 ncRNA genes that belong to rRNA (1985), tRNA (46), sRNA (9), and snRNA (3) types (**Supplementary Data 8**). For the category of *cis*- regulation nine genes have significant homology. In order to determine the functionality of these predicted non coding genes further studies are required.

DISCUSSION

Next generation sequencing has become an essential tool for studying the genomic constitution of living organisms, especially those with high genome complexity. Despite the relatively low cost of sequencing, *de novo* assembly of whole genomes without prior sequence information is still costly and is computationally resource intensive. Compared with whole genome sequencing, *de novo* transcriptome analysis has made it possible to understand the genetic architecture of those organisms without a reference genome, at a low cost, and it is also computationally less intensive. Thus, a *de novo* transcriptome study is a valuable tool for identifying new genes, molecular markers, regulatory elements and the expression profile of genes (Verk et al., 2013). Indeed, this advancement has made it possible to study less well characterized species, for example, mosses

and liverworts that have not been extensively studied compared with higher land plants such as rice (Zhang et al., 2010), poplar (Qiu et al., 2011) sesamum (Wei et al., 2011) as well as yeasts (Nagalakshmi et al., 2008) and animals (Feldmeyer et al., 2011). A *de novo* approach has previously been used for the moss species *S. caninervis* (Gao et al., 2014) and *B. argenteum* (Gao et al., 2015). Therefore, a *de novo* assembled transcriptome can be used to quantify the expression of genes and identify new genes.

Radula marginata (Liverworts), have been identified to contain a diverse array of secondary metabolites of high structural and medicinal properties for over the past two decades, such as perrottetinene and its acid (Toyota et al., 1994, 2002). However, there are only a few studies on liverwort species and most of those are related to biodiversity and molecular clock to infer the evolutionary relationships to land plants (Alaba et al., 2015; Singh et al., 2015; Honkanen et al., 2016). While *Radula* is an important plant order, it has not been investigated at the molecular and genomic level. Until now, only 34 nucleotide and 13 proteins have been reported for this liverwort species within the NCBI public database. Here, we report for the first time a *de novo* assembled transcriptome of *R. marginata.* In the RNA-seq analysis, we have identified all the upstream genes of the central precursor in cannabinoid biosynthesis. We have also validated the presence of CBGA-analogous by using HPLC ESI/MS-MS and GC-MS approach. We could identify all the prerequisite enzymes for the precursor which are likely to be conserved. However, we have identified stilbene synthase (SA) instead of olivetolic acid (OA), one of the CBGA precursor in *C. sativa*. Additionally, transcription factors (TFs), microsatellite markers (SSRs) and ncRNAs have also been identified. This *de novo* assembled reference transcriptomic dataset would provide a resource for exploring the *R. marginata.*

The sequence contents and quality of *de novo* assembled transcriptomes essentially depend on input material and further bioinformatics processing of the data. Assembly of the raw reads into contigs is the first step in the *de novo* transcriptomic study. We assembled 217,702 contigs from 30,981,418 raw reads. Because of splice site variation, assembly from Illumina sequencing generates many transcripts of different lengths for a single gene, where selection and identification of a single full-length transcript is a critical step in *de novo* assembly (Schliesky et al., 2012; Steijger et al., 2013). To remove this transcript redundancy (see section Materials and Methods) we finally obtained 87,460 unigenes. The total number of contigs and unigenes identified in our study was higher than the number of contigs (106,066) and unigenes (57,247) described in previously *de novo* assembled moss species (*B. argenteum*). This further illustrates the quality and accuracy of our assembly. Furthermore, the average length, GC contents and N50 value of these unigenes were comparable with the results from previous studies (Liang et al., 2013).

To estimate the quality of *de novo* assembly, raw reads were mapped back to the contigs generated from sequence assembly using bowtie2 with default parameters. Seventy-two percent of reads were successfully mapped back to the contigs and these mapping results are congruent with previous studies. However, the parameters used to check the quality of *de novo* assembly provide only rough estimates. Therefore, it is important to select all the transcripts that have coding potential and annotate all these transcripts to assign them a putative gene function. For this purpose, we used BLAST algorithms, such as blastn, blastx, and blastp with databases including NCBI non-redundant protein (nr), nucleotide (nt), Swiss-Prot (UniProtKB), cosmos (v1.6, *P. patens* proteins) and *M. polymorphia* genomes using a cut off value for homology search of at least 65% of identity and a stringent *E*-value of 1e-05. This allowed for 78% protein-coding transcripts with different degrees of homology.

Functional characterization of unigenes provides insights into the active BP, MF, and CC within an organism. To determine the functionality of genes, regarding regulation and expression, annotation is desirable. Therefore, all the unigenes were blasted with InterProScan (IPR), Gene ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. In contrast to BLAST, 74% of the unigenes were mapped with the protein signature from 14 databases allied with IPR. Out of the total assembled unigenes scanned for GO terms, 56% of the unigenes had at least one GO which was more than the identified GO terms reported in the bryophyte moss species *S. caninervis* (48%) and in the non-model transcriptome of *Bacopa monnieri* (47%) (Jeena et al., 2017). However, *B. argenteum* and *P. patens* (Rensing et al., 2008; Zimmer et al., 2013) have 64 and 58% of GOs which is little higher than in our study. This might be due to an incomplete transcriptome or different cut-off values for transcriptome assembly processes.

To determine the interaction of functionally annotated unigenes of *R. marginata*, we performed KEGG pathway analysis (Kanehisa et al., 2006). This analysis elucidated the potential signal transduction biosynthesis pathways that were abundantly represented from our transcriptomic data. Out of 49,167 annotated unigenes, 11,421 were mapped with 145 signal transduction pathways which were greater than in other bryophyte species. For instance, *D. hirsute*, *S. caninervis,* and *B. argenteum* revealed the presence of 95, 119, and 127 pathways, respectively (Gao et al., 2014, 2015; Singh et al., 2015). Our results are also in line with those for other non-model organisms like *Sesamum indicum L.*,*Camelina sativa,* and *Callosobruchus maculatus* which showed 116, 119, and 119 pathways identified from their respective transcriptomes (Wei et al., 2011; Liang et al., 2013; Sayadi et al., 2016). The biosynthesis of the terpenoid backbone was the most significant pathway in the metabolism of terpenoids and polyketides with 84 unigenes identified that encoded 12 enzymes. Overall, network analysis revealed the presence of almost all the pathways involved in metabolism, of which carbohydrate metabolism was the most dominant with 323 enzymes followed by amino acid metabolism (275), metabolism of co-factors and vitamins (135) and lipid metabolism with 114 enzymes mapped to all of the pathways for each. However, we could not identify some of the secondary metabolic pathways, for example, zeatin, tetra cycline, and brassino steroid biosynthesis. It might be because the bryophyte transcriptomes have greater metabolic versatility than land plants which might have favored alternative metabolic pathways during evolution (Rensing et al., 2002; Oliver et al., 2004; Lang et al., 2005; Wood and Duff, 2009).

We have also identified the precursor genes for the synthesis of SA and GPP. Alkylation of OA with GPP results in the formation of CBGA which is the primary precursor for the cannabinoid biosynthesis in *C. sativa* (Fellermeier and Zenk, 1998). In contrast, we have identified SA in *R.* marginata and it is likely that SA acts in *R.* marginata as a first intermediate in place of OA to form CBGA-analogous. A metabolomics study also revealed the identification of the CBGA analogous and confirms our hypothesis that SA acts as the first intermediate where stilbene is the backbone of the identified compound. It is worth noting that the THCA like compound identified in *R. marginata* (Toyota et al., 2002) also has the stilbene backbone.

In our analysis, we were unable to annotate 44% of the transcriptome, and these were named as unknown genes. A similar percentage of unknown genes has also been observed in other moss species such as *S. caninervis* (41%), *B. argenteum* (36%) as well as other closely related liverwort species like *M. polymorpha* (43%) (Sharma et al., 2014). These high numbers of unknown genes might be due to the unavailability of sufficient genomic information for mosses in general and *R. marginata* specifically. To date, the only available reference genome for bryophytes is the moss species *P. patens* which also has 42% of unknown genes (Rensing et al., 2008; Ortiz-Ramírez et al., 2016). Therefore, we speculate that these unknown genes might play a role in gene regulation since they were generated from the expressed transcripts (mRNA). These unknown genes might be orphan genes and may have binding sites for TF, SSRs or ncRNAs.

The identified unigenes that have potential binding sites for 1088 TFs are consistent with those in other moss species such as *S. caninervis* with 778 and *B. argenteum* with 778 TFs, while *P. patens* has 1,156 TFs. However, the numbers of identified TF families are less than in the other related moss species. This variable number of TFs might be due to speciation events during evolution which explains the phenomenon of evolutionary expansion/contraction of TFs. We found six TF families that are unique to *R. marginata* which may be a result of expansion and three TF families that could not be detected from comparative analysis with related moss species. Although the number of transcription factors within a family varies, the sequence tends to remain conserved because of the conserved nature of TFs.

Our analysis of SSRs also revealed a large number of di- to hexa- repeats, of which 90% belongs to the di- and tri- repeat types. These data are correlated with *P. patens* where 91% of the SSRs comprised of di- and tri-nucleotide repeats. Some plant species such as sunflower, sesame, canola, *Arabidopsis*, peanut, sugar beet, cabbage, soybean, sweet potato, pea, and grape have a higher number of dinucleotide repeat motifs while trinucleotide type repeats are more frequent SSR motifs in cereals such as barley, rice, and wheat (Kumpatla and Mukhopadhyay, 2005; La Rota et al., 2005; Wei et al., 2011). Although repeat units were observed with a minimum of five to a maximum of 46 times, significant proportions of the repeat types were associated with five repeat units, as also observed in moss species.

CONCLUSION

We describe the first high quality *de novo* assembled transcriptome, and have annotated the highest number of genes to date in *R. marginata*. Moreover, identification of the precursor genes in the cannabinoid biosynthesis pathway suggests that the cannabinoid pathway is likely to be conserved in lower and high land plants with the exception of first intermediate. These findings require further experimental work to confirm this proposed novelty. Overall, this study would serve as a new transcriptomic resource among the bryophyte species and also proposes *R. marginata* as an alternate to *C. sativa* for cannabinoid-like compounds.

MATERIALS AND METHODS

Liverworts Collection

Radula marginata (liverworts) samples were collected from their basal habitat at Waitakere Ranges Regional Park, New Zealand. For metabolite extraction, material was transported in sealed plastic zip lock bags in liquid nitrogen to the Technical University Dortmund, Germany. For RNA extraction, samples were collected in liquid nitrogen and then stored at −80°C at The New Zealand Institute for Plant & Food Research Limited (PFR), Auckland, 1142, New Zealand.

Extraction of RNAs

Total RNA was extracted from six samples using RNeasy™ Plant Mini Kit (Qiagen N.V., The Netherlands) according to the manufacturer's recommendations. To eliminate possible DNA contamination On-column DNaseI (Qiagen) digestion was added to the extraction protocol. RNA concentration was measured with Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and quality was checked on agarose gel electrophoresis using Bioanalyzer agilant 2100 (Agilent Technologies, Santa Clara, CA, USA) was used to determine the integrity by RNA integrity number (RIN). Enrichment of mRNA by removing rRNA from total RNA was carried out by oligo (d) T beads (Qiagen, Hilden Germany). Purified mRNA was then transported in RNAStable plate (Biomatrica, San Diego, USA) to IGA Technology Services Udine, Italy.

cDNA Library Preparation and Illumina NGS Sequencing

The total RNA from each of six samples was pooled and 2.5 μg of the pooled RNA was used to prepare cDNA according to the instructions of Illumina® (Illumina, San Diego, USA). The further steps in the preparation such as adapter insertion, interruption of the fragment, size selection and PCR amplification were performed at IGA Technology, Italy. Sixty million paired end reads of a length 250 bp were generated using Illumina HiSeq-2000 platform.

Quality Control

Illumina Hiseq-2000 sequencing platform (Illumina, San Diego, CA, USA) was used to sequence the transcriptome. FASTQC

(version 0.11.24) (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to determine the quality of the raw read. The base quality score was set to (Q20 = 1%) and all the reads below Q20 were considered as low quality (Andrews). Trimmomatics (version 0.36) was used to trim the low quality reads from 3′ and 5′ end using sliding window, leading and trailing to 4, 5, and 5 bases, respectively (Bolger et al., 2014). Adapter sequences present in the sequences were removed with cutadapt (version 1.9.1) (http://code.google.com/p/cutadapt/) (Martin, 2011). Read lengths with <50 bp and reads with ambiguous "N" >5% were also dropped.

Data Processing and *de Novo* Transcriptomic Assembly

The sequencing quality of raw reads was determined with FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) software package. To avoid any false positive gene prediction adapters, low-quality base (higher than 20% nucleotides having a quality value ≤10) and reads (ambiguous nucleotide >5%) were removed with cutadapt and trimmomatic, respectively (Martin, 2011; Bolger et al., 2014). *In silico* read normalization was done before the assembly and then filtered normalized reads were *de novo* assembled using Trinity software package with default k-mer size of 25 (Haas et al., 2013). Inchworm, Chrysalis and Butterfly are the three independent modules for the assembly in Trinity. At the first step of assembly, Inchworm, raw reads are used to form a k-mer catalog followed by Chrysalis, using de Bruijn graph approach (based on distance and relation) to constructs the contigs. Finally, Butterfly connects all possible de Bruijn graph to construct the full-length transcripts as well as their splice variants as a result of alternative splicing.

To remove redundancy, assembled reads were clustered using CD-HIT-EST tool (Fu et al., 2012). To determine the protein coding region each transcript (putative gene) was *in silico* translated using TransDecoder (http://transdecoder.github.io). Among all the predicted frames, ORFs (open reading frames) peptide candidates were selected. Subsequently, the quality of the transcriptome was accessed by mapping raw reads back to the assembled reads using Bowtie2 (Langmead and Salzberg, 2012). Since *de novo* assembly is challenging for accurate mapping quantification, RNA-seq by Expectation-Maximization (RSEM) was used to determine the relative abundance of transcripts (Li and Dewey, 2011) which quantifies assembled transcripts with a high degree of accuracy by first preparing a reference set followed by the read mapping and abundance estimation. Afterwards, gene expression was calculated and reported as Transcripts per million (TPM) and fragment per kilobase of transcript per million mapped reads (FPKM) expected count numbers. Finally, the clustering of different isoforms of the transcripts (genes), selection of the longest putative ORF peptide candidate and the transcripts with at least TPM value > 1 were defined as a unigene.

Functional Annotation and Characterization of Unigenes

All the unigenes were searched against different public databases. BLASTx was used to determine the similarity against the NCBI non-redundant protein database (nr), Swiss-Prot, NCBI nucleotide database (nt) with stringent parameters (*E*-value 1e-05). For the functional characterization of unigenes, ORFs were scanned against InterProScan (IPR) release 60 (Finn et al., 2017). IPR suite scans unigenes against protein signatures (predictive models) from 14 affiliated databases resulting in a comprehensive protein annotation that includes protein superfamily classification, specific protein domains, repeats, and sites. Based on the homology results from nr databases, GO terms were mapped to each unigene using a comprehensive annotation suite BLAST2GO (Conesa et al., 2005). IPR protein annotation signatures were also mapped to GO terms and finally an integrated annotation was obtained after merging it with blast-derived GO annotation. Metabolic pathway analysis was done by similarity searching using blastx against the KEGG database (Kanehisa et al., 2004). The resulting unigenes and enzyme code were assigned and mapped to their respective pathway map.

Identification of Transcription Factors (TFS)

Unigenes were scanned using Hidden Markov Model approach (HMM) in the HMMER software package against the Pfam, PlnTFDB, and PlantTFDB databases (Eddy, 2011; Finn et al., 2011, 2016). TFs were predicted according to the family assessment rule as described in PlantTFDB (Pérez-Rodríguez et al., 2009; Jin et al., 2014). The TFs retention criterion was set to an *E*-value less than 1e-05 and percentage identity of 50%.

Identification of Simple Sequence Repeats (SSRs)

To identify SSRs, we used microsatellite identification tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/). All the unigenes were screened to identify potential sites for SSRs. Criteria were set such as minimum repeat unit number of 5 for di- to hexa-nucleotide repeats and a minimum repeat overlapping length of 100 bp. Mono nucleotide repeats were not included because of the higher error rate of homo polymer variation. Compound SSRs were also predicted with additional parameters as described above.

Quantitative Real-Time PCR Analysis

For quantitative real time PCR (qRT-PCR), the total RNA was extracted from four biological replicates of *R. marginata*. cDNA was synthesized according to the manufacturer's recommendations (QuantiTect® Reverse Transcription, QIAGEN Group). qRT-PCR assay was performed, using the LightCycler® SYBR Green I Master (Roche Diagnostics, Germany). Four technical replicates were used for each sample and quantified based on relative expression levels and normalized against the housekeeping gene *Actin* of *R. marginata,* using LightCycler® Software (Roche; version 1.5). Gene specific primers used for the qRT-PCR are listed in the **Supplementary Table 9**.

HPLC ESI-MS/MS

One hundred milligrams of dried *R. marginata* material was grinded in 1 mL 80% methanol using a glass homogenizer. After centrifugation of the cell debris at 13,100 g at room temperature, the supernatant was analyzed by HPLC-ESI MS/MS analysis. Separation of the liverwort extract was performed by RP-HPLC (DAD) (Agilent 1260 Infinity HPLC, Waldbronn, Germany) equipped with a Poroshell 120 EC-C18, 2.1 × 100 mm, 2.7 µm column (Agilent, Waldbronn, Germany) with a flow rate of 0.7 mL min^{-1} at 40°C under isocratic conditions [50% (v/v) H_2O with 0.1% (v/v) FA, 50% (v/v) ACN]. The identity of the liverwort cannabinoids was confirmed by mass and tandem mass spectra using a Bruker compact™ ESI-Q-TOF (Bruker, Bremen, Germany) operating at a positive ionization mode. Fragmentation was achieved at collision energy of 20.0 eV. The methodology was described in detail (Zirpel et al., 2017).

GC-MS Analysis

Fifty mg of the liquid nitrogen-ground *Radula* tissue was suspended in 100 µL of hexane, vortexed vigorously, and sonicated for 30 min. The samples were centrifuged at 16,000 g, for 10 min. The resulting supernatant was collected and analyzed. The GC/MS analysis was performed using TRACE 1310 gas chromatograph connected to a TSQ8000 triplequad MS (both Thermo Scientific). A DB-5 bonded-phase fused-silica capillary column (30 m length, 0.25 mm inner diameter, 0.25 µm film thickness) (J&W Scientific Co., USA) was used for separation. The GC oven temperature program was as follows: 2 min at 70°C, raised by 10°C/min to 300°C, and held for 10 min at 300°C. The total time of GC analysis was 36 min. Helium was used as the carrier gas at a flow rate of 1 mL/min. One microliter of each sample was injected in splitless mode. The initial injector temperature was 40°C for 0.1 min and after that time raised by 600°C/min to 350°C. The septum purge flow rate was 3 mL/min and the purge was turned on after 60 s. The transfer line and ion source temperatures were set to 250°C. Ion-source fragmentation was performed with 70 eV energy. Mass spectra were recorded in the mass range 35–650 m/z.

Data acquisition, automatic peak detection, mass spectrum deconvolution, retention index calculation, and library search were done by XCalibur and AMDIS software. The metabolites were automatically identified by library search (NIST library); the analyte was considered as identified when it passed a quality threshold: i.e., similarity index (SI) above 700 and matching retention index ± 10. The artifacts (alkanes, column bleed, plasticizers, MSTFA, and reagents) were identified analogously, and then excluded from further analysis. To obtain accurate peak areas for the deconvoluted components, unique quantification masses for each component were specified and the samples were reprocessed. The obtained profiles were normalized against the sum of chromatographic peak area (using the TIC approach).

Structural Analysis

A STS structural model was built with SWISS-MODEL by comparative approach using PDB ID code 3awk and 2p0u as a template (Arnold et al., 2006; Biasini et al., 2014).

AUTHOR CONTRIBUTIONS

TH and OK: Wrote the first draft; TH, RE, and OK: Designed the experiment; TH and BP: Performed the experiments; TH and OK: Analyzed the data. TH, ME, RE, and OK: Wrote the manuscript. All authors contributed in the final version of the manuscript.

ACKNOWLEDGMENTS

This study was supported by the DISCO project that received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under the grant agreement 613513. Computational analyses were performed using Linux-HPC-Cluster at Technical University Dortmund (LiDong). We thank Mr. Sven Buijssen for his valuable help and repeated support regarding numerous aspects while using the compute cluster LiDong partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 49560908. We also thank Dr. Felix Stehle for ESI-MS/MS and Dr. Pawel Rodziewicz for GC/MS analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00537/full#supplementary-material

Supplementary Table 1 | Filtering summary of raw reads.

Supplementary Table 2 | Isoform per gene summary.

Supplementary Table 3 | Read mapping summary.

Supplementary Table 4 | Predicted ORFs summary.

Supplementary Table 5 | Enzymes identified related to secondary metabolites and "terpenoids and polykettides."

Supplementary Table 6 | Enzyme distribution according to subclass category.

Supplementary Table 7 | Comparative summary of identified TFs.

Supplementary Table 8 | Frequency of repeat number.

Supplementary Table 9 | Oligonucleotide primers used for qPCR.

Supplementary Data 1 | BLAST analysis summary.

Supplementary Data 2 | IntrerProScan protein signature list.

Supplementary Data 3 | List of gene ontologies (GO) identified from three categories.

Supplementary Data 4 | List of identified enzymes associated with genes.

Supplementary Data 5 | List of GC-MS detected compounds.

Supplementary Data 6 | List of transcription factors (TFs).

Supplementary Data 7 | List of simple sequence repeats (SSRs).

Supplementary Data 8 | List of non-coding RNAs (ncRNAs).

Supplementary Figure 1 | (A) Association of protein signatures for a specific transcript among PANTHER, GENE3D SMART, TIGERFAM, and SIGNALP protein databases. **(B)** Distribution of gene ontologies annotation from GO-levels 2 to 15 for each category of biological process (BP) molecular function (MF) and cellular components (CC). **(C)** Gene ontology distribution among three main categories of

molecular function (MF), biological process (BP), and cellular components (CC). Y-axis represents the number of genes for their respective function/process/component as on X-axis.

Supplementary Figure 2 | (A) Tandem mass spectra of the bibenzyl analog of cannabigerolic acid with m/z 395.214 **(B)** GC-MS chromatogram of *Radula marginata*.

REFERENCES

Alaba, S., Piszczalka, P., Pietrykowska, H., Pacak, A. M., Sierocka, I., Nuc, P. W., et al. (2015). The liverwort Pellia endiviifolia shares microtranscriptomic traits that are common to green algae and land plants. *New Phytol.* 206, 352–367. doi: 10.1111/nph.13220

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Andre, C. M., Hausman, J. F., and Guerriero, G. (2016). *Cannabis sativa*: the plant of the thousand and one molecules. *Front. Plant Sci.* 7:19. doi: 10.3389/fpls.2016.00019

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22, 195–201. doi: 10.1093/bioinformatics/bti770

Asakawa, Y., Hashimoto, T., Takikawa, K., Tori, M., and Ogawa, S. (1991). Prenyl bibenzyls from the liverworts *Radula perrottetii* and *Radula complanata*. *Phytochemistry* 30, 235–251. doi: 10.1016/0031-9422(91)84130-K

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Austin, M. B., Bowman, M. E., Ferrer, J. L., Schröder, J., and Noel, J. P. (2004). An aldol switch discovered in stilbene synthases mediates cyclization specificity of type III polyketide synthases. *Chem. Biol.* 11, 1179–1194. doi: 10.1016/j.chembiol.2004.05.024

Austin, M. B., and Noel, J. P. (2003). The chalcone synthase superfamily of type III polyketide synthases. *Nat. Prod. Rep.* 20, 79–110. doi: 10.1039/b100917f

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42, 252–258. doi: 10.1093/nar/gku340

Bohlmann, F., and Hoffmann, E. (1979). Cannabigerol-ähnliche verbindungen aus *Helichrysum umbraculigerum*. *Phytochemistry* 18, 1371–1374. doi: 10.1016/0031-9422(79)83025-3

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bouman, A. C., Dirkse, G. M., and Yamada, K. (1998). *Radula jonesii* spec. nov. (Hepaticae), a new species from Tenerife. *J. Bryol.* 15, 161–164. doi: 10.1179/jbr.1988.15.1.161

Bowman, J. L., Floyd, S. K., and Sakakibara, K. (2007). Green genes-comparative genomics of the green branch of life. *Cell* 129, 229–234. doi: 10.1016/j.cell.2007.04.004

Broun, P., Liu, Y., Queen, E., Schwarz, Y., and Leibman, A. (2006). Importance of transcription factors in the regulation of plant secondary metabolism and their relevance to the control of terpenoid accumulation. *Phytochem. Rev.* 5, 27–38. doi: 10.1007/s11101-006-9000-x

Buck, W. R. (1986). Introduction to bryology. *Brittonia* 38, 94–95. doi: 10.2307/2807430

Cloutier, S., NiuRaju, Z., and Duguid, D. (2009). Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). *Theor. Appl. Genet.* 119, 53–63. doi: 10.1007/s00122-009-1016-3

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Croteau, R., and Purkett, P. T. (1989). Geranyl pyrophosphate synthase: characterization of the enzyme and evidence that this chain-length specific prenyltransferase is associated with monoterpene biosynthesis in Sage (*Salvia officinalis*). *Arch. Biochem. Biophys.* 271, 524–535. doi: 10.1016/0003-9861(89)90304-4

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Eddy, S. R., and Hughes, H. (2001). Non-coding RNA genes. *Genetics* 2, 919–929. doi: 10.1038/35103511

Espley, R. V., Hellens, R. P., Putterill, J., Stevenson, D. E., Kutty-Amma, S., and Allan, A. C. (2007). Red colouration in apple fruit is due to the activity of the MYB transcription factor, MdMYB10. *Plant J.* 49, 414–427. doi: 10.1111/j.1365-313X.2006.02964.x

Feldmeyer, B., Wheat, C. W., Krezdorn, N., Rotter, B., and Pfenninger, M. (2011). Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC Genomics* 12:317. doi: 10.1186/1471-2164-12-317

Fellermeier, M., Eisenreich, W., Bacher, A., and Zenk, M. H. (2001). Biosynthesis of cannabinoids. *Eur. J. Biochem.* 268, 1596–1604. doi: 10.1046/j.1432-1327.2001.02030.x

Fellermeier, M., and Zenk, M. H. (1998). Prenylation of olivetolate by a hemp transferase yields cannabigerolic acid, the precursor of tetrahydrocannabinol. *FEBS Lett.* 427, 283–285. doi: 10.1016/S0014-5793(98)00450-5

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., et al. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190–D199. doi: 10.1093/nar/gkw1107

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, 29–37. doi: 10.1093/nar/gkr367

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344

Flemming, T., Muntendam, R., Steup, C., and Kayser, O. (2009). "Chemistry and biological activity of tetrahydrocannabinol and its derivatives," in *Topics in Heterocyclic Chemistry*, ed R. R. Gupta (Berlin; Heidelberg: Springer), 1–42. doi: 10.1007/7081_2007_084

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gachet, M. S., Schubert, A., Calarco, S., Boccard, J., and Gertsch, J. (2017). Targeted metabolomics shows plasticity in the evolution of signaling lipids and uncovers old and new endocannabinoids in the plant kingdom. *Sci. Rep.* 7:41177. doi: 10.1038/srep41177

Gao, B., Zhang, D., Li, X., Yang, H., Zhang, Y., and Wood, A. J. (2015). *De novo* transcriptome characterization and gene expression profiling of the desiccation tolerant following remoss *Bryum argenteum* hydration. *BMC Genomics* 16:416. doi: 10.1186/s12864-015-1633-y

Gao, B., Zhang, D., Li, X., Yang, H., and Wood, A. J. (2014). *De novo* assembly and characterization of the transcriptome in the desiccation-tolerant moss *Syntrichia caninervis*. *BMC Res. Notes* 7:490. doi: 10.1186/1756-0500-7-490

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J. F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12:245. doi: 10.1186/1471-2164-12-245

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hallingbäck, T., and Hodgetts, N. (2001). *Mosses, Liverworts and Hornworts*. IUCN/SSC Bryophyte Specialist Group.

Happyana, N., Agnolet, S., Muntendam, R., Van Dam, A., Schneider, B., and Kayser, O. (2013). Analysis of cannabinoids in laser-microdissected trichomes of medicinal *Cannabis sativa* using LCMS and cryogenic NMR. *Phytochemistry* 87, 51–59. doi: 10.1016/j.phytochem.2012.11.001

Hisano, H., Sato, S., Isobe, S., Sasamoto, S., Wada, T., and Matsuno, A. (2008). Characterization of the soybean genome using EST-derived microsatellite markers. *DNA Res.* 14, 271–281. doi: 10.1093/dnares/dsm025

Honkanen, S., Jones, V. A. S., Morieri, G., Champion, C., Hetherington, A. J., Kelly, S., et al. (2016). The mechanism forming the cell surface of tip-growing rooting cells is conserved among land plants. *Curr. Biol.* 26, 3238–3244. doi: 10.1016/j.cub.2016.09.062

Jeena, G. S., Fatima, S., Tripathi, P., Upadhyay, S., and Shukla, R. K. (2017). Comparative transcriptome analysis of shoot and root tissue of Bacopa monnieri identifies potential genes related to triterpenoid saponin biosynthesis. *BMC Genomics* 18:490. doi: 10.1186/s12864-017-3865-5

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* 42, 1182–1187. doi: 10.1093/nar/gkt1016

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., et al. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–D357. doi: 10.1093/nar/gkj102

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, 277–280. doi: 10.1093/nar/gkh063

Kumpatla, S. P., and Mukhopadhyay, S. (2005). Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48, 985–998. doi: 10.1139/g05-060

Lang, D., Eisinger, J., Reski, R., and Rensing, S. A. (2005). Representation and high-quality annotation of the *Physcomitrella patens* transcriptome demonstrates a high proportion of proteins involved in metabolism in mosses. *Plant Biol.* 7, 238–250. doi: 10.1055/s-2005-837578

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

La Rota, M., Kantety, R. V., Yu, J. K., and Sorrells, M. E. (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 6:23. doi: 10.1186/1471-2164-6-23

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, C. F., Zhu, Y., Yu, Y., Zhao, Q., Wang, S., Wang, X., et al. (2015). Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (*Camellia sinensis*). *BMC Genomics* 16:560. doi: 10.1186/s12864-015-1773-0

Li, M. Y., Wang, F., Jiang, Q., Ma, J., and Xiong, A. S. (2014). Identification of SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens* L.) by deep transcriptome sequencing. *Hortic. Res.* 1:10. doi: 10.1038/hortres.2014.10

Liang, C., Liu, X., Yiu, S. M., and Lim, B. L. (2013). *De novo* assembly and characterization of *Camelina sativa* transcriptome by paired-end sequencing. *BMC Genomics* 14:146. doi: 10.1186/1471-2164-14-146

Liu, J., Osbourn, A., and Ma, P. (2015). MYB transcription factors as regulators of phenylpropanoid metabolism in plants. *Mol. Plant* 8, 689–708. doi: 10.1016/j.molp.2015.03.012

Losada-Lima, A., Rodriguez-Nunez, S., and Dirkse, G. M. (2007). Bibliographical references on the bryophyte flora of the Canary Islands. *Arch. Biol.* 24, 1–27. Available online at: http://www.archive-for-bryology.com/Archive%2024.pdf

Ludwiczuk, A., and Asakawa, Y. (2008). Distribution of terpenoids and aromatic compounds in selected southern hemispheric liverworts. *Fieldiana Bot.* 47:37. doi: 10.3158/0015-0746-47.1.37

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* 17:10. doi: 10.14806/ej.17.1.200

Mattick, J. S., and Mattick, J. S. (2010). The functional genomics of noncoding RNA. *Science* 1527, 1527–1529. doi: 10.1126/science.1117806

Metzgar, D., Bytof, J., and Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* 10, 72–80. doi: 10.1101/gr.10.1.72

Moazed, D. (2009). Small RNAs in transcriptional gene silencing and genome defence. *Nature* 457, 413–420. doi: 10.1038/nature07756

Morgante, M., Hanafey, M., and Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30, 194–200. doi: 10.1038/ng822

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1350. doi: 10.1126/science.1158441

Oliver, M. J., Dowd, S. E., Zaragoza, J., Mauget, S. A., and Payton, P. R. (2004). The rehydration transcriptome of the desiccation-tolerant bryophyte *Tortula ruralis*: transcript classification and analysis. *BMC Genomics* 5:89. doi: 10.1186/1471-2164-5-89

Ortiz-Ramírez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., et al. (2016). A transcriptome atlas of *Physcomitrella patens* provides insights into the evolution and development of land plants. *Mol. Plant* 9, 205–220. doi: 10.1016/j.molp.2015.12.002

Park, B. H., and Lee, Y. R. (2010). Concise synthesis of perrottetinene with bibenzyl cannabinoid. *Bull. Korean Chem. Soc.* 31, 2712–2714. doi: 10.5012/bkcs.2010.31.9.2712

Pérez-Rodríguez, P., Riaño-Pachón, D. M., Corrêa, L. G., Rensing, S. A., Kersten, B., and Mueller-Roeber, B. (2009). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.* 38, 822–827. doi: 10.1093/nar/gkp805

Qiu, Q., Ma, T., Hu, Q., Liu, B., Wu, Y., Zhou, H., et al. (2011). Genome-scale transcriptome analysis of the desert poplar, *Populus euphratica*. *Tree Physiol.* 31, 452–461. doi: 10.1093/treephys/tpr015

Qiu, Y., Li, L., Wang, B., Chen, Z., Dombrovska, O., Lee, J., et al. (2007). A nonflowering land plant phylogeny inferred from nucleotide sequences of seven chloroplast, mitochondrial, and nuclear genes. *Int. J. Plant Sci.* 168, 691–708. doi: 10.1086/513474

Renner, M. A. M., Devos, N., Brown, E. A., Orme, A., Elgey, M., Wilson, T. C., et al. (2013). Integrative taxonomy resolves the cryptic and pseudo-cryptic. *PhytoKeys* 113, 1–113. doi: 10.3897/phytokeys.27.5523

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., et al. (2008). The physcomitrella genome reveals evolutionary insights into the conquest of land by plants stefan. *Science* 319, 64–69. doi: 10.1126/science.1150646

Rensing, S. A., Rombauts, S., Hohe, A., Lang, D., Duwenig, E., Rouze, P., et al. (2002). The transcriptome of the moss *Physcomitrella patens*: comparative analysis reveals a rich source of new genes. *Plant Biotech* 1–18. Available online at: http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf

Rubinstein, C. V., Gerrienne, P., de la Puente, G. S., Astini, R. A., and Steemans, P. (2010). Early middle ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytol.* 188, 365–369. doi: 10.1111/j.1469-8137.2010.03433.x

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., et al. (2015). A higher level classification of all living organisms. *PLoS ONE* 10:e0119248. doi: 10.1371/journal.pone.0119248

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., and Burleigh, J. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* 14:23. doi: 10.1186/1471-2148-14-23

Russo, E. B. (2016). Beyond cannabis: plants and the endocannabinoid system. *Trends Pharmacol. Sci.* 37, 594–605. doi: 10.1016/j.tips.2016.04.005

Sayadi, A., Immonen, E., Bayram, H., and Arnqvist, G. (2016). The *de novo* transcriptome and its functional annotation in the seed beetle callosobruchus maculatus. *PLoS ONE* 11:e158565. doi: 10.1371/journal.pone.0158565

Schliesky, S., Gowik, U., Weber, A. P., and Bräutigam, A. (2012). RNA-seq assembly – are we there yet? *Front. Plant Sci.* 3:220. doi: 10.3389/fpls.2012.00220

Sharma, N., Jung, C. H., Bhalla, P. L., and Singh, M. B. (2014). RNA sequencing analysis of the gametophyte transcriptome from the liverwort, Marchantia polymorpha. *PLoS ONE* 9:e97497. doi: 10.1371/journal.pone.0097497

Singh, H., Rai, K. M., Upadhyay, S. K., Pant, P., Verma, P. C., Singh, A. P., et al. (2015). Transcriptome sequencing of a thalloid bryophyte; *Dumortiera hirsuta* (Sw) Nees: assembly, annotation, and marker discovery. *Sci. Rep.* 5:15350. doi: 10.1038/srep15350

Sirikantaramas, S., Morimoto, S., Shoyama, Y., Ishikawa, Y., Wada, Y., Shoyama, Y., et al. (2004). The gene controlling marijuana psychoactivity. *J. Biol. Chem.* 279, 39767–39774. doi: 10.1074/jbc.M403693200

Soethoudt, M., Grether, U., Grim, T. W., Fezza, F., Perret, C., Gils, N., et al. (2017). Cannabinoid CB_2 receptor ligand profiling reveals biased signalling and off-target activity. *Nat. Commun.* 8:13958. doi: 10.1038/ncomms13958

Spyropoulou, E. A., Haring, M. A., and Schuurink, R. C. (2014). RNA sequencing on Solanum lycopersicum trichomes identifies transcription factors that activate terpene synthase promoters. *BMC Genomics* 15:402. doi: 10.1186/1471-2164-15-402

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Abril, J. F., Akerman, M., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi: 10.1038/nmeth.2714

Stotler, R., and Crandall-Stotler, B. (1977). A checklist of the liverworts and hornworts of North America. *Bryologist* 80, 405–428. doi: 10.2307/3242017

Stout, J. M., Boubakir, Z., Ambrose, S. J., Purves, R. W., and Page, J. E. (2012). The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in *Cannabis sativa* trichomes. *Plant J.* 71, 353–365. doi: 10.1111/j.1365-313X.2012.04949.x

Styrczewska, M., Kulma, A., Ratajczak, K., Amarowicz, R., and Szopa, J. (2012). Cannabinoid-like anti-inflammatory compounds from flax fiber. *Cell. Mol. Biol. Lett.* 17, 479–499. doi: 10.2478/s11658-012-0023-6

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

Tisserand, R., and Young, R. (2014). *Essential Oil Safety*. Churchill Livingstone Elsevier. Available online at: https://www.elsevier.com/books/essential-oil-safety/tisserand/978-0-443-06241-4

Toyota, M., Kinugawa, T., and Asakawa, Y. (1994). Bibenzyl cannabinoid and bisbibenzyl derivative from the liverwort *Radula perrottetii*. *Phytochemistry* 37, 859–862. doi: 10.1016/S0031-9422(00)90371-6

Toyota, M., Shimamura, T., Ishii, H., Renner, M., Braggins, J., and Asakawa, Y. (2002). New bibenzyl cannabinoid from the New Zealand liverwort *Radula marginata*. *Chem. Pharm. Bull.* 50, 1390–1392. doi: 10.1248/cpb.50.1390

Tropf, S., Lanz, T., Rensing, S. A., Schröder, J., and Schröder, G. (1994). Evidence that stilbene synthases have developed from chalcone synthases several times in the course of evolution. *J. Mol. Evol.* 38, 610–618. doi: 10.1007/BF00175881

Verk, M. C., Van Hickman, R., and Van Wees, S. C. M. (2013). RNA-Seq: revelation of the messengers. *Trends Plant Sci.* 18, 175–179. doi: 10.1016/j.tplants.2013.02.001

Vom Endt, D., Kijne, J. W., and Memelink, J. (2002). Transcription factors controlling plant secondary metabolism: what regulates the regulators? *Phytochemistry* 61, 107–114. doi: 10.1016/S0031-9422(02)00185-1

Wei, W., Qi, X., Wang, L., Zhang, Y., Hua, W., Li, D., et al. (2011). Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics* 12:451. doi: 10.1186/1471-2164-12-451

Willis, K. J., and McElwain, J. (2002). *The Evolution of Plants*. Oxford: Oxford University Press.

Wood, A. J., and Duff, R. J. (2009). The aldehyde dehydrogenase (ALDH) gene superfamily of the moss *Physcomitrella patens* and the algae *Chlamydomonas reinhardtii* and *Ostreococcus tauri*. *Bryologist* 112, 1–11. doi: 10.1639/0007-2745-112.1.1

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., et al. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* 20, 646–654. doi: 10.1101/gr.100677.109

Zhang, J., and Huang, L. (2016). *De novo* transcriptome analysis and molecular marker development of two hemarthria species. *Front. Plant Sci.* 7:496. doi: 10.3389/fpls.2016.00496

Zhu, X., Leng, X., Sun, X., Mu, Q., Wang, B., Li, X., et al. (2015). Discovery of conservation and diversification of genes by phylogenetic analysis based on global genomes. *Plant Genome* 8, 1–11. doi: 10.3835/plantgenome2014.10.0076

Zimmer, A. D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., et al. (2013). Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genomics* 14:498. doi: 10.1186/1471-2164-14-498

Zirpel, B., Degenhardt, F., Martin, C., Kayser, O., and Stehle, F. (2017). Engineering yeasts as platform organisms for cannabinoid biosynthesis. *J. Biotechnol.* 259, 204–212. doi: 10.1016/j.jbiotec.2017.07.008

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.