Expression of presumptive Microbial terpene synthase like genes (MTPSLs) from *Lophocolea bidentata* (liverwort) in moss *Physcomitrella patens*

Author
Satish Kumar Kodiripaka

Supervisors

Henrik Toft Simonsen
Associate Professor

Mette Lübeck
Associate Professor

In collaboration with
Department of Biotechnology and Biomedicine
Technical University of Denmark

Thesis submitted in partial fulfillment of the requirements for the degree of
MSc in Engineering (Sustainable Biotechnology)

Aalborg University (Copenhagen)
Section for Sustainable Biotechnology

June 2020

Abstract

Terpenoids are structurally and functionally diverse natural compounds that are voluminous and more divergent in plants and less abundant in microbes. Terpene synthases are the enzymes that proprietarily produce broad range of terpene metabolites in nature by catalyzing the conversion of prenyl diphosphate precursor molecules. Putative microbial terpene synthase like genes (MTPSLs) were identified predominantly from non-seed plants but not in seed plants. This study is aimed at heterologous expression and biochemical characterization of presumptive microbial terpene synthase like genes (LbMTPSL1,3,4and 5) from untapped liverwort species *Lophocolea bidentata* in model plant *Physcomitrella patens* by exploiting its indispensable natural in vivo homologous recombination machinery. A new promising protocol using a blend of enzymes Cellulase R-10 and Macerozyme R-10 was utilized for protoplast preparation from *P. patens*. Homology based structural models for LbMTPSL1,4 have closest structural similarity to Epi-isozizaene synthase from *Streptomyces coelicolor* and LbMTPSL3,5 have identical stereoview as that of Selinadiene synthase from *Streptomyces pristinaespiralis* both belonging to C_{15} subclass, sesquiterpene synthases. Structure based function analysis predicted the active site binding affinity to Farnesyl diphosphate (FPP) for LbMTPSL1,3,5 and its analogue FsPP (Farnesyl thiopyrophosphate) for LbMTPSL4. The three basic residues along with phenol containing amino acid involved in substrate recognition motif and three aromatic residues with cation-π interactions for stabilizing carbocation intermediates were identified from the active site contour of all enzymes. It is believed that this knowledge on structural and chemical biology of LbMTPSLs built on the basis of computational data analysis could be sourceful for future mechanistic characterization and subsequent manipulations of FPP cyclization trajectory by rational design approach to alter the product profile significantly.

Table of Contents

List of Tables

List of Figures

Abbreviations

IPP	Isopentenyl diphosphate
DMAPP	Dimethylallyl diphosphate
MVA	Mevalonic acid
MEP	Methyl erythritol 4 phosphate
GPP	Geranyl diphosphate
GPS	Geranyl phosphate synthase
FPP	Farnesyl diphosphate
FsPP	Farnesyl thiopyrophosphate
FPS	Farnesyl phosphate synthase
GGPP	Geranyl geranyl diphosphate
GGPPS	Geranyl geranyl phosphate synthase
DXS	1-deoxy-D-xylulose-5-phosphate synthase
DXR	1-Deoxy-D-xylulose 5-phosphate reductoisomerase
HDR	Hydroxymethylbutenyl diphosphate reductase
HMGR	3-hydroxy-3-methyl-glutaryl-CoA reductase
TPSs	Terpene synthases
SHC	Squalene hopene cyclase
CPS	Copalyl diphosphate synthase
MTPSLs	Microbial terpene synthase like genes/proteins
CRE	Cis-regulatory elements
CaMV	Cauli flower mosaic virus
OCS	Octopine synthase
Pfu	*Pyrococcus furiosus*
PCR	Polymerase chain reaction
MES	2-(N-Morpholino)-ethanesulfonic acid
PEG	Polyethylene glycol
SPME	Solid phase microextraction
PDB	Protein data bank
RMSD	Root mean square deviation
GC-MS	Gas chromatography-Mass spectrometry
HS-SPME	Head space-solid phase microextraction
RT	Room temperature
I-TASSER	Iterative threading assembly refinement
dsDNA	Double stranded deoxyribonucleic acid

1. INTRODUCTION

The origin and evolution of green plants on earth have introduced new renewable energy in the form of organic compounds that constitutes principle energy source in aerobic and anaerobic energy metabolisms of most terrestrial biotic species. Plants are photoautotrophs that fix atmospheric CO_2 in the presence of light and water into primary metabolites required for growth and development. The emergence of terrestrial plants has transformed the earths biosphere into complex ecosystems comprising ecological interactions with other organisms-animals, microorganisms and other parasitic plants which exquisitely depend on plants for their primary metabolites concurrently plants through symbiotic association acquire optimal growth and able to complete life cycles. As part of selection in order to cope with various biotic and abiotic stresses that the plants underwent during their interactions, they have evolved an array of biochemical pathways that can synthesize secondary metabolites to combat with specific ecological problems. Since different plant species make trophic levels in their own ecological niches, the secondary metabolites are not commonly shared but rather are unique and lineage specific. Majority of terpenes are classified as secondary metabolites due to their aid in adaptation.

Terpenoids

Terpenoids are structurally and functionally diverse natural compounds that are voluminous and more divergent in plants and less abundant in microbes. These secondary metabolites were not detected in the charophyte green algae, a common ancestor of embryophytes and are widely distributed in all land plant lineages signifying their key role in adaptation of embryophytes to the terrestrial habitat (Chen et al., 2018). The heterogeneity in the distribution of terpenoids across plant kingdom is rational by their abundance & diversity in seed plants (angiosperms & gymnosperms) and in liverworts of non-seed plants (Jia et al., 2016). As plants in the process of adaptive selection are confronted with new adaptive responses from other organisms to already existing secondary metabolites, they in turn produce more promising secondary metabolites that confer increased fitness. As a result a significant number of 58,091 terpenoids encompassing mono, di and sesquiterpenoids were registered hitherto (Banerjee et al., 2019). These organic compounds in plants confer resistance against environmental and biotic stresses, attract pollinators and useful insects

through volatile molecules, serve as repellants to herbivores and pathogens (**Figure 1**;Weitzel & Simonsen, 2015). These large group of plant secondary metabolites enriched with bioactive properties were exploited by food, cosmetic and pharmaceutical industries commercially as flavours, fragrances and drugs (Zhan et al., 2014). Terpenoids or Isoprenoids are classified based on the number of isoprene(C_5) units as Hemiterpenes(C_5), Monoterpenes(C_{10}), Sesquiterpenes(C_{15}), Diterpenes(C_{20}) (Rodríguez-Concepción, 2014). Functional classification categorize Isoprenoids in to primary metabolites which includes phytohormones, photosynthetic pigments and components, while the second group constitutes a myriad of Isoprenoids called secondary metabolites that serves in defense, attract pollinators and cope with stress tolerance in plants (Rodríguez-Concepción, 2014). Metabolite synthesis of these organic molecules is either sequestered in specialized structures like oil glands and resin ducts or synthesized ubiquitously in all tissues in some plants (Weitzel & Simonsen, 2015). Since most of the terpene metabolites confront with stress, herbivores and pathogens, it is imperative that the gene expression can be induced with stress and regulation seems to take place at transcriptional level(Weitzel & Simonsen, 2015).

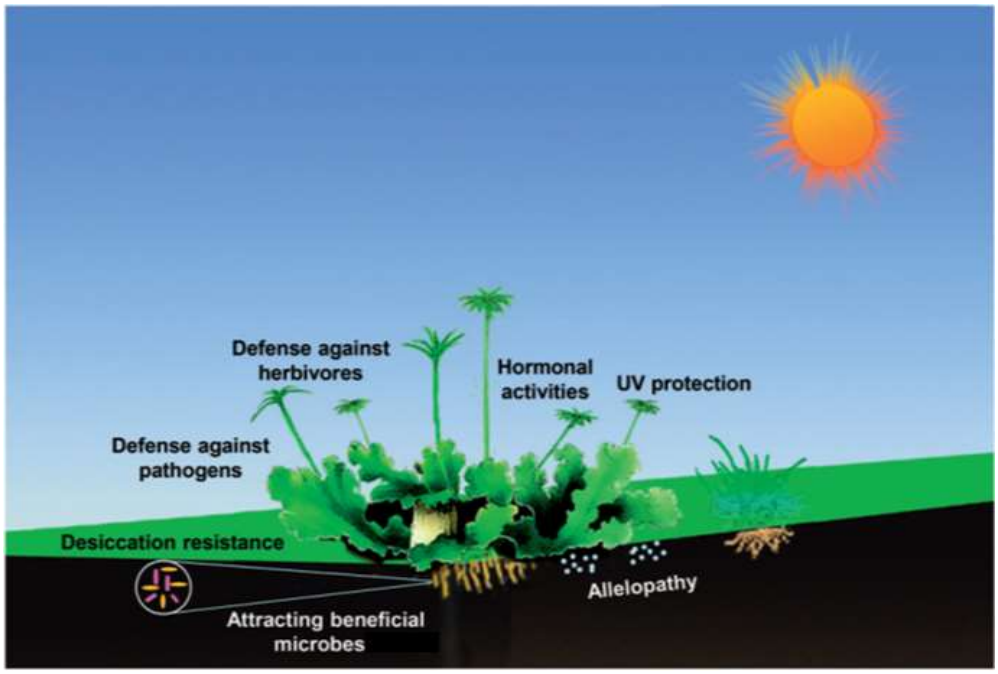

Figure 1. Biological functions of terpenoids in bryophyte-environment interactions. (Chen et al., 2018)

Biosynthesis of terpenoids-precursors & substrates

Despite of huge structural and functional diversity of isoprenoids in nature, It is ironical that the biosynthesis of terpenoids is originated from a single C_5 precursor bio bricks (isoprene units) called Isopentenyl diphosphate (IPP) and its double bond isomer Dimethylallyl diphosphate (DMAPP) (Ikramt et al., 2015). Two localized biosynthetic cascades for IPP were deciphered in different compartments namely cytosolic Mevalonic acid pathway(MVA) in fungi, animals and plants and plastidic Methyl erythritol 4 phosphate pathway(MEP) in most eubacteria and plants (Rodrı, 2002). The terpenoid biosynthesis begins with a head to tail condensation of one or more IPP and/or DMAPP biobricks in prenyl transferase reactions catalyzed by specific prenyl transferase enzymes (Chen et al., 2011). The condensation of one DMAPP with one IPP gives Geranyl diphosphate (GPP) by the activity of Geranyl phosphate synthase (GPS). The GPP fuses with one IPP catalyzed by Farnesyl phosphate synthase (FPS) to form Farnesyl diphosphate unit (FPP). The FPP is condensed with one IPP in the presence of enzyme Geranyl geranyl phosphate synthase (GGPPS) to form one Geranyl geranyl diphosphate molecule **(Figure 2)**. Less abundant longer prenyl diphosphate molecules of increased chain length are also synthesized. All the above prenyl diphosphate molecules in their trans configuration serve as starting material for the production of diverse terpenoids(Chen et al., 2011). GPP is the precursor for monoterpenes made of two isoprene units (C_{10}). Likewise, sesquiterpenes (C_{15}) and diterpenes (C_{20}) are derivatives of FPP and GGPP respectively. Due to the localization of isoprenoid biogenesis pathways, the MEP pathway enzymes transcribed from genomic DNA are transited to plastids and MVA pathway enzymes are present in cytosol along with other subcellular organelles like endoplasmic reticulum and peroxisomes (Rodríguez-Concepción, 2014). Similarly, prenyl transferases are also present innately in different compartments e.g. In Arabidopsis most of GPP and GGPP synthases are plastidial while few functional GGPP synthases are also seen in endoplasmic reticulum and mitochondria where as FPP synthases are encompassed in cytosol and mitochondria conversely peroxisomal and plastidic isoforms are reported in other plants (Rodríguez-Concepción, 2014). Some labelling experiments have reported that there has been transport and exchange of isoprene units and few downstream metabolites sparsely across compartments resulting in hybrid isoprenoid metabolites polymerized of isoprene biobricks derived from both MVA and MEP pathways (Lichtenthaler, 1999).

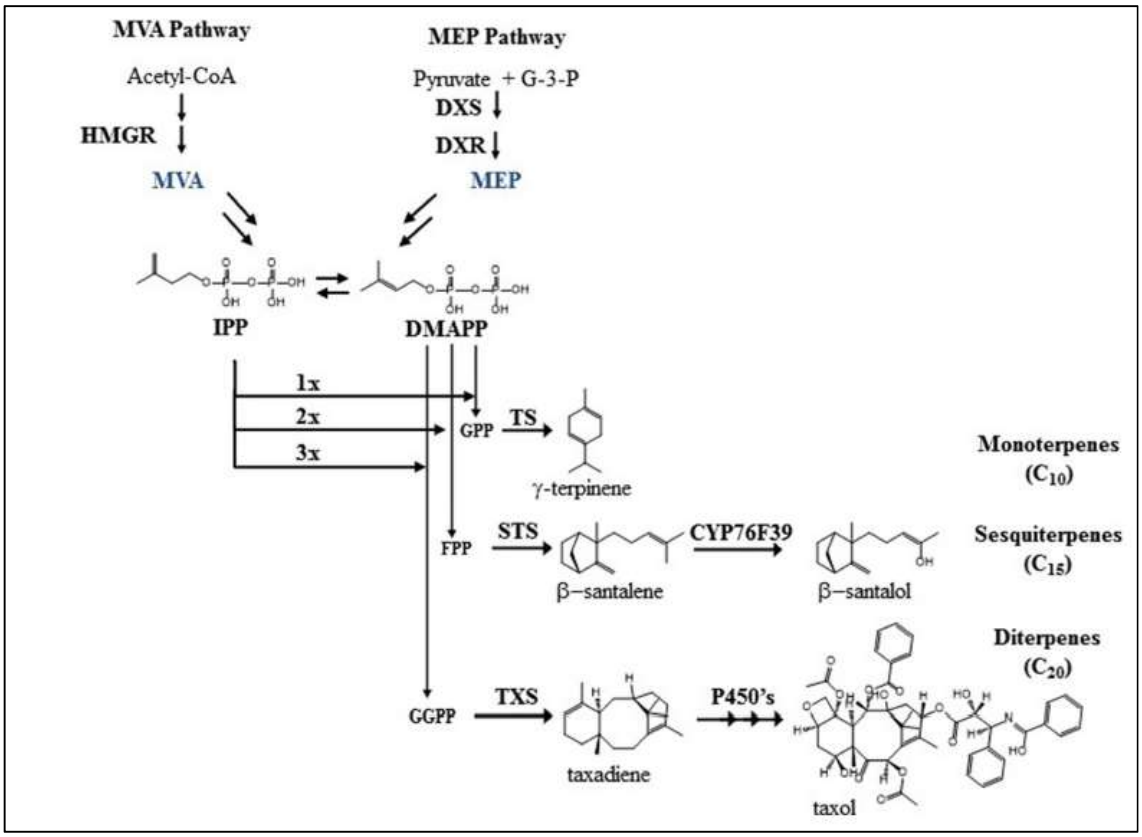

Figure 2. Elucidation of MVA & MEP pathways, synthesis of prenyl diphosphate substrates & some subsequent terpenoids.(Ikramt et al., 2015)

Metabolic flux regulation for terpenoid synthesis

Over expression studies have unraveled the rate determining genes which helped to increase the isoprene flux towards the precursors of specialized metabolites through pathway engineering. For the MEP pathway, the enzyme 1-deoxy-D-xylulose-5-phosphate synthase (DXS) and 1-Deoxy-D-xylulose 5-phosphate reductoisomerase (DXR) were found to have a limiting role. The results support universal regulatory role for DXS while the contribution of DXR to the control of flux tend to be less clear and depend on species, organ and/or developmental stage (Rodrı, 2002). Among the other enzymes of MEP pathway, the upregulation studies of Hydroxymethylbutenyl diphosphate reductase (HDR) encoding gene deciphered its major role in controlling the production of MEP derived isoprene units for plastid originated terpenoid biosynthesis (Botella-Pavía et al., 2004). For the MVA pathway, the upregulation of 3-hydroxy-3-methyl-glutaryl-CoA reductase (HMGR) gene led to

increased production of prenyl diphosphate substrate molecues which is evident by 5 fold increase in amorphodiene levels in engineered yeast (Ro et al., 2006) and a significant accumulation of 50 fold amorphodienes when an additional HMGR is integrated in to a yeast chromosome (Ro et al., 2006). Down regulation or knock down of genes that encode enzymes like squalene synthase that compete for prenyl diphosphate precursors is another approach to prevent sterol biosynthetic pathways (Engels et al., 2008; Ro et al., 2006). Engineering promoters and transcription factors that facilitate fine tuning of gene expression is also an emerging strategy. Upregulation of prenyl transferase genes like GPP, FPP, GGPP synthases also contribute in increasing terpenoid production by supplying specific precursor molecules (Ro et al., 2006; Takahashi et al., 2007).

Terpene synthases-Catalytic domains

Terpene synthases (TPSs) are the enzymes that proprietarily produce terpene metabolites in nature by catalyzing the conversion of prenyl diphosphate precursor molecules into a broad range of terpenoids. Terpene synthases is a midsize gene family in plants with the number varying approximately from 20 to 150 corresponding genes in several sequenced and annotated plant genomes though the model genome of *P. patens* with one TPS gene is not in consensus (Chen et al., 2011). The Terpene synthase family is categorized in to monoterpene, sesquiterpene and diterpene synthases based on their catalytic functions of producing cyclic and acyclic monoterpene, sesquiterpene and diterpene hydrocarbons and alcohols respectively (Chen et al., 2011; Jia et al., 2018). Recent gene and enzyme characterization experiments have demonstrated the existence of multi substrate terpene synthases in nature and the first multi substrate utilization capacity of TPSs was discerned from *Mentha x pipperita* Farnesene synthase, a Sesquiterpene synthase that uses FPP as substrate can also use GPP to produce cyclic and acyclic monoterpenes like limonene and congruently there are also evidences of sesquiterpenes in plastids (Pazouki & Niinemetst, 2016).

The first X-ray crystal structures of Terpene synthases have decoded three types of catalytic domains α, β & γ. In nature, TPSs exist in different combinations of these domains like only α, αβγ, αβ, βγ leading to a domain diversity (**Figure 4**;Christianson, 2017). Typical plant terpene synthases are composed of an assembly of either αβγ type or αβ type domains. Based on the reaction mechanisms employed by terpene synthases during initial carbocation

formation, they are categorized into class I, class II and bifunctional enzymes. Class I TPSs catalyze cyclization of prenyl diphosphates through metal ion dependent ionization induced carbocation formation, conversely class II terpene cyclases induces protonation of terminal C-C double bond of isoprenoid substrate yielding tertiary carbocation which undergoes a cascade of cyclizations and termination by deprotonation (Chen et al., 2011; Christianson, 2017). Bifunctional enzymes use both strategies. These carbocation intermediates of transition states are stabilized by weak polar interactions with spatially oriented amino acid sidechains and cation-π interactions with aromatic side chains of phenylalanine, tyrosine and tryptophan residues in the active site (Christianson, 2017). The active site of class I enzymes is localized in the middle of clove of α-helices and is characterized by the presence of two conserved metal binding motifs in the C-terminal α domain, the first being aspartate rich DDXXD (**Figure 3**) motif on D-helix and the second is NSE/DTF motif on H-helix (Chen et al., 2011; Christianson, 2017; Jia et al., 2018). The active site of class II enzymes is tailored at the interface between β and γ domains in an intact βγ and αβγ domain architectures and is characterized by the presence of aspartate rich DXDD motif (**Figure 3**;Cao et al., 2010; Chen et al., 2011; Christianson, 2017). From the above classification, monofunctional enzymes are found as class Iα, class I αβ, class I αβγ architectures or class II βγ, class II αβγ cyclases and bifunctional enzymes are found with αα (class I-class I) and αβγ (class I-class II) domain architectures (Christianson, 2017). It has been reported that bifunctional enzyme Abitadiene synthase (ABS) with αβγ domain architecture is only functional when intact and correctly folded while the domains failed to retain function when expressed individually unlike βγ or α domain containing terpene cyclases (**Figure 3B**;Cao et al., 2010)**.**

Terpene synthases of bacteria and fungi are also composed of α, β and γ domains with similar catalytic functions to those of plant TPSs. Fungi possess αβγ type, α only type and α plus IDS (Isopentenyl diphosphate synthase) type domains while bacteria is also poised with three types of TPSs: βγ type, α only type, α-α type (Jia et al., 2018). The most prevalent is α only type in both bacteria and fungi with characteristic DDxxD/E and NSE motifs (**Figure S1**). Though the structural homology, similar catalytic and reaction mechanisms of bacterial and fungal TPSs with plant TPSs is conspicuous, it is astounding that they share very low sequence similarity and can be phylogenetically separated in to two diverse groups (Jia et al., 2018).

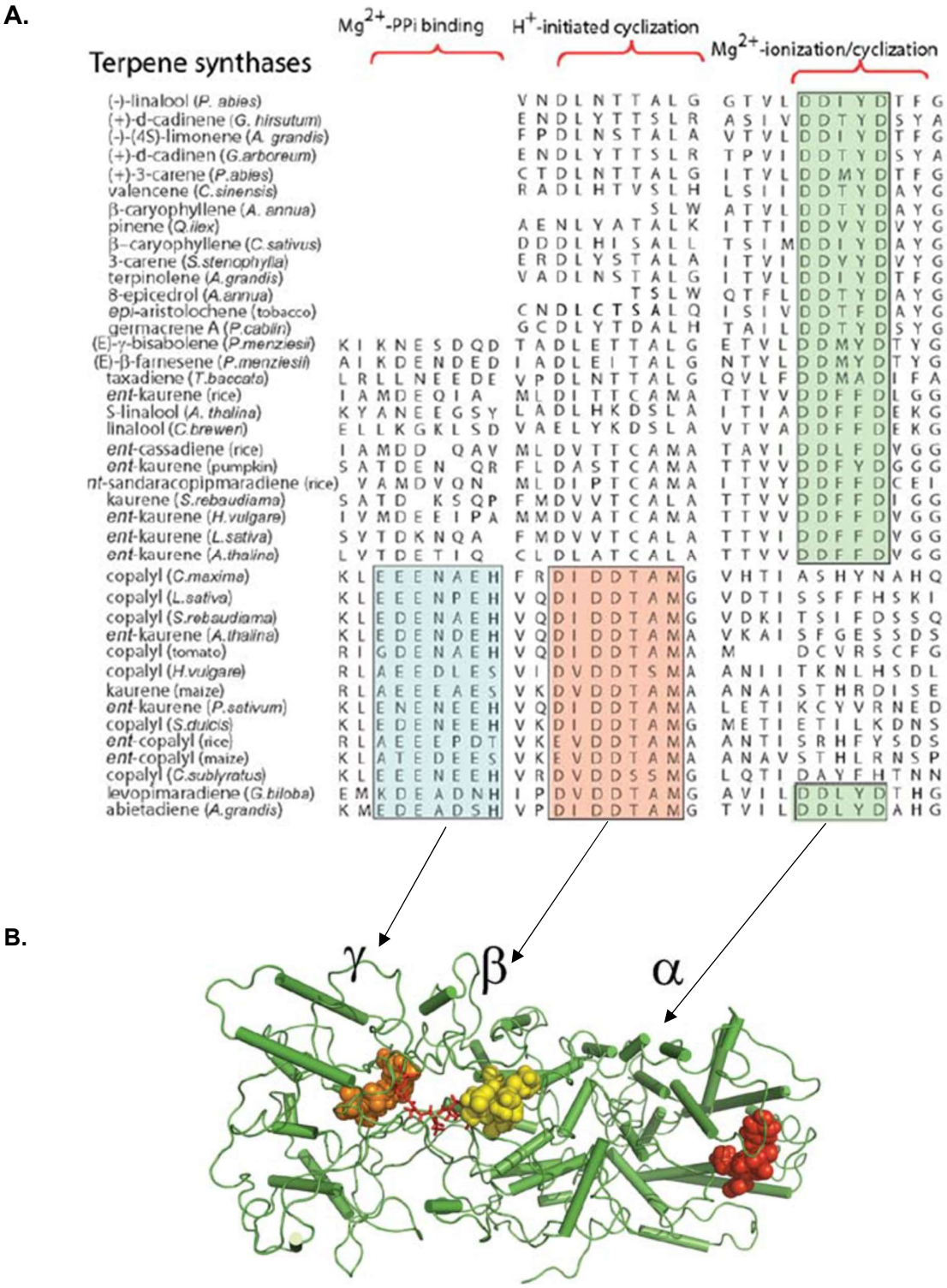

Figure 3. (A) Partial sequence alignment of terpene synthases displaying DDxxD(green) motif in the α domain, DxDD(orange) residues in general in β domain and D/E(cyan) rich region in γ domain. (B) Structural model of *Abies grandis* ABS displaying DDXXD (red), DXDD (yellow), EDxxD (orange) motifs with substrate GGPP (stick structure). (Cao et al., 2010)

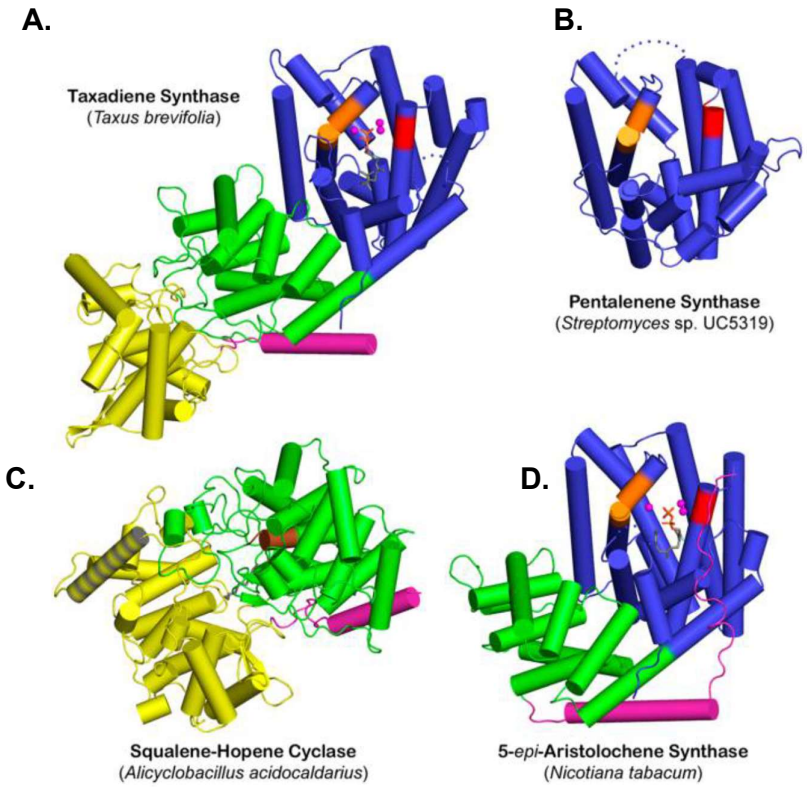

Figure 4. Domain diversity: Terpene synthases in various combinations of α (blue), β (green), γ (yellow) domains. (A) αβγ domain assembly in taxadiene synthase from Pacific Yew. (B) The single α domain featured with class I terpenoid synthase fold in bacterial Pentalenene synthase. (C) Bacterial Squalene hopene cyclase representing βγ domain assembly with class II terpenoid synthase active site. (D) αβ domain architecture shown in tobacco epi-Aristolochene synthase. (Christianson, 2017)

Evolution of terpene synthases

Elucidation of key structural features of ancient terpene synthases using experimental and computational methods have helped in rational understanding of evolution of modern TPSs. It has been proposed that the βγ domain architecture of class II bacterial di terpene synthases is similar to bacterial tri terpene cyclases based on properties: (1) the aspartate rich DXDD catalytic motif in active site (2) the number of α helices (~21±2) and β loops (~23±4) (3) presence of highly conserved identical QW repeats such as QxxDGGWG on aligning sequences (4) susceptibility to enzyme inhibitors (Cao et al., 2010). These obsevations are also applicable to plant diterpene synthases. Triterpene cyclases tend to be primitive to diterpene cyclases as triterpene cyclase mediated Squalene hopene cyclase (SHC) products

and hopanoids are detected in ancient sediments while diterpene products appear more recently in timescale of chemical fossils from which it can be inferred that ancestral class II triterpene cyclases gave rise to modern class II diterpene cyclases (Cao et al., 2010). The origin of class II diterpene synthases in plants could be by acquisition of genes from soil dwelling bacteria like *Bacillus japonicum* and *Rhizobium* that are in close association with plants through multiple horizontal gene transfer events and the acquired DNA is integrated in to the host genome (Cao et al., 2010). The ancient terpene synthases in early land plants are of tri domain αβγ architecture type *eg.* bifunctional Copalyl diphosphate synthase (CPS) enzyme in the bryophyte *P. patens* is a αβγ cyclase. It is hypothesized that the primitive αβγ cyclase in early plants is formed by the fusion of ancient α and βγ terpene synthases (Cao et al., 2010) and the downstream evolution of simplified diterpene, sesqui and mono terpene synthases is explained by the exon loss of γ domain, loss of transit peptide and subsequent recombinations (**Figure 5**;Cao et al., 2010; Pazouki & Niinemetst, 2016).

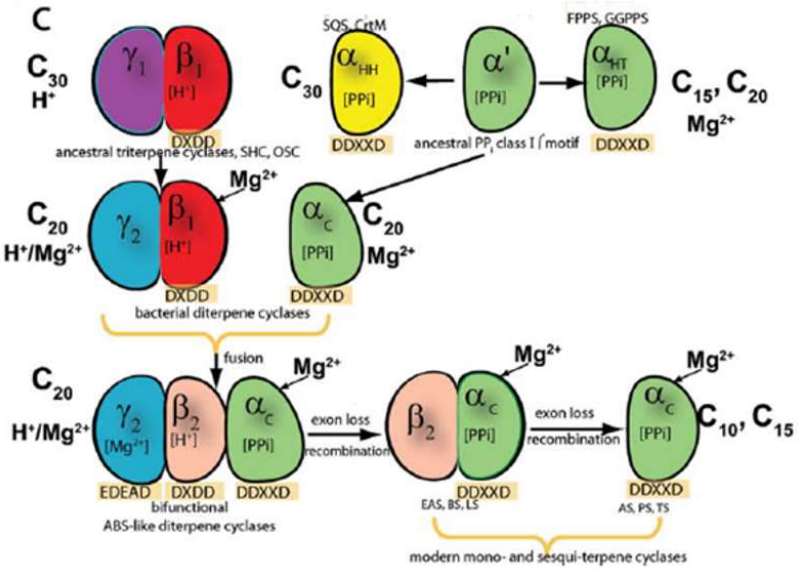

Figure 5. Schematic representation of hypothesis for evolution of terpene synthases in plants. (Cao et al., 2010)

Microbial terpene synthase like proteins/genes (MTPSLs)

Microbial terpene synthase like proteins are newfound class of terpene synthases in plants and the nomenclature is attributed to their phylogenetic close relationship with microbial terpene synthases. Though terpene synthases from bacteria and fungi are distantly related

to typical plant TPSs, MTPSL genes were identified in the genomes of early land plants. Genes similar to microbial terpene synthase genes were first identified in the lycophyte *Selaginella moellendorffii* whose encoded product portfolio included monoterpenes and sesquiterpenes (Jia et al., 2016). The comprehensive study of distribution of MTPSLs across plant kingdom carried out by transcriptome analysis have revealed that 99.2% of MTPSLs found are from non-seed plants (Jia et al., 2016). The five different lineages of non-seed plants include liverworts, hornworts, mosses, monilophytes and lycophytes. Among 166 species of non-seed plants, MTPSLs have been identified with precision from transcriptomes of 143 species encompassing 24 liverworts species, 30 moss species, 3 hornworts species, 21 lycophytes species and 65 species of ferns while intriguingly their existence is extremely low in seed plants (only 2 off 779 species), charophytes (only 1 of 47 species) and chlorophytes (none of 111 species) (Jia et al., 2016). The presumptive MTPSL genes obtained from transcriptome analysis of these non-seed plants have originated from innate plant genomes rather than from closely associated microbes is supported by three lines of evidence: (1) The endogenous nature of putative MTPSLs were determined by genomic DNA isolation from axenic cultures, amplification using PCR and sequence analysis by sequencing. (2) Plant origin of MTPSLs is verified by the identification of immediate authentic neighbor genes, PCR amplification of coding sequence of MTPSLs with the neighbor gene and full sequencing of the cloned product. (3) The innate presence of MTPSLs in plants is strengthened from their phylogenetic relationships which align in the same order as that of evolutionary relationships of plants from which they are obtained (Jia et al., 2016).

Typical plant TPSs and MTPSLs differ in number of features like structural configuration, genomic organization, variable catalytic motifs. Based on intron-exon conservation pattern, typical plant TPSs can be divided in to three classes: 12-14 introns, 9 introns or six introns (Trapp & Croteau, 2001). There is no conserved intron-exon pattern and tend to be highly variable in newly identified MTPSL genes. For example in *Marchantia polymorpha* four MTPSL genes have localized conservation of intron-exon pattern with three introns, three MTPSL genes lack introns completely while two MTPSL genes have four introns with position of three downstream introns conserved with those of other MTPSL genes (Santosh Kumar et al., 2016). In contrary to typical plant TPSs which are made of either αβγ type or αβ type domain configurations, MTPSLs are of only α-domain type as witnessed in bacteria and fungi (Jia et al., 2018, 2016). MTPSLs are much smaller (~350 residues) than typical plant TPSs

(~550-800 residues). Along with canonical aspartate rich catalytic motif DDxxD, MTPSLs also show non-canonical motifs like DDxxxD and DDxxx whose mode of actions during catalysis remains to be determined (Jia et al., 2018)

Many typical plant terpene synthases possess multi substrate activity depending on the availability of substrate due to perturbations in the plant metabolisms. The multi substrate utilizing capacity of MTPSLs have also been determined by their characterization. Some of the recently characterized MTPSLs and their invitro substrates are listed in (**Table 1**;Jia et al., 2018). The typical plant TPSs of two non-seed plants, the moss *P.patens* and the lycophyte *S.moellendorffii* are diterpene synthase type. Since most of the MTPSLs in *S.moellendorffii* have been elucidated to express mono and sesquiterpene synthases, it is in harmony that they function in this manner (Jia et al., 2016).

Table 1. Characterized MTPSLs for catalysis along with their invitro substrates. (Jia et al., 2018)

Species	MTPSLs	Substrates
Selaginella moellendorffii	SmMTPSL1	*(E,E)*-FPP
	SmMTPSL13	None
	SmMTPSL17	*(E,E)*-FPP
	SmMTPSL22	*(E,E)*-FPP, GPP
	SmMTPSL26	*(E,E)*-FPP
	SmMTPSL30	None
Marchantia polymorpha	MpMTPSL1	None
	MpMTPSL2	NPP
	MpMTPSL3	*(E,E)*-FPP
	MpMTPSL4	*(E,E)*-FPP
	MpMTPSL5	*(E,E)*-FPP
	MpMTPSL6	NPP
	MpMTPSL7	*(E,E)*-FPP
	MpMTPSL8	None
	MpMTPSL9	*(E,E)*-FPP
Scapania nemorea	Liv-IRBN-MTPSL2	GPP, *(E,E)*-FPP, *(Z,E)*-FPP
	Liv-IRBN-MTPSL4	*(E,E)*-FPP, *(Z,E)*-FPP, *(Z,Z)*-FPP, *(E,E,E)-GG*PP
Anthoceros punctatus	Hon-ApMTPSL7	GPP, *(E,E)*-FPP, *(Z,E)*-FPP, *(Z,Z)*-FPP
Sphagnum lescurii	Mos-GOWD-MTPSL2	GPP
Pseudotaxiphyllum elegans	Mos-QKQO-MTPSL3	*(E,E)*-FPP, *(Z,E)*-FPP, *(Z,Z)*-FPP
Anomodon rostratus	Mos-VBMM-MTPSL3	*(E,E)*-FPP, *(Z,E)*-FPP, *(Z,Z)*-FPP, *(E,E,E)-GG*PP
Myriopteris eatonii	Mon-GSXD-MTPSL3	*(E,E)*-FPP, *(Z,E)*-FPP
Pityrogramma trifoliata	Mon-UJTT-MTPSL4	GPP, *(E,E)*-FPP, *(Z,E)*-FPP, *(Z,Z)*-FPP
Woodsia scopulina	Mon-YJJY-MTPSL1	*(E,E)*-FPP, *(Z,E)*-FPP, *(Z,Z)*-FPP

Physcomitrella patens- a green production host

P. patens (moss) is a primitive, non-seed, non-vascular land plant belonging to bryophyte group originated around 500 Ma years ago and diverged along with sister clades (liverworts and hornworts) from vascular plants at about 450 Ma years ago in the time scale of land plant evolution (Morris et al., 2018; Simonsen et al., 2009). Evolutionarily mosses are placed halfway between charophytes and vascular seed plants due to which they are used as model plants to characterize phylogenetic, development and physiological properties of plants (Simonsen et al., 2009).

P. patens has been established as a model species to perform functional genomics-based endeavors and an ideal candidate for production of plant biopharmaceuticals with its innate key beneficiary features and recent developments of invitro tissue culture techniques and whole genome characterization. The key features include: (1) The lifecycle of moss as shown in **Figure 7** has a dominant haploid gametophytic phase with photoautotrophic energy metabolism and a short diploid heterotrophic sporophyte phase nourishing on gametophyte (Schaefer & Zry, 2001; Simonsen et al., 2009). This dominant haploid phase is exploited for gene targeting to generate subsequent knock out mutants of genes to unravel their functions without a need for time inefficient back cross, an inevitable breeding method in diploid organisms (Reski et al., 2018). (2) *P. patens* can propagate vegetatively and has the ability to produce increased amounts of biomass in protonemal filamentous state. All *Physcomitrella* tissue forms originated during different developmental stages like haploid protonema, gametophore and diploid sporophyte upon mechanical disruption can regenerate chloronemal apical cells that regenerate new filamentous protonemal network from the affected areas whose property is exploited to maintain the wild type and mutant cultures in single cell protonemal state indefinitely while stable genetically (Prigge & Bezanilla, 2010; Simonsen et al., 2009). (3) *Physcomitrella* can be cultured invitro on simple inorganic media without any organic supplements like carbon source, phytohormones and vitamins and can be maintained in both sterile solid and liquid media in plant tissue culture flasks, Erlenmeyer flasks and photobioreactors under regulated light and dark cycles (**Figure 6**;Reski et al., 2018; Simonsen et al., 2009). Submerged state cultivation enables scaling up to several thousands of litres in photobioreactors and with media comprising only minerals and water, offers benefits like minimum risk of contamination and low-cost high-volume production. (4)

Availability of *Physcomitrella* full sequenced genome of around 500 Mbp size distributed on 27 chromosomes in http://www.cosmoss.org and published expression levels (transcriptomic) of sets of genes provides valuable information on genetic networks that control specific biological processes which enables transcriptome analysis and studying the protein functions in vivo (Reski et al., 2018) (5) Heterologous expression of genes in *Physcomitrella* can be achieved by constructing selectable expression cassettes using well characterized inducible (induced by temperature, light & chemicals) and constitutitve promoters from bacteria, moss and seed plants (Reski et al., 2018). Availability of characterized synthetic promoters constructed by random assembly of CREs from *Zea mays* constitutive promoters facilitates expression of multiple genes in *Physcomitrella* (Peramuna et al., 2018).

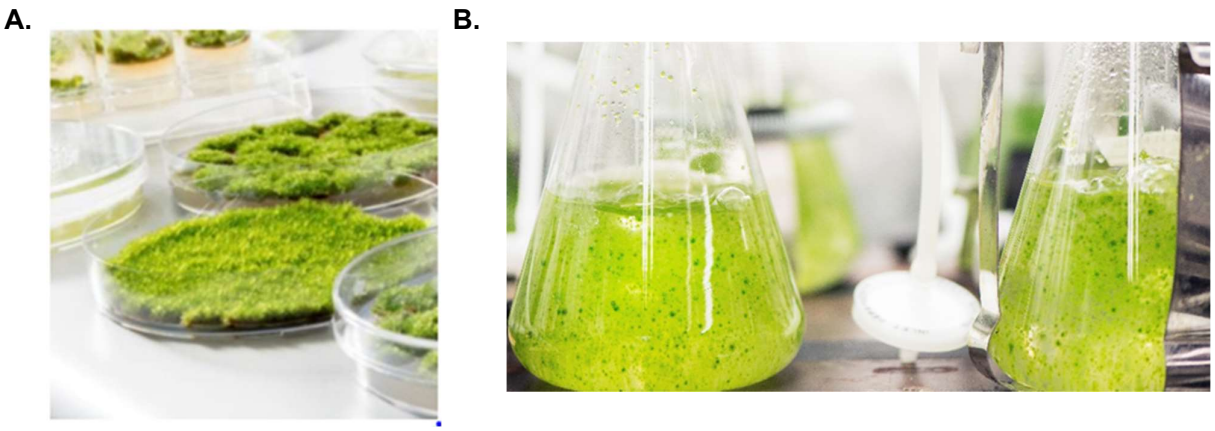

Figure 6. (A) Culture of *P. patens* on solid media. (B) Submerged cultures of *P. patens* in protonema stage grown in Erlenmeyer flasks.

Physcomitrella can incorporate transformed DNA fragments and constructs in to genomic DNA at target loci by homologous recombination and is rendered with high efficiency similar to yeast, the only renowned eukaryote that performs in vivo DNA assembly of multiple fragments with homologous regions (King et al., 2016; Prigge & Bezanilla, 2010; Reski et al., 2018; Schaefer & Zry, 2001). This exceptional feature seen only in *Physcomitrella* in contrary to other plants enables gene function studies in transgenic mosses in a reverse genetics approach.

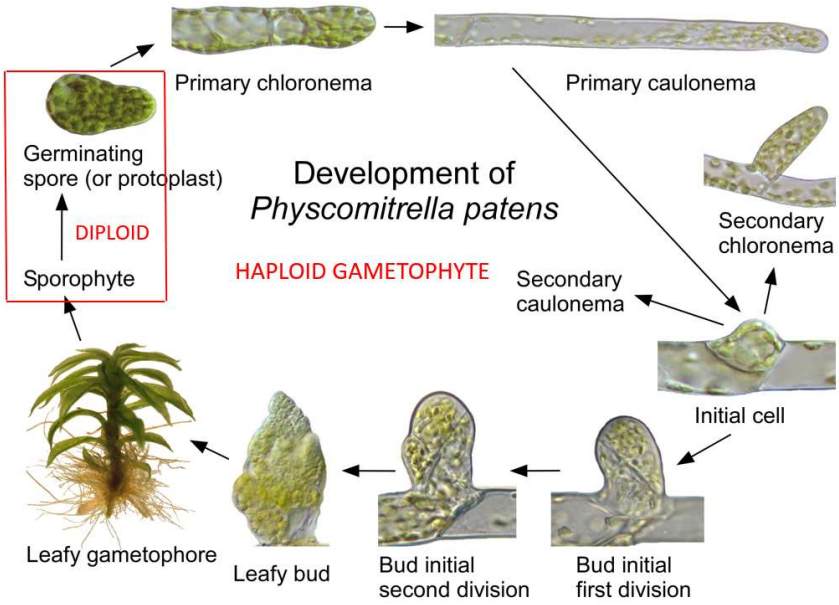

Figure 7. Illustration of developmental stages during *Physcomitrella* life cycle. (Roberts et al., 2012)

P. patens, a promising photoautotrophic expression system is a competing choice over microbial hosts for sustainable production of high value terpenoid metabolites due to innate tolerance levels to exogenous terpenoid accumulation, simple metabolic background with only one endogenous functional diterpene synthase gene, the copalyl-diphosphate/kaurene synthase whose disruption does not cause any change phenotypically (Zhan et al., 2015), established molecular mechanism for genome editing through homologous recombination and especially provides a natural environment for expressing plant cytochrome P450 enzymes, the challenge that has to be confronted in microbial systems particularly in bacteria with typical approaches.

2. HYPOTHESIS

It is hypothesized that the Microbial terpene synthase like genes in non-seed plants have their ancestral origin in bacteria and fungi. The wide distribution of these genes in non-seed lineages of plant kingdom is presumably by means of multiple horizontal gene transfer events during ecological and physiological interactions of microorganisms with plants. Evolution of

MTPSLs in bryophytes may be associated with strengthening the defense ability during transition of plants from aquatic to terrestrial habitats.

3. OBJECTIVES

Liverworts are recognized as copious producers of vast diversity of terpenoids among non-seed plants that have been suggested to function in drought resistance and herbivore defence (Jia et al., 2018). The functional studies of MTPSLs from this group might provide an insight into their possible role in structural diversity of terpenes and their contribution to combat biotic and abiotic stresses caused by terrestrial habitats. MTPSL genes from *L.* bidentata, an untapped liverwort species collected from tree barks of northern zealand forest of Denmark were chosen for this functional genomics study.

The objectives of this project are enumerated in a sequence as below.

1. Individual expression of four presumptive MTPSL genes obtained through transcriptome analysis from *L. bidentata* in moss *P. patens.*

   Gene nomenclature: LbMTPSL 1
   LbMTPSL 3
   LbMTPSL 4
   LbMTPSL 5

2. Extraction of volatile & nonvolatile metabolites, analysis and determination of biological function of metabolites produced.
3. Structural and functional characterization of corresponding terpene synthases
4. Phylogenetic analysis to decipher evolution of MTPSL proteins under study.

4. STRATEGY

The strategy employed was summarized in a flow diagram in (**Figure 8**) which in a sequential manner involves the sequence data acquisition from transcriptome analysis of gametophytic tissue of *L. bidentata*. The raw DNA sequences were analyzed in a bioinformatics platform CLC Main work bench 8.1.3 for determining ORFs, joining sequences to make constructs

and designing primers. Synthetic genes and DNA fragments were cloned, and PCR amplified. In parallel the moss was cultured invitro, generated protoplasts followed by PEG mediated transformation with DNA fragments containing overlapping regions facilitating DNA assembly and targeted integration in to moss genome at *Pp*108 locus on chromosome 20 via homologous recombination. The transformants can be screened either by genetic or metabolite screening methods to verify the presence of heterologous genes. There after volatile and nonvolatile metabolites should be extracted and sampled for identification by GC-MS analysis. Structure and catalytic domains of MTPSL enzymes can be characterized based on terpenoid profile and subsequently evolutionary origin of MTPSL genes can be determined through phylogenetic analysis.

Due to unforeseen circumstances of Denmark lockdown by covid-19 outbreak, the intended experimental strategy could not be completed. It was only possible to carry experimental work up to and including transformation procedure as shown in green color in flow chart (**Figure 8**). The obtained transformants could not be assessed for heterologous genes either by genetic screening or metabolite screening and inevitably further enzyme characterization studies were sequence dependent and insilico by means of bioinformatics approaches as displayed in blue color loop in flow chart which is finally concluded with the sequence dependent phylogenetic analysis.

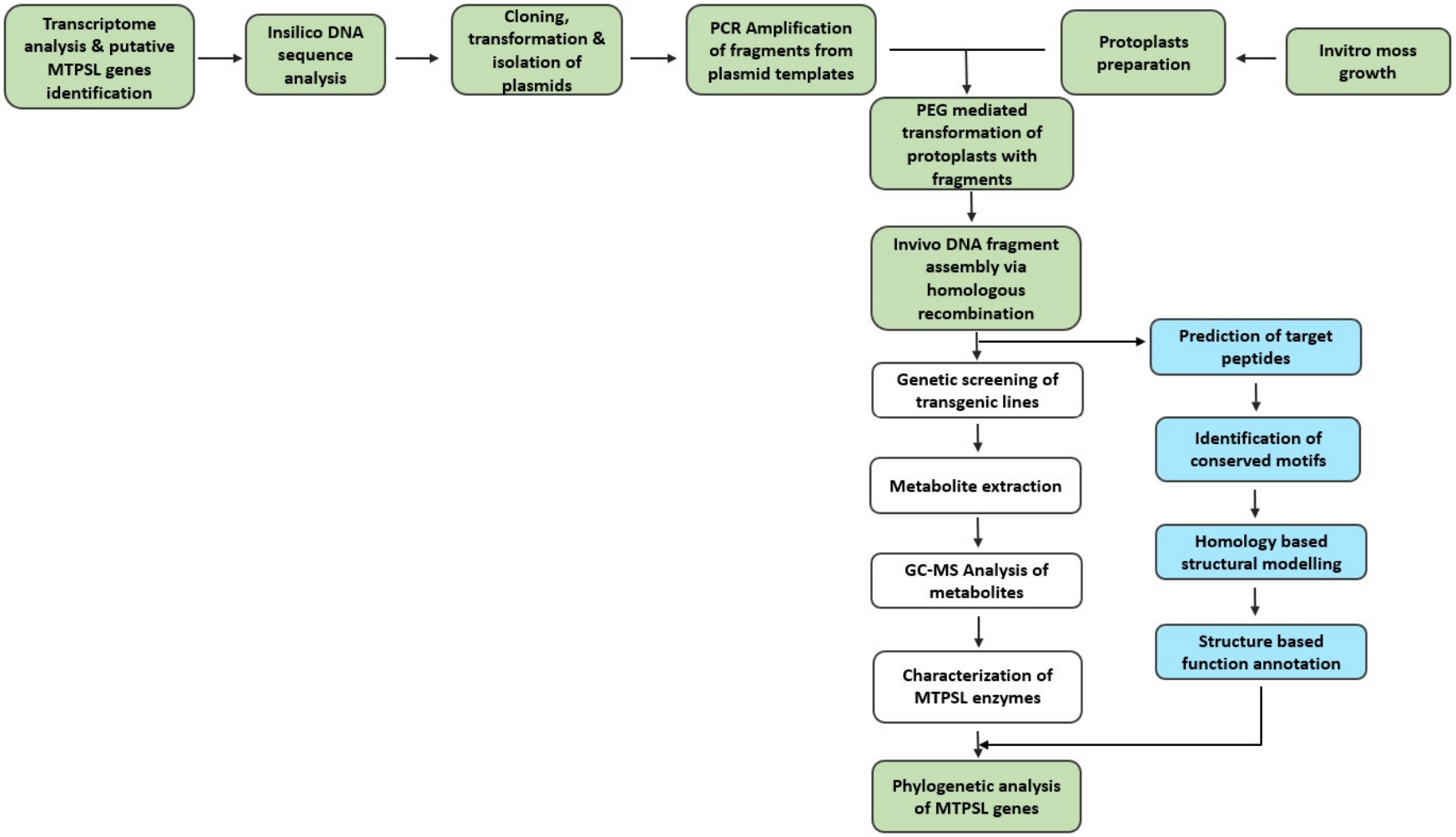

Figure 8. Illustration of strategy in a flow diagram

Genome targeting – In-vivo DNA assembly & integration via homologous recombination.

Unlike most eukaryotes in which gene targeting occurs through illegitimate recombination by random insertions, *P. patens* is the only recognized multicellular eukaryote among plant kingdom that delivers precise gene targeting through homologous recombination (HR). The HR efficiency of this bryophyte is comparable to that of unicellular eukaryotic microbial model for functional genomics *Saccharomyces cerevisiae* which renders in vivo assembly and targeted integration with 90% efficiency (Schaefer & Zrÿd, 1997). The targeted insertion sites for gene targeting are neutral loci whose disruption does not lead to any morphological changes. The three standard neutral loci discovered much earlier were *Pp108, Pp213 and Pp420* among which *Pp108* located on chromosome 20 (**Figure 9**) was employed for the current endeavor due to its high transformation tendency (Banerjee et al., 2019; Schaefer & Zrÿd, 1997). It has been demonstrated that a minimum of 12bp homology regions can ensure episomal recombination events in a stable and reproducible manner (Murén et al., 2009). With reference to king et al., 2016 who demonstrated efficiently the in vivo assembly and integration of DNA fragments in the moss genome making use of short 12-20 bp overlaps with adjacent fragments and greater than 500bp homology regions to genomic locus, a 20-25bp homology region was used in this project to achieve in vivo DNA assembly and greater than 500bp isogenous portions were incorporated for exogenous DNA integration in to *Pp108* locus on chromosome 20 of moss genome. As portrayed in **Figure 10**, the four fragments comprising of specific DNA elements (explained in section 5) can be assembled into a single construct by means of homologous recombination machinery and integrated in to moss genome at *Pp*108 locus. Thus, the transforming DNA facilitates gene targeting by single copy allele replacement of targeted locus.

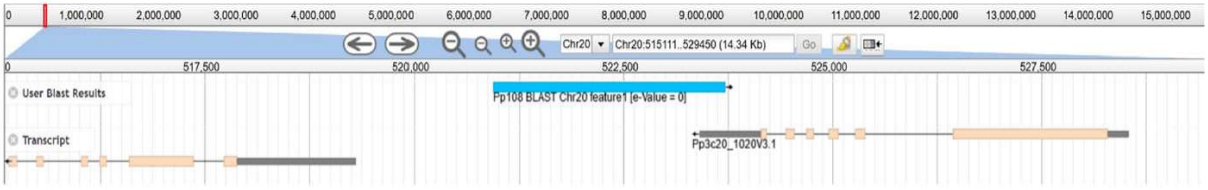

Figure 9. Pp108 locus is positioned between coordinates 520000 and 525000 on chromosome 20. (Banerjee et al., 2019)

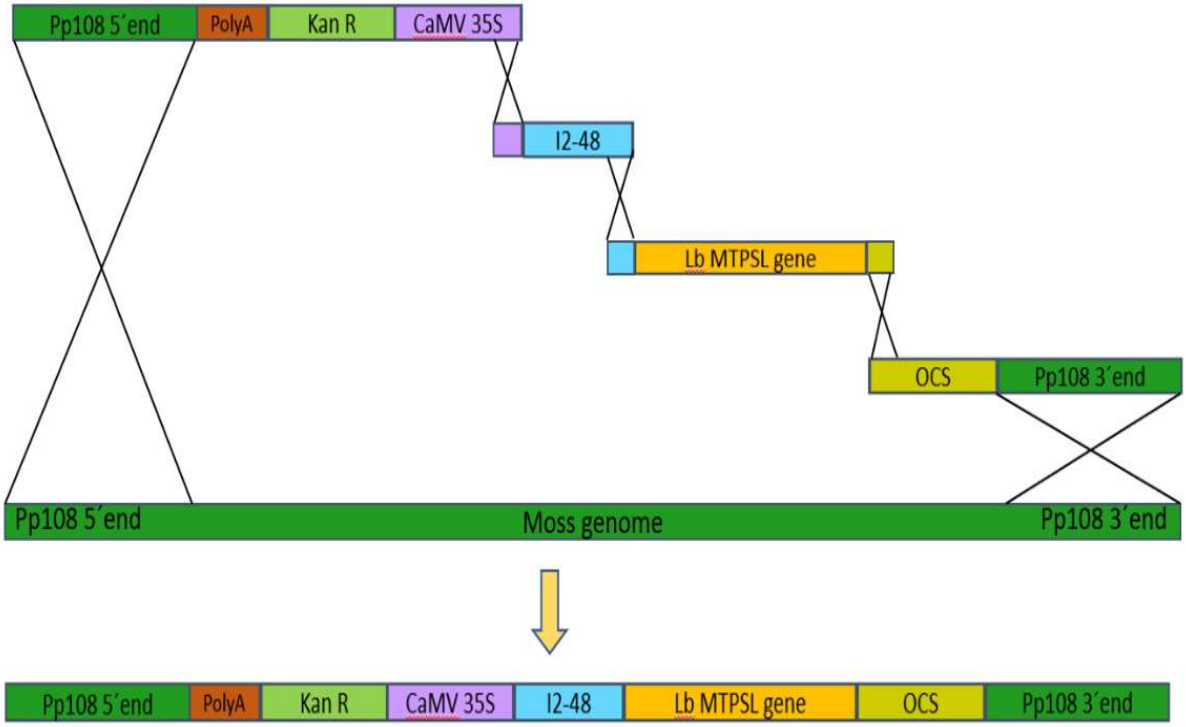

Figure 10. Schematic of four fragments for transformation, their assembly and integration in to Pp108 locus of moss genome via homologous recombination.

5. MATERIALS AND METHODS

Plant material, growth media and cultivation

Wild type *P. patens* (Grandsen ecotype) was obtained from International Moss stock center at the University of Freiburg (http://www.moss-stock-center.org/). It was cultivated and propagated on solid (0.7% agar) and liquid BCD media (**Table 2)**. The sub culturing and maintenance of wild type *P. patens* in protonema state was ensured by blending the moss tissue using IKA tissue homogenizer (T25 digital ULTRA-TURRAX) for every 2-3 weeks and inoculating into fresh BCD media. All cultures including both mutant and wild types were grown in a controlled growth chamber with 16h light and 8h dark cycles. A temperature of 25°C and light intensities of 20 to 50 W/m^2 were used as standard growth conditions (Bach et al., 2014)

Table 2. Composition of BCD media

Solution	Composition	Volume	Reference
B	25 g/L $MgSO_4.7H_2O$	10 mL	(Bach et al., 2014)
C	25 g/L KH_2PO_4 (Ph 6.5 with 4M KOH)	10 mL	
D	101 g/L KNO_3	10 mL	
TES	614 mg H_3BO_3 389 mg $MnCl_2.4H_2O$ 110 mg $Alk(SO_4).12H_2O$ 55 mg $CoCl_2.6H_2O$ 55 mg $CuSO_4.5H_2O$ 55 mg $ZnSO_4.7H_2O$ 28 mg KBr, 28 mg KI 28 mg LiCl 28 mg $SnCl_2.2H_2O$	1 mL	
	1M $CaCl_2$	1 mL	
	Distilled water	To 1000 mL	

DNA parts, Vectors

The necessary DNA elements for moss transformation were prepared in the form of four blocks or fragments. As shown in **Figure 11** moving from left to right, the first fragment (fragment 1) with 2688bp size consists of 5´108 *Pp* locus genome targeting sequence of 1220bp and CaMV 35S promoter driven Kanamycin resistance gene with CaMV poly (A)

transcription termination signal. This 2.688kb region was amplified from pRH004 plasmid (**Figure 12A**). The second fragment (fragment 2) is 357bp synthetic promoter I2-48 developed by Peramuna et al., 2018. It was blunt end ligated in to pJET1.2/blunt plasmid (**Figure 12B)** using the CloneJET PCR cloning kit (Thermo scientific) in order to use as a template for PCR amplification. The fourth fragment (fragment 4) is 2093bp region comprising OCS terminator and 1352bp 3´108 *Pp* locus homologous flanking region and was amplified from pRH004 plasmid (**Figure 12A**).

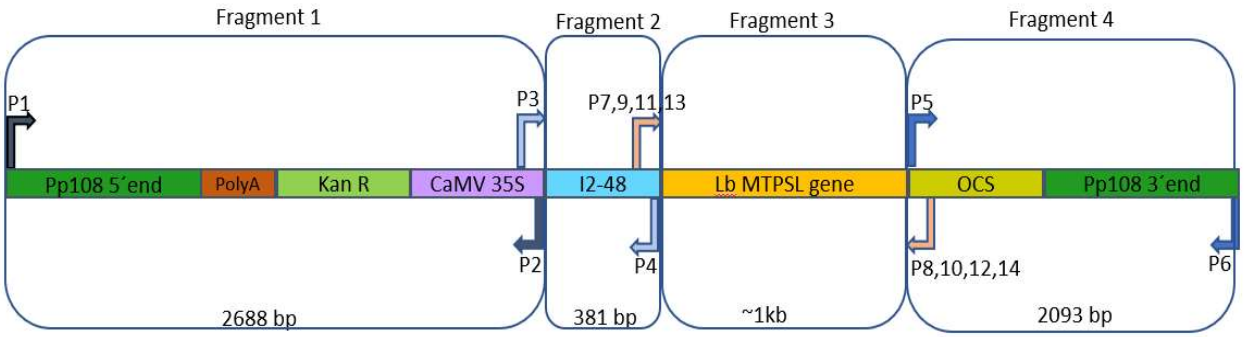

Figure 11. Functional DNA elements incorporated in to four different PCR fragments to transform LbMTPSL genes in to Pp108 neutral loci. Primer pairs for amplification of fragments are shown as identical colour bent arrows.

All plasmids were propagated by transformation of *Escherichia Coli* DH5α strain and plasmids were isolated from recombinant *E. Coli* using Gen Elute Plasmid Miniprep Kit (Sigma-Aldrich) as per manufacturers protocol

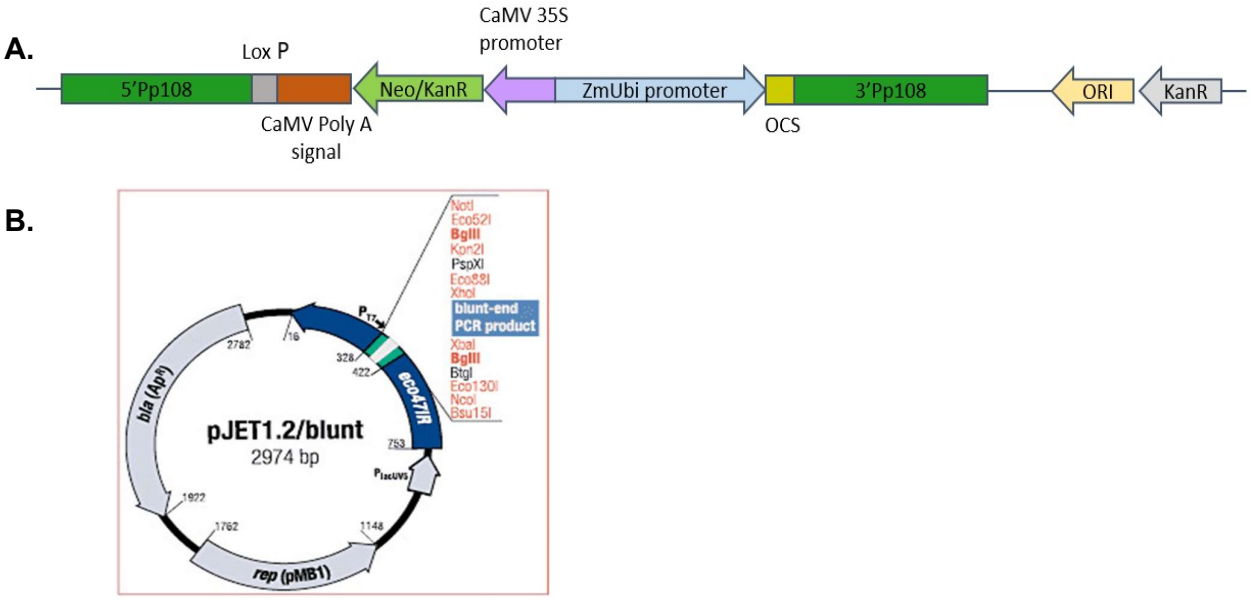

Figure 12. (A) Vector map of linearized pRH004 vector.(Peramuna et al., 2018) (B) Vector map of linearized pJET1.2/blunt Cloning vector.

Acquisition of putative LbMTPSL genes

RNA extraction and sequencing

The total RNA was extracted from the gametophytic tissue of *L. bidentata* using the Spectrum™ Plant Total RNA Kit (Sigma, STRN250). This yielded 300 µg total RNA as determined by nano-drop. RNA integrity was initially confirmed by agarose gel electrophoresis and the visualization of intact ribosomal RNA bands. Subsequent RNA quality control was carried out on a 2100 Bioanalyzer (Agilent Technologies, Hørsholm, Denmark) and each sample received an RNA integrity numbers (RIN) of greater than 8.5. The total RNA of biological triplicates was pooled to make one technical sample, and total RNA samples were submitted to Macrogen (Seoul, Korea) for stranded mRNA library preparation using an Illumina Truseq Stranded mRNA library prep kit. The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation. Alternatively, "tagmentation" combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified. For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. When cluster generation is complete, the templates are ready for sequencing. The sequencing is performed by Novaseq 150bp paired-end sequencing providing 30-40 million reads per sample. Illumina SBS technology utilizes a proprietary reversible terminator-based method that detects single bases as they are incorporated into DNA template strands. As all 4 reversible, terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies. Sequencing data is converted into raw data for the analysis.

Trinity assembly and transcriptome annotation

A *de novo* transcriptome was assembled as a reference for read mapping and Differential Expression (DE) analysis using Trinity v2.4.0 (Haas et al., 2013). As recommended in the Trinity protocol, one single Trinity assembly was generated by combining all reads across samples as input to ease following downstream analysis. Quality trimming and adapter

removal was performed using trimmomatic (Bolger et al., 2014) with default parameters. Transcript abundance was estimated using the alignment-based quantification method RSEM that uses Bowtie2 (Langmead & Salzberg, 2012) as an alignment method. Transcript and gene expression matrices were generated, and the numbers of expressed genes were calculated. Finally, differential expression analysis was performed at the gene level using edgeR (Robinson et al., 2009) with a dispersion rate of 0.1. Extractions and clustering of differentially expressed genes were performed with combinations of P-value cutoff for FDR of 1e-3 and fold-change values of 2, 4, 16, 64 and 256. Functional annotation of the transcriptome was performed using the annotation suite Trinotate. The functional annotation includes homology searches to BLAST, SwissProt, PFAM and various annotation databases such as eggNOG/GO/Kegg and were designated as LbMTPSL1,3,4 and 5.

The third block of construct (**Figure 11**) includes either of LbMTPSL1, LbMTPSL3, LbMTPSL4 and LbMTPSL5 genes which are of 1128bp, 1287bp, 1155bp and 1164bp sizes respectively. LbMTPSL genes were synthesized by Twist bioscience. These synthetic genes were ligated in to pJET1.2 plasmid (**Figure 10B)** using the CloneJET PCR cloning kit (Thermo scientific) and were labelled as pJET1.2_LbMTPSL1; pJET1.2_LbMTPSL3; pJET1.2_LbMTPSL4 and pJET1.2_LbMTPSL5 which serves as templates for PCR amplification of specific LbMTPSL genes. The FASTA format of all four gene sequences were provided in the Appendix (**Supplimental data 1)**.

Polymerase chain reaction

The multi fragment transformation of moss in this work requires 20µg of total DNA for one transformation event encompassing equi-molar concentrations of all 4 fragments mentioned earlier (King et al., 2016). In order to accumulate required quantity of fragments for a single transformation, PCR was carried out with 50µl reaction mixture in 16 tubes for each fragment. A classical three step cycle (**Table 3)** of standard PCR for 34 times was run for all amplifications with either Phusion^{R} High-Fidelity DNA Polymerase (New England Biolabs) or PfuX7 polymerase (Nørholm, 2010). T100 Thermal cycler (BIO-RAD) was used for all PCR amplifications.

Table 3. PCR steps and conditions used.

Steps	Temperature (°C)	Time (s)
Initial denaturation	98	120
	Start cycle	
Denaturation	98	15
Annealing	Variable	15
Extension	72	Variable
	End cycle (34x)	
Final extension	72	300

All primers used were synthesized by Integrated DNA Technologies (IDT) and were listed in **Table 4**. The first block was amplified from pRH004 plasmid using primer pair P1 & P2. The second block containing synthetic promoter I2-48 was amplified from pJET1.2_I2-48 plasmid using forward primer P3 containing 21bp overhang homologous to fragment 1 and reverse primer P4. The third fragment contains either of four LbMTPSL 1,3,4,5 genes which were amplified from gene specific pJET1.2_LbMTPSL 1,3,4,5 plasmids with primer pairs P7,P8; P9,P10; P11,P12; P13,P14 respectively. All forward and reverse oligos contain 20-25bp flanking regions homologous to fragments 2 and 4. The fourth fragment is also amplified from pRH004 plasmid with primer pair P5 & P6.

Table 4. List of primers used for amplification of all fragments.

Name	Sequence	Description
P1	CCACATCCTTCTCCGGCTTC	Forward primer to fragment 1
P2	TGAGACTTTTCAACAAAGGGTAATT	Reverse primer to fragment 1
P3	ACCCTTTGTTGAAAAGTCTCATGTGCTCGGA CCTGTAGATGCTAG	Forward primer to fragment 2 with 21bp overhang
P4	GGTTCTATCTCCTTCGGATCCTCGAGCGT	Reverse primer to fragment 2
P5	CTGCTTTAATGAGATATGCGAGACG	Forward primer to fragment 4
P6	ACGAAGGCCGTTCTTCCCTG	Reverse primer to fragment 4
P7	GGATCCGAAGGAGATAGAACCATGGAGGTG CCAGAAACGAAGGAG	Forward primer to LbMTPSL 1 with 21bp overhang
P8	GTCTCGCATATCTCATTAAAGCAGCTAGGTG AAACGTTTGCTTGT	Reverse primer to LbMTPSL 1 with 24bp overhang
P9	GGATCCGAAGGAGATAGAACCATGGTACGT GATATGAATTCCGCT	Forward primer to LbMTPSL 3 with 21bp overhang
P10	CGTCTCGCATATCTCATTAAAGCAGTTAACC GAAAAGGCCTGAAA	Reverse primer to LbMTPSL 3 with 25bp overhang
P11	GGATCCGAAGGAGATAGAACCATGGGAGCG TTAGAAGGTGATGAG	Forward primer to LbMTPSL 4 with 21bp overhang
P12	CGTCTCGCATATCTCATTAAAGCAGCTAGTC GAACCGTTTGTTTG	Reverse primer to LbMTPSL 4 with 25bp overhang
P13	GATCCGAAGGAGATAGAACCATGGCTGCTG CTGAAGCAATTCCT	Forward primer to LbMTPSL 5 with 20bp overhang
P14	CGTCTCGCATATCTCATTAAAGCAGCTACAA ATAGCGCACAGATTT	Reverse primer to LbMTPSL 5 with 25bp overhang

The fragments 1,3 and 4 were PCR amplified in 50 µL reaction using PfuX7 polymerase. The concentrations and volumes of individual reaction components for these fragments were shown in **Table 5**. The synthetic promoter I2-48 was amplified in 40 µL reaction with Phusion GC buffer and polymerized by compatible Phusion^{R} High-Fidelity DNA Polymerase (**Table 6**)

Table 5. 50 µL PCR reaction components for fragments 1,3 & 4

Component	Volume
10X CXL buffer	10 µL
10 mM dNTPs	2 µL
10 µM Forward Primer	0.5 µL
10 µM Reverse Primer	0.5 µL
30 ng/µL Template DNA	Variable
Pfu X7 Polymerase	0.5 µL
Milli-Q water	To 50 µL

Table 6. 40 µL PCR reaction components for I2-48 Synthetic promoter

Component	Volume
5X Phusion GC buffer	8 µL
10 mM dNTPs	4 µL
10 µM Forward Primer	2 µL
10 µM Reverse Primer	2 µL
30 ng/µL Template DNA	Variable
Phusion DNA Polymerase	0.3 µL
Milli-Q water	To 40 µL

DNA purification and Concentration

The amplified PCR products were purified using Nucleospin Gel and PCR clean-up Kit (Macherey-Nagel), according to manufacturer's protocol. The purified PCR products of all fragments were concentrated up to ~1 µg/ µL by allowing 300-400 µL volumes of DNA vials kept in heat block for evaporation at 37°C overnight under vacuum hose. The concentrations of dsDNA were determined by measuring the absorbance maximum at 240nm using NanoDrop2000 (Thermo Fisher Scientific). All DNA fragments were stored in freezer at -18°C until used.

Preparation and PEG mediated transformation of protoplasts

A new unpublished protocol for preparation of protoplasts from protonema by using a blend of cell wall degrading enzymes with Cellulase, Hemicellulase and Pectinase activities was utilized in this project. Elaborately, 1.0-1.2 g of 5-day old moss protonemal tissue that was distributed in 4 petri plates was digested with 10 mL of 1.5% Cellulase R-10 (Duchefa Biochemie) and 0.5% Macerozyme R-10 (Duchefa Biochemie) enzyme solution in protoplast medium (PM) (8.5% mannitol, 20mM MES, 10mM $CaCl_2$). The protonema tissue in enzyme solution was incubated for 3 hours at 30°C. The degraded tissue suspension was filtered through 70µm mesh followed by 40µm pore size mesh and was distributed equally among 10 mL round bottomed glass tubes. Pellet the protoplasts by centrifuging at 200 x g with acceleration speed 4 and deceleration speed 2. Decanting the supernatants, the pellets were collected with 1 mL protoplast wash (PW) solution (8.5% mannitol, 4mM MES, 10mM $CaCl_2$)

into single tube and was made up to 10 mL final volume. After centrifugation at 200 x g for 10 minutes with lower breaking speeds, the pellets were resuspended in 3 mL PW solution. The protoplast density was estimated with a hemocytometer. The protoplast suspension was recentrifuged as before and resuspended in MMM solution (1M $MgCl_2$, 0.5M MES, 8.5% mannitol) to obtain a final concentration of 1.6×10^6 protoplasts/mL

In this multi-fragment transformation, a total of 20µg DNA per transformation which includes all fragments in equimolar ratios (**Table S1-S4**) with in 30 µL volume were transferred in to 15 mL falcon tube in to which 250 µL protoplast suspension and 300 µL PEG-MCT solution (PEG, 0.1M Ca(NO3)2.4H2O, 0.4M Tris HCl, 8.5% mannitol) were added. The mixture was incubated in a 45°C water bath followed by 5 minutes at RT. There after transformation mixtures were diluted with 300 µL PW solution for 5 times with intervals of 1 minute between dilutions followed by second round of dilutions with 1 mL PW solution for 5 times with 1 minute holding times between dilutions. The transformed protoplasts were pelleted by centrifugation and the supernatant was discarded. The protoplast pellet was resuspended in 500 µL PW solution and 2.5 mL of Protoplast regeneration media (top layer; PRMT). This mixture was distributed 1 mL each among three plates containing cellophane overlaid Protoplast regeneration media (bottom layer; PRMB).

The protoplasts were allowed to recover overnight in dark followed by incubation in light for 5 days. Subsequently the protoplasts bearing cellophane disks were transferred to BCD media plates for Kanamycin selection and allowed to grow in continuous light at 25°C for two weeks. The cellophane disks with recovered transformants were moved to standard BCD media for two more weeks to allow loss of any unintegrated DNA. Those moss lines that have survived the selection process have to be verified for presence of heterologous genes by PCR amplification from genomic DNA.

Genetic Screening

Since it is possible to obtain the non-transformant moss lines despite of selection, it is recommended to perform genetic screening to verify the existence of exogenous DNA in the

genome of transgenic lines. For this PCR analysis of genomic DNA has to be performed. The genomic DNA of transgenic moss can be purified and extracted using Wizard genomic DNA purification kit (Promega) using manufacturers protocol. Briefly, grind approximately 40 mg of leaf tissue in liquid nitrogen, add 600 µl of nuclei lysis solution and incubate at 65°C for 15 min, then add 3 µL of RNase solution ad incubate at 37°C for 15 min followed by cooling sample to room temperature (RT) for 5 min. Add 200 µl protein precipitation solution and vortex, centrifuge the contents at 13000 x *g* for 3 min and transfer the supernatant to clean tube containing 600 µl RT isopropanol. Mix by inversion and centrifuge at 13000 x *g* for 1 min. Decant the supernatant and add 600 µl RT 70% ethanol followed by centrifugation at 13000 x *g* for 1 min. Aspirate the ethanol and air-dry pellet. Add 100 µl DNA rehydration solution to rehydrate at 65°C for 1 hour or overnight at 4°C. Use 50 ng/µl genomic DNA for PCR analysis

Metabolite Analysis

Volatile metabolites

For profiling volatile metabolites of transgenic moss lines, HS-SPME (Head space – Solid phase microextraction) and GC-MS (Gas Chromatography – Mass Spectrometry) analysis method (Bach et al., 2014) was intended to employ. Briefly, place a lump of moss on top of solidified BCD media (approximately 4ml) filled in 20ml GC vials. Incubate capped cultures for 1-4 weeks in the growth chamber. Head space volatiles will be sampled by incubating SPME fiber in the vials at room temperature for 30 minutes and analyze by GC-MS

Non-volatile terpenoids

For non-volatile compound analysis, a solvent extraction method previously described in (Bach et al., 2014) would be useful. Transfer a moss aggregate in to 2ml GC glass vials with 0.5-1 ml of organic solvent like n-hexane including an approximate internal standard with a concentration of 0.2-1 mg/ml. Cap and vortex samples briefly. Extract at room temperature for 1 hour while mixing. Transfer the extract to a new vial for GC-MS analysis.

Prediction of target peptides

The presence of transit peptides in the LbMTPSL1,3,4,5 peptide sequences (**Supplimental data** 2) were predicted using TargetP-2.0 server (http://www.cbs.dtu.dk/services/TargetP/) by detecting the presence of N-terminal transit sequences like signal peptides, mitochondrial transit peptide, chloroplast transit peptide or thylakoid luminal transit peptides (Armenteros et al., 2019). TargetP-2.0 uses bidirectional long short-term memory (BiLSTM) network and multi-attention mechanism which enables the network to predict both the type of peptide and position of cleavage site

Identification of Conserved motifs

Canonical and non-canonical conserved sequence motif logos of LbMTPSL protein sequences (**Supplimental data 2**) were identified using Weblogo, a sequence logo generator (Crooks et al., 2004). A multiple sequence alignment FASTA file of all 4 LbMTPSLs obtained using COBALT tool in NCBI was submitted in weblogo program which generated the sequence logo data. The overall height of each stack indicates the sequence conservation at that position (measured in bits), whereas the height of symbols within the stack reflects the relative frequency of the corresponding amino acid at that position. The amino acids that compose the stacks display default colors according to their chemical properties.

Homology based structural modelling

Homology based structural models were generated using Iterative threading assembly refinement (I-TASSER) server(Yang & Zhang, 2015; Zhang, 2009). I-TASSER generates structure assembly simulation for each target. The significance in threading alignments is measured by Z-score where a normalized Z-score >1 indicates a good alignment and vice versa. The confidence in each model was quantified using C-score, typically in the range (-5,2) was calculated based on significance of threading template alignments and convergence parameters of structure assembly simulations. Higher C-score signifies model with higher confidence. After the structure assembly simulation, I-TASSER uses the TM-align structural alignment program to match the first I-TASSER model to all structures in the PDB library and reports the top 10 proteins from PDB that have closest structural similarity. TM score like

Root mean square deviation (RMSD) is a scale for measuring structural similarity. A TM score higher than 0.5 signifies high homology while scores below 0.17 correspond to random similarity. RMSD is an average distance of all residue pairs in two structures. The structural models were visualized using PyMOL Molecular Graphics system, version 2.0 Schrödinger, LLC.

Protein function prediction

Structure based function annotation was deduced by COFACTOR algorithm using structure comparison and protein-protein networks. COFACTOR will thread the query through the BioLiP protein function database by local and global structure matches to identify functional sites and homologies. COACH is a meta-server approach that combines multiple function annotation results (on ligand-binding sites) from the COFACTOR, TM-SITE and S-SITE programs which recognize ligand-binding templates from the BioLiP protein function database by binding-specific substructure and sequence profile comparisons to generate final ligand binding site prediction (Roy et al., 2012; Yang & Zhang, 2015). C-score is the confidence score of the prediction which ranges (0-1) where a higher score indicates more confidence.

Phylogenetic analysis

For phylogenetic analysis full length multiple sequence alignment of LbMTPSL1,3,4,5 proteins and 26 other characterized MTPSL proteins of different non-seed plant species along with 5 bacterial and 2 fungal TPSs were carried out using CLUSTALW algorithm for assembling sequence alignments and a boot strap phylogenetic tree was constructed using neighbor joining statistical method in Molecular evolution genetic analysis computing platform (MEGA-X) (Kumar et al., 2018). MEGA provides tools for comparative analysis of molecular sequences to identify functional and adaptive genomic differences. MEGA software has a large repository of programs to conduct evolutionary analysis. Boot strap values were calculated with 500 replications and a cutoff of 60% was imposed on values for the phylogenetic reconstruction tree in this study.

6. RESULTS

Experimental results

PCR Amplification of DNA fragments and genes

Gene constructs for LbMTPSL1,3,4,5 genes were designed to procure four individual transformations. Each construct assembled in vivo and integrated in to *Pp*108 neutral locus was made of four fragments. The overlapping DNA fragments used for transformations in equimolar ratios were amplified by means of primers containing 20-25bp overhangs homologous to adjacent fragments. The PCR products of amplified fragments were verified for their size on 1% Agarose gel electrophoresis for fragments 1 and 4 and on 2% agarose gel for fragment 2. **Figure 13** shows bands of fragments 1,2 and 4 on agarose gels. Fragment 1 gave a band of 2688bp with primer pair P1 and P2. Fragment 2 gave a 378bp band which includes 21bases homologous to fragment 1 at its 5´end amplified with primers P3 and P4. Fragment 4 have shown 2093bp size which was amplified with primer pair P5 and P6.

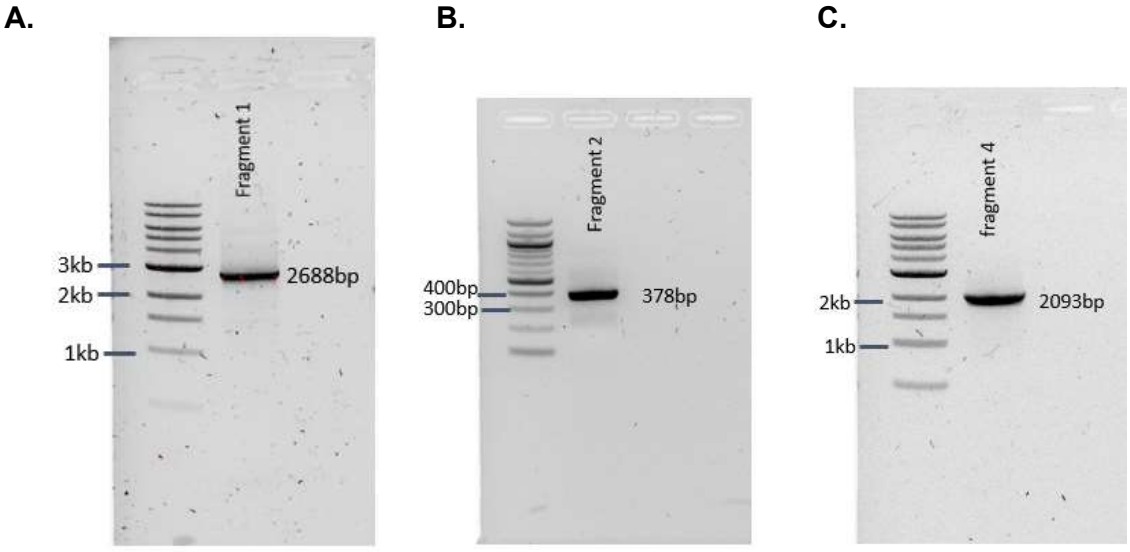

Figure 13. Agarose gels showing DNA fragments used for transformation experiments in equimolar ratios. (A) Fragment 1 visible as 2688bp band on gel. (B) Fragment 2 verified as 378bp size including region of overlap. (C) Fragment 4 levels 2kb band of ladder indicating 2093bp length.

Fragment 3 of each construct includes either of LbMTPSL1,3,4,5 genes that were aimed for biochemical characterization in this project. The purified PCR products of genes were verified for their size using 1% agarose gel electrophoresis as shown in **Figure 14**. LbMTPSL1 gene

amplified with primers P7 and P8 gave a size of 1173bp that includes flanking regions at both 5´ and 3´ ends homologous to fragment 2 and 4. LbMTPSL3 gene amplified with P9 and P10 primers have shown 1333bp band due to 21 bases and 25 bases overlapping regions at 5´ and 3´ ends respectively. The PCR product of LbMTPSL4 gene gave 1201bp band which also includes homologous recombination flanking regions at both ends added with PCR primers P11 and P12. Similarly, the PCR product of LbMTPSL5 amplified with primer pair P13 and P14 have shown size of 1209 bases due to joined flanking sequences of 20 bases and 25 bases homology to fragments 2 and 4 respectively.

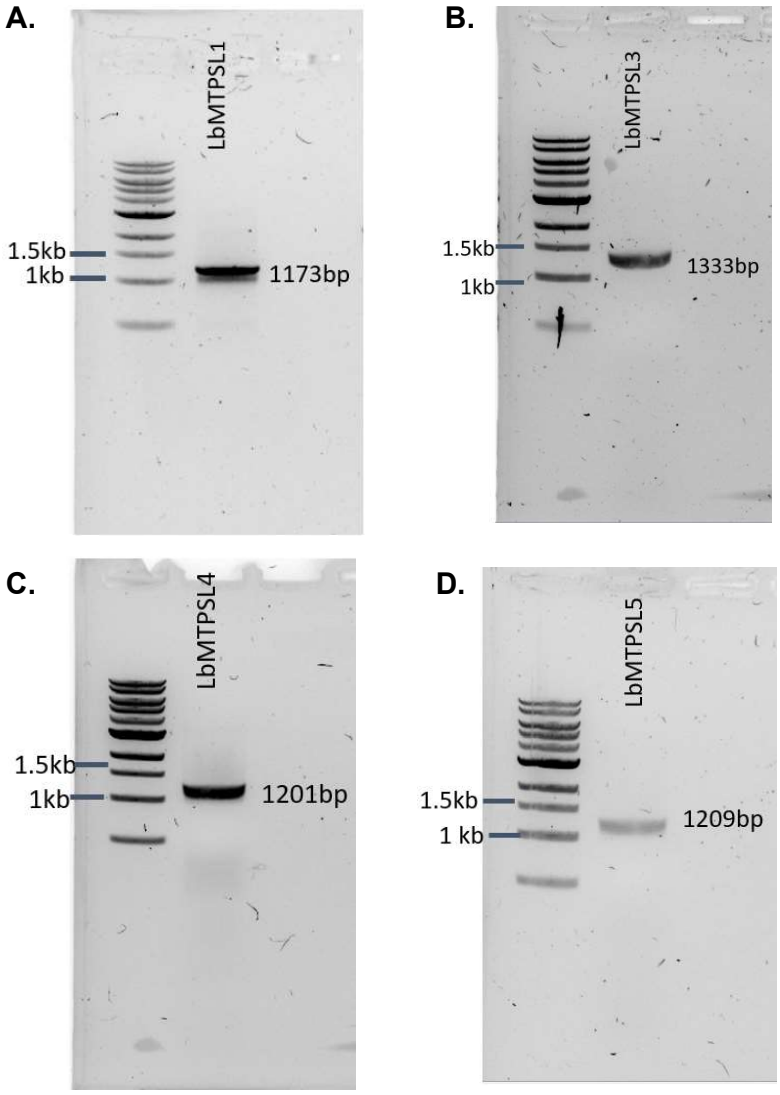

Figure 14. Agarose gels showing DNA bands of purified PCR products of LbMTPSL genes with adjoined flanking regions homologous to fragment 2 and fragment 4. (A) LbMTPSL1 gene matching 1173bp size. (B) LbMTPSL3 gene verified as 1333bp fragment. (C) LbMTPSL4 gene shows a band of 1201bp size. (D) LbMTPSL5 gene visible as 1209bp fragment.

PEG Mediated transformation of protoplasts

Estimated number of protoplasts in the solution:

$$protoplasts\ on\ grid\ area\ \times\ 10^4\ \times total\ volume\ in\ mL = number\ of\ protoplasts$$

$60 \times 10^4 \times 3 = 1.8 \times 10^6$ protoplasts

Final volume of MMM-protoplast suspension:

$$\frac{number\ of\ protoplasts}{1.6 \times 10^6} = final\ volume\ in\ mL$$

$1.8 \times 10^6 \div 1.6 \times 10^6 = 1.125$ mL

Four individual transformations involving fragments for all LbMTPSL gene constructs were performed with the protoplast suspension obtained. After selection on antibiotic the protoplasts were seen regenerated with spreading protonema network as green zones. These cultures after incubation on BCD growth media for two weeks yielded possible transgenic colonies as shown on a transformation plate in **Figure 15**.

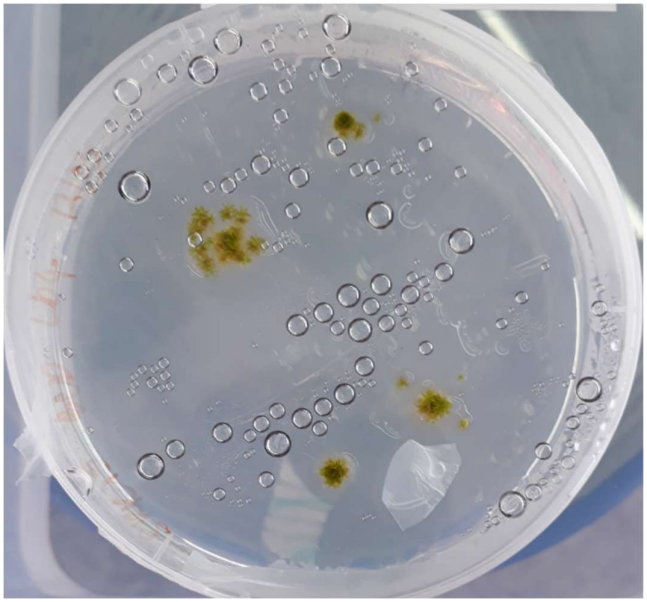

Figure 15. BCD Plate showing possible transformant colonies containing LbMTPSL4 gene construct.

Bioinformatics analysis

Prediction of target peptides

None of the LbMTPSL enzyme sequences were predicted to contain target peptides to any intracellular organelles. All the enzymes have much higher likelihood probability that the sequence does not have any kind of signal peptide (**Table 7**). This infers that LbMTPSL1,3,4 and 5 enzymes could be localized in the cytosol with the probable sesquiterpene synthase activities.

Table 7. Likelihood probabilities of LbMTPSLs to have different sorting peptides predicted by TargetP 2.0.

Protein	Signal peptide	Mitochondrial transfer peptide	Chloroplast transfer peptide	Thylakoid luminal transfer peptide	Other
LbMTPSL1	0.0001	0	0	0	0.9999
LbMTPSL3	0.0006	0.0002	0	0	0.9992
LbMTPSL4	0.0001	0	0	0	0.9999
LbMTPSL5	0.0002	0.0007	0.0007	0	0.999

Identification of conserved motifs

Sequence motif logos of LbMTPSL1,3,4,5 proteins were made using Weblogo for partial sequences as shown in **Figure 16**. As described earlier the substrate binding in Terpene synthases is facilitated by two highly conserved motifs: the DDxxD/E and NSD/DTE motifs. While the NSD/DTE motif is highly conserved in MTPSLs, the aspartate rich DDxxD/E motif tend to exist in other noncanonical forms (Jia et al., 2016). In accordance with this phenomenon, LbMTPSL1,3,4,5 proteins show two variations among aspartate rich motifs. LbMTPSL1,4 proteins have conserved DDxxE motif and LbMTPSL3,5 proteins displayed canonical DDxxD motif. LbMTPSLs are congruent with other MTPSLs with respect to second aspartate rich conserved NDxxSxxxD/E motif. The residues of DDxxD/E motif are positioned within the residue numbers 98 and 113 in LbMTPSL1,3,4,5 enzymes (**Table 8**) and remarkably NDxxSxxxD/E motif is positioned exactly 182 residues downstream to first aspartate rich DDxxD/E motif in all LbMTPSL proteins (**Table 9**).

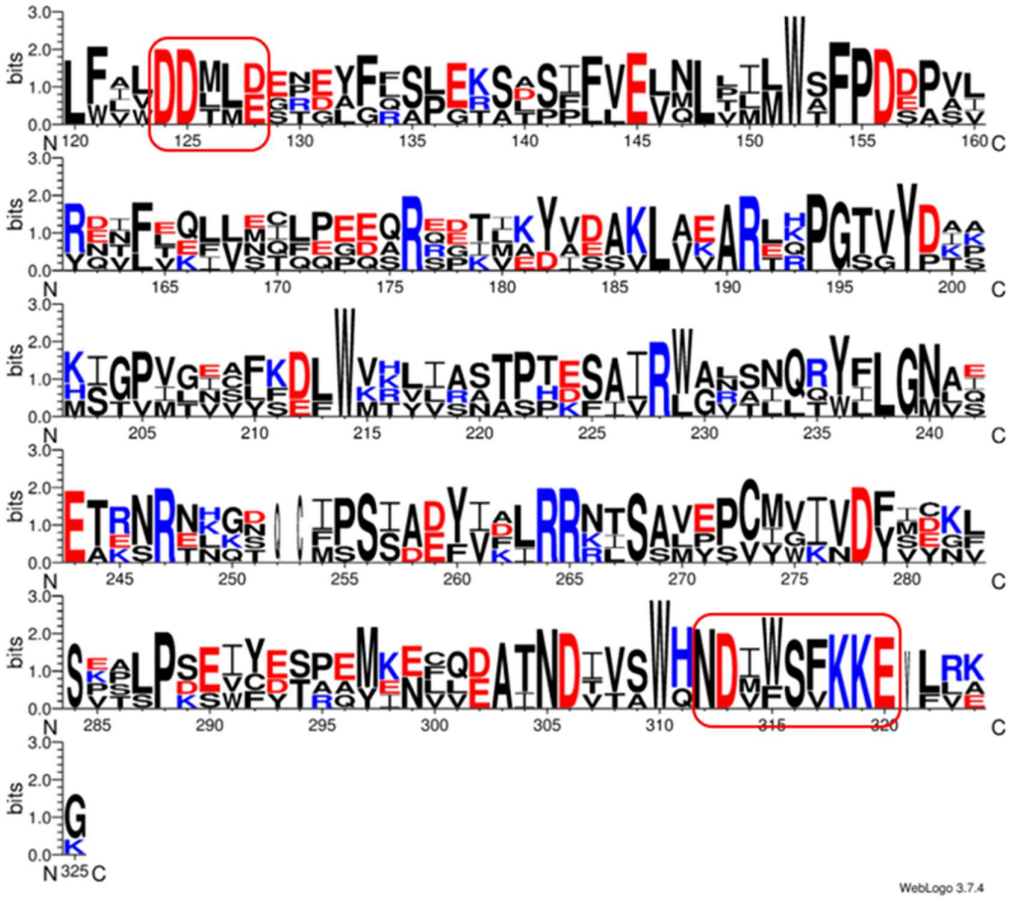

Figure 16. Sequence motif logos of LbMTPSL 1,3,4,5 proteins made using weblogo 3.7.4 showing canonical conserved DDXXD/E and NSE motifs (surrounded by red line) of microbial type terpene synthases.

Table 8. Conserved aspartate rich motifs in LbMTPSLs

Protein	DDxxD/DDxxE motif	Positions
LbMTPSL1	DDMLE	98-102
LbMTPSL3	DDLLD	122-126
LbMTPSL4	DDMLE	107-111
LbMTPSL5	DDTMD	108-113

Table 9. Conserved NSE motifs in LbMTPSLs

Protein	NSE motif	Positions
LbMTPSL1	NDMWSFKKE	284-292
LbMTPSL3	NDIFSVKKE	308-316
LbMTPSL4	NDIWSFKKE	293-301
LbMTPSL5	NDVWSFKKE	295-303

Homology based structural Models

Full length structural models were generated by I-TASSER to understand the structural features of LbMTPSL1,3,4 and 5 proteins. Predicted models of all LbMTPSLs had good C-score, TM-score and RMSD values supporting the correct topology of models (**Table 10**).

Table 10. Prediction quality scores of all LbMTPSL models generated by I-TASSER

Protein model	C-score	TM-score	RMSD (Å)
LbMTPSL1	-0.67	0.63±0.14	8.2±4.5
LbMTPSL3	-2.02	0.47±0.15	11.8±4.5
LbMTPSL4	-0.82	0.61±0.14	8.6±4.5
LbMTPSL5	-1.01	0.59±0.14	9.0±4.6

The predicted structure for LbMTPSL1(**Figure 17A**) has closest structural similarity to Epi-Isozizaene synthase (PDB ID: 3kb9) from *S. coelicolor* (strain ATCC BAA-471 / A3(2) / M145). The aspartate rich motif DDxxE and second metal binding motif NSE were located in the middle of α helices. The superposition of LbMTPSL1 model with X-ray structure of Epi-isozizaene synthase have shown that DDxxE motif was located at similar position as that of DDxxD motif in three-dimensional structure of Epi-isozizaene synthase (**Figure 17B**). Partial sequence alignment of LbMTPSL1 with three hits from PDB containing conserved first and second metal binding motifs were shown in **Figure 17C** along with their positions in homology model (**Figure 17D**) and are consistent with the conservations of class I terpene cyclases.

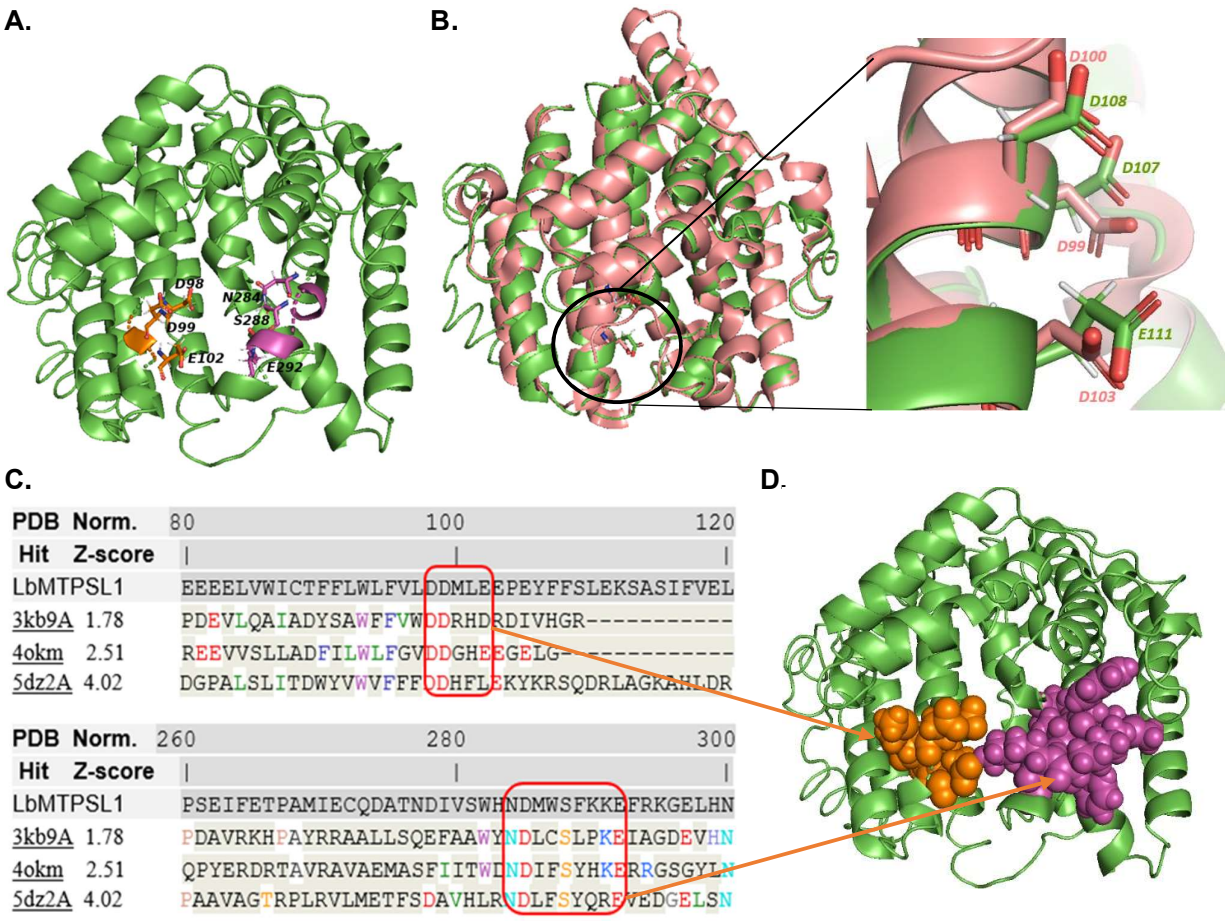

Figure 17. Structural model of LbMTPSL1 obtained by using I-TASSER program showing metal binding motifs DDxxE (orange); NSE (magenta). (B) Superposition of LbMTPSL1 (green) model with X-ray structure of top hit template Epi-isozizaene synthase (salmon) (PDB ID: 3kb9); The DDxxE region is in the circle and the inset shows the closeup view of DDxxE/DDxxD superposition (C) Alignment of LbMTPSL1 enzyme with epi-isozizaene synthase (3kb9), selinadiene synthase (4okm), geosmin synthase (5dz2) by using I-TASSER program (D) LbMTPSL1 model showing DDxxE (orange) and NSE (magenta) motifs in space filling.

The computational structural model of LbMTPSL3 (**Figure 18A**) was deduced by matching predicted model with Selinadiene synthase (PDB ID: 4okm) from *S. pristinaespiralis* (strain ATCC 25486 / DSM 40338 / CBS 914.69 / JCM 4507 / NBRC 13074 / NRRL 2958 / 5647). Superposition of LbMTPSL3 model with X-ray structure of its analogue shows that canonical DDxxD motif of LbMTPSL3 and non-canonical DDxxx motif of Selinadiene synthase occupy similar stereoscopic location with the first two aspartates in same orientations (**Figure 18B**). According to these results, the canonical aspartate rich DDxxD motif of model share similar catalytic function to that of non-canonical DDxxx motif of Selinadiene synthase. Partial sequence alignment of LbMTPSL3 with three PDB hits containing identical first and second

metal binding motifs were shown in **Figure 18C** along with their positions in homology model (**Figure 18D**).

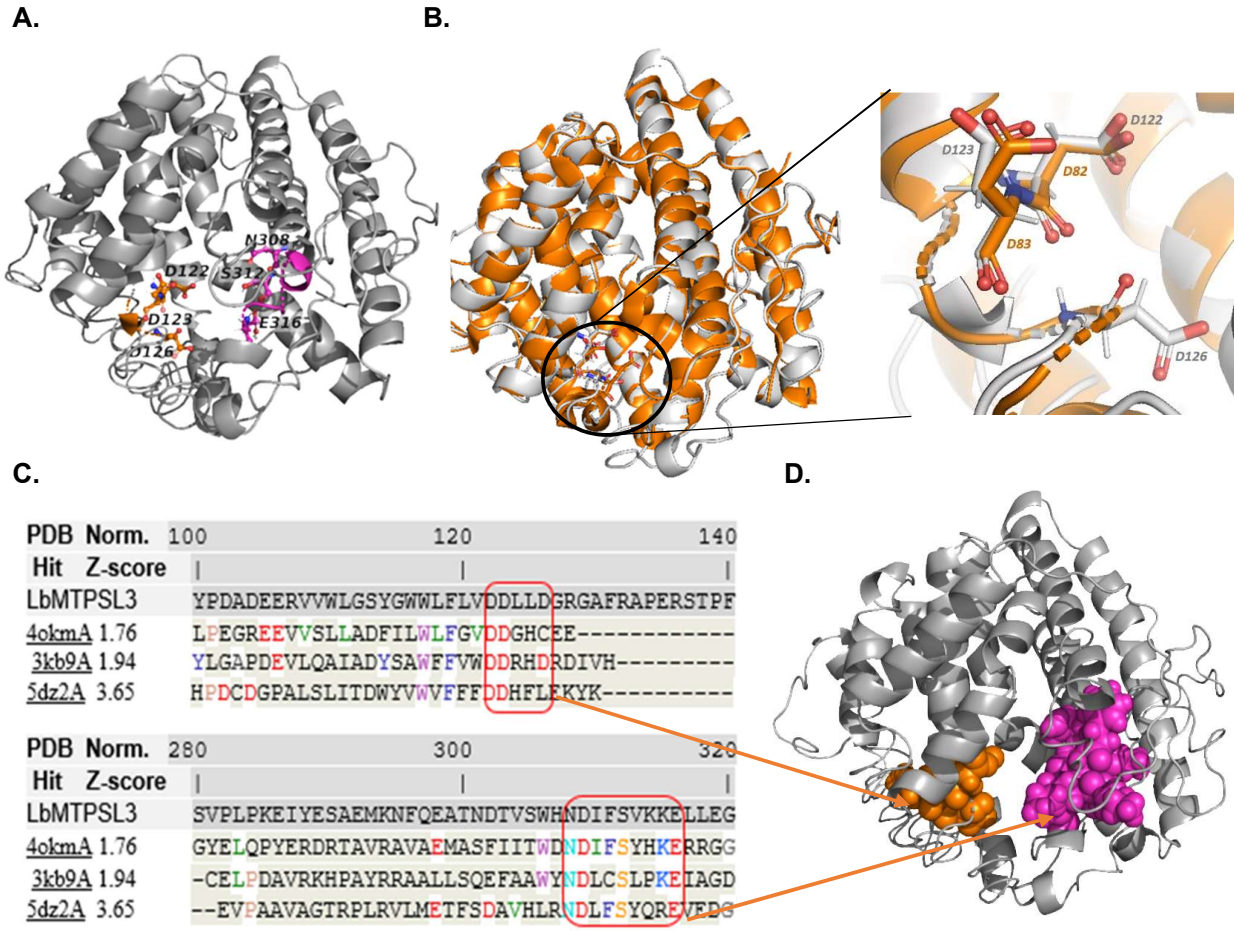

Figure 18. Structural model of LbMTPSL3 obtained by using I-TASSER program showing metal binding motifs DDxxD (orange); NSE (magenta). (B) Superposition of LbMTPSL3 (grey) model with X-ray structure of top hit template Selinadiene synthase (orange) (PDB ID: 4okm); The DDxxD region is in the circle and the inset shows the closeup view of DDxxD/DDxxx superposition (C) Alignment of LbMTPSL3 enzyme with epi-isozizaene synthase (3kb9), selinadiene synthase (4okm), geosmin synthase (5dz2) by using I-TASSER program (D) LbMTPSL3 model showing DDxxD (orange) and NSE (magenta) motifs in space filling.

The predicted homology model for LbMTPSL4 (**Figure 19A**) was identical to LbMTPSL1. The aspartate rich metal binding motifs of both enzymes share same conservation of residues DDxxE which was superposed well with DDxxD motif in three-dimensional structure of Epi-isozizaene synthase (**Figure 19B)**. Only minor differences in the spatial orientations of aspartate sidechains were visible in LbMTPSL1 and LbMTPSL4 when aligned with their template.

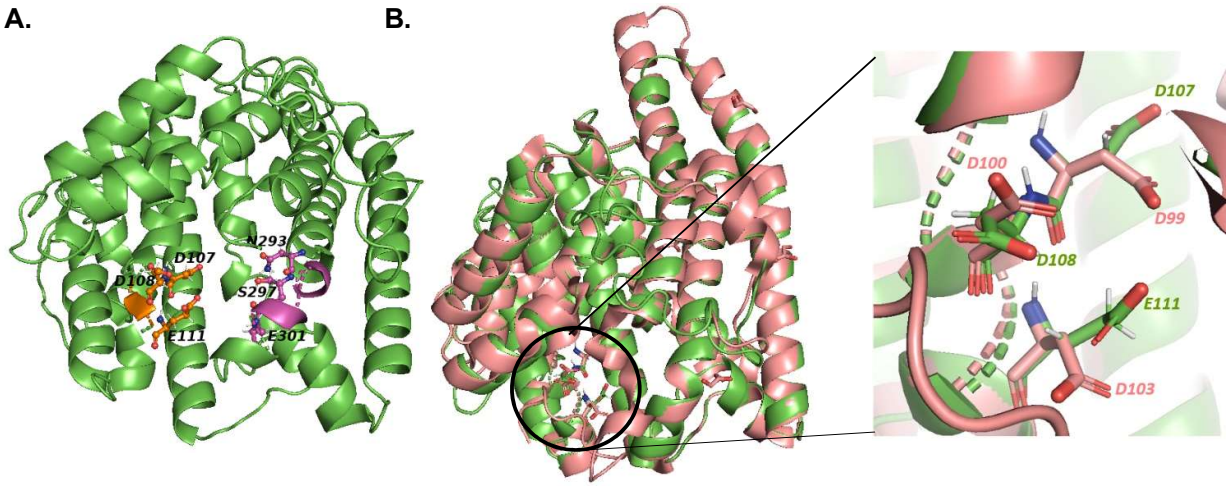

Figure 19. Structural model of LbMTPSL4 obtained by using I-TASSER program showing metal binding motifs DDxxE (orange); NSE (magenta). (B) Superposition of LbMTPSL4 (green) model with X-ray structure of top hit template Epi-isozizaene synthase (salmon) (PDB ID: 3kb9); The DDxxE region is in the circle and the inset shows the closeup view of DDxxE/DDxxD superposition.

The predicted molecular model for LbMTPSL5 (**Figure 20A**) had identical stereoview as that of LbMTPSL3. Though the first metal binding motif of these enzymes was canonical DDxxD, the first two aspartates were well aligned with non-canonical DDxxx motif of template Selinadiene synthase when superposed (**Figure 20B**). The spatial distribution of aspartate sidechains were slightly variable when these enzymes were aligned with their template.

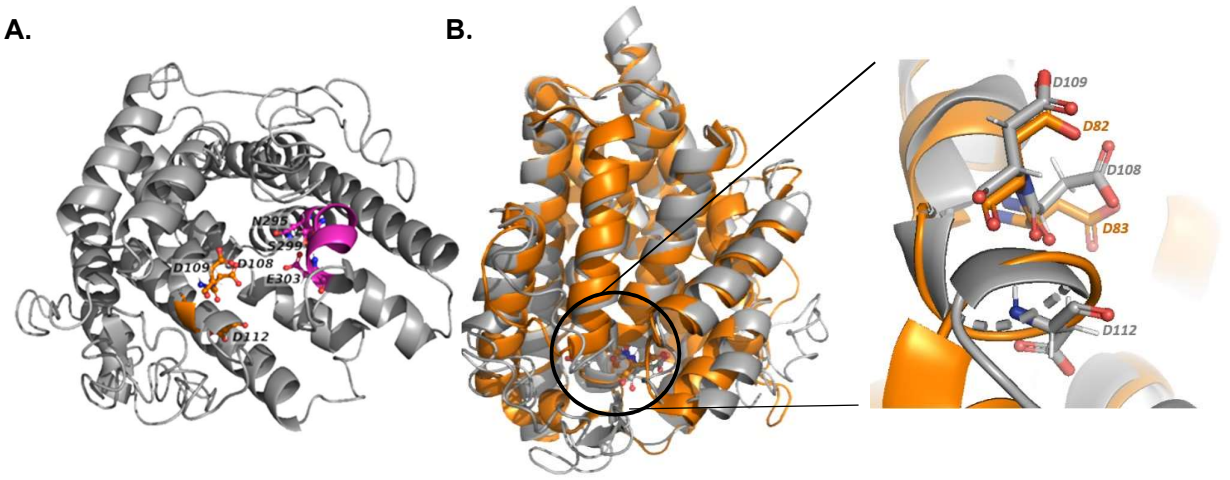

Figure 20. Structural model of LbMTPSL5 obtained by using I-TASSER program showing metal binding motifs DDxxD (orange); NSE (magenta). (B) Superposition of LbMTPSL5 (grey) model with X-ray structure of top hit template Selinadiene synthase (orange) (PDB ID: 4okm); The DDxxD region is in the circle and the inset shows the closeup view of DDxxD/DDxxx superposition

Structure based function prediction

As stated before, LbMTPSL1 & LbMTPSL4 share similar homology based structural model and the predicted prenyl phosphate binding ligands were structural analogues Farnesyl diphosphate (FPP) and Farnesyl thiopyrophosphate (FsPP) respectively. The active site of top hit template Epi-isozizaene synthase is complexed with 3 Mg^{+2} ions, inorganic pyrophosphate and benzyltriethylammonium cation (**Figure 21A**). The overview of the ligand and conserved residues involved in Ligand binding sites for LbMTPSL1 and 4 were depicted in **Figure 21B & 21C** respectively.

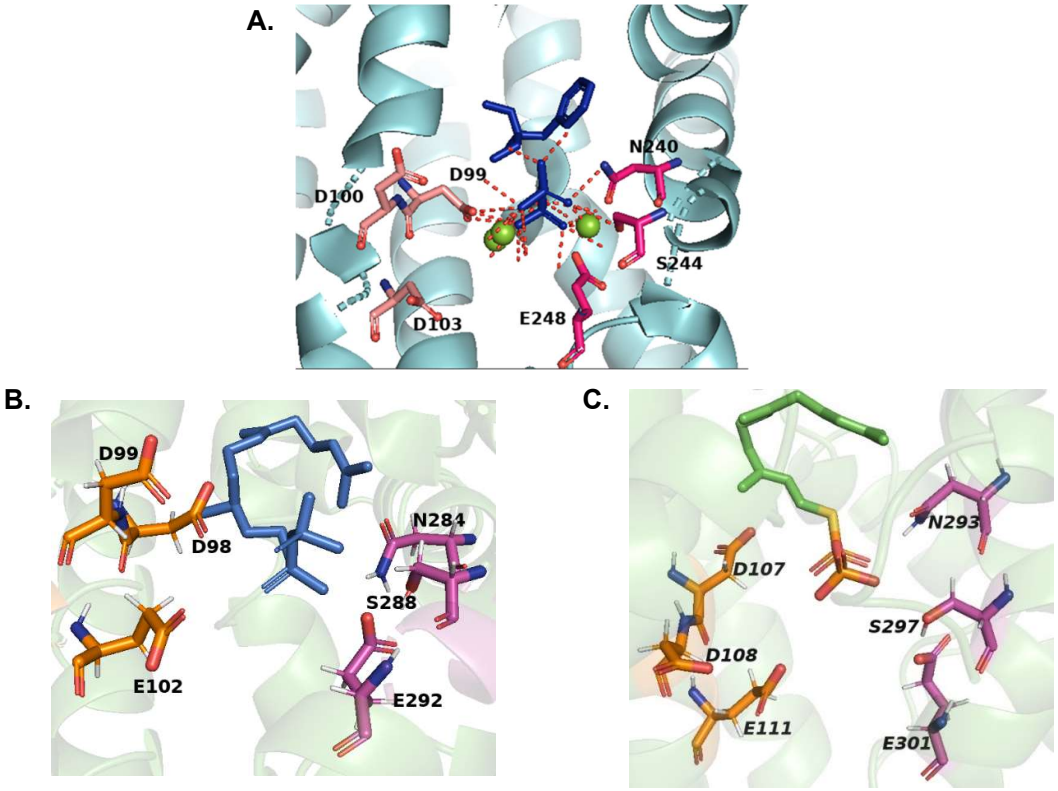

Figure 21. (A) Stereoview of Epi-isozizaene synthase complex with 3 Mg^{+2} ions (green spheres), inorganic pyrophosphate (blue stick) and benzyltriethylammonium cation (blue ringed stick) with metal coordination and H-bond interactions (red dashed lines). (B) Stereoview of LbMTPSL1 active site with metal binding motifs and liganded with FPP (Mg^{+2} ions and bond interactions not shown). (C) Structure of modelled LbMTPSL4 active site complexed with FsPP (Mg^{+2} ions and bond interactions not shown).

The spatial configuration of residues in active site of Epi-isozizaene synthase led to the understanding that 3 Mg^{+2} ions, 3 basic groups R194, K247, R338 and phenolic hydroxyl group of Y339 constitute the molecular recognition complex for diphosphate group and contributes required forces for substrate ionization (**Figure 22A)** and all class I terpene

cyclases contain 3 metal ions and 3 basic residues in their active site phosphate recognition motifs (Christianson, 2017). The aromatic residues F95, F96 and F198 have been shown to be involved in cation-π interactions for stabilizing transition states of epi-isozizaene synthase (**Figure 22B;** Christianson, 2017). Most of the terpene synthases tend to possess two or three aromatic amino acids in their active sites allocated with the function of stabilizing carbocation intermediates (Christianson, 2017). Residues involved in molecular recognition motifs and cation-π interactions were predicted in active site contour of LbMTPSL1 and 4 (**Table 11**)

Table 11. Predicted active site residues of LbMTPSL1 and 4 involved in substrate recognition motif and transition state stabilization

Protein	Molecular recognition motif residues	Transition state stabilizing residues
Epi-isozizaene synthase	R194, K247, R338, Y339	F95, F96, F198
LbMTPSL1	R236, K291, R373, Y210	F91, F95, F374
LbMTPSL4	R245, K300, R382, Y219	F104, F376, F383

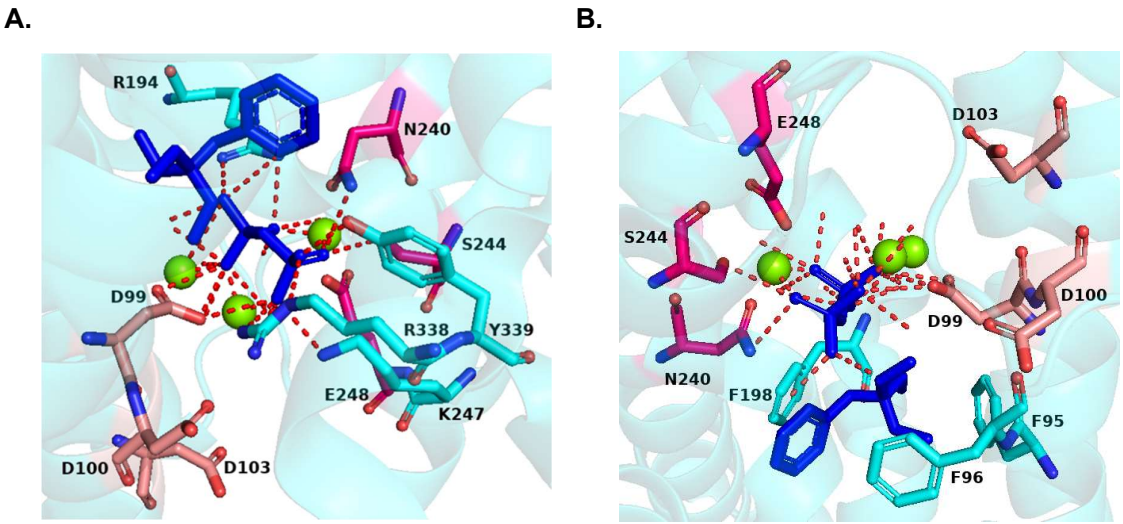

Figure 22. (A)Three Mg^{+2} ions and three basic residues (R194, K247, R338) along with Y339 in cyan forming diphosphate recognition motif in Epi-isozizaene synthase. (B)Representation of carbocation stabilizing aromatic amino acid residues (F95, F96 & F198 in cyan) in the active site of Epi-isozizaene synthase.

LbMTPSL3 & LbMTPSL5 models share structural similarity and analogous to same template Selinadiene synthase with the possible predicted ligand for both proteins as Farnesyl diphosphate (FPP). The ligand and conserved residues involved in Ligand binding sites of LbMTPSL3 and 5 were depicted in **Figure 23B & 23C** respectively with reference to active site of Selinadiene synthase (**Figure 23A**) complexed with inorganic pyrophosphate**.**

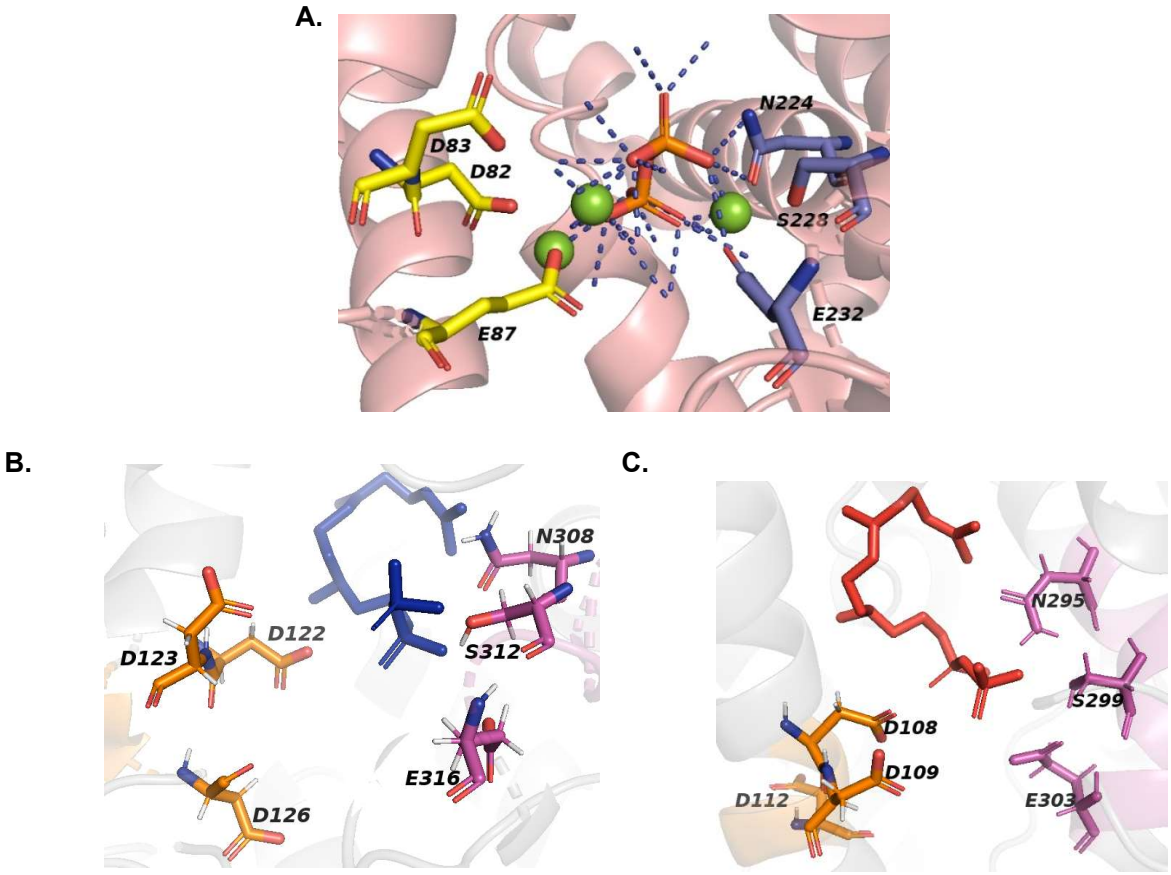

Figure 23. (A) Stereoview showing 3 Mg^{+2} (green spheres) metal coordination and H-bond interactions (blue dashed lines) in the active site of Selinadiene synthase-Inorganic pyrophosphate complex. (B) Stereoview of LbMTPSL3 active site bound to FPP (Mg^{+2} ions and bond interactions not shown). (C) Spatial view of LbMTPSL5 active site liganded with FPP (Mg^{+2} ions and bond interactions not shown).

Aromatic residues F55 and F79 were found to have a role in stabilizing carbocation intermediates through cation-π interactions within active site of Selinadiene synthase (Christianson, 2017). Diphosphate moiety recognition and carbocation stabilization functions were presumably attributed to basic and aromatic residues respectively in active sites of LbMTPSL3 and 5 (**Table 12**)

Table 12. Predicted active site residues of LbMTPSL 3 and 5 engaged in transition state stabilization and diphosphate recognition

Protein	Transition state stabilizing residues	Diphosphate recognition motif residues
Selinadiene synthase	F55, F79	R178, K231, R310, Y311
LbMTPSL3	F119, Y234	R260, K315, R398, Y399
LbMTPSL5	F81, W220	R248, K302, R385, Y386

The similarity of the positions and orientations of conserved aminoacids in the active sites of LbMTPSL protein models with their templates (**Figure 21; Figure 23**) suggests that the first metal binding motif DDxxD/E coordinates to two Mg^{+2} ions and second metal binding motif chelates one Mg^{+2} ion. Along with active site residues indulged in substrate recognition and cation-π interactions, there are also other actual residues that might be significant for three-dimensional shape of active site pocket as predicted by I-TASSER for all LbMTPSLs (**Table 13**). The C-scores are reasonably good for ligand predictions (**Table 13**).

Table 13. Possible ligands and overall residues of predicted Ligand binding sites of Lb MTPSLs.

Protein	Ligand	C-Score	Overall residues of predicted Ligand binding sites
LbMTPSL1	FPP	0.21	71,94,95,98,102,210,236,240,241,245,284,288,291,292,373,374
LbMTPSL3	FPP	0.41	95,118,119,122,234,260,264,265,269,308,312,315,316,398,399
LbMTPSL4	FPS	0.42	84,100,104,107,219,245,249,254,293,297,300,301,376,382,383
LbMTPSL5	FPP	0.28	81,104,105,108,113,220,248,252,253,254,257,291,295,299,302,303,372,385,386

Phylogenetic analysis

Phylogenetic analysis was performed for deciphering the evolutionary relationship of LbMTPSLs with the other characterized MTPSLs of five different lineages of Bryophytes and microbial TPSs. These incude 5 MTPSLs each of *Anthoceros agrestis* and *Anthoceros punctatus* species that belong to hornworts, 5 MTPSLs from *M. polymorpha* which is a liverwort, 5 MTPSLs from *S. moellendorfii* a lycophyte, 3 TPSs from mosses, 3 TPSs from different monilophytes and TPSs from bacteria and fungi (**Table S5)**. The phylogenetic reconstruction tree was inferred through comparative analysis of provided MTPSL molecular sequences by MEGA-X software as shown in **Figure 24**. The topology of the tree indicates that the bacterial TPSs and MTPSLs from non-seed plants were clustered in to two distinct groups. According to the phylogenetic tree, the distribution pattern of MTPSLs from non-seed plants is lineage specific. Nevertheless, some of the MTPSL genes are not placed in species specific evolutionary branch. The two fungal TPSs were embedded among the clusters containing hornworts with mosses and lycophytes with monilophytes. LbMTPSLs are found to be in nearest neighborhood with MTPSLs of another liverwort *M. polymorpha* in a single cluster. The orthologous nature of AaMTPSLs and ApMTPSLs was empirical and can be observed in the tree. All clades were supported by good bootstrap values.

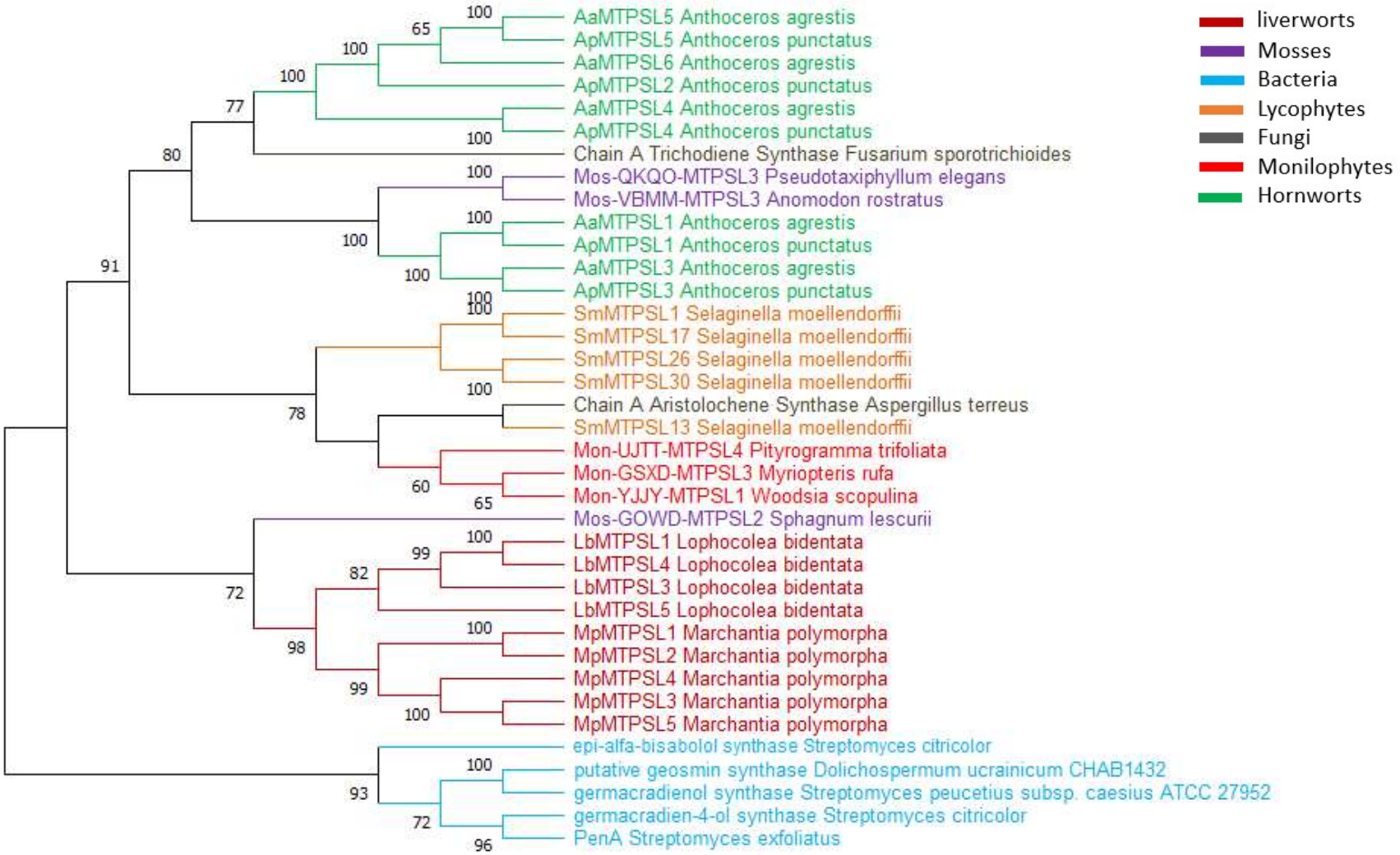

Figure 24. Neighbor joining phylogenetic tree of characterized MTPSLs from non-seed plants including putative LbMTPSLs along with bacterial and fungal TPSs

7. DISCUSSION

The first unforeseen obstacle was complete depletion of Driselase enzyme due to temporary halt in supply from manufacturer (Sigma-aldrich). The optimized protocol for protoplast preparation from *P. patens* protonemal tissue using Driselase (Bach et al., 2014) could not be employed due to the prevailed situation. In order to overcome this problem, attempts were made to produce protoplasts using different combinations of cell wall degrading enzymes comprising cellulase, hemicellulase and pectinase activities. A new unpublished protocol using a mixture of Cellulase-R10 and Macerozyme R-10 enzymes was used for this endeavor. Though not as time efficient as Driselase protocol the new mixture of enzymes can produce considerable number of protoplasts that can be transformed efficiently to produce transgenic lines (**Figure 15**).

Sequence based characterization studies of LbMTPSLs sequences including protein targeting withTargetP-2.0 (**Table 7**), similarity-based structure prediction and structure-based function annotation (I-TASSER) have attributed cytosolic localization and its corresponding sesquiterpene synthase activity for all enzymes with either FPP or FsPP as predicted ligands. In this context it is noteworthy to recollect that the typical plant sesquiterpene synthases in seed plants are cytosol centric and have access to FPP (substrate) that share same intracellular compartment for biosynthesis (Jia et al., 2018, 2016). The sesquiterpene synthase dominance of LbMTPSLs is in accordance with the previous characterized MTPSLs of non-seed plants like *S. moellendorfii*, a lycophyte shows predominant sesquiterpene synthase activities than monoterpene synthase function (Li et al., 2012). Hornworts *A. agrestis* and *A. punctatus* (Xiong et al., 2018), liverwort *M. polymorpha* (Jia et al., 2018) and other non-seed plants (Jia et al., 2016) were consistent with this pattern.

The direct evidence was provided for accumulation of sesquiterpenes in oil bodies of *M. polymorpha* and also some specialized metabolites were observed in oil bodies of other liverworts (Tanaka et al., 2016). As Oil bodies were found in the liverworts belonging to order Marchantiales and Jungermanniales (Tanaka et al., 2016), it has been assumed that sesquiterpene metabolites produced by MTPSLs of *L. bidentata* belonging to order Jungermanniales accumulate in oil bodies when expressed heterologously in *P. patens.* It

will be interesting to increase the number of oil bodies in the vegetative structures of *P. patens* and observe the localization of metabolites. Overexpression of oil body associated proteins such as oleosin, seipin and fibrillin significantly increased the number of oil bodies in the protonemal cells of *P. patens* (Bae et al., 2016)

The homology models of LbMTPSL1 and 4 predicted by I-TASSER is based on structural similarity to Epi-isozizaene synthase, a sesquiterpene synthase from *S. coelicolor* (**Figure 17 & 19**). Likewise, predicted models of LbMTPSL3 and 5 have same fold as Selinadiene synthase from *S. pristinaespiralis* (**Figure 18 & 20**). The characteristic α only domain configuration and other structural similarities significantly the same locus of aspartate rich DDxxD/E motifs resemble LbMTPSLs to terpene synthases of bacteria. Furthermore plant TPSs are characterized by the presence of a 252 amino acid highly conserved N-terminal α helical domain (PF01397) in the form of α barrel and not being detected with any specific biochemical function (Yamada et al., 2015). By contrast LbMTPSLs lack the characteristic highly conserved N-terminal domain which is also a recognizable feature in bacterial TPSs (Yamada et al., 2015). This resemblance of LbMTPSL enzymes to bacterial terpene synthases is also backed by phylogenetic analysis (**Figure 24**). The topology of phylogenetic tree shows a monophyletic relationship of MTPSLs from *M. polymorpha* and *L. bidentata*, the two liverworts species. The plausible explanation could be that the common ancestor of these liverworts might have acquired TPS genes from bacteria (*Streptomyces*). The genus *Streptomyces* which belongs to family Actinomycetes is a plant growth promoting soil bacteria inhabiting rhizosphere and rhizoplane as well as colonize inner tissues of host plants as endophyte (Olanrewaju & Babalola, 2019) providing a physiological means for Horizontal gene transfer of TPS genes. Since typical plant TPSs are only distantly related to MTPSLs (Jia et al., 2018) and remarkably high confidence in structural similarity to bacterial TPSs strongly supports the hypothesis that the non-seed plants might have acquired the genes from plant associated microbes through lateral transfer.

The local alignment method such as BLASTp algorithm when applied to deduced amino acid sequences of LbMTPSL1,3,4 and 5 proteins (**Supplimental data 2**), the results (**Figure S1**) have shown a Conserved domain database (CDD) hit with terpene_cyclase_nonplant_C1(accession:cd00687) conserved domain that belong to

Isoprenoid_Biosyn_C-I superfamily characterized by the presence of two conserved metal binding DDxxD/E and NSE motifs. Web logo, a sequence logo generator has shown this conservation within the multiple sequence alignment of all LbMTPSLs (**Figure 16**). It is speculated that LbMTPSL1,3,4 and 5 enzymes with the two metal binding motifs in the active site coordinate Mg^{+2} bridged binding of diphosphate moiety of FPP and induces substrate ionization to form reactive carbocation intermediates stabilized by aromatic aminoacids of active site contour (Chen et al., 2011; Christianson, 2017).

Despite of homology structures provided by the structure and function prediction program (I-TASSER) and improvement in the understanding of reaction mechanism based on the type of interactions among residues in the active site, it is unrealistic to predict the number and cyclized sesquiterpene products of LbMTPSLs. This can be explained by the intriguing feature of class-I terpene synthases that the indeterminate nature of bond rearrangements like cyclizations, hydride transfers, methyl migrations etc. undergone by the carbocation intermediate results in diverse array of carbon skeletons forming multiple sesquiterpene products from single substrate (Chen et al., 2011; Christianson, 2017). Adding more complexity, the cytochrome P450 enzymes which are species specific and highly substrate specific (Weitzel & Simonsen, 2015) catalyze further modifications through oxidations of terpenoids contributing to structural diversity of terpenes. Nevertheless, there are evidences for the substrate promiscuity of possible cytochrome P450 enzymes in *P. patens* (Ikram et al., 2019). Emphasizing on cytochrome P450 functionality, it could be an interesting approach to coexpress along with LbMTPSL genes, the candidate genes associated with sesquiterpene oxidation obtained from transcriptome data and compare the terpenoid metabolite profile with respect to endogenous cytochrome P450 arsenal of *P. patens*.

According to structure based binding ligand prediction by I-TASSER program, the predicted prenyl diphosphate substrates are FPP for LbMTPSL1 (**Figure 21B**),3,5 (**Figure 23B&C**) and its structural analogue FsPP for LbMTPSL4 (**Figure 21C**). As discussed earlier, a terpene synthase complexed with single substrate is able to generate an array of natural metabolites with identical number of C_5 units. It is also notable that there are evidences that several TPSs have binding affinity to more than one substrate depending on the substrate availability and capable of synthesizing terpenes with varying number of C_5 units (Pazouki & Niinemetst,

2016). The multi-substrate utilization capacity of LbMTPSLs can be unleashed by *in vitro* functional characterization of expressed enzymes with a range of substrates.

Structure guided active site characterization of LbMTPSL1,3,4 and 5 enzymes unraveled the key amino acid residues with relevant chemical nature and hydrophobicity in the vicinity of substrate binding pocket that are possibly indulged in diphosphate recognition and stabilizing carbocation intermediates (**Table 11&12**). This deduced structural information can serve as a guide for mechanistic characterization of LbMTPSLs and concomitant manipulation of cyclization pattern to alter product spectra. For this, the role of putative active site residues in crafting the dynamics of FPP cyclization should be tested by site directed mutagenesis as demonstrated earlier by substitution of residues present in metal binding motifs (Vedula et al., 2005), point mutagenesis in the residue layer of diphosphate recognition motif (Greenhagen et al., 2006; Vedula et al., 2005), combinatorial mutagenesis of multiple residues buried in active site contour to significantly increase the product promiscuity of enzymes (Greenhagen et al., 2006).

8. CONCLUSION

In this study a new protocol which makes use of blend of Cellulase R-10 and Macerozyme R-10 enzymes was found to be promising for preparation of protoplasts from protonema of *P. patens*. The fact that the LbMTPSL genes are acquired from gene expression analysis data and their characterization through bioinformatics approaches along with phylogenetic methods produced data which supports that the genes are functional microbial terpene synthase like genes. Homology based modelling studies predicted sesquiterpene synthase activity for LbMTPSL1,3,4,5 enzymes and backed the hypothesis that the *L. bidentata*, a liverwort (non-seed plant) might have acquired MTPSL genes from bacteria (*Streptomyces* species*)* through horizontal gene transfer by means of endophytic associations which is also in line with the outcome of sequence based phylogenetic analysis. The hypothesis that the MTPSL genes could have equipped the non-seed plants with fitness metabolites could only be tested by determining the biological functions of genes through heterologous expression and terpene profiling in planta host *P. patens*, the otherwise objective of the project. The data-based understanding of structural and chemical biology of LbMTPSL proteins opens an

avenue for protein engineering of these enzymes through Site directed mutagenesis to exploit catalytic plasticity of TPSs and produce a range of cyclized products with possible commercial value.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr.Henrik Toft Simonsen for having chosen me for this project while ensuring professional environment with constant supervision. Special thanks to my internal supervisor Dr.Mette Lubeck for prompt response in times of need. I would like to extend my thanks to Dr.Rituraj batth for technical assistance in lab. Research facility at DTU Bioengineering, Section for synthetic biology was used for this entire study.

Bibliography

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, *2*(5), 1–14. https://doi.org/10.26508/lsa.201900429

Bach, S. S., King, B. C., Zhan, X., Simonsen, H. T., & Hamberger, B. (2014). Heterologous Stable Expression of Terpenoid Biosynthetic Genes Using the Moss Physcomitrella patens. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1153, pp. 257–271). https://doi.org/10.1007/978-1-4939-0606-2_19

Banerjee, A., Arnesen, J. A., Moser, D., Motsa, B. B., Johnson, S. R., & Hamberger, B. (2019). Engineering modular diterpene biosynthetic pathways in Physcomitrella patens. *Planta*, *249*(1), 221–233. https://doi.org/10.1007/s00425-018-3053-0

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Botella-Pavía, P., Besumbes, Ó., Phillips, M. A., Carretero-Paulet, L., Boronat, A., & Rodríguez-Concepción, M. (2004). Regulation of carotenoid biosynthesis in plants: Evidence for a key role of hydroxymethylbutenyl diphosphate reductase in controlling the supply of plastidial isoprenoid precursors. *Plant Journal*, *40*(2), 188–199. https://doi.org/10.1111/j.1365-313X.2004.02198.x

Cao, R., Zhang, Y., Mann, F. M., Huang, C., Mukkamala, D., Hudock, M. P., … Oldfield, E. (2010). Diterpene cyclases and the nature of the isoprene fold. *Proteins: Structure, Function and Bioinformatics*, *78*(11), 2417–2432. https://doi.org/10.1002/prot.22751

Chen, F., Ludwiczuk, A., Wei, G., Chen, X., Crandall-Stotler, B., & Bowman, J. L. (2018). Terpenoid Secondary Metabolites in Bryophytes: Chemical Diversity, Biosynthesis and Biological Functions. *Critical Reviews in Plant Sciences*, *37*(2–3), 210–231. https://doi.org/10.1080/07352689.2018.1482397

Chen, F., Tholl, D., Bohlmann, J., & Pichersky, E. (2011). The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant Journal*, *66*(1), 212–229. https://doi.org/10.1111/j.1365-313X.2011.04520.x

Christianson, D. W. (2017). Structural and Chemical Biology of Terpenoid Cyclases. *Chemical Reviews*, *117*(17), 11570–11648. https://doi.org/10.1021/acs.chemrev.7b00287

Crooks, G., Hon, G., Chandonia, J., & Brenner, S. (2004). NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. *Genome Res*, *14*, 1188–1190. https://doi.org/10.1101/gr.849004.1

Engels, B., Dahm, P., & Jennewein, S. (2008). Metabolic engineering of taxadiene biosynthesis in yeast as a first step towards Taxol (Paclitaxel) production. *Metabolic Engineering*, *10*(3–4), 201–206. https://doi.org/10.1016/j.ymben.2008.03.001

Greenhagen, B. T., O'Maille, P. E., Noel, J. P., & Chappell, J. (2006). Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9826–9831. https://doi.org/10.1073/pnas.0601605103

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

Ikram, N. K. K., Kashkooli, A. B., Peramuna, A., Van Der Krol, A. R., Bouwmeester, H., & Simonsen, H. T. (2019). Insights into heterologous biosynthesis of Arteannuin B and artemisinin in physcomitrella patens. *Molecules*, *24*(21). https://doi.org/10.3390/molecules24213822

Ikramt, N. K. B. K., Zhan, X., Pan, X. W., King, B. C., & Simonsen, H. T. (2015). Stable heterologous expression of biologically active terpenoids in green plant cells. *Frontiers in Plant Science*, *6*(MAR). https://doi.org/10.3389/fpls.2015.00129

Jia, Q., Köllner, T. G., Gershenzon, J., & Chen, F. (2018). MTPSLs: New Terpene Synthases in Nonseed Plants. *Trends in Plant Science*, *23*(2), 121–128. https://doi.org/10.1016/j.tplants.2017.09.014

Jia, Q., Li, G., Köllner, T. G., Fu, J., Chen, X., Xiong, W., … Chen, F. (2016). Microbial-type terpene synthase genes occur widely in nonseed land plants, but not in seed plants. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(43), 12328–12333. https://doi.org/10.1073/pnas.1607973113

King, B. C., Vavitsas, K., Ikram, N. K. B. K., Schrøder, J., Scharff, L. B., Hamberger, B., … Simonsen, H. T. (2016). In vivo assembly of DNA-fragments in the moss, Physcomitrella patens. *Scientific Reports*, *6*, 2–6. https://doi.org/10.1038/srep25030

Kumar, Santosh, Kempinski, C., Zhuang, X., Norris, A., Mafu, S., Zi, J., … Chappell, J. (2016). Molecular diversity of terpene synthases in the liverwort marchantia polymorpha. *Plant Cell*, *28*(10), 2632–2650. https://doi.org/10.1105/tpc.16.00062

Kumar, Sudhir, Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, G., Köllner, T. G., Yin, Y., Jiang, Y., Chen, H., Xu, Y., … Chen, F. (2012). Nonseed plant Selaginella moellendorffii has both seed plant and microbial types of terpene synthases. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), 14711–14715. https://doi.org/10.1073/pnas.1204300109

Lichtenthaler, H. K. (1999). the 1-Deoxy-D-Xylulose-5-Phosphate Pathway of Isoprenoid Biosynthesis in Plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, *50*(1), 47–65. https://doi.org/10.1146/annurev.arplant.50.1.47

Morris, J. L., Puttick, M. N., Clark, J. W., Edwards, D., Kenrick, P., Pressel, S., … Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(10), E2274–E2283. https://doi.org/10.1073/pnas.1719588115

Murén, E., Nilsson, A., Ulfstedt, M., Johansson, M., & Ronne, H. (2009). Rescue and characterization of episomally replicating DNA from the moss Physcomitrella. *Proceedings of the National Academy of Sciences of the United States of America*,

106(46), 19444–19449. https://doi.org/10.1073/pnas.0908037106

Nørholm, M. H. H. (2010). A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnology*, *10*. https://doi.org/10.1186/1472-6750-10-21

Olanrewaju, O. S., & Babalola, O. O. (2019). Streptomyces: implications and interactions in plant growth promotion. *Applied Microbiology and Biotechnology*, *103*(3), 1179–1188. https://doi.org/10.1007/s00253-018-09577-y

Pazouki, L., & Niinemetst, U. (2016). Multi-substrate terpene synthases: Their occurrence and physiological significance. *Frontiers in Plant Science*, *7*(JULY2016), 1–16. https://doi.org/10.3389/fpls.2016.01019

Peramuna, A., Bae, H., Rasmussen, E. K., Dueholm, B., Waibel, T., Critchley, J. H., … Simonsen, H. T. (2018). Evaluation of synthetic promoters in Physcomitrella patens. *Biochemical and Biophysical Research Communications*, *500*(2), 418–422. https://doi.org/10.1016/j.bbrc.2018.04.092

Prigge, M. J., & Bezanilla, M. (2010). Evolutionary crossroads in developmental biology: Physcomitrella patens. *Development*, *137*(21), 3535–3543. https://doi.org/10.1242/dev.049023

Reski, R., Bae, H., & Simonsen, H. T. (2018). Physcomitrella patens, a versatile synthetic biology chassis. *Plant Cell Reports*, *37*(10), 1409–1417. https://doi.org/10.1007/s00299-018-2293-6

Ro, D. K., Paradise, E. M., Quellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., … Keasling, J. D. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, *440*(7086), 940–943. https://doi.org/10.1038/nature04640

Roberts, A. W., Roberts, E. M., & Haigler, C. H. (2012). Moss cell walls: Structure and biosynthesis. *Frontiers in Plant Science*, *3*(JUL), 1–7. https://doi.org/10.3389/fpls.2012.00166

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rodríguez-Concepción, M. (2014). Plant Isoprenoids: A General Overview. In M. Rodríguez-Concepción (Ed.), *Plant Isoprenoids* (pp. 1–5). https://doi.org/10.1007/978-1-4939-0606-2_1

Rodrı, M. (2002). Elucidation of the Methylerythritol Phosphate Pathway for Isoprenoid Biosynthesis in Bacteria and Plastids. *Plant Physiology*, *130*(November), 1079–1089. https://doi.org/10.1104/pp.007138.ISOPRENOID

Roy, A., Yang, J., & Zhang, Y. (2012). COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, *40*(W1), 471–477. https://doi.org/10.1093/nar/gks372

Schaefer, D. G., & Zry, J. (2001). The Moss. *Society*, *127*(December), 1430–1438. https://doi.org/10.1104/pp.010786.1430

Schaefer, D. G., & Zrÿd, J. P. (1997). Efficient gene targeting in the moss Physcomitrella patens. *Plant Journal*, Vol. 11, pp. 1195–1206. https://doi.org/10.1046/j.1365-

313X.1997.11061195.x

Simonsen, H. T., Drew, D. P., & Lunde, C. (2009). Perspectives on using Physcomitrella patens as an alternative production platform for thapsigargin and other terpenoid drug candidates. *Perspectives in Medicinal Chemistry*, *2009*(3), 1–6. https://doi.org/10.4137/pmc.s2220

Takahashi, S., Yeo, Y., Greenhagen, B. T., McMullin, T., Song, L., Maurina-Brunker, J., … Chappell, J. (2007). Metabolic engineering of sesquiterpene metabolism in yeast. *Biotechnology and Bioengineering*, *97*(1), 170–181. https://doi.org/10.1002/bit.21216

Tanaka, M., Esaki, T., Kenmoku, H., Koeduka, T., Kiyoyama, Y., Masujima, T., … Matsui, K. (2016). Direct evidence of specific localization of sesquiterpenes and marchantin A in oil body cells of Marchantia polymorpha L. *Phytochemistry*, *130*, 77–84. https://doi.org/10.1016/j.phytochem.2016.06.008

Trapp, S. C., & Croteau, R. B. (2001). Genomic organization of plant terpene synthases and molecular evolutionary implications. *Genetics*, *158*(2), 811–832.

Vedula, L. S., Rynkiewicz, M. J., Pyun, H. J., Coates, R. M., Cane, D. E., & Christianson, D. W. (2005). Molecular recognition of the substrate diphosphate group governs product diversity in trichodiene synthase mutants. *Biochemistry*, *44*(16), 6153–6163. https://doi.org/10.1021/bi050059o

Weitzel, C., & Simonsen, H. T. (2015). Cytochrome P450-enzymes involved in the biosynthesis of mono- and sesquiterpenes. *Phytochemistry Reviews*, *14*(1), 7–24. https://doi.org/10.1007/s11101-013-9280-x

Xiong, W., Fu, J., Köllner, T. G., Chen, X., Jia, Q., Guo, H., … Chen, F. (2018). Biochemical characterization of microbial type terpene synthases in two closely related species of hornworts, Anthoceros punctatus and Anthoceros agrestis. *Phytochemistry*, *149*, 116–122. https://doi.org/10.1016/j.phytochem.2018.02.011

Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-ya, K., Omura, S., Cane, D. E., & Ikeda, H. (2015). Terpene synthases are widely distributed in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 857–862. https://doi.org/10.1073/pnas.1422108112

Yang, J., & Zhang, Y. (2015). I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Research*, *43*(W1), W174–W181. https://doi.org/10.1093/nar/gkv342

Zhan, X., Bach, S. S., Hansen, N. L., Lunde, C., & Simonsen, H. T. (2015). Additional diterpenes from Physcomitrella patens synthesized by copalyl diphosphate/kaurene synthase (PpCPS/KS). *Plant Physiology and Biochemistry*, *96*, 110–114. https://doi.org/10.1016/j.plaphy.2015.07.011

Zhan, X., Zhang, Y. H., Chen, D. F., & Simonsen, H. T. (2014). Metabolic engineering of the moss physcomitrella patens to produce the sesquiterpenoids patchoulol and α/β-santalene. *Frontiers in Plant Science*, *5*(NOV), 1–10. https://doi.org/10.3389/fpls.2014.00636

Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins: Structure, Function and Bioinformatics*, *77*(SUPPL. 9), 100–113. https://doi.org/10.1002/prot.22588

Appendix

Supplimental data 1. LbMTPSL1,3,4,5 gene sequences in FASTA format

>LbMTPSL1_TRINITY_DN210373_c0_g1_i3_selection-1_LbMTPSL1_gene len=1128

ATGGAGGTGCCAGAAACGAAGGAGGAAAGCTGTGACAAAGTACAATCCTACAAGGTG
CCAGAGTTCATCTCGCCTTATCCAGCAAGGCAGAACGCTGTCGCTCCAAGAGTAGGC
CCTCGGTGCCAGACCTGGTTTGACGAACAATCACTCTCCACAGTCTTCTCTAACCCGA
GCGACTGGCAGTTCATTCTGGAGTGCAGAGTTTACGACTTGCCCACTTGGATCTTCGT
TGATGCTGAGGAGGAAGAGCTGGTGTGGATCTGTACCTTTTTCCTGTGGTTGTTCGTG
CTCGATGATATGCTCGAGGAGCCTGAGTACTTCTTCTCGCTGGAGAAGTCGGCCTCGA
TCTTCGTGGAGCTCAACTTGCTCATCATGTGGACCTTTCCTGACGACCCCTCCATTCG
CGAAGTCTTCACAAAGTTACTCATGACCCAACAGCCTGAGCAGCGATCAGACACCATA
AAATATGTCGACGCCAAGCTTGTAGAGGCCAGATTACAGCCTGGCACAGTATATGACA
TTGCAAAAATTGGGCCAGTCGGCATAGCATTGAAGGATCTGTGGGTGAGACTCATAAG
TGCAACGCCCACCAAATCGGCCATCCGATGGGCCGTCACGCTGCTACGCTATATTCTC
GGCAATGCTGAAGAGACTAGGAACCGGAATAAGAAAACGTTCCCTAGCTCTGCTGACT
ATGTTGCTCTCCGCCGAAATATCTCGGCTGTGGAACCATGCTATGTGATAGTGGATTT
CATGGACAAGGTGAGCGAAGCTCTGCCTAGCGAGATTTTCGAAACTCCTGCAATGATC
GAGTGTCAGGACGCCACCAACGACATAGTGTCTTGGCATAACGACATGTGGTCTTTCA
AGAAGGAGTTTCGAAAAGGAGAACTGCATAACCTGGTGTACATAACCAGCCAAGAACG
AGGATGCTCTTTCTCGGAAGCTGGGGATTTGGTCCTCGAGATGATTTATAGAAGACTG
GAAGATCTTGTCCAAGTCTTTGCAGACCTGGAGAAGATGACACCGCTTGAGCACCAGC
AAGGTGTTGCAGGGTACATCAAGGCATCGAAGTTCTGGATTTCGGGTACTCATCAGTT
TCATAGGACAAGCAAACGTTTCACCTAG

>LbMTPSL3_TRINITY_DN174979_c0_g1_i1_selection_selection len=1287

ATGGTACGTGATATGAATTCCGCTGGCGCTGGAGAAGTTGCAAATGCTCAGTTCCCAG
AATTTCCCCCTGCGTTGTTTGCCGGGAGAACGAATGAACAGATGATGGCAGAAATCAA
CTCTCTCAAGCCGCCGAAGTTTTACCCTCCATATCCATCGAGACGGAACCGAAATGCA
ACAAAGGCAAACGTAGAGAGCATCGCCTGGTTACATGAGTATAAGGTGGGAACGGTTT
TCGAGGATCCGAAGCGATGGGAACACTTCCAGCGGATGAAGTTGAGCGAGATTATGA
CACGGGTCTACCCCGACGCGGACGAGGAGCGGGTGGTGTGGCTCGGCTCCTATGGG
TGGTGGCTGTTTTTGGTCGACGATTTGCTCGATGGGCGTGGGGCGTTTCGTGCACCG
GAAAGGTCCACTCCATTCTTCGTGGAGGTCAATCTGACGATGCTGTGGTCCTTTCCCG
ACAGTCCCGTTCTCTACAATACATTTGTAGAGATAGTGGAGTGCTTTCCTGAGGACCA
GCGACGTGGGATCTTGGAGGATGTCAGCGCCAAGCTTGTCAAGGCCAGAGAACATCC
TGGCTCAGTTTACGATACTTCCCATTCCGGGCCGGTGCTTGAGTGTTTCAAAGATTTAT
GGGTGAAACTGCTGGCATCGACGCCCCACGAATCTATCATTAGATGGGGAAATTCGAT
CCAGACTTACCTCTTAGGGAACGTCATTGAGGCCAAGAACCGAACTCACGGCAACATA
CCTTCTATCGCCGAGTACATCGACATACGCAGAAATACGTCTTCTATGTACCCGTGCAT
GTGGATAGTCGACTACTCAGAGAATTTCAGCGTTCCTTTACCGAAGGAGATCTACGAG
AGTGCTGAAATGAAAAACTTTCAGGAGGCTACCAATGATACTGTGTCGTGGCACAACG
ACATCTTTTCAGTCAAGAAGGAACTGCTGGAGGGCGAAGTGCACAACCTGGTAACTGT
GGTCAGCCACGAGCGTAAGTGCTCATTTCAAGAAGCCATGTGGATTGCCGTGGGGAT

GCTTCACGACAGGCTGCAGGACTTGGACCGAGCCGTTCTTGACTTGGAGGCAATCAC
ACCGCCAGAGCACTCGCAGATGGTTGCAGGTTACGTGAAAACTGCCCACTGCTGGTT
CTCCGGATCACACGATTTTACTGTCGTGAGTAATGAACGGTACTCGTGGGACATGTCC
GCATCCATTCATTGCCTCCTTTGTCACAACGCACCGCCGAAGGAAGCGAAAGTTATTT
CAGGCCTTTTCGGTTAA

>LbMTPSL4_TRINITY_DN216420_c0_g1_i8_selection_LbMTPSL4_gene len=1155

ATGGGAGCGTTAGAAGGTGATGAGGTTCTGATGCCACTGTCTCTCACCAAGGAAGAGA
AACAGCAGATCGAATTGTTCAAAATTCCAGCATTCATCTCTCCCTATCCAGTGATGAGA
AATGCTCTCATAGACAATGTAGAATCAAAATGCCTGCTGTGGTTTGAAAAACAGTCACT
ACACACAGTCTTCGCTGATCCGAAGGACTGGAATCGCTTGATGGCATGCAAACTGTAT
GTGATTCCCGGCTACATCTTCATGGACGCTGAGGAGGAGGAGCTGTTGTGGGGTGCT
CTTTATACGGTGTGGTTGTTTGCGCTCGACGATATGCTCGAAGAGAATGACTACTTCCA
GTCGCTGGGCAAGTCAGATTCGATTTTCGTCGAGCTCATGCTGGTGATCATGTGGGCT
TTTCCCGATGAGGCTGCCGTCCGTGACATCTTTTTACAGTTGGTGTCTATATTGCCTGA
GGAGAGCCGAGAACCCACCATGAAATACGCCGATGCCAAGCTCGCTGAGGCCAGATT
ACGCCCCGGCACAGGATATGATGCTAAGAAGATCGGGCCGATAGGGGTTTCTTTTAGT
GATCTCTGGATGACTTACGTTAGATCAACCCCTACAGATTCGGCCATCCGGTGGGGAC
TCTCGAATCAGCGTTATTTTCTCGGAAACGCAAGCGAGACCAGGAACCGAAATCTTCA
GTCCATCCCTTCCTCTGCTGATTTTGTTTTTCTTCGTCGAAAACTCTCGGCTGTGGAAC
CGTGCATGGTCAAAGTGGACTACATATGCAAGCTGAGCCCCAGCCTGCCCAGCGAGG
TATACGACACTCCTCAAATGGAAGAATTGCTTGACGCGACCAACGACATAGTGGCCTG
GCATAACGACATCTGGTCTTTCAAGAAGGAGTTGAGGAAAGGAGAATTGCACAACTTG
GTCTTCATAACTAGCCGCGAGAGAGGGTGCTCGTTCTCGGCGGCTGCAGAAGTTGTC
ATGGACAAGGTTTACAGCCGACTTCGAGATCTCGCCCAAAGCTTTGTGGACCTGGAGA
AGATTACACCGCCAGAGCACCACCACGCCAGTGCCATGTACATAAAAACCGCCAAATT
ATGGGTTTCGGGAACACATCAATTTCACTCCACAAACAAACGGTTCGACTAG

>LbMTPSL5_TRINITY_DN216972_c0_g2_i1_selection_LbMTPSL5_gene len=1164

ATGGCTGCTGCTGAAGCAATTCCTGCTGGCACGTCGGCATTTTCGAGCTCAACCGACA
ACGAGTTTGTTAAGACTTTCAGACCTCCATTGCTTGACAGTTCGTATCCTCTCAACATC
CATCCCAAGTTTTCAAGCTCGGAAGCCAGGGACACAATCGAAAAATGGATGCAGCTGC
ACAAGGTGGATGGCATTTTCACTCCGGAGGGGTACAAGCTGCTCTTGGATATGGACAT
CCCTGCATTCGGGGGCCGAATCTTTACGGAGGCTCCTGAGGAAGGGCTGGAATGGG
GTATCAAGTTCTTGTTCATGCTCTGGATTTGGGATGACACCATGGACTCCACTGAACTA
GGGCTATCCCCCGAGACGGCTCTCTCCCCGCTGCTGGAAGTGCAGCTGCCGTTGCTG
TGGTCCTTCCCTGATGATCCGGTCTTGCGCCAAAACTTGGAGCAGTTTCTGAACCAAT
TGGAGGGCCAGGCGCGCCAGGAGAAAGTAGCATACATCGAGTCTGTGTTGGCGGTAG
CCAGAACAAAACCGGGCACAGTCTATCCCAAGCCAATGTCCACGGTGATGACGAATGT
GTACTTCGAGTTCTGGAAGCATGTAATTGCAAACGCATCACCTGAGTTCGCCGTAAGA
TTAGCTCGTGCAAACCAGCAGTGGTTCCTGGGTATGCTCCAAGAGACGGAGAGCCGC
GAAAATGGAGATCAATGCATGTCCTCCATTGACGAGTACATCAAGCTACGCAGGAGGA
CATCAGCTCTTCCCAGCGTCATTGGTATTAATGATTTCGTCTACGGTCTTAAGACGTCA
CCCGACAGTTGGTGTTACTCTCGTGAGTATAAAAACGTTGTGGAAGCAATCAACGATG
TCACTTCTTGGCAGAATGATGTTTGGTCCTTCAAAAAGGAAGTGTTGGTAGCAAAGGAT
CCCTACAACATGGTTCTGCATGTCAGCGTCCACCGCAAAGTGTCATACACGGAAGCCG

CAATGATCACAAATCAGATGATCCAAGACCGGATTTTGGATCTCGAGAAAGCCGCAAA
GGAATTGGAATCGATTACACCTCCAGAGTGCCAACGAAATTTTGAAGTACTCCTTTTGA
CCGGCCGAAACATTGTATCCGGCGGGGAATATTTTTACTCAAAATCTGTGCGCTATTTG
TAG

Supplimental data 2. LbMTPSL1,3,4,5 protein sequences in FASTA format

>LbMTPSL1_-_TRINITY_DN210373_c0_g1_i3_(-2) A liverwort

MEVPETKEESCDKVQSYKVPEFISPYPARQNAVAPRVGPRCQTWFDEQSLSTVFSNPSD
WQFILECRVYDLPTWIFVDAEEEELVWICTFFLWLFVLDDMLEEPEYFFSLEKSASIFVEL
NLLIMWTFPDDPSIREVFTKLLMTQQPEQRSDTIKYVDAKLVEARLQPGTVYDIAKIGPV
GIALKDLWVRLISATPTKSAIRWAVTLLRYILGNAEETRNRNKKTFPSSADYVALRRNIS
AVEPCYVIVDFMDKVSEALPSEIFETPAMIECQDATNDIVSWHNDMWSFKKEFRKGELHN
LVYITSQERGCSFSEAGDLVLEMIYRRLEDLVQVFADLEKMTPLEHQQGVAGYIKASKFW
ISGTHQFHRTSKRFT*

>LbMTPSL3_-_TRINITY_DN174979_c0_g1_i1_(-3)

MVRDMNSAGAGEVANAQFPEFPPALFAGRTNEQMMAEINSLKPPKFYPPYPSRRNRNAT
KANVESIAWLHEYKVGTVFEDPKRWEHFQRMKLSEIMTRVYPDADEERVVWLGSYGWWL
FLVDDLLDGRGAFRAPERSTPFFVEVNLTMLWSFPDSPVLYNTFVEIVECFPEDQRRGILE
DVSAKLVKAREHPGSVYDTSHSGPVLECFKDLWVKLLASTPHESIIRWGNSIQTYLLGNVI
EAKNRTHGNIPSIAEYIDIRRNTSSMYPCMWIVDYSENFSVPLPKEIYESAEMKNFQEAT
NDTVSWHNDIFSVKKELLEGEVHNLVTVVSHERKCSFQEAMWIAVGMLHDRLQDLDRAVL
DLEAITPPEHSQMVAGYVKTAHCWFSGSHDFTVVSNERYSWDMSASIHCLLCHNAPPKEA
KVISGLFG*

>LbMTPSL4_-_TRINITY_DN216420_c0_g1_i8_(-2)

MGALEGDEVLMPLSLTKEEKQQIELFKIPAFISPYPVMRNALIDNVESKCLLWFEKQSLH
TVFADPKDWNRLMACKLYVIPGYIFMDAEEEELLWGALYTVWLFALDDMLEENDYFQSLG
KSDSIFVELMLVIMWAFPDEAAVRDIFLQLVSILPEESREPTMKYADAKLAEARLRPGTG
YDAKKIGPIGVSFSDLWMTYVRSTPTDSAIRWGLSNQRYFLGNASETRNRNLQSIPSSAD
FVFLRRKLSAVEPCMVKVDYICKLSPSLPSEVYDTPQMEELLDATNDIVAWHNDIWSFKK
ELRKGELHNLVFITSRERGCSFSAAAEVVMDKVYSRLRDLAQSFVDLEKITPPEHHHASA
MYIKTAKLWVSGTHQFHSTNKRFD*

>LbMTPSL5_-_TRINITY_DN216972_c0_g2_i1_(-3)

MAAAEAIPAGTSAFSSSTDNEFVKTFRPPLLDSSYPLNIHPKFSSSEARDTIEKWMQLHK
VDGIFTPEGYKLLLDMDIPAFGGRIFTEAPEEGLEWGIKFLFMLWIWDDTMDSTELGLSP
ETALSPLLEVQLPLLWSFPDDPVLRQNLEQFLNQLEGQARQEKVAYIESVLAVARTKPGT
VYPKPMSTVMTNVYFEFWKHVIANASPEFAVRLARANQQWFLGMLQETESRENGDQCM
SSIDEYIKLRRRTSALPSVIGINDFVYGLKTSPDSWCYSREYKNVVEAINDVTSWQNDVWS
FKKEVLVAKDPYNMVLHVSVHRKVSYTEAAMITNQMIQDRILDLEKAAKELESITPPECQR
NFEVLLLTGRNIVSGGEYFYSKSVRYL*

Table S1. Equimolar volumes of DNA fragments for LbMTPSL1 gene construct

Fragment	Size(kb)	concentration (ng/ul)	pmol	Amount(ng)	Volume(µl)
1	2.688	977	4.75	8426.88	8.625261
2	0.378	190	4.75	1185.03	6.237
LbMTPSL1	1.173	844	4.75	3677.355	4.357056
4	2.093	814	4.75	6561.555	8.060878
			TOTAL	19850.82	27.2802

Table S2. Equimolar volumes of DNA fragments for LbMTPSL3 gene construct

Fragment	Size(kb)	concentration (ng/ul)	pmol	Amount(ng)	Volume(µl)
1	2.688	977	4.75	8426.88	8.625261
2	0.378	190	4.75	1185.03	6.237
LbMTPSL3	1.333	750	4.75	4178.955	5.57194
4	2.093	814	4.75	6561.555	8.060878
			TOTAL	20352.42	28.49508

Table S3. Equimolar volumes of DNA fragments for LbMTPSL4 gene construct

Fragment	Size(kb)	concentration (ng/ul)	pmol	Amount(ng)	Volume(µl)
1	2.688	977	4.75	8426.88	8.625261
2	0.378	190	4.75	1185.03	6.237
LbMTPSL 4	1.201	1294	4.75	3765.135	2.909687
4	2.093	814	4.75	6561.555	8.060878
			TOTAL	19938.6	25.83283

Table S4. Equimolar volumes of DNA fragments for LbMTPSL5 gene construct

Fragment	Size(kb)	concentration (ng/ul)	pmol	Amount(ng)	Volume(µl)
1	2.688	977	4.75	8426.88	8.625261
2	0.378	190	4.75	1185.03	6.237
LbMTPSL 5	1.209	792.6	4.75	3790.215	4.782002
4	2.093	814	4.75	6561.555	8.060878
			TOTAL	19963.68	27.70514

Table S5. List of MTPSL proteins used in phylogenetic analysis

Protein	Protein ID or GenBank accession numbers
AaMTPSL1	MF417641
AaMTPSL3	MF417642
AaMTPSL4	MF417643
AaMTPSL5	MF417644
AaMTPSL6	MF417645
ApMTPSL1	MF417637
ApMTPSL2	MF417647
ApMTPSL3	MF417638
ApMTPSL4	MF417639
ApMTPSL5	MF417640
MpMTPSL1	APP91786
MpMTPSL2	APP91787
MpMTPSL3	APP91788
MpMTPSL4	APP91789
MpMTPSL5	APP91790
SmMTPSL1	J9R1J8
SmMTPSL13	J9R393
SmMTPSL17	D8RLD3
SmMTPSL26	J9QS25
SmMTPSL30	D8S255

Mon-UJTT-MTPSL4	KX230842.1
Mon-GSXD-MTPSL3	KX230841.1
Mon-YJJY-MTPSL1	KX230843.1
Epi-alfa-bisabolol synthase	BAL14867.1
Germacradien-4-ol synthase	BAL14866.1
Putative geosmin synthase	AEAO3338.1
PenA	ADO85594.1
Germacradienol synthase	ABY50951.1
Mos-GOWD-MTPSL2	KX230837.1
Mos-QKQO-MTPSL3	KX230838.1
Mos-VBMM-MTPSL3	KX230839.1
Aristolochene synthase	2OA6_A
Trichodiene synthase	2AEK_A

Figure S1. Full length sequence alignment of α-only domain type microbial terpene synthases highlighting Substrate-Mg^{+2} binding motifs. Obtained from conserved domain database (CDD)

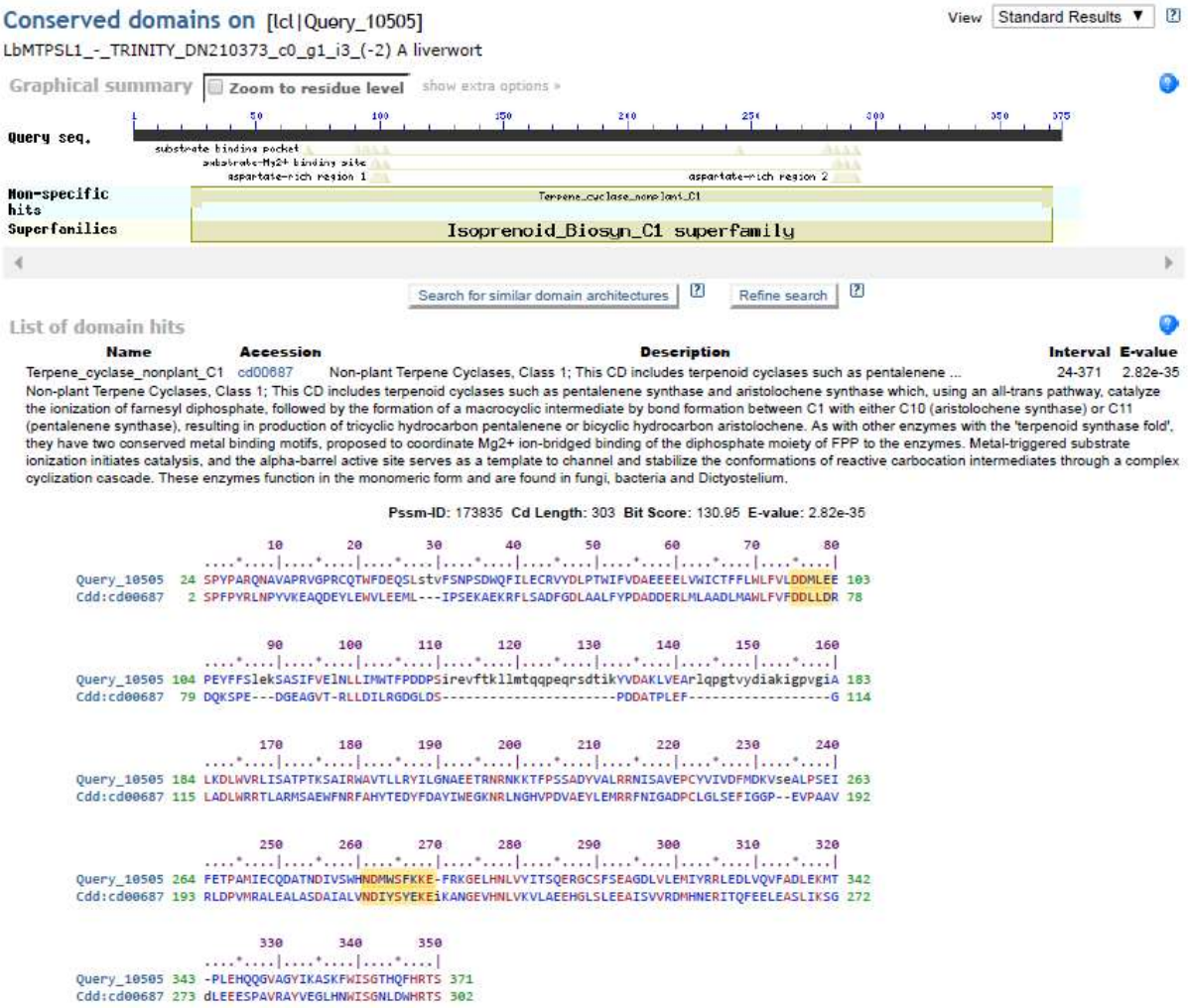

Figure S2. BLASTp results of LbMTPSL1 showing domain hit to Terpene_cyclase_nonplant_C1 (cd00687) conserved domain.