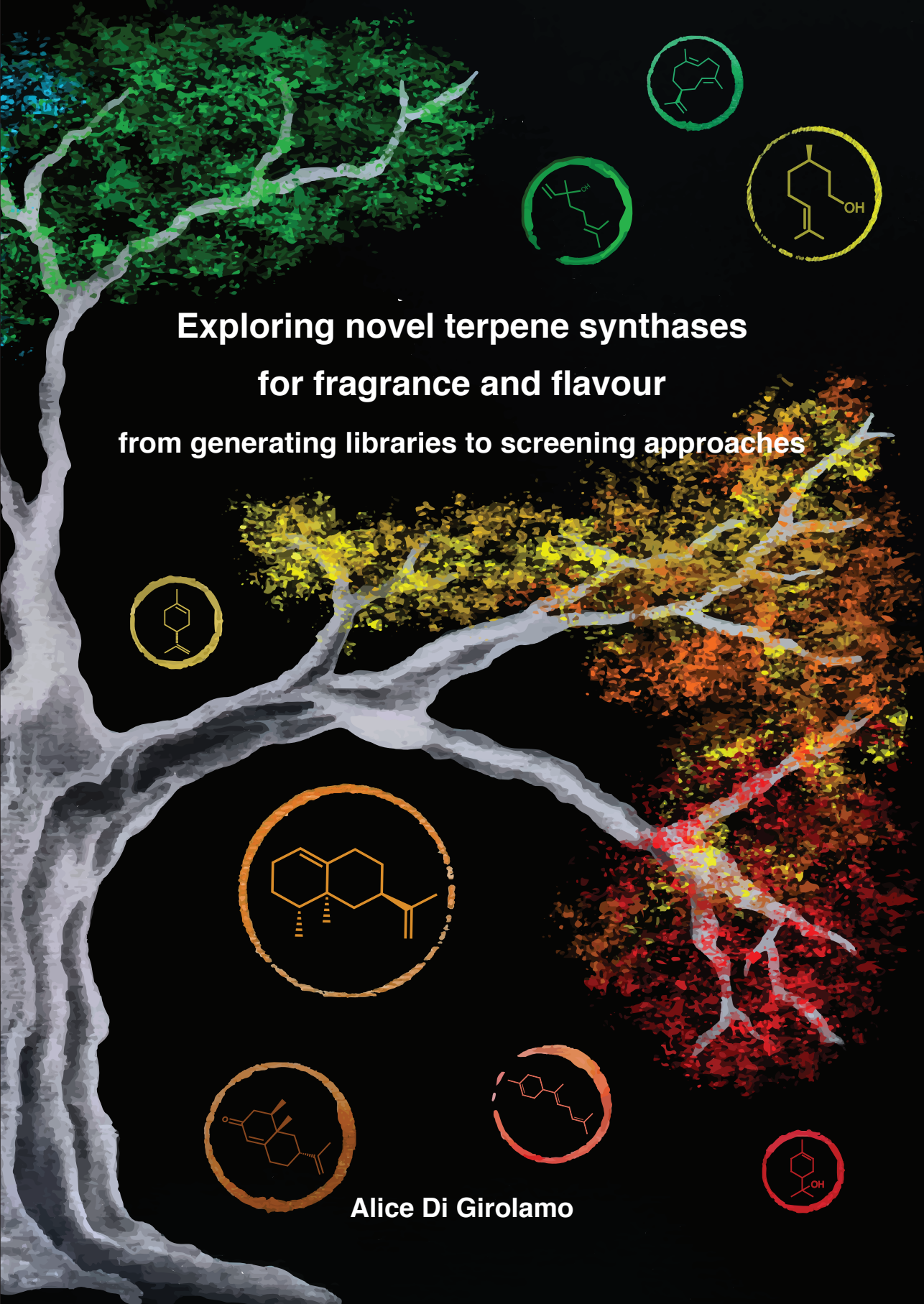



# Exploring novel terpene synthases for fragrance and flavour

## from generating libraries to screening approaches

Alice Di Girolamo

**Propositions**

1. High-throughput screening in biotechnology is only necessary until rational protein design becomes unquestionable.
   (this thesis)

2. The development of novel biotechnological solutions requires concerted approaches based on prediction tools and directed protein engineering.
   (this thesis)

3. Consensus in taxonomical nomenclature is essential to maintain consistency in scientific literature.

4. Nowadays, knowing where to find reviewed scientific facts is more important than remembering notions.

5. Easily accessible databases are the most valuable resources for the advancement of scientific knowledge.

6. Philosophical and ontological debates are fundamental scientific practices.

7. Smashing the patriarchy is as crucial for men as it is for women.

Propositions belonging to the thesis entitled

Exploring novel terpene synthases for fragrance and flavour - from generating libraries to screening approaches.

Alice Di Girolamo

Wageningen, 21 October 2022

# Exploring novel terpene synthases for fragrance and flavour

from generating libraries to screening approaches

**Alice Di Girolamo**

**Thesis committee**

**Promotors**
Prof. Dr D. de Ridder
Personal chair at Bioinformatics
Wageningen University & Research

Prof. Dr H.J. Bouwmeester
Personal chair at Plant Hormone Biology
Swammerdam Institute for Life Sciences, University of Amsterdam

**Co-promotor**
Dr J. Beekwilder
CTO at Isobionics B.V.
Geleen

**Other members**
Prof. Dr R. van Kranenburg, Wageningen University & Research
Prof. Dr M. de Mey, Ghent University, Belgium
Dr P.M. Bleeker, University of Amsterdam
Dr R. Mumm, Wageningen University & Research

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences.

# Exploring novel terpene synthases for fragrance and flavour

from generating libraries to screening approaches

**Alice Di Girolamo**

**Thesis**

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 21 October 2022
at 4 p.m. in the Omnia Auditorium.

Alice Di Girolamo
Exploring novel terpene synthases for fragrance and flavour: from generating libraries to screening approaches
185 pages

PhD thesis, Wageningen University, Wageningen, The Netherlands (2022)
With references, with summary in English

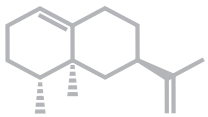
DOI: https://doi.org/10.18174/573954
ISBN: 978-94-6447-312-4

"I was made for another planet altogether, I mistook the way."

- Simone de Beauvoir

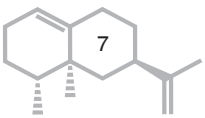
To my mother

# Table of contents



# List of Abbreviations

| | |
|---|---|
| **CiCaMS** | *Cinnamomum camphora* Monoterpene Synthase |
| **CiCaSSy** | *Cinnamomum camphora* Santalene Synthase |
| **CnVS** | *Callitropsis nootkatensis* Valencene Synthase |
| **CoQ10** | Coenzyme Q10 |
| **CPS** | Counts per seconds |
| **DBD** | DNA-Binding Domain |
| **DEG** | Differentially expressed gene |
| **DMADP** | Dimethylallyl Diphosphate |
| **FDP** | Farnesyl Diphosphate |
| **FHP** | Farnesyl Hydroxyphosphonate |
| **FDR** | False Discovery Rate |
| **GC-MS** | Gas Chromatography Mass Spectrometry |
| **GDP** | Geranyl Diphosphate |
| **HMM** | Hidden Markov Model |
| **HTH** | Helix-turn-helix |
| **HTS** | High Throughput Screening |
| **IDP** | Isopentenyl diphosphate |
| **KI** | Kovats retention index |
| **KO** | Knock out |
| **LBD** | Ligand-Binding Domain |
| **MDS** | Multi-dimentional Scaling |
| **MEP** | 2-C-methyl-D-erythriol-4-phosphate |
| **MTS** | Monoterpene Synthase |
| **MVA** | Mevalonate |
| **NCPS** | Normalized counts per seconds |
| **PCA** | Principal Component Analysis |
| **PTR-MS** | Proton Transfer Mass Spectrometry |
| **PTR-Qi-ToF-MS** | Proton Transfer Quadrupole interface Time of Flight Mass Spectrometry |
| **RNA-seq** | RNA sequencing |
| **STS** | Sesquiterpene Synthase |
| **TEAS** | Tobacco 5-epi-aristolochene synthase |
| **TPM** | Transcripts per million |
| **TR** | Transcriptional Regulator |
| **TRBS** | Transcriptional Regulator Binding Site |
| **TS** | Terpene Synthase |
| **VOC** | Volatile Organic Compound |



# 1

# General Introduction

Alice Di Girolamo[1]

[1] Laboratory of Plant Physiology, Department of Plant Sciences, Wageningen University, Netherlands

Terpenoids are one of the largest known classes of natural compounds. With over 80,000 different members[1], these specialized metabolites have been part of human culture for centuries. Traditionally, essential oils were extracted from plants for a wide variety of applications, from cosmetics to medicine[2–4]. In recent years, the increasing demand for these compounds has triggered a race for the development of novel production methods. The flavour and fragrance industry has been at the forefront of this race over the past decade, seeking to reduce overexploitation of endangered plant species and competition for resources with the food industry[5,6]. To address this issue, biotech industry has worked to implement the production of terpenoids using microbial platforms to ensure a regular supply[7–9].

There have been two major challenges in the development of microorganisms as production platforms: first, optimising substrate supply[5,10,11] and second, identifying and implementing the enzymes responsible for the production of terpenes, the terpene synthases (TSs)[12,13]. Currently, the main bottleneck in terpenoid production is the efficiency and specificity of the synthases. In order to find new synthases and improve those already in use, it is crucial to gain new insights into the biochemistry of these enzymes.

## 1.1 Ecological roles of terpenoids and terpenome diversity in plants

Throughout evolution, plants have developed a wide variety of metabolites to carry out their biological functions. With an estimated total production ranging from 100,000 to 1 million unique compounds, plants are able to synthesize the highest diversity of metabolites of all organisms[14]. Unravelling the regulation and production of these metabolites led to the birth of metabolomics, a discipline focused on the study of metabolomes[15]. Terpenoids represent a subset of plant specialized metabolites, and their specific product repertoire is defined as the terpenome[16]. Given the large number of terpenoids, one of the many challenges that remain to be investigated is the assessment of the ecological role of these compounds.

Over the past years, great effort has been made to investigate the interactions between plants and their environment and it has become clear that terpenoids play a fundamental role as signalling molecules[17–20]. However, in addition to their ecological role, several promising compounds have also been studied for their commercial applications. The exploration of these fascinating compounds is far from complete,

as a large part of terpene functions and their related synthases have not yet been discovered. A few examples will give an overview of the complexity and the diversity of the functions of terpenoids.

Terpenoids have been identified not only in plants, but also in fungi[21,22], bacteria[23,24] and marine organisms[25]. Sesquiterpenes, an important subclass of terpenoids (see Section 1.3), are known for their diverse roles in ecosystems. They have been described to have repellent activity against insects[26], as well as antimicrobial and antifungal activity[27,28]. These characteristics make sesquiterpenes a potential alternative to harmful pesticides in agriculture. In many cases, the same compound is responsible for a variety of responses in different organisms, like in the case of polygodial[29]. This drimane sesquiterpene has been described to be not only an antimicrobial and insect repellent compound[28,30], but also an elicitor of hot taste and an anti-hyperalgesic in humans[31].



Sesquiterpenes produced in flowers and fruits have been described to attract pollinators and herbivores[32,33], as well as predators[34]. A well-studied example is valencene, together with its oxidized form, nootkatone, which can be found in oranges and grapefruits and is responsible for the typical orange smell[35,36]. Nootkatone also shows an antibacterial effect against gram-positive bacteria[37]. Valencene and nootkatone are among the most appreciated volatile compounds present in essential oils used in perfumery and food flavouring. The study of the ecological role of these compounds and the regulation of their production has led to the discovery and characterization of the enzymes responsible for their synthesis[38,39].

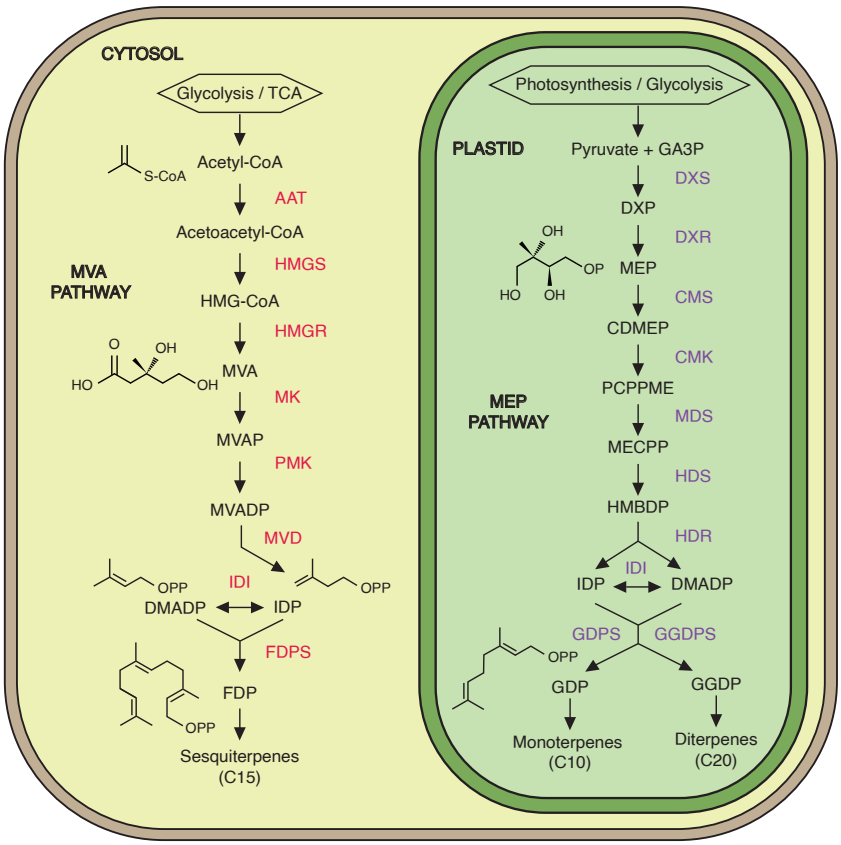
## 1.2 It all starts with the substrate: MEP and Mevalonate pathway

The metabolic pathways underlying the production of isoprene units have been studied in depth. Isoprene is a hemiterpene with formula $C_5H_8$ and is considered the building block for several other organic molecules[40,41]. Isoprene can be produced via two independent pathways that have been widely described: the methylerythritol phosphate (MEP) pathway and the mevalonate (MVA) pathway[42] (**Figure 1.1**). The MEP pathway is active in bacteria, green algae and plants, while the mevalonate pathway is present in archaea, some bacteria and most eukaryotes[43,44]. Both pathways produce isopentenyl diphosphate (IDP) and dimethylallyl diphosphate (DMADP), through different intermediates steps. While the MVA pathway converts acetyl-CoA to IDP in six steps, the MEP pathways includes seven reaction steps for the conversion of glyceraldehyde-3-phosphate (GA3P) and pyruvate to IDP and DMADP. In both pathways, an isopentenyl diphosphate isomerase (IDI) is responsible for the isomerization of IDP into DMADP[5,45].

Next, prenyltransferases condensate isoprene units into allylic phosphate substrates, resulting in the production of geranyl diphosphate (GDP), farnesyl diphosphate (FDP) and a range of other isoprenoids precursors, with addition of further isoprene units[46].

In plants, the MEP and MVA pathways are present in different cellular compartments. The production of monoterpenes and diterpenes generally takes place in plastids, while sesquiterpenes are generally synthesised in the cytosol[47], although different studies have shown a number of exceptions to this general rule. For instance, in some species, the production of both monoterpene and sesquiterpenes seems to be sustained by the MEP pathway, and multi-substrate enzymes have been characterized[48–51].

Different strategies have been employed to improve the production of isoprenoid precursors in microbial platforms. In some cases, highly efficient production was reached by the implementation of both pathways within the same microbial platform[5,52], while in other cases, the native MEP pathway was replaced with an exogenous MVA pathway, obtaining a better performing production system[53].




**MVA PATHWAY**
**Metabolites**:
HMG-CoA: Hydroxymethylglutaryl-CoA
MVA: (R)-mevalonate
MVAP: Mevalonate-5-phosphate
MVADP: Mevalonate diphosphate

**Enzymes**:
AAT: Acetyl-CoA acetyltransferase
HMGR: HMG-CoA reductase
HMGS: HMG-CoA synthase
MK: MVA kinase
PMK: MVAP kinase
MVD: MVADP kinase

**MEV PATHWAY**
**Metabolites**:
GA3P: D-glyceraldehyde 3-phosphate
DXP: 1-deoxyxylulose-5-phosphate
MEP: 2-C-methyl-D-erythriol-4-phosphate
CDMEP: 4-(cytidine 5'-diphospho)-2-C-methyl-D-erytriol
PCPPME: 2-phospho-4-(cytidine-5'-diphospho)-2-C-methyl-D-erytriol
MECPP: 2-C-methyl-D-erythriol 2,4-cyclodiphosphate
HMBPP: 1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate

**Enzymes**:
DXS: DXP synthase
DXR: DXP reductoisomerase
CMS: CDMEP synthase
CMK: CDMEP kinase
MDS: MECPP synthase
HDS: HMBPP synthase
HDR: HMBPP reductase

**SYNTHESIS OF ISOPRENOIDS**
**Metabolites**:
IDP: Isopentenyl diphosphate
DMADP: Dimethylallyl diphosphate
FDP: Farnesyl diphosphate
GDP: Geranyl diphosphate
GGDP: Geranylgeranyl diphosphate

**Enzymes**:
IDI: IDP isomerase
FDPS: FDP synthase
GDPS: GDP synthase
GGDPS: GGDP synthase

**Figure 1.1** Schematic representation of the MEP and MVA pathways in a plant cell. Precursors of FDP are generally produced in the cytosol through the MVA pathway, while precursors of GDP and GGDP are produced in the plastids via the MEP pathway. Relevant chemical structures are included in the figure.

## 1.3 Terpene synthases: structure-function relationships

While the MEP/MVA substrate-supplying pathways have been optimised and can be widely implemented in industrial platforms, the efficiency of TSs represents the main bottleneck to overcome. As for each new product a new synthase must be optimized, it is crucial to develop high throughput screening (HTS) methods that allow an efficient evaluation of potential novel candidates.



TSs mediate the first dedicated step in terpenoid biosynthesis. Their catalytic pocket functions as a "template" that binds the flexible isoprenoid substrate and induces the generation of a reactive carbocation. The interactions between the substrate and the residues present in the catalytic pocket are directly responsible for the formation of new carbon-carbon bonds that determine the final structure of the product, and therefore its properties[54].

TSs use as substrate different isoprenoids, that result from the condensation of two or more isoprene units, IDP and DMADP. These substrates are classified based on the number of C atoms and, consequently, TSs are classified based on the C number of the substrate they use[47,55]. These isoprenoid substrates are GDP, FDP, GGDP, as illustrated in **Figure 1.1**. In turn, the synthases are classified as monoterpene synthases, using GDP (MTSs, C10); sesquiterpene synthases, using FDP (STSs, C15); and diterpene synthases, using GGDP (DTSs, C20). TSs can be described as "high fidelity" when able to generate one exclusive product; however, most synthases have a more "promiscuous" yield and generate multiple products. It has been hypothesized that this promiscuity is a consequence of an imprecise template that allows alternative carbon-carbon interactions. Structural studies on the 3D structures of TSs provide useful insights on fidelity and promiscuity[16,56].

TSs have a common fold, consisting of interacting alpha helices (**Figure 1.2**). These structures have been described as $\alpha$, $\beta$ and $\gamma$ domains. Typically, in class I TSs, the active site is located within the $\alpha$-domain, while the functions of the $\beta$ and $\gamma$ domains are mainly structural. In the $\alpha$-domain the aspartate-rich motif DDxx(D,E) and NSE/DTE metal-binding motifs are found, responsible for the coordination of divalent $Mg^{2+}$ or $Mn^{2+}$ ions and formation of the carbocation[57,58]. The role of the divalent ions is to ionize the substrate to the carbocation, removing the inorganic diphosphate. The allylic cation undergoes a cascade of rearrangements which generates intermediate cations. Through a

specific series of methyl shifts, hydride shifts and alkyl shifts, the final product is cyclized. The amino acids present in the active site are crucial agents in these steps, as they direct the cascade of rearrangements. In class I synthases, to which both MTSs and STSs belong[47,] the active site pocket is composed of four $\alpha$-helices. The DDxx(D/E) motif is localized on helix D, while the NSE/DTE motif is localized on helix H[54]. It has been shown that mutations of these residues severely impair the functionality of the enzymes, especially when the mutations occur in the first aspartate of the DDxx(D/E) motif[59]Another consensus motif has been identified in the TS-a and TS-b subfamilies[60] of class I. The RxR motif is localized 35 amino acids upstream of the DDxx(D/E) motif. In STSs it interacts with the diphosphate moiety, preventing the nucleophilic attack on the forming carbocation[61]. A variation of this motif, RxQ, has been observed in several nerolidol synthases and might have a role in the stereo-selection of the initial carbocation[62]. These consensus motifs are important for the functionality of the synthases but do not carry sufficient information to predict the product that will be formed[47,54].



Diterpene synthases can be found in both class I and class II TSs. The mechanisms of product formation are different in the two classes and are beyond the scope of this thesis to describe in detail. Briefly, for class II TSs, the active site is at the interface between $\beta$ and $\gamma$ domains. The catalytically active amino acid is part of the DxDD motif, which is aspartate rich but not related to the motif of class I. It has been shown that class II enzymes do not always require the presence of bivalent metal ions for their activity[63,64]. DTSs can be found in both class I and class II. As mentioned, the residues present in the active site are responsible for the several cyclization steps that lead to the product formation. However, recent studies showed that there might be an interaction between the N-term and the C-term of the enzyme that determines the 3D structure of the catalytic site and its activity[65,66]. Knowledge on the role of specific residues in the cyclization cascade could facilitate the design of mutants when aiming to optimize synthase efficiency or product specificity.

This is made particularly challenging due to the lack of identity between synthases that generate the same compound(s) but belong to different species. It has been shown that sequence similarity is higher between synthases from the same species producing different terpenoids than between synthases producing the same terpenoid in different species[62]. Although the amino acid sequence thus cannot be used as a predictor for product formation, a more promising approach is represented by the study of the tertiary structure of the enzymes[67]. Understanding the

relationship between the structure of synthases and the precursor cation formed in the cascade seems a very promising way to predict the final product(s) of uncharacterized synthases[65].



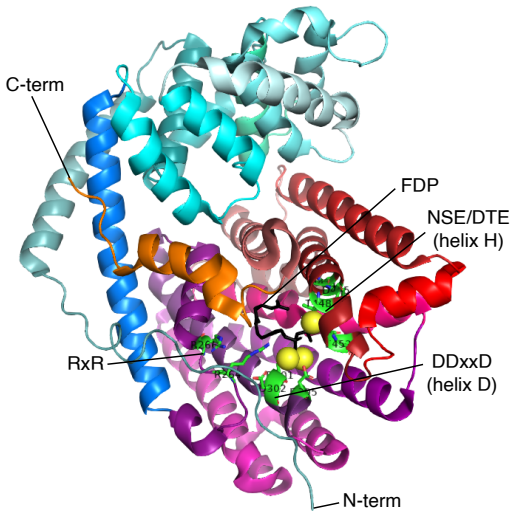



**Figure 1.2** Crystal structure of Tobacco 5-epi-aristolochene synthase (TEAS, PDB: 5IK0) in complex with FDP (depicted in black). The α-helices in the N-term domain (β domain) are indicated in light blue, while orange, red and magenta denote the α-helices in the C-term domain (α domain). The three yellow spheres are the $Mg^{2+}$ present in the active site, while the residues in green belong to the conserved motifs RxR, DDxxD and NSE/DTE.

## 1.4 Terpene synthases in industry

The interest in the development of highly efficient TSs is tightly connected to their application in industrial production. Given the limited availability of natural resources, biotech industry has worked to implement the production of terpenoids using microbial platforms, to guarantee regular supply and prevent the overexploitation of natural resources[6]. A well-known example is the sandalwood oil from *Santalum album* and *Santalum spicatum*[68,69]. The main components of this oil are the cyclic sesquiterpene alcohols α- and β-santalol and trans-α-bergamotol[70], responsible for the woody-floral scent. Valencene and its oxidized form nootkatone represent two other sesquiterpenes of particular interest for the flavour and fragrance industry. These compounds are found in citrus fruits and are responsible for the characteristic orange fragrance[13,71]. The identification of the TSs responsible for the production of these compounds in plants unlocked the possibility for the biotechnological production of santalenes[66,72], valencene[73,74] and many other compounds, offering a concrete alternative to the exploitation of the natural resources. The characterization of the valencene synthase from *Callitropsis nootkatensis*[13] and *Vitis vinifera*[75] has been pivotal for the biotechnological production of valencene and nootkatone in a number

of microbial hosts[71,73,74]. However, large differences are observed in the activity of native enzymes and in their product profiles, and optimisation is often required before implementation in production systems.

Terpenoids have also been of interest for their medicinal applications, as exemplified by artemisinin, a sesquiterpene lactone derived from *Artemisia annua*[76]. The discovery of antimalarial properties of this compound led to a worldwide implementation of artemisinin in the treatment of the *Plasmodium falciparum* infections[77]. The metabolic pathway for the biosynthesis of amorphadiene, the artemisinin precursor, has been implemented in microbial platform[78], allowing a large-scale production of semi-synthetic artemisinins[79], including novel artemisinin-derived compounds[80]. In another example, taxol is a diterpene well known for its microtubule-stabilizing activity in the treatment of different types of cancer[81]. The biosynthesis of taxol is complex, requiring 19 steps, and it has not been possible so far to implement in microbial platforms[82]. However, promising studies have shown that precursors of taxol can be successfully produced, opening a new perspective on industrial production of this important chemotherapy drug[83].



These few examples are only a small fraction of the terpenes currently produced biotechnologically. Still, there are many more compounds of interest that cannot yet be produced due to a lack of viable synthases. Native enzymes often present sub-optimal activity or synthesize a mixture of different compounds that might hinder their applicability in industry. For this reason, the discovery of novel synthases should go head-to-head with the improvement of the characterized ones and the development of reliable HTS methods.

## 1.5 Generation of terpene synthases mutant libraries

As mentioned above, engineering characterized synthases is a critical step towards the improvement of the synthase towards a competitive production of the compounds of interest. The generation of enzyme mutant libraries has been one of the best approaches to investigate and identify functional residues. In order to study the role of the residues in the active site, several strategies have been developed. Single and double mutations have been generated for the enzymes in analysis to demonstrate the function of individual residues. This approach provides essential insights but is time consuming and often restricted to small libraries, due the limits imposed by the screening methods. A more thorough approach is to generate larger libraries of mutants,

including more variation and combination of different mutations. This would, in turn, require a more efficient and reliable screening method to assess the effects of the mutations in the efficiency of the enzyme. The next subsections provide a small overview of the methods employed for the generation of TSs mutant libraries and their application.



### 1.5.1 Domain swapping

Domain swapping is a well-known phenomenon observed in nature and has been described as a strategy for protein oligomerization. The phenomenon was first observed for the diphtheria toxin[84]. 3D domain swapping is described as the exchange between two or more proteins of corresponding structural elements[85]. In other words, monomers of different proteins can combine with each other to generate oligomers with different functional properties[86,87]. This mechanism was widely exploited to engineer proteins, performing directed domain swaps[88] to investigate the roles of the different monomers. However, domain swaps have been used as engineering tool not only to exchange monomers, but also to investigate functional domains of different families of enzymes.

One of the first examples of domain swapping applied to the study of STSs is described in Back *et al.* (1996)[89]. In this work, different domains of the tobacco 5-epi-aristolochene synthase (TEAS)[90] were swapped with the corresponding domains of the *Hyoscyamus muticus* vetispiradiene synthase (HVS)[91]. A library of 14 chimeric genes was generated to identify the catalytic domain of each synthase and to characterize its role in the production of various sesquiterpenes. By using argentation-thin layer chromatography (TLC) and gas chromatography-mass spectrometry (GC-MS), they were able to discriminate between the products specific for each synthase and quantify the relative amounts produced. Different relative product amounts were observed for the different chimeric variants, allowing to identify the domains directly involved in product formation. Later studies allowed to assess the role of single residues of HVS by modelling their positions based on the TEAS crystal structure[92].

Domain swapping has also been employed for the study of catalytic domains of monoterpenes[93] and diterpenes[94]. While domain swapping is a useful tool for the study of the roles of the specific domain in different enzymes, there are limitations. In fact, most domain combination result in non-functional proteins due to misfolding. To generate functional chimeric proteins, it is often necessary to choose enzymes with high sequence identity to ensure the correct folding and the availability of

such synthases is generally limited.

### 1.5.2 Random and site-saturation mutagenesis

Random mutagenesis has been widely employed as a technique for enzyme development and to assess the plasticity of enzymes. This approach is based on the generation of mutant libraries through an iterative process, without a rational design. Mutations are achieved by "error-prone" PCR or by site-saturation mutagenesis and the selection of the improved enzyme requires reliable HTS methods[95].



For STSs, more targeted approaches are often preferred, in the absence of suitable HTS methods for large libraries. The residues chosen for mutagenesis are usually located in key positions in the catalytic site or in its surroundings. Considering the key role of the metal-binding residues of the DDxxD and NSE/DTE consensus motifs[54] and the fact that few substitutions are typically permissible without disrupting the activity of the synthase, site-saturation mutagenesis is often preferred over more traditional random mutagenesis. An interesting example is provided by the γ-humulene synthase from *Abies grandis*[96], known to produce 52 different sesquiterpenes. Being so versatile, it has been employed as a model to study plasticity by choosing the 19 residues of the catalytic pocket for site-saturation mutagenesis[97]. With this approach, several different less promiscuous enzymes were obtained. More recently, site-directed mutagenesis has been combined with computational approaches, aiming to improve the efficiency of the process and reduce the generation of non-functional enzymes[92,98].

### 1.5.3 Directed mutagenesis: generation of smart libraries

Directed mutagenesis is used when the role of a specific subset of residues must be investigated. To successfully identify the crucial residues to substitute and the best substitution to perform, different computational approaches have been developed and employed. An early attempt at generating and screening mutant libraries based on a computational approach was performed by Diaz *et al.* in 2011[99]. With their statistical, computationally assisted design strategy (SCADS) they aimed to identify potentially crucial sites for mutagenesis in TEAS. Twelve residues were selected as they were predicted to be good candidates to improve solubility and thermostability of TEAS. This approach yielded a synthase with increased thermo-stability, but lower activity compared to the wild type. Other studies provided insights on the roles of specific residues in the cyclization cascade[100,101] and in the

role of residues in product specificity[59,102,103].

Directed mutagenesis is probably the most widely used method for engineering of TSs. The residues selected for mutagenesis are often those surrounding the catalytic site. Mutations in the catalytic pocket are often not tolerated and can cause loss of function, while substitutions in the surrounding $\alpha$-helices are generally most promising. There may be potential for improvement in the N-term of the TSs, where more mutations are tolerated without disrupting the enzyme activity, which could improve the activity of the enzyme by indirectly changing the shape of the catalytic pocket[66].



## 1.6 Development of High Throughput Screening methods

As presented above, the generation of large libraries of mutants requires the development of efficient HTS methods to identify improved enzymes. In this section, examples of such methods will be presented, with attention on the challenges and complexities related to each method.

### 1.6.1 Gas Chromatography methods

Sesquiterpenes are hydrophobic volatile compounds, which makes it particularly challenging to accurately detect their concentration without recurring to gas chromatography (GC) analysis. Bacterial cultures are grown in rich media and an organic overlay such as dodecane is often used to trap and extract the sesquiterpene produced. The most widely utilized method for the detection and quantification of mono- and sesquiterpenes GC-MS. This technique provides data about relative concentrations of different compounds. Although the accuracy of GC-MS is widely recognized, it does not allow a direct measurement of the sample within the growth medium. The organic overlay containing the terpenoids must be extracted from the growth medium by centrifugation, diluted and then prepared for analysis. These steps are not only time consuming but also render measurement of different time points from the same sample impossible. Furthermore, depending on the temperature program, each sample requires between 5 and 40 minutes to be measured. These characteristics make traditional GC methods non-suitable for the investigation of large libraries of mutants.

A recent study, however, implemented a GC-MS-based automated pipeline utilizing a liquid handling robotic platform for the HTS of MTS libraries[104]. In this study, Leferink *et al.* (2019) generated a library of (-)-$\alpha$-pinene synthase mutants, targeting 16 amino acids[104]. 96-deepwell

plates were used for biphasic cultures (growth medium-dodecane) and all steps were controlled by the robotic liquid handling platform. 1,000 colonies were initially screened for improved function. The GC- MS was equipped with a 96-well plate auto-sampler. By sequencing, 94 unique variants were identified among the 350 active enzymes detected in the pipeline. Of these 94 variants, 27 turned out to be false positives, as no monoterpene was detected in the shake-flask cultures. Overall, the data obtained using the pipeline were comparable with data obtained with traditional shake-flasks cultures, showing consistency in detected production. Although some limitations of the pipeline were evident, such as exact production titers and the masking of products due to solvent overlap, this method offers a rapid initial screen to enrich libraries of mutants of new valuable variants in a single measurement, with a total measured sample run time of 4.6 minutes. This method has the potential to be applied not only to the production of monoterpenes but to all the compounds produced in biphasic cultures.



### 1.6.2 Chemical-based methods

Chemical-based methods rely on the use of different synthetic chemicals that allow to detect the enzyme activity. These compounds can be used in chemical assays testing the presence of antioxidant compounds (like 2,2-diphenyl-1-picrylhydrazyl, DPPH) or more broadly specific for hydrophobic compounds (such as Nile Red). The activity of the enzymes can be directly measured, when the chemical is reacting with the compound produced, or indirectly, when the chemical reacts with other compounds present in the reaction due to enzyme activity (e.g., the phosphate released by the substrate). The methods will be here described as direct and indirect, based on this criterion.

Chemical-based screenings present the advantage of having broad specificity and can potentially be employed for the screening of a wide range of synthases without requiring genetic engineering. However, the sensitivity and the reactivity of the chemical agents in different conditions may limit on the applicability.

#### 1.6.2.1 Direct methods

An example of direct chemical-based screening is described in the work of Behrendorff *et al.* (2013)[105]. In this study, DPPH is employed for the screening of MTSs activity. DPPH is a stable radical with a strong absorbance at 517nm when dissolved in methanol. The absorbance decreases proportionately with the loss of the radical,

which corresponds to a change in colour from purple to yellow. The proposed method was established by using DPPH in different solution of monoterpenes in dodecane. The results showed that this method provides a semi-quantitative assay for monoterpene biosynthesis, where the sensitivity depends on the monoterpene of interest: myrcene and γ-terpinene were detectable at concentrations as low as 10μM, while limonene above 100μM. β-pinene was tested as well, providing no clear detection. To test the applicability of this method in a production system, two different limonene synthases were expressed in *S. cerevisiae* for limonene production, using three different growth media. The screening of the cultures was performed by harvesting the dodecane overlay after 120h growth. The results revealed a significant difference in production between two limonene synthases tested in *Saccharomyces cerevisiae*, confirmed by the GC-MS analysis performed on the same dodecane overlay. Although this method could be easily applied on overlays extracted from fermentations once optimized for the system of interest, it does not serve a true HTS function, as for each sample it is still necessary to perform a fermentation, followed by dodecane extraction. Therefore, there is only a marginal advantage over the use of GC-MS analysis. Moreover, DPPH shows different reactivity with different compounds and monoterpene concentrations do not directly correlate with the measured absorbance. Therefore, it does not seem possible to use DPPH with a broad range of compounds and it may provide low accuracy in the analysis of enzymes producing a mixture of different monoterpenes.

#### 1.6.2.2 Indirect methods

A malachite green (MG) assay for the detection of inorganic pyrophosphate was implemented for screening TS activity by Vardakou *et al.* (2014)[106]. The use of MG was previously established for the detection of phosphate[107], but not yet for the study of TSs. This method can be considered indirect, although the measurement of the inorganic pyrophosphate provides a direct correlation to the amount of terpene produced. The MG assay is an *in vitro* assay and can be performed using 96 well plates after prior purification of the enzymes to test. It can be used to determine the kinetic properties and the efficiency of a given enzyme and was observed to be highly reproducible. Vardakou *et al.* (2014) tested 88 mutants for higher activity, together with 8 wild type controls. Out of this library, 13 mutants were isolated and further characterized for production via GC-MS, confirming improved activity.

This MG assay was successfully adapted for the study of TSs activity, with a sensitivity in the linear range of FDP concentrations. However, the need for good quality purified enzymes is the biggest limiting factor of the method. In fact, it requires individual colonies to be grown for production of proteins, which then also need to be purified individually before any indication of the activity can be assessed. Therefore, while being an effective method, the throughput will remain limited.



A non-conventional chemical-based HTS method is presented in the work of Lauchli *et al.* (2013)[108]. A non-natural isoprenoid (vinyl methyl ether) was employed as a substrate for TSs. This substrate is used instead of FDP as an *in vitro* enzyme assay and allows the generation of methanol as by-product of terpene formation. This methanol was converted into formaldehyde by a treatment with alcohol oxidase and detected using Purpald[109]. The method was used to test 2,800 mutants for heat resistance and productivity on FDP was compared with the productivity on the synthetic substrate, confirming the correlation between methanol produced and sesquiterpene produced. In the first round the T50 (T50 being the incubation temperature at which 50% of productivity on the substrate remained after 10 min of heat treatment) was increased from 42°C to 47°C. The best performing enzyme was chosen, and a second round of mutagenesis was performed, obtaining 1,800 mutants that were screened. A variant having a T50 of 54°C while maintaining the same product specificity, expression, and activity of the parental enzyme was then identified. Thus, a heat resistant variant was selected in only 2 rounds of mutagenesis without prior structural knowledge of the parental enzyme. The major strength point of this method is its wide applicability, as the synthetic substrate can be used by all STSs, and the screening does not require advanced equipment. However, as for other chemical-based screenings, the requirement of performing a cell lysate prior to the assay represents a constraint in the throughput of the method, limiting application to medium-sized libraries.

Among the bioanalytical tools available, a recent study describes the implementation of an oil-free picodrop bioassay for the detection of hydrophobic compounds. Siltanen *et al.* (2018) show how to apply the concept of Printed Droplet Microfluidics (PDM) in the development of an array supporting picolitre droplets[110]. With this method, it is possible to print picolitre droplets containing single yeast cells, in arrays containing a maximum of 20 droplets. A microfluidic droplet sorter dispenses the droplets on a customizable microwell array, where they are trapped

into nanolitre wells. It was shown that it is possible to detect differences in farnesene production between two strains of yeast using a Nile Red solution to stain the lipid droplets containing farnesene. Although this method shows a good potential for high throughput, the use of Nile Red as a staining solution does not offer specificity for terpenoids, as it stains lipids unselectively. The fluorescence measured must be normalized for biomass, which introduces errors as the measurement cannot be considered a direct representation of the compound produced, but rather a quantification of the lipid droplets in the cells. The sensitivity of this method might therefore be limited, as relatively low improvements in enzyme activity may not be detectable. Applicability might also depend on the organism chosen as production platform. While yeast produces lipid droplets containing farnesene, bacteria might exploit a different strategy, making the Nile Red staining non suitable for the analysis.

### 1.6.3 Genetic-based methods

Genetic-based methods are based on the coupling of different regulatory elements for the development of a system sensitive to the production of the compound(s) of interest. The main advantage of these methods is to offer an *in vivo* platform as HTS method, reducing, in principle, the number of steps required for the first assessment of the improved enzymes.

Several approaches have been presented to develop quick and reliable screening methods for improved synthases. A method based on substrate competition was described by Furubayashi *et al.* (2014)[111]. The screening strain contained a carotenoid synthase using the same substrate as the terpene synthase of interest. The strain was co-transformed with both the carotenoid synthase and the terpene synthase of interest and, as the substrate supply remained constant, the best performing enzyme should prevail in production. As a result, efficient terpene synthases were supposed to generate white-looking bacterial colonies, while poor terpene synthases would "lose" the substrate competition and yield orange/yellow-looking bacterial colonies, as the carotenoid synthase prevailed. A follow-up study applied this method to select for an improved pinene synthase[112]. A library of 16,000 mutants was screened, resulting in the identification of a mutant displaying a 60% production improvement compared to the wild type. This method seems more suitable for monoterpenes rather than sesquiterpenes, although fine-tuning the system's kinetics in relation to the properties of the synthase of interest might be sufficient to widen its applicability.

The design of genetic biosensors is a widely exploited strategy to obtain reliable and specific screening methods. A mevalonate biosensor was developed by Pfleger *et al.* (2007) using an *E. coli* auxotroph of mevalonate harbouring a plasmid containing a constitutively expressed green fluorescent protein (GFP)[113]. This biosensor strain was able to grow only when mevalonate was supplemented, so its growth and thus the amount of GFP served as proxy for the concentration of mevalonate in the medium. Ultimately, the biosensor strain was grown with a mevalonate producing strain harbouring a library of enzyme mutants, allowing to select for the highest mevalonate production.

A different biosensor design based on the use of transcriptional regulators (TRs) is described by de Paepe *et al.* (2017)[114]. Their genetic biosensor was based on the combination of an "effector module", containing a reporter gene regulated by the TR DNA binding site; and a "detector module", containing the gene encoding the TR under a constitutive promoter. The strain harbouring this biosensor system was used to detect varying concentrations of naringenin in the medium. Knowing all the elements of a regulatory circuit is the most favourable starting point for its implementation in a biosensor system. Different approaches are required when only part of the circuit is known. In such cases, it might be necessary to resort to the use of chimeric TRs, especially to investigate the interaction between the TR and the ligand[115,116]. Other biosensors for flavonoids[117] and quercetin[118] have been described, and more TRs are known to be sensitive to flavonoids[119–121]. However, no TRs have so far been described to recognize sesquiterpenes as ligands. One TetR-like TR has been described in *Pseudomonas aeruginosa* to be a regulator of the acyclic monoterpene utilization gene operon[122]. To develop a genetic biosensor for terpenes it is therefore essential to first identify a TR able to bind the terpene, or group of terpenes, of interest.

One of the main advantages of biosensor methods is the possibility to screen large libraries without the need for performing fermentation experiments. However, there are challenges to consider, such as the reliable quantification of the reporter gene and the need for approaches tailored to the compound of interest.

As there is no perfect HTS method, it is important to take into account the advantages and downsides of each approach for every new challenge. In this work, we explored the possibilities for implementation of two HTS methods for sesquiterpene synthases, in an attempt to provide a novel approach for the identification of improved synthases.

## 1.7 Thesis overview

This thesis explores the generation of smart libraries for terpene synthases (TSs) and the development of novel approaches for high throughput screening (HTS) methods of TS mutants.



To collect the information available on STSs, in **Chapter 2** we create a database of characterized STSs and provide an analysis based on protein sequence. We observe that phylogenetic relations affect sequence similarities more than product specificity. An example of this is explored in **Chapter 3**, where a case study of two synthases from *Cinnamomum camphora* having high sequence similarity but different product specificity is presented. We aim to identify residues crucial for the sesquiterpene synthase activity by employing both domain swapping and site-directed mutagenesis and evaluating the effects of the mutations on product profile and product specificity. In **Chapter 4** we propose the use of proton transfer reaction mass spectrometry (PTR-MS) as a high throughput screening method for a library of valencene synthase mutants. We assess the applicability of PTR-MS in a small-scale fermentation system using 96 deep-well plates. Finally, in **Chapter 5** we aim to identify a candidate valencene-sensitive TR, through a comparative expression analysis of *Rhodobacter sphaeroides*. RNA-seq analysis was performed on a wild type *R. sphaeroides* supplemented with a valencene synthase against a wild type *R. sphaeroides* harbouring an empty vector.

**Chapter 6** concludes this thesis with an overview of the achievements of this work and an outlook on the future perspectives of HTS methods for enzyme development. Given the exponential growth of data produced, I emphasize the increasing need for a synergistic effort between bioinformatics and biotechnology to optimally support biotechnological production of valuable TSs.

## References


1. Sparkman, O. D. Identification of essential oil components by gas chromatography / mass spectroscopy Robert P. Adams. *J. Am. Soc. Mass Spectrom.* **8**, 671–672 (1997).
2. Foote, J. On the Use of the Essential Oil of Lemons in Various Inflammations of the Eye; with Cases. *London Med. Phys. J.* **15**, 23–33 (1833).
3. Cannon, A. The Essential Oil Treatment of Cholera. *Br. Med. J.* **1**, 98 (1927).
4. S, X., L, K. & F, N. Effects of Different Hydroponic Substrate Combinations and Watering Regimes on Physiological and Anti-Fungal Properties of Siphonochilus Aethiopicus. *African J. Tradit. Complement. Altern. Med. AJTCAM* **14**, 89–104 (2017).
5. Martin, V. J. J., Piteral, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.* **21**, 796–802 (2003).
6. Kirby, J. & Keasling, J. D. Biosynthesis of plant isoprenoids: Perspectives for microbial engineering. *Annu. Rev. Plant Biol.* **60**, 335–355 (2009).
7. Yamada, Y. *et al.* Novel terpenes generated by heterologous expression of bacterial terpene synthase genes in an engineered Streptomyces host. *J. Antibiot. (Tokyo).* **68**, 385–394 (2015).
8. Alves Gomes Albertti, L. *et al.* Identification of the bisabolol synthase in the endangered candeia tree (eremanthus erythropappus (dc) mcleisch). *Front. Plant Sci.* **9**, (2018).
9. Lange, B. M. *et al.* Improving peppermint essential oil yield and composition by metabolic engineering. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16944–16949 (2011).
10. Chatzivasileiou, A. O., Ward, V., Edgar, S. M. B. & Stephanopoulos, G. Two-step pathway for isoprenoid synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 506–511 (2019).
11. Kang, A. *et al.* Isopentenyl diphosphate (IPP)-bypass mevalonate pathways for isopentenol production. *Metab. Eng.* **34**, 25–35 (2016).
12. O'Maille, P. E., Chappell, J. & Noel, J. P. Biosynthetic potential of sesquiterpene synthases: Alternative products of tobacco 5-epi-aristolochene synthase. *Arch. Biochem. Biophys.* **448**, 73–82 (2006).
13. Beekwilder, J. *et al.* Valencene synthase from the heartwood of Nootka cypress (Callitropsis nootkatensis) for biotechnological production of valencene. *Plant Biotechnol. J.* **12**, 174–182 (2014).
14. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends in Plant Science* vol. 24 83–98 (2019).
15. Rai, A., Saito, K. & Yamazaki, M. Integrated omics analysis of specialized metabolism in medicinal plants. *Plant J.* **90**, 764–787 (2017).
16. Christianson, D. W. Unearthing the roots of the terpenome. *Current Opinion in Chemical Biology* vol. 12 141–150 (2008).
17. Tholl, D. Biosynthesis and biological functions of terpenoids in plants. *Adv. Biochem. Eng. Biotechnol.* **148**, 63–106 (2015).
18. Pichersky, E. & Raguso, R. A. Why do plants produce so many terpenoid compounds? *New Phytol.* **220**, 692–702 (2018).
19. Zhou, F. & Pichersky, E. More is better: the diversity of terpene metabolism in plants. *Current Opinion in Plant Biology* vol. 55 1–10 (2020).
20. Bhardwaj, T. & Somvanshi, P. Plant Systems Biology: Insights and Advancements. in *PlantOmics: The Omics of Plant Science* 791–819 (Springer India, 2015). doi:10.1007/978-81-322-2172-2_28.
21. Keller, N. P., Turner, G. & Bennett, J. W. Fungal secondary metabolism - From biochemistry to genomics. *Nature Reviews Microbiology* vol. 3 937–947 (2005).
22. Agger, S., Lopez-Gallego, F. & Schmidt-Dannert, C. Diversity of sesquiterpene synthases in the basidiomycete Coprinus cinereus. *Mol. Microbiol.* **72**, 1181–1195 (2009).
23. Yamada, Y. *et al.* Terpene synthases are widely distributed in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 857–862 (2015).

24. Rico, J. *et al.* Exploring natural biodiversity to expand access to microbial terpene synthesis. *Microb. Cell Fact.* **18**, (2019).
25. Gross, H. & König, G. M. Terpenoids from marine organisms: Unique structures and their pharmacological potential. *Phytochemistry Reviews* vol. 5 115–141 (2006).
26. Jansen, B. J. M. & De Groot, A. Occurrence, biological activity and synthesis of drimane sesquiterpenoids. *Natural Product Reports* vol. 21 449–477 (2004).
27. Lunde, C. S. & Kubo, I. Effect of polygodial on the mitochondrial ATPase of Saccharomyces cerevisiae. *Antimicrob. Agents Chemother.* **44**, 1943–1953 (2000).
28. Fujita, K. I. & Kubo, I. Multifunctional action of antifungal polygodial against Saccharomyces cerevisiae: Involvement of pyrrole formation on cell surface in antifungal action. *Bioorganic Med. Chem.* **13**, 6742–6747 (2005).
29. Gershenzon, J. & Dudareva, N. The function of terpene natural products in the natural world. *Nature Chemical Biology* vol. 3 408–414 (2007).
30. Caprioli, V. *et al.* Insect antifeedant activity and hot taste for humans of selected natural and synthetic 1,4-Dialdehydes. *J. Nat. Prod.* **50**, 146–151 (1987).
31. Mendes, G. L. *et al.* Anti-hyperalgesic properties of the extract and of the main sesquiterpene polygodial isolated from the barks of Drymis winteri (Winteraceae). *Life Sci.* **63**, 369–381 (1998).
32. Dweck, H. K. M. *et al.* Olfactory preference for egg laying on citrus substrates in Drosophila. *Curr. Biol.* **23**, 2472–2480 (2013).
33. Liu, F. *et al.* A dual-target molecular mechanism of pyrethrum repellency against mosquitoes. *Nat. Commun.* **12**, (2021).
34. Pan, L. *et al.* CcOBP2 plays a crucial role in 3-carene olfactory response of the parasitoid wasp Chouioia cunea. *Insect Biochem. Mol. Biol.* **117**, (2020).
35. Elston, A., Lin, J. & Rouseff, R. Determination of the role of valencene in orange oil as a direct contributor to aroma quality. *Flavour Fragr. J.* **20**, 381–386 (2005).
36. Deterre, S. *et al.* Identification of key aroma compounds from bitter orange (Citrus aurantium L.) products: essential oil and macerate-distillate extract. *Flavour Fragr. J.* **27**, 77–88 (2012).
37. Yamaguchi, T. Antibacterial Properties of Nootkatone Against Gram-Positive Bacteria. *Nat. Prod. Commun.* **14**, (2019).
38. Abbas, F. *et al.* Volatile terpenoids: multiple functions, biosynthesis, modulation and manipulation by genetic engineering. *Planta* vol. 246 803–816 (2017).
39. Nagegowda, D. A. Plant volatile terpenoid metabolism: Biosynthetic genes, transcriptional regulation and subcellular compartmentation. *FEBS Letters* vol. 584 2965–2973 (2010).
40. Sharkey, T. D. Isoprene synthesis by plants and animals. *Endeavour* **20**, 74–78 (1996).
41. Sharkey, T. D., Wiberley, A. E. & Donohue, A. R. Isoprene Emission from Plants: Why and How. *Ann. Bot.* **101**, 5–18 (2007).
42. Vickers, C. E., Bongers, M., Liu, Q., Delatte, T. & Bouwmeester, H. Metabolic engineering of volatile isoprenoids in plants and microbes. *Plant. Cell Environ.* **37**, 1753–75 (2014).
43. Chandran, S. S., Kealey, J. T. & Reeves, C. D. Microbial production of isoprenoids. *Process Biochem.* **46**, 1703–1710 (2011).
44. Lange, B. M., Rujan, T., Martin, W. & Croteau, R. Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13172–13177 (2000).
45. Lichtenthaler, H. K. Non-mevalonate isoprenoid biosynthesis: enzymes, genes and inhibitors. *Biochem. Soc. Trans.* **28**, 785–789 (2000).
46. Liang, P. H., Ko, T. P. & Wang, A. H. J. Structure, mechanism and function of prenyltransferases. *Eur. J. Biochem.* **269**, 3339–3354 (2002).
47. Degenhardt, J., Köllner, T. G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70**, 1621–1637 (2009).
48. Dudareva, N. *et al.* The nonmevalonate pathway supports both monoterpene and sesquiterpene formation in snapdragon flowers. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 933–938 (2005).
49. Gutensohn, M. *et al.* Cytosolic monoterpene biosynthesis is supported by plastid-generated

geranyl diphosphate substrate in transgenic tomato fruits. *Plant J.* **75**, 351–363 (2013).
50. Johnson, S. R. *et al.* Promiscuous terpene synthases from Prunella vulgaris highlight the importance of substrate and compartment switching in terpene synthase evolution. *New Phytol.* **223**, 323–335 (2019).
51. Pazouki, L. & Niinemetst, U. Multi-substrate terpene synthases: Their occurrence and physiological significance. *Front. Plant Sci.* **7**, (2016).
52. Liu, H. *et al.* Combination of Entner-Doudoroff pathway with MEP increases isoprene production in engineered Escherichia coli. *PLoS One* **8**, e83290 (2013).
53. Orsi, E. *et al.* Functional replacement of isoprenoid pathways in Rhodobacter sphaeroides. *Microb. Biotechnol.* **13**, 1082–1093 (2020).
54. Christianson, D. W. Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).
55. Smit, A. & Mushegian, A. Biosynthesis of isoprenoids via mevalonate in archaea: The lost pathway. *Genome Res.* **10**, 1468–1484 (2000).
56. Kampranis, S. C. *et al.* Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: Structural insights into the evolution of terpene synthase function. *Plant Cell* **19**, 1994–2005 (2007).
57. Lesburg, C. A., Zhai, G., Cane, D. E. & Christianson, D. W. Crystal Structure of Pentalenene Synthase: Mechanistic Insights on Terpenoid Cyclization Reactions in Biology. *Science (80-. ).* **277**, 1820–1824 (1997).
58. Zhou, K. & Peters, R. J. Investigating the conservation pattern of a putative second terpene synthase divalent metal binding motif in plants. *Phytochemistry* **70**, 366–369 (2009).
59. Little, D. B. & Croteau, R. B. Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases δ-selinene synthase and γ-humulene synthase. *Arch. Biochem. Biophys.* **402**, 120–135 (2002).
60. Jiang, S.-Y., Jin, J., Sarojam, R. & Ramachandran, S. A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns. *Genome Biol. Evol.* **11**, 2078–2098 (2019).
61. Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi- aristolochene synthase. *Science (80-. ).* **277**, 1815–1820 (1997).
62. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).
63. Prisic, S., Xu, J., Coates, R. M. & Peters, R. J. Probing the role of the DXDD motif in class II diterpene cyclases. *ChemBioChem* **8**, 869–874 (2007).
64. Baunach, M., Franke, J. & Hertweck, C. Terpenoid biosynthesis off the beaten track: Unconventional cyclases and their impact on biomimetic synthesis. *Angewandte Chemie - International Edition* vol. 54 2604–2626 (2015).
65. Durairaj, J. *et al.* Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comput. Biol.* **17**, (2021).
66. Di Girolamo, A. *et al.* The santalene synthase from Cinnamomum camphora: Reconstruction of a sesquiterpene synthase from a monoterpene synthase. *Arch. Biochem. Biophys.* **695**, (2020).
67. Durairaj, J., Akdel, M., de Ridder, D. & van Dijk, A. D. J. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics* **36**, I718–I725 (2020).
68. Arunkumar, A. N., Dhyani, A. & Joshy, G. *Santalum album IUCN Red List Assessment. The IUCN Red List of Threatened Species* (2019).
69. Birkbeck, A. A. The synthesis of fragrant natural products from Santalum album L.: (+)-(Z)-α-Santalol and (-)-(Z)-β-Santalol. *Chimia (Aarau).* **71**, 823–835 (2017).
70. Diaz-Chavez, M. L. *et al.* Biosynthesis of Sandalwood Oil: Santalum album CYP76F Cytochromes P450 Produce Santalols and Bergamotol. *PLoS One* **8**, e75053 (2013).
71. Wriessnegger, T. *et al.* Production of the sesquiterpenoid (+)-nootkatone by metabolic engineering of Pichia pastoris. *Metab. Eng.* **24**, 18–29 (2014).
72. Zhang, X. *et al.* Identification and functional characterization of three new terpene synthase

genes involved in chemical defense and abiotic stresses in Santalum album. *BMC Plant Biol.* **19**, 115 (2019).
73. Frohwitter, J., Heider, S. A. E., Peters-Wendisch, P., Beekwilder, J. & Wendisch, V. F. Production of the sesquiterpene (+)-valencene by metabolically engineered Corynebacterium glutamicum. *J. Biotechnol.* **191**, 205–213 (2014).
74. Meng, X. *et al.* Metabolic engineering Saccharomyces cerevisiae for de novo production of the sesquiterpenoid (+)-nootkatone. *Microb. Cell Fact.* **19**, (2020).
75. Lücker, J., Bowen, P. & Bohlmann, J. Vitis vinifera terpenoid cyclases: Functional identification of two sesquiterpene synthase cDNAs encoding (+)-valencene synthase and ( - )-germacrene D synthase and expression of mono- and sesquiterpene synthases in grapevine flowers and berries. *Phytochemistry* **65**, 2649–2659 (2004).
76. Woodrow, C. J., Haynes, R. K. & Krishna, S. Artemisinins. *Postgraduate Medical Journal* vol. 81 71–78 (2005).
77. Liao, F. Discovery of artemisinin (Qinghaosu). *Molecules* vol. 14 5362–5366 (2009).
78. Zeng, Q., Qiu, F. & Yuan, L. Production of artemisinin by genetically-modified microbes. *Biotechnology Letters* vol. 30 581–592 (2008).
79. Kung, S. H., Lund, S., Murarka, A., McPhee, D. & Paddon, C. J. Approaches and recent developments for the commercial production of semi-synthetic artemisinin. *Frontiers in Plant Science* vol. 9 (2018).
80. Badshah, S. L., Ullah, A., Ahmad, N., Almarhoon, Z. M. & Mabkhot, Y. Increasing the strength and production of artemisinin and its derivatives. *Molecules* vol. 23 (2018).
81. Weaver, B. A. How Taxol/paclitaxel kills cancer cells. *Molecular Biology of the Cell* vol. 25 2677–2681 (2014).
82. Howat, S. *et al.* Paclitaxel: Biosynthesis, production and future prospects. *New Biotechnology* vol. 31 242–245 (2014).
83. Huang, J. J. *et al.* Microbial Cell Factory of Baccatin III Preparation in Escherichia coli by Increasing DBAT Thermostability and in vivo Acetyl-CoA Supply. *Front. Microbiol.* **12**, (2022).
84. Bennett, M. J., Choe, S. & Eisenberg, D. Domain swapping: Entangling alliances between proteins. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 3127–3131 (1994).
85. Liu, Y. & Eisenberg, D. 3D domain swapping: As domains continue to swap. *Protein Sci.* **11**, 1285–1299 (2002).
86. Upadhyay, A. K. & Sowdhamini, R. Genome-wide analysis of domain-swap predicted products in the genome of anti-stress medicinal plant: Ocimum tenuiflorum. *Bioinform. Biol. Insights* **13**, (2019).
87. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* vol. 29 105–153 (2000).
88. Li, H. *et al.* Domain swap approach reveals the critical roles of different domains of symrk in root nodule symbiosis in lotus japonicus. *Front. Plant Sci.* **9**, (2018).
89. Back, K. & Chappell, J. Identifying functional domains within terpene cyclases using a domain-swapping strategy. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 6841–6845 (1996).
90. Facchini, P. J. & Chappell, J. Gene family for an elicitor-induced sesquiterpene cyclase in tobacco. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 11088–11092 (1992).
91. Back, K. & Chappell, J. Cloning and bacterial expression of a sesquiterpene cyclase from Hyoscyamus muticus and its molecular comparison to related terpene cyclases. *J. Biol. Chem.* **270**, 7375–7381 (1995).
92. Greenhagen, B. T., O'Maille, P. E., Noel, J. P. & Chappell, J. Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9826–9831 (2006).
93. El Tamer, M. K. *et al.* Domain swapping of Citrus limon monoterpene synthases: Impact on enzymatic activity and product specificity. *Arch. Biochem. Biophys.* **411**, 196–203 (2003).
94. Qin, B. *et al.* An Unusual Chimeric Diterpene Synthase from Emericella variecolor and Its Functional Conversion into a Sesterterpene Synthase by Domain Swapping. *Angew. Chemie* -

*Int. Ed.* **55**, 1658–1661 (2016).
95. Labrou, N. Random Mutagenesis Methods for In Vitro Directed Enzyme Evolution. *Curr. Protein Pept. Sci.* **11**, 91–100 (2010).
96. Steele, C. L., Crock, J., Bohlmann, J. & Croteau, R. Sesquiterpene synthases from grand fir (Abies grandis) comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of $δ$-selinene synthase and $γ$-humulene synthase. *J. Biol. Chem.* **273**, 2078–2089 (1998).
97. Yoshikuni, Y., Ferrin, T. E. & Keasling, J. D. Designed divergent evolution of enzyme function. *Nature* **440**, 1078–1082 (2006).
98. Szymczyk, P., Szymańska, G., Lipert, A., Weremczuk-Jeżyna, I. & Kochan, E. Computer-Aided Saturation Mutagenesis of Arabidopsis thaliana Ent-Copalyl Diphosphate Synthase. *Interdiscip. Sci. Comput. Life Sci.* **12**, 32–43 (2020).
99. Diaz, J. E. *et al.* Computational design and selections for an engineered, thermostable terpene synthase. *Protein Sci.* **20**, 1597–1606 (2011).
100. Garms, S., Chen, F., Boland, W., Gershenzon, J. & Köllner, T. G. A single amino acid determines the site of deprotonation in the active center of sesquiterpene synthases SbTPS1 and SbTPS2 from Sorghum bicolor. *Phytochemistry* **75**, 6–13 (2012).
101. Huang, J. Q. *et al.* 1,10/1,11-Cyclization catalyzed by diverged plant sesquiterpene synthases is dependent on a single residue. *Org. Biomol. Chem.* **19**, 6650–6656 (2021).
102. Ker, D. S., Chan, K. G., Othman, R., Hassan, M. & Ng, C. L. Site-directed mutagenesis of β sesquiphellandrene synthase enhances enzyme promiscuity. *Phytochemistry* **173**, 112286 (2020).
103. Fang, X. *et al.* Systematic identification of functional residues of Artemisia annua amorpha-4,11-diene synthase. *Biochem. J.* **474**, 2191–2202 (2017).
104. Leferink, N. G. H. *et al.* An automated pipeline for the screening of diverse monoterpene synthase libraries. *Sci. Rep.* **9**, 1–12 (2019).
105. Behrendorff, J. B. Y. H., Vickers, C. E., Chrysanthopoulos, P. & Nielsen, L. K. 2,2-Diphenyl-1-picrylhydrazyl as a screening tool for recombinant monoterpene biosynthesis. *Microb. Cell Fact.* **12**, 76 (2013).
106. Vardakou, M., Salmon, M., Faraldos, J. A. & O'Maille, P. E. Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* **1**, e187–e196 (2014).
107. Feng, J. *et al.* An improved malachite green assay of phosphate: Mechanism and application. *Anal. Biochem.* **409**, 144–149 (2011).
108. Lauchli, R. *et al.* High-throughput screening for terpene-synthase-cyclization activity and directed evolution of a terpene synthase. *Angew. Chemie - Int. Ed.* **52**, 5571–5574 (2013).
109. Quesenberry, M. S. & Lee, Y. C. A rapid formaldehyde assay using purpald reagent: Application under periodation conditions. *Anal. Biochem.* **234**, 50–55 (1996).
110. Siltanen, C. A. *et al.* An Oil-Free Picodrop Bioassay Platform for Synthetic Biology. *Sci. Rep.* **8**, (2018).
111. Furubayashi, M. *et al.* A high-throughput colorimetric screening assay for terpene synthase activity based on substrate consumption. *PLoS One* **9**, e93317 (2014).
112. Tashiro, M. *et al.* Bacterial Production of Pinene by a Laboratory-Evolved Pinene-Synthase. *ACS Synth. Biol.* **5**, 1011–1020 (2016).
113. Pfleger, B. F., Pitera, D. J., Newman, J. D., Martin, V. J. J. & Keasling, J. D. Microbial sensors for small molecules: Development of a mevalonate biosensor. *Metab. Eng.* **9**, 30–38 (2007).
114. De Paepe, B., Peters, G., Coussement, P., Maertens, J. & De Mey, M. Tailor-made transcriptional biosensors for optimizing microbial cell factories. *J. Ind. Microbiol. Biotechnol.* **44**, 623–645 (2017).
115. Dimas, R. P. *et al.* Engineering DNA recognition and allosteric response properties of TetR family proteins by using a module-swapping strategy. *Nucleic Acids Res.* **47**, 8913–8925 (2019).
116. De Paepe, B., Maertens, J., Vanholme, B. & De Mey, M. Chimeric LysR-Type Transcriptional

Biosensors for Customizing Ligand Specificity Profiles toward Flavonoids. *ACS Synth. Biol.* **8**, 318–331 (2019).
117. Wang, R. *et al.* Design and Characterization of Biosensors for the Screening of Modular Assembled Naringenin Biosynthetic Library in Saccharomyces cerevisiae. *ACS Synth. Biol.* **8**, 2121–2130 (2019).
118. Siedler, S., Stahlhut, S. G., Malla, S., Maury, J. & Neves, A. R. Novel biosensors based on flavonoid-responsive transcriptional regulators introduced into Escherichia coli. *Metab. Eng.* **21**, 2–8 (2014).
119. Peck, M. C., Fisher, R. F. & Long, S. R. Diverse flavonoids stimulate NodD1 binding to nod gene promoters in Sinorhizobium meliloti. *J. Bacteriol.* **188**, 5417–5427 (2006).
120. Vargas, P., Felipe, A., Michán, C. & Gallegos, M. T. Induction of Pseudomonas syringae pv. tomato DC3000 mexAB-oprM multidrug efflux pump by flavonoids is mediated by the repressor PmeR. *Mol. Plant-Microbe Interact.* **24**, 1207–1219 (2011).
121. Hirooka, K. & Fujita, Y. Identification of Aromatic Residues Critical to the DNA Binding and Ligand Response of the Bacillus subtilis QdoR (YxaF) Repressor Antagonized by Flavonoids. *Biosci. Biotechnol. Biochem.* **75**, 1325–1334 (2011).
122. Förster-Fromme, K. & Jendrossek, D. AtuR is a repressor of acyclic terpene utilization (Atu) gene cluster expression and specifically binds to two 13 bp inverted repeat sequences of the atuA-atuR intergenic region. *FEMS Microbiol. Lett.* **308**, 166–174 (2010).

# 2

# An analysis of characterized plant sesquiterpene synthases

Alice Di Girolamo[1]*, Janani Durairaj[2]*, Harro J. Bouwmeester[3], Dick de Ridder[2], Jules Beekwilder[1,4], Aalt DJ. van Dijk[2,5]

[1] Laboratory of Plant Physiology, Department of Plant Sciences, Wageningen University, Netherlands
[2] Bioinformatics Group, Department of Plant Sciences, Wageningen University, Netherlands
[3] Swammerdam Institute for Life Sciences, University of Amsterdam, Netherlands
[4] Bioscience, Wageningen Plant Research, Wageningen University, Netherlands
[5] Biometris, Department of Plant Sciences, Wageningen University, Netherlands
* Authors contributed equally to the manuscript

Published in Phytochemistry 158, 2019

# abstract

Plants exhibit a vast array of sesquiterpenes, C15 hydrocarbons which often function as herbivore-repellents or pollinator-attractants. These in turn are produced by a diverse range of sesquiterpene synthases. A comprehensive analysis of these enzymes in terms of product specificity has been hampered by the lack of a centralized resource of sufficient functionally annotated sequence data. To address this, we have gathered 262 plant sesquiterpene synthase sequences with experimentally characterized products. The annotated enzyme sequences allowed for an analysis of terpene synthase motifs, leading to the extension of one motif and recognition of a variant of another. In addition, putative terpene synthase sequences were obtained from various resources and compared with the annotated sesquiterpene synthases. This analysis indicated regions of terpene synthase sequence space which so far are unexplored experimentally. Finally, we present a case describing mutational studies on residues altering product specificity, for which we analysed conservation in our database. This demonstrates an application of our database in choosing likely-functional residues for mutagenesis studies aimed at understanding or changing sesquiterpene synthase product specificity.

## 2.1 Introduction

The terpenome represents a huge, ancient and diverse family of natural products. In addition to terpenes, it also encompasses steroids and carotenoids, comprising more than 60,000 members[1]. These compounds all derive from the same 5-carbon precursor units, coupled together linearly and then cyclized, rearranged, and modified in various ways. Terpenes serve many roles in plants, for example as toxins against herbivores or pathogens, or as attractants for pollinators[2]. In turn, terpenes extracted from plants are used by mankind for a range of applications - as pharmaceutical agents, insecticides, preservatives, fragrances, and flavors[3]. Terpenes are built from 5-carbon isoprenoid units and they mainly exist as monoterpenes (C10), sesquiterpenes (C15) or diterpenes (C20), based on the number of such units used. In each case, a linear substrate loses a diphosphate group, usually cyclizes and then undergoes a variety of carbocation rearrangements. Though the exact number of sesquiterpenes found in nature is hard to determine[4], estimated computationally that the number of sesquiterpene intermediates far outnumber those of monoterpenes, due to the increase in chain length. Interestingly, sesquiterpenes found in nature can be divided into seven groups based on their parent cation and the first cyclization step in their formation[5]. Hence the extreme diversity of chemical compounds with desirable fragrances or medicinal properties is based on just seven initial carbocations. This makes the enzymes catalysing their formation both interesting and difficult to characterize functionally.



Each plant species is capable of synthesizing a number of sesquiterpenes using a specialized class of enzymes called sesquiterpene synthases (STSs). First, a farnesyl diphosphate synthase, produces the C15 substrate for STSs, farnesyl diphosphate (FDP), from the C5-unit isopentenyl diphosphate (IDP) and its isomer dimethylallyl diphosphate (DMADP)[6]. STSs then create the myriad of sesquiterpenes found in nature by catalysing carbocation formation from the linear FDP followed by a series of cyclizations and rearrangements (**Figure 2.1**). Products are formed from intermediate carbocations after deprotonation, phosphorylation, or hydration[4]. The STSs themselves represent a very diverse set of enzymes with a wide range of sequence similarities, despite having a common structural fold shared by plant, animal, fungal, and bacterial terpene synthases (TPSs)[7]. Hence, prediction of enzyme function from sequence is highly challenging in the case of STSs. Moreover, sequence diversity in STSs is not dependent on the products formed. This problem has been addressed so far by inspection of TPS structures[7] and by mutational analyses that attempt to change the

product of a synthase with the smallest number of residue changes[8]. The former, though an attractive approach, is limited especially in plants due to the sparsity of experimentally determined structures, while the latter often leads to unnatural enzymes with lower catalytic activity than their wild-type parents. Characterization of multiple TPSs from the same species by the same study has allowed for some small-scale sequence comparison of those synthases[9,10]. However, no previous attempts have been made to compare all experimentally characterized plant STS sequences according to the products that they form. We have collated a curated database of plant STSs with characterized products from literature. This database can be accessed at www.bioinformatics.nl/sesquiterpene/synthasedb.



With this database and aforementioned product grouping scheme, the active domain sequences of 262 plant STSs were analysed in terms of the precursor carbocations of their products. These were also compared with the many yet-uncharacterized putative TPS enzymes. Residues from previous product-changing mutational studies were mapped on our database of enzymes, indicating conservation of the corresponding positions across groups of sequences forming different product cations. This demonstrates the usefulness of our database in finding residues involved in STS product specificity.

## 2.2 Methods

### 2.2.1 Literature search for characterized STSs

To find potentially characterized STSs, an HMM search was performed using hmmer (version 3.1b2)[11] on the UniProt database[12] using the HMM of the C-terminal domain of TPSs from Pfam[13] (Pfam ID: PF03936). Protein sequences with a hit having an E-value < $10^{10}$ and a total protein length between 350 and 650 residues were selected. The Uniprot IDs of these sequences were then linked to Pubmed IDs, either directly through programmatic access of Uniprot if the Pubmed ID was present, or through a programmatic text search of the title and authors given in Uniprot, using the Pubmed API[14]. The Pubmed articles thus obtained were searched manually for evidence of experimental characterization of sesquiterpenes through in-vivo or in-vitro GC-MS studies, and the corresponding Uniprot IDs were collected. For each UniProt ID found, the major product described in the corresponding paper was stored. Minor products with GC-MS peaks at least quarter the height of the major product peak were stored as well.

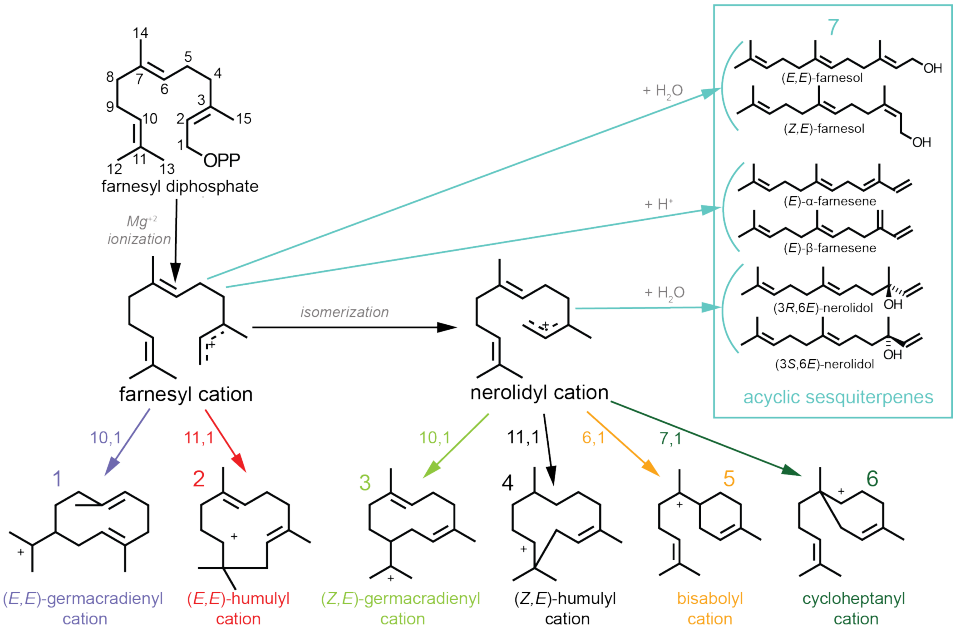



**Figure 2.1** The reaction mechanism of sesquiterpene production starts with farnesyl diphosphate (FDP). Loss of the diphosphate moiety (OPP) leads to farnesyl cation formation. The farnesyl cation can subsequently be converted to the nerolidyl cation. Possible cyclizations for both cations are indicated in the figure. The subsequently formed cyclic cations undergo further modifications and rearrangements to form sesquiterpenes. An alternative route is to form acyclic sesquiterpenes from either the farnesyl or the nerolidyl cation as indicated in the box. These different product precursors are used to classify the different sesquiterpenes and their synthases.

## 2.2.2 Measuring chemical similarities

The diagram of the sesquiterpene grouping scheme was made using ChemDoodle (version 9)[15]. The InChI strings for 165 sesquiterpenes were obtained from PubChem[16] using the python wrapper for the PubChem REST API[17], PubChemPy (version 1.0.4). To measure the similarity between different sesquiterpenes, rdkit (Release 2017.09.3) was used[18]. A circular chemical fingerprint, called the Morgan fingerprint, with a radius of 2 angstroms, as explained by Rogers and Hahn was obtained for each sesquiterpene[19]. The similarity between every pair of fingerprints was then calculated using Dice similarity[20]. The distance was given as 1 *similarity*. The distance matrix of all sesquiterpenes was then used to create a multi-dimensional scaling (MDS) plot using the Python scikit-learn library (version 0.19.1)[21], and then plotted using matplotlib (version 2.1.2)[22].



## 2.2.3 Aligning sequences

For characterized spermatophyte plant STS sequences, the C-terminal catalytically active portion and the N-terminal portion of the enzyme were found with hmmer HMM searches (version 3.1b2)[11] using the TPS C-terminal Pfam domain (Pfam ID: PF03936) and the TPS N-terminal Pfam domain (Pfam ID: PF01397) respectively. These were then separately aligned using Clustal Omega (version 1.2.4)[23], with all heuristic features off and the respective Pfam domains as a guide for alignment. From these separate alignments, a concatenated N+C alignment was formed, covering both domains.

For some of the nonseed plant STS sequences however, a C-terminal Pfam domain search returned < 200 residues instead of the usual 250-270. Aligning the full nonseed sequences using the spermatophyte C-terminal sub-sequence alignment as a profile showed the position of the C-terminal portion for these sequences, so this was used to extract the required C-terminal sub-sequences for nonseed plants. An alignment consisting of both seed and nonseed characterized C-terminal sub-sequences was constructed using Clustal Omega with the same parameters as above.

## 2.2.4 Phylogenetic tree construction

A phylogenetic tree was built and visualized for the characterized spermatophyte and nonseed plant enzymes in the database using the ete toolkit (version 3.1.1)[24]. The previously explained alignment of all C-terminal sub-sequences was used, with columns having > 50% gaps removed using

trimAL[25]. The best protein model from JTT, WAG, VT, LG and MtREV was chosen using ProtTest[26], and finally a RaxML maximum likelihood tree was built[27]. Similarly, a phylogenetic tree for the spermatophyte sequences was built with the same approach using the concatenated N+C alignment.

### 2.2.5 Finding mono-, di-, and uncharacterized TPSs

Characterized plant mono- and diterpene synthases were obtained from SwissProt[28] using a C-terminal TPS Pfam domain hmmer (verson 3.1b2)[11] HMM search followed by collecting the sequences from plant species for which the catalytic activity was mentioned. These were not manually checked. Uncharacterized TPS C-terminal sub-sequences were then obtained from plant species in TremBl[28], Ensembl Plants (release 38)[29], and the 1000 Plants Transcriptome Project[30] again using a Pfam domain search. Only those sequences where the search returned a sub-sequence having both DDXX(D,E) and (N,D)DXX(S,T,G)XXXE or two DDXX(D,E) motifs within it, and whose sub-sequence length was within two standard deviations of the mean C-terminal sub-sequence length of characterized STS enzymes were retained. In both sets, sequences from nonseed plant species were discarded.



### 2.2.6 Measuring sequence similarities

A distance matrix of all spermatophyte TPS C-terminal sub-sequences: characterized mono-, di- and sesquiterpene synthases as well as uncharacterized enzymes, was constructed using the pairwise sequence k-tuple measure described by Wilbur *et al.*(1983)[31], implemented in Clustal Omega (version 1.2.4)[23]. This distance matrix was then used to construct an MDS plot using scikit-learn (version 0.19.1)[21] and plotted using matplotlib (version 2.1.2)[22]. A cluster-map of sequence identities between characterized STS enzymes was made using the distance matrix of just these enzymes and complete hierarchical clustering using scipy (version 1.0.0)[32] and seaborn (version 0.8.1)[33].

### 2.2.7 Visualizing an STS structure

The 5EAT tobacco 5-epi-aristolochene synthase structure from the Protein Data Bank (PDB)[34] was used to visualize known TPS motifs, along with $Mg^{2+}$ ions and farnesyl hydroxyphosphonate (FHP) substrate analog. Visualization was done in Pymol 2.1[35].

## 2.3 Results and discussion

### 2.3.1 Database of characterized STSs

To obtain a comprehensive set of annotated STSs, our starting point was the SwissProt database, a subset of UniProt[28] in which a curated and annotated set of proteins is available. This provided a set of 104 STSs. In addition, we manually reviewed literature linked to enzymes with the characteristic TPS domain in TremBl, the uncurated subset of UniProt. In this way, the number of curated plant STS sequences with experimentally characterized product data in the database was more than doubled.



We present a database of 262 manually curated characterized plant STSs, shown in **Table 2.1**. The enzymes originate from a hundred different plant species and collectively account for the production of 117 different sesquiterpenes. Such a large number of possible products makes it difficult to find enough enzymes with the same product for a meaningful analysis of product specificity. To solve this, the sequences were divided into seven groups, making use of the sesquiterpene precursor carbocation scheme as specified by Degenhardt *et al.* (2009)[5], described in **Figure 2.1**. The reaction cascade of an STS is initiated by metal-mediated removal of the diphosphate anion in the FDP substrate, leading to the formation of a transoid (*2E, 6E*)-farnesyl cation (farnesyl cation) which can undergo cyclization either via *10-exo-trig* or *11-endo-trig* cyclizations on the C10-C11 double bond to the resulting cations 1 or 2 respectively However, the farnesyl cation can also isomerize to form a cisoid *(2Z, 6E)*-farnesyl cation (nerolidyl cation). The nerolidyl cation, in addition to a C1-attack (either via *10-exo-trig* or *11-endo-trig*) on the C10-C11 double bond to form cations 3 or 4, can also undergo cyclization at its C6-C7 double bond either via *6-exo-trig* or *7-endo-trig*, forming cations 5 or 6. These carbocations undergo multiple further skeletal rearrangements, cyclizations, hydride or methyl shifts, and other modifications to form the end products of the enzyme[5] Along with this myriad of cyclic products, acyclic sesquiterpenes can also be formed from either the farnesyl or the nerolidyl cation through proton loss or addition of water[5,36,37]. This schematic of carbocations derived from FDP can be used to divide sesquiterpenes produced by plants into seven groups both based on their parent cation (farnesyl or nerolidyl) and the first cyclization that occurs (by attack of the carbocation on the 10,1-; 11,1-; 6,1-; or 7,1- double bond; or acyclic). For an STS enzyme, the carbocation of its major product is then used to determine its group in **Table 2.1**.

This division of STSs is in general straightforward even when multiple products are formed by one enzyme. Specifically, of the 98 sequences which also have minor products (**Table S2.1**), only 17 have minor products whose precursor carbocation differs from the major product's. Nine of these produce acyclic products in addition to their major product. This could be the result of incomplete cyclization caused by premature termination of intermediates[38]. Eight enzymes in the database either produce (-)-germacrene D or they produce germacrene D and the chirality was not determined during the enzyme's characterization. (-)-germacrene D can be formed via a 10,1- or a 11,1- cyclization of the farnesyl cation (cation 1 or 2).



Though each enzyme is likely to only follow one cyclization route to form its product, this route has so far not been determined, so these sequences are shown separately in **Table 2.1** and in the remainder of the text. The existence of other sesquiterpenes which can be formed via different cyclization routes cannot be ruled out, however in our analysis we stick to the cyclization routes provided by IUBMB's *Enzyme Nomenclature* Supplement 24 (2018)[39] in order to determine the precursor carbocation for a given sesquiterpene.

The database contains 233 angiosperm STSs, 16 gymnosperm enzymes from coniferous species and 13 enzymes from nonseed plants such as mosses and ferns. As described by Jia *et al.* (2018)[40], the latter species have TPSs which are more related to microbial TPSs than those from spermatophytes. Information on each of the 262 enzymes, including the sequence, species, Uniprot ID, products (major and minor), product type, and Pubmed ID of the paper detailing its experimental characterization, is available as a web service at www.bioinformatics.nl/sesquiterpene/synthasedb. The service supports searching, sorting and downloading of all or subsets of the data.

On average, the enzymes comprise of 553 ± 56 residues. The tertiary structure of STS enzymes usually comprises of two alpha-helical domains[41]. The N-terminal domain is considered relictual in plant STSs and is not present at all in nonseed plant STSs[40], while the C-terminal domain, consisting of an $\alpha$-helical bundle, is catalytically active[7,42]. The hydrophobic active site pocket in this domain is formed by six $\alpha$-helices, closed by two loops. **Table S2.2** gives a list of plant STS structures from the Protein Data Bank (PDB)[34]. The C-terminal sub-sequences containing the active site are obtained from each enzyme in the database using information from Pfam[13], and consist

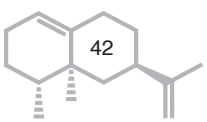
of 266 ± 7 residues. N-terminal sub-sequences were extracted only from the spermatophyte enzymes in the database, again using information from Pfam, and consist of 173 ± 12 residues. In spermatophyte STSs, residues distal to the active site have been shown to contribute to product specificity potentially by influencing active site geometry[43]. These residues may reside in the extremities of the C-terminal domain, or in the N-terminal domain. **Figure S2.1** shows the pairwise sequence identity scores for each pair of C-terminal domain sub-sequences for the enzymes in the database, hierarchically clustered and coloured by product cation type. It can be seen that many pairs of sequences have less than 40% sequence identity. Similarly, **Figure S2.2** shows the hierarchical clustering of concatenated N-terminal and C-terminal sub-sequences for spermatophyte enzymes. Both clusterings appear very comparable.



| Major Product Group | Cation/ Cyclization | No. of Sequences | | | | No. of Species | | | | No. of Products |
|---|---|---|---|---|---|---|---|---|---|---|
| | | A | G | N | Total | A | G | N | Total | |
| 1 | 10,1/farnesyl | 77 | 1 | 3 | 81 | 44 | 1 | 3 | 48 | 43 |
| 2 | 11,1/farnesyl | 42 | 3 | 3 | 48 | 32 | 3 | 3 | 38 | 11 |
| 3 | 10,1/nerolidyl | 19 | 1 | 1 | 21 | 16 | 1 | 1 | 18 | 20 |
| 4 | 11,1 nerolidyl | 0 | 4 | 0 | 4 | 0 | 4 | 0 | 4 | 3 |
| 5 | 6,1/nerolidyl | 44 | 3 | 2 | 49 | 23 | 3 | 2 | 28 | 32 |
| 6 | 7,1/nerolidyl | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 7 | acyclic | 43 | 4 | 3 | 50 | 23 | 4 | 3 | 30 | 6 |
| (–) | (–)-germacrene D | 8 | 0 | 0 | 8 | 6 | 0 | 0 | 6 | 1 |
| Total | | 233 | 16 | 13 | 262 | 84 | 8 | 9 | 101 | 117 |

**Table 2.1** Number of characterized plant STS sequences, species, and products covered in each product group. (–)-germacrene D synthases are shown separately as discussed in the text. A=Angiosperms, G=Gymnosperms, N=Nonseed.

The phylogenetic tree of C-terminal sub-sequences of all 262 enzymes (**Figure 2.2**) shows some grouping of spermatophyte enzymes based on their product precursor. In general, the neighbor of an enzyme is from the same or related species, and if there are enough examples from the same species then some product-based grouping is seen. For example, the clades containing mostly enzymes from *Zea mays* on the right are separated based on the product carbocation of the enzyme even while being grouped by the species. However, this is not a consistent trend; enzymes from *Vitis* and *Santalum* at the top of the tree group mainly by species and not by product type. In fact, the three *Santalum* synthase sequences marked in **Figure 2.2**, making products derived from three different cyclic carbocations, have more than 90% in common.



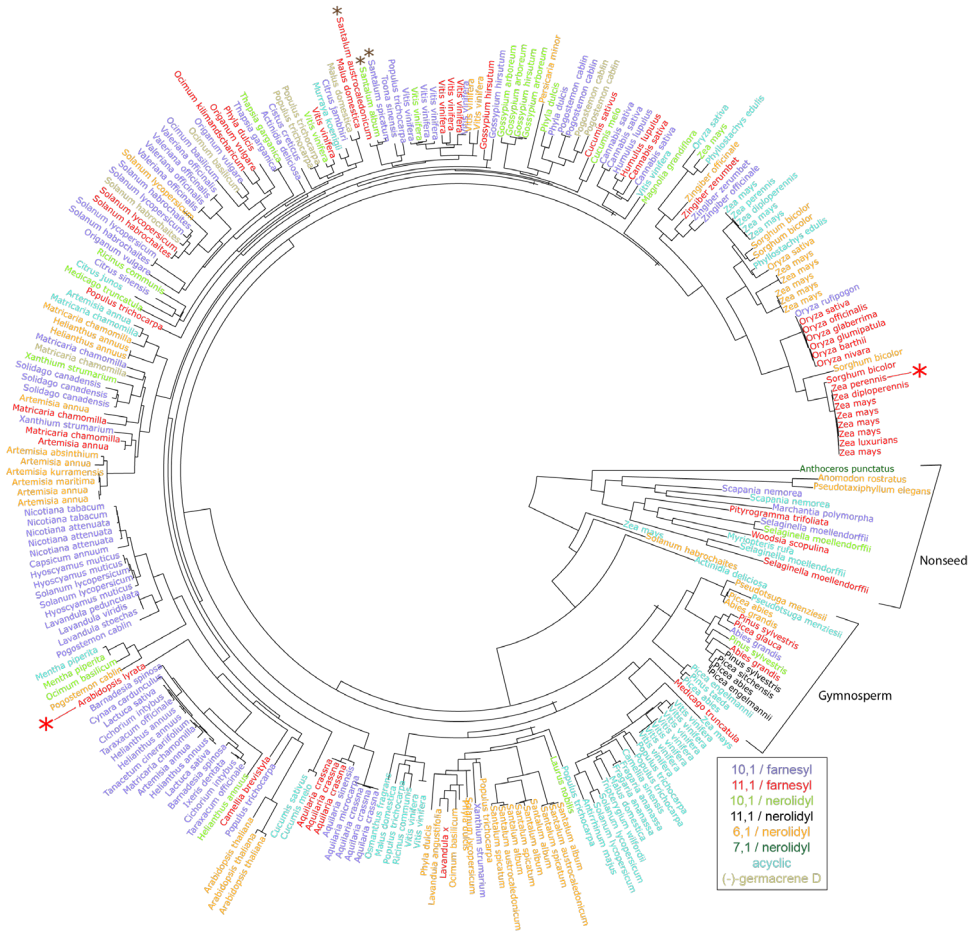



**Figure 2.2** Phylogenetic tree of C-terminal sub-sequences for characterized plant STSs, coloured according to the major product's initial carbocation (see **Figure 2.1**). Nonseed and gymnosperm clades are indicated separately. Red and brown asterisks mark cases discussed in the text: red - two β-caryophyllene synthases from *Arabidopsis lyrata* and *Zea perennis* which have less than 30% pairwise sequence identity; brown - three synthases from Santalum with higher than 90% sequence identity.

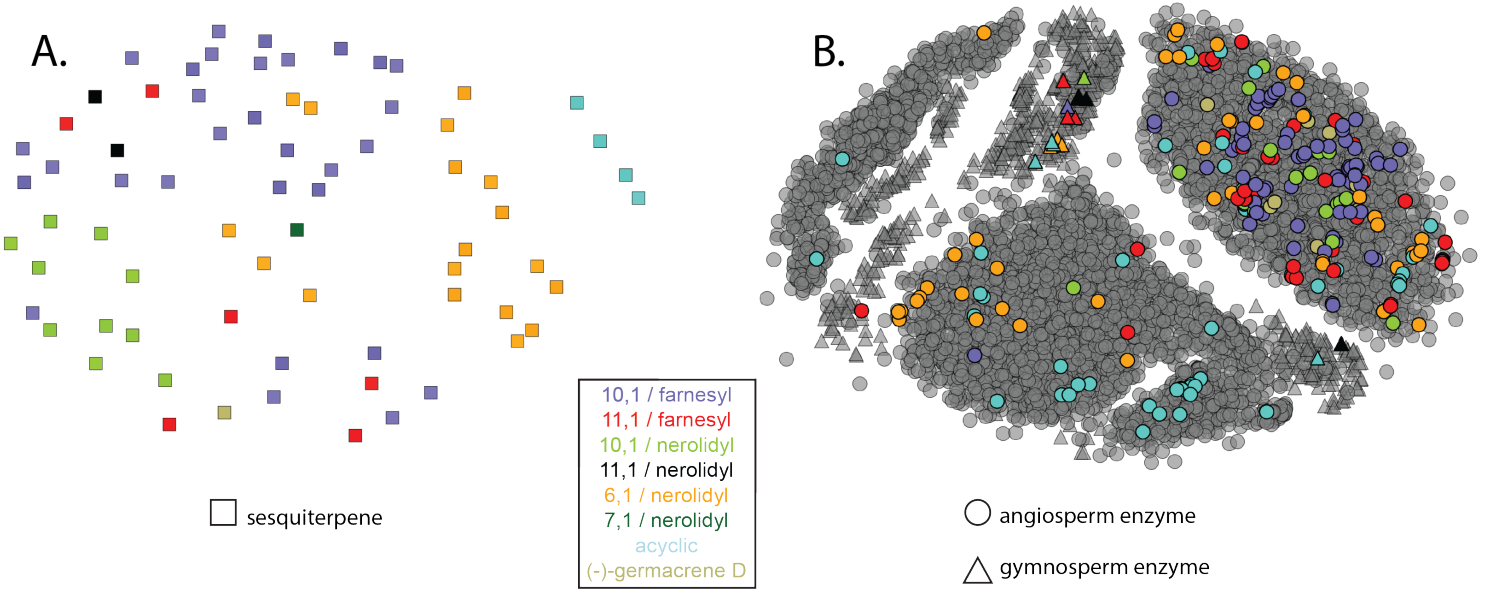



**Figure 2.3** (A). MDS plot of 165 sesquiterpenes found in nature, based on chemical fingerprint similarities. Each square represents a sesquiterpene and the more chemically similar two sesquiterpenes are, the closer they are placed in the plot. Colours are based on the sesquiterpene's precursor carbocation. (B). MDS plot of TPS C-terminal domain sub-sequences with colouring based on STS major product carbocation. Unknown proteins which are likely to be TPSs are shown in gray. The more similar two sequences are, the closer they are in the plot.

In any case, the product group of an enzyme from a species not present in the tree is nearly impossible to predict, while enzymes from species which are less represented in the tree can also be difficult to classify. In addition, clades forming predominantly one product carbocation are seen in many different parts of the tree, showing that strongly varying sequences can catalyse the same cyclization reaction and even produce the same product, such as the two marked β-caryophyllene synthases from *Arabidposis lyrate* and *Zea perennis* which have a sequence identity less than 30%. Hence phylogenetic analysis is biased and cannot be an accurate predictor of TPS product specificity. Figure S2.3, shows a similar tree considering both N-terminal and C-terminal sub-sequences concatenated together, for spermatophyte STS sequences only. N-terminal domain information again does not seem to effect the structure of the tree. Even though this does not rule out the possibility that residues in the N-terminal domain influence product specificity, it indicates that including the N-terminal domain in the large-scale sequence analysis that we perform does not add information compared to using only the C-terminal domain. Since product and intermediate formation occur in the active site pocket, it may be easier to find sequence-function determinants in the C-terminal domain. Hence, from this point on we concentrate on the C-terminal sub-sequences of TPSs.



The clade containing all the nonseed plant STSs in **Figure 2.2** is clearly separate from the spermatophyte sequences. The enzyme from *Anthoceros punctatus*, a bryophyte, is the only sequence in the database producing a 7,1/nerolidyl-derived product (β-acoradiene) and is hence an out-group both in terms of species as well as product carbocation. Comparing nonseed plant sequences to the more typical plant TPS sequences would be futile, both due to their homology with microbial enzymes and their low numbers in the database, hence they are excluded from the remainder of the analysis.

### 2.3.2 Chemical similarities between sesquiterpenes

Each of the seven possible sesquiterpene precursors (**Figure 2.1**) usually undergoes a wide range of further rearrangements, cyclizations, and modifications, catalysed by the STS enzyme, to finally result in a sesquiterpene product. To start exploring the enzyme grouping scheme, we initially investigated whether similarities between the final sesquiterpene chemical structures would reflect the parent carbocations involved in their production. To this end, chemical similarities between sesquiterpenes with the same parent cation were compared to similarities between those without. Chemical similarities were measured using Dice similarity[20]

between extended connectivity fingerprints, as described by Rogers and Hahn[19]. Similarities between 165 sesquiterpenes are plotted using multi-dimensional scaling (MDS), in **Figure 2.3A** with the colour representative of the precursor cation. These 165 compounds collectively represent every enantiomer of the 117 sesquiterpenes produced by the enzymes in our database, since many of the experimental characterization studies used to build the database did not resolve the chirality of the STS's product. MDS is a technique used to visualize the level of similarity of individual objects in a dataset using a distance matrix, such that the between-object distances are preserved as well as possible. Therefore, two objects appearing close to each other in the MDS plot represent sesquiterpenes which likely have a high chemical similarity, while those further away have lower similarity. Acyclic sesquiterpenes are clearly distinguishable in the plot, as they are linear in nature. Interestingly, many products derived from the 6,1-cyclized cation (cation 5) are also distinct from those derived from 10,1- or 11,1-cyclized cations despite further cyclizations and rearrangements after this first step. They cluster midway between the acyclic and other cyclic products, which makes sense given the presence of an acyclic tail portion in cation 5. The sesquiterpenes formed from the other cyclic cations seem less distinguishable.



## 2.3.3 Characterized sequence space

Though a manual literature search gave us access to more functionally characterized TPS sequences, there is a large and steadily growing number of protein sequences present in various databases which have not been characterized at all. Many of these proteins are potential TPSs which contain the characteristic, catalytic site containing, C-terminal domain. Comparing uncharacterized and characterized enzymes may give indications of the nature of an uncharacterized enzyme, in particular about the cyclization route it is likely to take, thereby assisting in the setup of experiments for functional characterization.

To explore this, an MDS plot was made of C-terminal sub-sequences of the 249 spermatophyte enzymes in our database with those of 6278 other spermatophyte TPS-like sequences, obtained from sequenced genomes and transcriptomes. These 6278 sequences are, to the best of our knowledge, uncharacterized. **Figure 2.3B** shows this plot where the colours represent the product precursor carbocation of characterized STSs and the uncharacterized sequences are shown in grey. Similar sequences are depicted closer together in the plot.

**Figure 2.3B** has a few commonalities with the MDS plot of chemical similarities between sesquiterpenes, **Figure 2.3A**. Many sequences catalysing acyclic products as well as those derived from cation 5 cluster separately from the others. In fact, the enzymes making nerolidol, an acyclic sesquiterpene, cluster separately at the bottom right of the plot (light blue), leading us to hypothesize that perhaps many of the other uncharacterized STSs in this area also catalyse the formation of nerolidol. A second similarity is that enzymes forming products derived from 10,1- and 11,1- cyclized cations are difficult to distinguish. This again confirms, as was seen in the phylogenetic tree (**Figure 2.2**), that overall sequence similarity by itself cannot be an accurate guide to product specificity.



The uncharacterized sequences depicted in **Figure 2.3B** could be mono-, di-, or sesquiterpene synthases. **Figure S2.4** shows 57 monoterpene synthases and 20 diterpene synthases from SwissProt, along with the 249 STSs in our database. Despite the skewed numbers, a separation between mono- and sesquiterpene synthases can be seen, indicating areas of the sequence space where more STSs are likely to be found.

Product specificity is even harder to identify in the case of gymnosperm synthases, as insufficient data is available to separate enzymes with different product cations. It has been noted before that gymnosperm TPSs resemble each other more than they do their angiosperm counterparts, regardless of catalytic activity[44,45]. The enzymes from these species may be more informative if analysed separately but this would require more gymnosperm sequences to be functionally annotated.

## 2.3.4 Comparing known TPS motif across sequences

A database such as ours allows for a comparison of residues in previously studied structural elements across many STS sequences. A thorough study of TPS structures has led to the identification of several motifs important for catalytic activity[7]. In the case of STSs, the hydrophobic moiety of the STS substrate, FDP, is directed into the active site cavity, to undergo the cyclizations and rearrangements described in **Figure 2.1**. Studies on STS structures have proposed that the diphosphate moiety is captured by the motif RxR and divalent metal ions like $Mg^{2+}$ or $Mn^{2+}$ which are themselves bound by motifs DDXXD and NSE/DTE, at the entrance of the active site[46]. Here, we compare these three motifs across the sequences in our database. **Figure 2.4A** shows the motifs discussed below on a tobacco aristolochene synthase structure[46]. **Figure 2.4B** shows each motif on a

schematic representation of the alignment of all C-terminal sub-sequences in the database.

#### 2.3.4.1 Aspartate-rich DDXXD motif conserved in plant STSs

The most conserved motif of TPSs is the metal binding aspartate-rich motif found both in plant and microbial TPSs as well as in isoprenyl diphosphate (IDP) synthases. Numerous studies performed on this motif, both site-directed mutagenesis and X-ray crystallography analysis, show that it is involved in binding the divalent metal ions in the active site entrance[47]. The canonical form of the motif, **D**DXX**(D,E)**, where bold-faced residues indicate those proposed to bind $Mg^{2+}$ or $Mn^{2+}$ is found in 247 of the 249 spermatophyte enzymes. Of the remaining two, one is a (+)-germacrene-D synthase from *Solidago canadensis* with an Asn replacing the first A[48]. The other is a bicyclogermacrene synthase from *Matricaria chamomilla* with an Asn replacing the second Asp[49]. These examples indicate that either one of the first two Aspartates may be sufficient for maintaining catalytic activity.

#### 2.3.4.2 Expanded NSE/DTE motif found in most sequences

The opposite site of the active site entry is also involved in metal-binding, due to the presence of a second, less-defined motif, termed the NSE/DTE motif[50]. An early form of this motif, as detailed by Christianson *et al.* (2006)[50] had a consensus of (L,V)(V,L,A)**(N,D)**D(L,I,V)X**(S,T)**XXX**E**, where the residues in bold coordinate $Mg^{2+}$ ions. However, searching for a motif with this consensus only captured 38 of the 249 spermatophyte sequences in our database, indicating that it may be too restrictive given the current knowledge of sequences.

When only the metal-binding portion of the motif is considered, the consensus sequence **(N,D)**DXX**(S,T,G)**XXX**E** covers 219 spermatophyte sequences in the database. The possibility of Gly in the second metal-binding position is justified by Zhou and Peters (2009)[51], with the proposal that Gly may allow a water molecule to substitute for the hydroxyl group of Ser/Thr. Some TPSs however, are known to have a second, catalytically active, aspartate rich motif instead of the NSE/DTE motif[52–54] with the same consensus as the first, **D**DXX**(D,E)**. This occurs in 20 sequences. **Table 2.2** shows the distribution of the sequences over the different versions of the second motif. A highly conserved Arg is found 3 residues upstream of all versions of the NSE/DTE motif or second aspartate-rich motif, in all of the spermatophyte sequences in the database. All 6278 uncharacterized spermatophyte TPS sequences also have an arginine in this position.

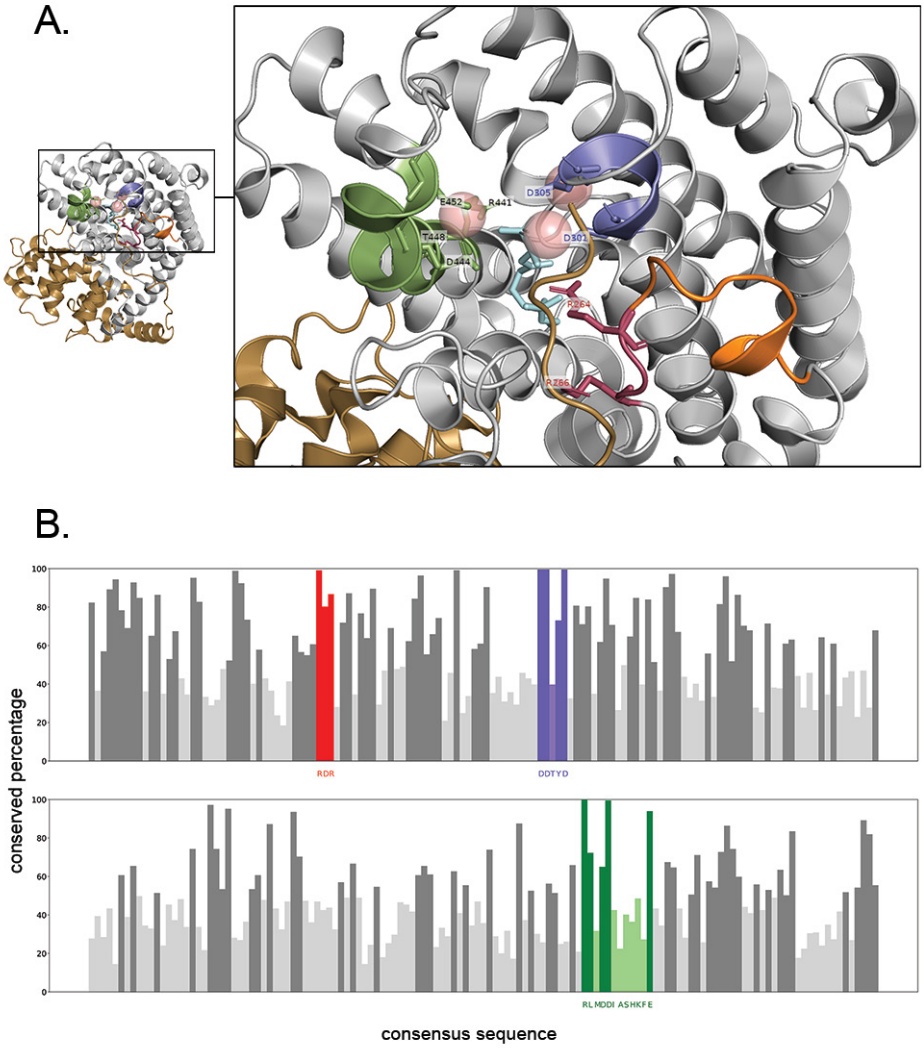



**Figure 2.4** A. Known TPS motifs RXR (red), DDXXD (purple) and NSE/DTE (green) shown on the structure of tobacco 5-epi-aristolochene synthase (PDB ID: 5EAT). The C- terminal domain is in gray while the N-terminal domain is in brown. Pink spheres represent Mg+2 ions. A substrate analogue, farnesylhydroxyphosphonate (FHP) is in blue. The A-C loop is coloured in orange. The two conserved Arginines in the RXR motif are shown along with the metal-binding residues in the DDXXD(DDXX(D,E)) and NSE/DTE motifs (RXX(N,D) DXX(S,T,G)XXXE). The Arginine in the expanded NSE/DTE motif is also shown and is found to be very conserved in spermatophyte TPSs. B. The same motifs shown on a schematic of the alignment of all spermatophyte C-terminal sub-sequences from the database. Each bar represents the percentage conservation of the consensus amino acid in the corresponding position of the alignment. Lighter coloured bars represent positions where the consensus amino acid is < 50% conserved.

Hence, an extended form of the motif may be more relevant for spermatophyte STSs, with the consensus RXX**(N,D)**DXX**(S,T,G)**XXXE or RXX**D**DXX**(D,E)**.

| Motif | No. of sequences |
|---|---|
| **D**DXX**T**XXX**E** | 57 |
| **D**DXX**S**XXX**E** | 55 |
| **N**DXX**S**XXX**E** | 44 |
| **D**DXX**G**XXX**E** | 25 |
| **N**DXX**T**XXX**E** | 22 |
| **N**DXX**G**XXX**E** | 16 |
| **D**DXX(**D**,**E**) | 20 |
| Other | 11 |

**Table 2.2** Division of the different versions of the second metal-binding motif among characterized spermatophyte STS sequences. Sequences with motifs not covered by either motif consensus sequence (N,D)DXX(S,T,G)XXXE or DDXX(D,E) are classified as "Other".



### 2.3.4.3 RXR motif not conserved in nerolidol synthase

The RxR motif is found about 35 amino acids upstream of the DDXXD motif, located on a flexible loop in the structure, termed the A-C loop. This loop has been shown to become ordered upon FDP binding[46]. The two Arg residues in the motif were proposed to be involved in the complexation of diphosphate after ionization of the substrate, thereby preventing nucleophilic attack on any of the carbocationic intermediates[46]. 215 of the 249 spermatophyte plant sequences have the canonical RXR motif while 18 of the remaining have an altered RXQ motif in the same region. Interestingly, these 18 enzymes all catalyze the formation of nerolidol, an acyclic sesquiterpene. This indicates that RXQ may be unable to capture diphosphate to the same extent as RXR, causing a premature quenching of an intermediate carbocation by water before cyclization has occurred[5].

### 2.3.4.4 Comparing residues involved in product specificity across sequences

Many studies have addressed the importance of specific residues located in the active site of TPSs via mutational analyses. Some of the best characterized TPSs derive from *Artemisia annua*, which is the source of many medicinal terpenes. Some of the STSs from *A. annua* have served as examples to identify residues involved in critical steps in the cyclization cascade. In this section three examples of *A. annua* STSs are described, for which residues involved in product specificity were experimentally

investigated. We use these as a case-study to illustrate how the large set of characterized STSs that we make available can potentially be used to guide such experimental investigations. These examples are:

1. Salmon *et al.* (2015)[55] tested a wide library of mutants for the β-farnesene synthase (UniProt: Q9FXY7) from *A. annua*, an STS catalyzing the formation of an acyclic product. They discovered that a single substitution, Tyr402Leu, confers to the synthase a cyclase activity, resulting in zingiberene and β-bisabolene as the most abundant products. Both these sesquiterpenes derive from cation 5. In sequences catalysing the formation of 10,1 and 11,1 cyclized products (cations 1, 2, 3 and 4), this position is highly conserved (88-100%) in the database as a Tyr, and Leu does not occur. However, STSs producing cation 5 and those producing acyclic products have relatively lower conservation in this position (70% Tyr and 53% Phe respectively) and Leu is found 14% of the time in cation 5. Thus, conservation patterns in this position are indicative of the corresponding residue's contribution to product specificity.

2. In another study, Li *et al.* (2013)[8] studied the effect of mutations on the cyclization reaction of the bisabolol synthase from *A. annua* (UniProt: M4HZ33). A possible reaction mechanism involves formation of a nerolidyl cation, followed by the formation of cation 5 by a 1,6 ring closure, and deprotonation to produce the final product bisabolol[56]. The authors identified a mutation that interfered with this 1,6 ring closure and showed that the substitution Leu399Thr changed the product specificity, to γ-humulene, derived from cation 2, a 11,1 cyclization of the farnesyl cation[8]. Interestingly, a Leu at this position is quite rare; it is present in only four sequences in the database, all four of which belong to the group of sequences producing cation 5. Instead, this position is highly conserved ( > 95%) as either a Ser or a Thr in the database.

3. Amorpha-4,11-diene is a bicyclic sesquiterpene produced from the 6,1-cyclized bisabolyl cation, cation 5 in **Table 2.1**. The cyclization reaction of the amorpha-4,11-diene synthase (ADS) involves the formation of a nerolidyl cation, followed by a bisabolyl cation, then by a 1,3-hydride shift, a 1,10-ring closure and finally deprotonation to produce the final product amorpha-4,11-diene[57,58]. Li *et al.* (2016)[59]

did a mutational analysis of the amorpha-4,11-diene synthase from *A. annua* (UniProt: Q9AR04), and showed that the residue Thr296 can cause a loss of cyclization activity when mutated. This residue is 82% conserved as either a Ser or a Thr in cyclic STSs. Importantly, in acyclic STSs the most common amino acid is a Tyr, with a conservation of 38%. Acyclic STSs even have amino acids such as Gln, Gly and Ile in this position, never seen in the cyclic STSs in the database. The variability and low conservation score indicates that changing this position in cyclic STSs away from a Ser or Thr could result in the formation of acyclic products, as shown by Li *et al.* (2016)[59].



In summary, analysis of these *A. annua* examples of residues involved in the first cyclization step in STSs indicates that conservation patterns across all the annotated enzymes are consistent with the functional roles of these residues. This suggests it would be possible to obtain residues potentially involved in product specificity from this database. Such a data-driven approach is in contrast to how these mutational studies have traditionally been guided, i.e. by comparison of two or three sequences from the same or related species. Therefore, a potential application of our database is in guiding site-directed mutagenesis studies in a way which avoids species bias and hence may reveal additional residues involved in product specificity. One such residue position obtained by studying conservation patterns has been discussed above in Section 2.3.4.3, namely the second arginine in the RXR motif. This position was found to be glutamine in most nerolidol synthases, something not seen in any of the cyclic synthases. Mutating this residue in cyclic synthases and monitoring for acyclic products, and vice versa, could confirm the residue's role in the cyclization of sesquiterpene products.

## 2.4 Conclusions

We compiled a manually curated set of experimentally characterized plant STSs along with their major products. This database is the largest centralized resource of annotated plant STSs to date and allows for thorough sequence-based analysis of these diverse enzymes. The enzymes in the database are grouped according to the carbocationic origin and cyclization of their major product. Such a division alleviates the task of functional analysis and comparison between the enzymes. Using the database we were able to extend and find variants of existing STS motifs. In addition, residues from previous mutational studies, when mapped onto the enzymes in the

database, were found to have detectable conservation patterns that differed from group to group. Such properties of residues can be extrapolated and used to guide further mutational studies. The database as a whole helps to understand the current state of STS sequence space characterization, and provides a starting point for future efforts to predict product specificity.

## Acknowledgements

This work is part of the research programme Novel Enzymes for Flavour and Fragrance with project number TTW 15043 which is financed by the Netherlands Organisation for Scientific Research (NWO).

## Appendix A

Supplementary data to this article can be found online at

https:// doi.org/10.1016/j.phytochem.2018.10.020.

## References


1. Buckingham, J. Dictionary of Natural Products, Supplement 4. *CRC Press* 11 (1997).
2. Gershenzon, J. & Dudareva, N. The function of terpene natural products in the natural world. *Nature Chemical Biology* vol. 3 408–414 (2007).
3. Schempp, F. M., Drummond, L., Buchhaupt, M. & Schrader, J. Microbial Cell Factories for the Production of Terpenoid Flavor and Fragrance Compounds. *Journal of Agricultural and Food Chemistry* vol. 66 2247–2258 (2018).
4. Tian, B., Poulter, C. D. & Jacobson, M. P. Defining the Product Chemical Space of Monoterpenoid Synthases. *PLoS Comput. Biol.* **12**, (2016).
5. Degenhardt, J., Köllner, T. G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70**, 1621–1637 (2009).
6. Ogura, K., Koyama, T. & Sagami, H. Polyprenyl diphosphate synthases. *Sub-cellular biochemistry* vol. 28 57–87 (1997).
7. Gao, Y., Honzatko, R. B. & Peters, R. J. Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat. Prod. Rep.* **29**, 1153–1175 (2012).
8. Li, J.-X. X. *et al.* Rational engineering of plasticity residues of sesquiterpene synthases from Artemisia annua: product specificity and catalytic efficiency. *Biochem. J.* **451**, 417–426 (2013).
9. Martin, D. M. *et al.* Functional annotation, genome organization and phylogeny of the grapevine (Vitis vinifera) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* **10**, 226 (2010).
10. Martin, D. M., Fäldt, J. & Bohlmann, J. Functional characterization of nine Norway spruce TPS genes and evolution of gymnosperm terpene synthases of the TPS-d subfamily. *Plant Physiol.* **135**, 1908–1927 (2004).
11. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
12. Consortium, U. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158--D169 (2016).
13. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138--D141 (2004).
14. Wheeler, D. L. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **35**, D5--D12 (2006).
15. Todsen, W. L. ChemDoodle 6.0. *J. Chem. Inf. Model.* **54**, 2391–2393 (2014).
16. Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. PubChem: integrated platform of small molecules and biological activities. in *Annual reports in computational chemistry* vol. 4 217–241 (Elsevier, 2008).
17. Kim, S., Thiessen, P. A., Bolton, E. E. & Bryant, S. H. PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res.* **43**, W605--W611 (2015).
18. Landrum, G. & others. RDKit: Open-source cheminformatics. (2006).
19. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
20. Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).
21. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
22. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. \& Eng.* **9**, 90–95 (2007).
23. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
24. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
25. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

26. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
27. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
28. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* vol. 31 365–370 (2003).
29. Kersey, P. J. *et al.* Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **46**, D802--D808 (2017).
30. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17 (2014).
31. Wilbur, W. J. & Lipman, D. J. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci.* **80**, 726–730 (1983).
32. Jones, E., Oliphant, T., Peterson, P. & others. {SciPy}: Open source scientific tools for {Python}.
33. Waskom, M. *et al.* mwaskom/seaborn: v0.8.1 (September 2017). (2017) doi:10.5281/zenodo.883859.
34. Bernstein, F. C. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542 (1977).
35. DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* **40**, 82–92 (2002).
36. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
37. Christianson, D. W. Structural and chemical biology of terpenoid cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).
38. Köllner, T. G., Gershenzon, J. & Degenhardt, J. Molecular and biochemical evolution of maize terpene synthase 10, an enzyme of indirect defense. *Phytochemistry* **70**, 1139–1145 (2009).
39. Webb, E. C. & others. *Enzyme nomenclature 1992. Recommendations of the nomenclature committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* (Academic Press, 1992).
40. Jia, Q., Köllner, T. G., Gershenzon, J. & Chen, F. MTPSLs: New Terpene Synthases in Nonseed Plants. *Trends Plant Sci.* **23**, 121–128 (2018).
41. Cao, R. *et al.* Diterpene cyclases and the nature of the isoprene fold. *Proteins Struct. Funct. Bioinforma.* **78**, 2417–2432 (2010).
42. Joly, A. & Edwards, P. A. Effect of site-directed mutagenesis of conserved aspartate and arginine residues upon farnesyl diphosphate synthase activity. *J. Biol. Chem.* **268**, 26983–26989 (1993).
43. Greenhagen, B. T., O'Maille, P. E., Noel, J. P. & Chappell, J. Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9826–9831 (2006).
44. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011).
45. Trapp, S. C. & Croteau, R. B. Genomic organization of plant terpene synthases and molecular evolutionary implications. *Genetics* **158**, 811–832 (2001).
46. Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi- aristolochene synthase. *Science (80-. ).* **277**, 1815–1820 (1997).
47. Aaron, J. A. & Christianson, D. W. Trinuclear metal clusters in catalysis by terpenoid synthases. *Pure Appl. Chem.* **82**, 1585–1597 (2010).
48. Prosser, I. *et al.* Enantiospecific (+)-and (-)-germacrene D synthases, cloned from goldenrod, reveal a functionally active variant of the universal isoprenoid-biosynthesis aspartate-rich motif. *Arch. Biochem. Biophys.* **432**, 136–144 (2004).
49. Son, Y.-J., Kwon, M., Ro, D.-K. & Kim, S.-U. Enantioselective microbial synthesis of the indigenous natural product (-)-α-bisabolol by a sesquiterpene synthase from chamomile (Matricaria recutita). *Biochem. J.* **463**, 239–248 (2014).
50. Christianson, D. W. Structural Biology and Chemistry of the Terpenoid Cyclases. *Chem. Rev.* **106**, 3412–3442 (2006).

51. Zhou, K. & Peters, R. J. Investigating the conservation pattern of a putative second terpene synthase divalent metal binding motif in plants. *Phytochemistry* **70**, 366–369 (2009).
52. Gennadios, H. A. *et al.* Crystal structure of (+)-δ-cadinene synthase from Gossypium arboreum and evolutionary divergence of metal binding motifs for catalysis. *Biochemistry* **48**, 6175–6183 (2009).
53. Steele, C. L., Crock, J., Bohlmann, J. & Croteau, R. Sesquiterpene synthases from grand fir (Abies grandis) comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of δ-selinene synthase and γ-humulene synthase. *J. Biol. Chem.* **273**, 2078–2089 (1998).
54. Little, D. B. & Croteau, R. B. Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases δ-selinene synthase and γ-humulene synthase. *Arch. Biochem. Biophys.* **402**, 120–135 (2002).
55. Salmon, M. *et al.* Emergence of terpene cyclization in Artemisia annua. *Nat. Commun.* **6**, 6143 (2015).
56. Benedict, C. R. *et al.* The cyclization of farnesyl diphosphate and nerolidyl diphosphate by a purified recombinant δ-cadinene synthase. *Plant Physiol.* **125**, 1754–1765 (2001).
57. Picaud, S. *et al.* Amorpha-4, 11-diene synthase: mechanism and stereochemistry of the enzymatic cyclization of farnesyl diphosphate. *Arch. Biochem. Biophys.* **448**, 150–155 (2006).
58. Kim, S.-H. *et al.* Cyclization mechanism of amorpha-4, 11-diene synthase, a key enzyme in artemisinin biosynthesis. *J. Nat. Prod.* **69**, 758–762 (2006).
59. Li, Z. *et al.* The T296V Mutant of Amorpha-4, 11-diene Synthase Is Defective in Allylic Diphosphate Isomerization but Retains the Ability To Cyclize the Intermediate (3 R)-Nerolidyl Diphosphate to Amorpha-4, 11-diene. *Biochemistry* **55**, 6599–6604 (2016).

# 3

# The santalene synthase from *Cinnamomum camphora*: Reconstruction of a sesquiterpene synthase from a monoterpene synthase

Alice Di Girolamo[1]*, Janani Durairaj[2]*, Adèle van Houwelingen[3], Francel Verstappen[1], Dirk Bosch[3], Katarina Cankar[3], Harro J. Bouwmeester[4], Dick de Ridder[2], Aalt DJ. van Dijk[2,5], Jules Beekwilder[1,3]

[1] Laboratory of Plant Physiology, Department of Plant Sciences, Wageningen University, Netherlands
[2] Bioinformatics Group, Department of Plant Sciences, Wageningen University, Netherlands
[3] Bioscience, Wageningen Plant Research, Wageningen University, Netherlands
[4] Swammerdam Institute for Life Sciences, University of Amsterdam, Netherlands
[5] Biometris, Department of Plant Sciences, Wageningen University, Netherlands
* Authors contributed equally to the manuscript

Published in Archives of Biochemistry 695, 2020

3

# abstract

Plant terpene synthases (TSs) can mediate formation of a large variety of terpenes, and their diversification contributes to the specific chemical profiles of different plant species and chemotypes. Plant genomes often encode a number of related terpene synthases, which can produce very different terpenes. The relationship between TS sequence and resulting terpene product is not completely understood. In this work we describe two TSs from the Camphor tree *Cinnamomum camphora* (L.) Presl. One of these, CiCaMS, acts as a monoterpene synthase (MTS), and mediates the production of myrcene, while the other, CiCaSSy, acts as a sesquiterpene synthase (STS), and catalyses the production of α-santalene, β-santalene and trans-α-bergamotene. Interestingly, these enzymes share 97% DNA sequence identity, and differ only in 22 amino acid residues out of 553. To understand which residues are essential for the catalysis of monoterpenes resp. sesquiterpenes, a number of hybrid synthases were prepared, and supplemented by a set of single-residue variants. These were tested for their ability to produce monoterpenes and sesquiterpenes by in vivo production of sesquiterpenes in *E. coli*, and by in vitro enzyme assays. This analysis pinpointed three residues in the sequence which could mediate the change in product specificity from a monoterpene synthase to a sesquiterpene synthase. Another set of three residues defined the sesquiterpene product profile, including the ratios between sesquiterpene products.

## 3.1 Introduction

Plant species have highly specific chemical profiles, which are often determined by the presence of different terpenes and other secondary metabolites. These profiles serve the plant to function in its ecological niche, but have also been widely employed in human applications, including pharmaceuticals, nutraceuticals, food, and cosmetics[1]. In some cases, such applications lead to a high demand for the natural source of the metabolites and put increasing pressure on the conservation of the plant species from which they are derived[2–4]. An example of such an endangered species, which has been over-exploited for the extraction of their essential oils, is the Sandalwood tree. Its heartwood oil is predominantly composed of sesquiterpenes with desirable odour characteristics, and has been used for fragrances and perfumes.



Terpenes, such as found in sandalwood oil, belong to a large class of metabolites and can be distinguished by their number of carbon atoms. Monoterpenes (C10 compounds) and sesquiterpenes (C15) are synthesized by the condensation of two or three isoprene units respectively. Plants produce monoterpenes in plastids, while sesquiterpenes are generally synthesised in the cytosol. These two cellular compartments have different pathways responsible for the production of isoprene units. In the cytosol of plants, the mevalonate (MVA) pathway is active, while the 2-C-methyl-D-erythriol-4-phosphate (MEP) pathway operates in plant plastids[5–7]. Both these pathways produce isopentenyl diphosphate (IDP) and dimethylallyl diphosphate (DMADP), which are further condensed into allylic diphosphate substrates such as geranyl diphosphate (GDP) and farnesyl diphosphate (FDP)[8,9]. Subsequently, terpene synthases (TSs) convert these substrates to monoterpenes or sesquiterpenes, respectively. Precursors for terpenes such as IDP, DMADP, FDP and GDP can be efficiently produced in industrial microorganisms, using similar pathways[10]. Formation of terpenoids from these precursors can also be achieved in microbes, and relies on TSs, often derived from plants[11,12].

A wide variety of synthases have been described for mono- and sesquiterpenes[6,13]. Plant genomes often encode 10-50 different synthases, which determine the terpene profile in the essential oil of the species, and often different chemotypes arise by diversification of terpene synthases. However, it is still challenging to predict the specific terpene produced by a synthase, and the roles of the residues present in the catalytic pocket of the synthase and their involvement in determining the product profile.

Overall sequence identity is higher between enzymes with different product specificity, belonging to the same species, compared to enzymes from different species producing the same compound[1]. A better understanding of the relationship between individual residues the primary sequence of a terpene synthase and its product profile will facilitate a prediction of uncharacterized synthases[14].

In this study, we set out to isolate a santalene synthase. Sandalwood oil has high value for perfumery and is traditionally extracted by steam distillation of *Santalum album* trees older than 15 years[15]. *S. album* has been listed as a vulnerable species in the IUCN Red list of Threatened species[16] and the use of this tree for sandalwood oil extraction has been strictly regulated. The four main compounds present in sandalwood oil are α-, β-, and epi-β-santalol and α-bergamotol, which are the hydroxylated analogues of α-, β-, and epi-β-santalene and α-bergamotene respectively. Hydroxylation of these compounds to their alcohols is mediated by cytochrome P450s[17]. The terpene synthase enzymes responsible for the production of santalenes have been identified in two *Santalum* species: *S. album* and in *S. spicatum*[18–20]. In the current work we isolate a santalene synthase from a completely unrelated tree, *Cinnamomum camphora*. One of the chemotypes of this tree also produces santalenes in its essential oil[21]. The gene encoding the santalene synthase was isolated together with a highly related gene encoding a monoterpene synthase. These two enzymes, displaying different substrate and product specificity despite their high sequence identity, were used to study the role of individual residues in determining substrate and product specificity.



We demonstrate how few residue positions are responsible for substrate specificity, allowing a MTS to acquire STS activity, without losing its original function. With the results, new insights on functional residues were obtained, contributing to the larger framework of TS substrate and product specificity prediction.

## 3.2 Methods

### 3.2.1 Identification of santalenes in *C. camphora*

A *C. camphora* plant of the cineole type was purchased from Planfor (France). Leaves, stems and roots were dissected, and 0.5 g of plant material was weighted in a pre-cooled glass tube and suspended in 2 mL dichloromethane. The samples were vortexed for 1 min, sonicated for 5 min in an ultrasonic bath and centrifuged at 1500 g at room temperature

to separate the plant material from the supernatant. 1 g $Na_2SO_4$ columns were used to dry the obtained supernatant. About 2 µL was analysed by gas chromatography mass spectrometry (GC-MS) as previously described [22,23]. All compounds were identified using the mass spectra deposited in the NIST library and confirmed using their retention index, or by comparison to an original standard, when available (**Figure 3.1**). Santalenes were further confirmed by comparison of retention times and mass spectra of a sandalwood oil standard (Merck, Germany).

### 3.2.2 RNA extraction from root tissue

To extract RNA from the root material of *C. camphora*, an extraction buffer was prepared (2% hexadecyl-trimethylammonium bromide. 2% polyvinylpyrrolidinone K 30, 100 mM Tris-HCl (pH 8.0), 25 mM EDTA, 2.0 M NaCl, 0.5 g/L spermidine and 2% β-mercaptoethanol). 3 g of ground tissue was mixed to 15 mL of pre-warmed (65°C) extraction buffer. The mixture was extracted twice with an equal volume of chloroform:isoamylalcohol (1:24), and ¼ volume of 10 M LiCl was mixed to the supernatant. The RNA was precipitated overnight at 4 °C and harvested by centrifugation at 10000 g for 20 min. The pellet was dissolved in 500 µL of STE buffer [1.0 M NaCl, 0.5% SDS, 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0)] and extracted once with an equal volume of chloroform: isoamylalcohol. Two volumes of ethanol were added to the supernatant, incubated for at least 2 h at -20 °C, centrifuged at 13000 rpm and the supernatant removed. The pellet was air-dried and resuspended in water. Total RNA (60 µg) was shIpped to Vertis Biotechnology AG (Freising, Germany). PolyA+ RNA was isolated, and cDNA was synthesized using a randomized N6 adapter primer and M-MLV H-reverse transcriptase (Sigma). TS sequences were identified by a cDNA sequencing approach, as described in detail by Beekwilder *et al.* (2014)[14].

### 3.2.3 Isolation of CiCaMS and CiCaSSy

Full length open reading frames of putative TSs were amplified from the cDNA of *C. camphora*. Specific primers (CCH_TS23_fw and CCH_TS23_re, see **Table S3.1**) were designed to amplify total ORFs with a 6-His tag fused at the N-terminus in the plasmid pCDF-Duet1 (Novagen, Merck Chemicals B.V., Amsterdam, the Netherlands). Two variants, namely pCDF-CiCaMS or pCDF-CiCaSSy, were cloned using the same primer pair using BamHI and NotI restriction enzymes. Amplification of cDNA ends (5′ RACE; Clontech cDNA RACE kit) experiments were performed to isolate longer versions of both genes, but no cDNA ends with more upstream start codons

could be identified. Sequences were deposited in Genbank under accession numbers MN756611 (CiCaMS) and MN756612 (CiCaSSy).

### 3.2.4 Cloning of CiCaMS/CiCaSSy hybrids and single mutants

To obtain the hybrid proteins of the parental enzymes CiCaMS and CiCaSSy, a library of fragments was designed. When possible, separate fragments were amplified using pCDF-CiCaMS or pCDF-CiCaSSy as template (minimum fragment length was 150bp). Each region was designed to contain 3-6 amino acid substitutions. In **Figure 3.3A** a schematic representation of the design is shown. **Table S3.2** reports the fragment composition of each hybrid, while the primer used for amplification and sequencing are listed in **Table S3.1A**. All primers were supplied by IDT (Leuven, Belgium). For fragment amplification, including the vector backbone (derived from Novagen commercial plasmid pACYC-Duet1), Q5 High Fidelity polymerase by NEB was used, following the protocol provided by the supplier. For hybrids assembly, the Circular Polymerase Extension Cloning (CPEC) method was used[24]. For fragments smaller than 150bp and for single mutants, the QuickChange site-directed mutagenesis protocol described by Xia *et al.*[25] was adapted for the use of Q5 HiFi polymerase, as described in the NEB protocol (Q5 site-directed mutagenesis kit protocol, E0554). All single mutants are listed in **Table S3.3**, including the primers used. Before transformation in DH5a, the products obtained from QuickChange were digested with DpNI to eliminate the traces of template. All constructs were confirmed by sequencing from Macrogen. All plasmids used in this study are listed in **Table S3.1B**.

### 3.2.5 Heterologous expression of CiCaMS and CiCaSSy in *E. coli* BL21DE3

To analyse the sesquiterpene product profile of the enzymes, an *E. coli* expression strain BL21DE3 containing an additional plasmid expressing all genes necessary for the synthesis of FDP (pBbA5c-MevT-MBIS-NPtll) was used. This plasmid is a variant of plasmid pBbA5c-MevT(CO)-MBIS(CO, IspA)[10,26] in which the chloramphenicol resistance marker has been exchanged for a kanamycin resistance marker (Nptll). Another variant, with a different origin of replication (colE1) was also used in the experiments (pBbE5k-MevT(CO)-MBIS(CO)). Fermentations were performed using 20 mL of 2xYT medium (16 g/L tryptone, 10 g/L NaCl, 10 g/L yeast extract) in 100 mL glass flasks. Overnight cultures were diluted to an OD600 of 0.150 and incubated at 37°C 250 rpm until an OD600 of 0.4-0.5 was reached, then IPTG 1 mM and 2 mL dodecane were added,

followed by 24 hours incubation at 28°C 250 rmp. A concentration of 50 μg/ml kanamycin and 50 μg/ml chloramphenicol was used to maintain the plasmids in the system. The 2 mL dodecane was then recovered for GCMS analysis by centrifugation at 3600 rpm for 15 min. For the GC-MS analysis, 20-80 mg dodecane were weighted and diluted in 2 mL ethyl acetate. This solution was dried over a $Na_2SO_4$ column before analysis.

To confirm the results obtained with the fermentation analysis and to assess the monoterpene activity of the enzymes, in vitro enzyme assays were also performed. The BL21DE3 *E. coli* expression strain was used for protein production. Overnight cultures were diluted to an $OD_{600}$ of 0.150 in 20 mL 2xYT and incubated for at 37°C 250 rpm until an OD600 of 0.6-0.8 was reached. A concentration of 1 mM IPTG was added, and the cultures were grown at 18°C 250 rpm overnight. Cells were then harvested by centrifugation (10 min 3600 rpm), medium was removed, and cells were resuspended in 1 mL Resuspension buffer (50 mM Tris-HCl pH = 7.5, 1.4 mM β-mercaptoethanol; 4°C). Cells were disrupted by shaking 2 times for 10 seconds with 0.2 g zirconium sand in a Fastprep machine at speed 6.5. Insoluble particles were subsequently removed by centrifugation (10 min 13000 rpm, 4°C). Soluble protein was immediately used for enzyme assays or stored in a 12.5% glycerol solution.



For enzyme assays, both farnesyl diphosphate and geranyl diphosphate (10 mM, Sigma FDP dry-evaporated and dissolved in 50% ethanol) were used as substrates. In a glass tube, a mix was made of 800 μL of MOPSO buffer (15 mM MOPSO (3-[N-morpholino]-2-hydroxypropane sulphonic acid) pH = 7.0, 12.5% glycerol, 1 mM $MgC1_2$, 0.1% tween 20, 1 mM ascorbic acid, 1 mM dithiothreitol). 100 μL of crude enzyme extract and 5 μL of FDP or GDP and 20 μL Na-orthovanadate 250 mM. 1 mL pentane was added to the mix to extract the terpenes. This mix was incubated at 30°C with mild agitation for 2-4 hours. Subsequently, the mix was centrifuged at 1200 xg to recover the pentane, which was dried over a $Na_2SO_4$ column and analysed by GC-MS.

### 3.2.6 GCMS analysis

The GC-MS analysis was performed on an Agilent Technologies system, comprising a 7980A GC system, a 597C inert MSD detector (70 eV), a 7683 auto-sampler and injector and a Phenomenex Zebron ZB-5ms column of 30m length x 0.25 mm internal diameter and 0.25 μιη stationary phase, with a Guardian precolumn (5 m). In this system, 1 μL of the sample was

injected. The injection chamber was at 250°C, the injection was splitless, and the ZB5 column was maintained at 45°C for 2 minutes after which a gradient of 10°C per minute was started, until 300°C. Peaks were detected in chromatograms of the total ion count. Compounds were identified by their retention index and by their mass spectrum in combination with comparison of the mass spectrum to libraries (NIST and in-house). The data obtained with the GCMS analysis were processed in order to obtain the average relative concentration of products produced, with the summed area of all peaks representing 100%, and the function "st.dev.s." of Excel was used to calculate the standard deviation. Average relative concentrations and standard deviation are reported in **Table S3.4**.

For concentrations, dodecane samples from three independent cultures were diluted 1:10 in acetone, and analysed by GCMS, using split 10 injection. Concentrations were calculated by comparison of peak areas of selected ions (m/z 69, 93, 94, 119, 122, 204) in samples and in standard curves of santalene oil ingredients (kindly provided by Celina Vossen). For total sesquiterpene concentrations, values for $\alpha$ santalene, $\beta$ santalene and trans $\alpha$ bergamotene were added.



### 3.2.7 Expression of CiCaMS and CiCaSSy in plants

Transient expression in *Nicotiana benthamiana* was employed to determine the product spectrum of the enzymes in a plant heterologous system. CiCaSSy and CiCaMS coding regions were amplified from the pACYC-constructs, using primers CCHattB1-FW and CCHattb2-RE (**Table S3.1**). The genes were then cloned using Gateway technique[27], into pBINplus [28]. pBINplus was taken along as a negative control. The obtained plasmids were confirmed by sequencing and transformed to A. tumefaciens AGL0 via electroporation. For agroinfiltration, the transformed *A. tumefaciens* AGL0 cultures were grown overnight in LB medium at 28°C, 300 rpm. 200 µL of O/N culture was transferred in 10 mL LB + 10 mM MES + 40 µM acetosyringone and grew again for 16 hours 28°C, 300 rpm. The cultures were then centrifuged at 4000 rpm for 10 minutes, and the pellets were resuspended in a 10 mM $MgCl_2$ solution, at a final OD600 of 1.5. Acetosyringone was added at a final concentration of 200 µM. The suspension was left at room temperature with no shaking for at least 3 hours before performing the agroinfiltration. Young leaves from 3-4 weeks old *N. benthamiana* were selected for agroinfiltration.

The experiment was performed using biological and technical replicates. Each leaf was infiltrated with about 1mL A. tumefaciens suspension. Trapping of headspace volatiles was performed as described[23] with following modifications: headspace sampling was performed in a climate room (20±2°C, 56% RH) with LED lighting (adjusted at 100% white, 10% deep red, 100% far red and 5% blue light). Volatiles were trapped by sucking air out of the jar at a rate of 100 mln/min (inlet flow at 150 mln/min) for 4h.Trapped headspace volatiles were analysed using a Thermo TraceGC Ultra connected to a Thermo TraceDSQ quadrupole mass spectrometer (Thermo Fisher Scientific, Waltham, USA). Settings as described[22], with the following modification: volatiles were injected on the analytical column at split ratio 300. Products were identified using original standards (myrcene standard and sandalwood oil, Sigma, Amsterdam, Netherlands), according to their retention time and mass spectra.

### 3.2.8 SDS-PAGE gel preparation



SDS-Polyacrylamide Gel Electrophoresis was used for protein visualization. The protocol was adapted from Sambrook and Russell, (2006)[29]. 10% resolving SDS gels were prepared using 4 mL water, 3 mL 30% acrylamide mix (Bio-Rad), 2.5 mL Tris-HCl (1.5 M, pH 8.8), 0.1 mL 10% SDS, 0.1 10% ammonium persulfate and 4 μL TEMED (Bio-Rad) for a total of 10 mL solution. Stacking gels were prepared as 5% SDS, using 0.68 mL water, 0.17 mL 30% acrylamide mix, 0.13 mL Tris-HCl (1 M, pH 6.8), 10 μL 10% SDS, 10 μL 10% ammonium persulfate and 1 μL TEMED (Bio-Rad) for a total of 1 mL solution. The protein content in the crude extracts were quantified using Bradford reagent and loaded in equal amount in 5X SDS-gel loading buffer. Precision Plus Protein (Bio-Rad) marker was used and the gels were stained in a 30% ethanol, 8% acetic acid and 0.05% (w/v) coomassie brillant blue for 2-6 hours. Destaining was performed in water overnight.

### 3.2.9 Modelling of a 3D structure for CiCaSSy

A homology model of CiCaSSy was created using multi-template modelling. The templates used were the *Hyoscyamus muticus* premnaspirodiene synthase (PDB ID: 5JO7), *Mentha spicata* limonene synthase (PDB ID: 2ONH), and *Citrus sinensis* limonene synthase (PDB ID: 5UV2); these were selected based on their high sequence similarity with the two *Cinnamonum* synthases. MODELLER[30] was used to create 500 models using the default automodel approach, and the model with the best N-DOPE score was chosen for further analysis. Furthermore, the position of an analogue of FDP, trifluorofarnesyl diphosphate (FFF) in the model was obtained by

superposing the 5-epi-aristolochene synthase from *Nicotiana tabacum* (PBD ID: 5EAU), using the align command of PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. Residues differing between the two *Cinnamonum* synthases were visualized using PyMOL. To address the quality of the model, the same command was also used to align the crystal structure of *Santalum album* santalene synthase (SaSSy, PDB ID:5ZZJ) to the modelled CiCaSSy. The structure alignment of CiCaSSy with SaSSy reveals a high structure similarity between the two enzymes, despite their low sequence identity. The CiCaSSy model can be superposed to SaSSy with a root-mean-square deviation (RMSD) of 1.09 angstrom over 420 residues.

### 3.2.10 Phylogenetic tree construction

A phylogenetic tree was constructed to identify the TS subfamily in which CiCaMS and CiCaSSy belong. 268 STSs from our previously assembled database[13], along with 59 MTSs, and 27 diTSs from SwissPROT[31], were clustered into groups of up to 70% sequence identity using CD-HIT[32]. Representative sequences were taken from each cluster and aligned along with CiCaMS and CiCaSSy using Clustal Omega[33] with the Pfam[34] domains Terpene synth (Pfam ID: PF01397) and Terpene_synth_C (Pfam ID: PF03936) as guides for the alignment. The alignment was preprocessed with trimAl[35] such that columns with over 50% gaps were discarded. The tree was constructed using the ETE3 python library[36] with the pmodeltest-ultrafast and RAxML[37] options, and visualized using iTOL[38]. The TS subfamily assignment was done as in Chen *et al.*[39]. The phylogenetic tree is represented in **Figure S3.4**.

## 3.3 Results

### 3.3.1 Isolation and characterization of CiCaMS and CiCaSSy

The cineole chemotype of *C. camphora* has been observed to contain santalenes [40,41], while other chemotypes (camphor, linalool) have not been reported to contain santalenes[21,42,43]. Extracts from different parts of a plant from the cineole chemotype were analyzed by GC-MS. Roots, leaves and stem of *C. camphora* appeared to contain compounds that correspond to α-santalene, trans-α-bergamotene, and β-santalene, among other compounds, as shown in **Figure 3.1**. The concentration of santalenes was highest in the roots, therefore this tissue was further selected for RNA extraction and cDNA sequencing. Among the root cDNAs found to correspond to TSs, one sequence was identified as a putative santalene synthase. When this synthase-encoding sequence was amplified from root cDNA, two different sequences were cloned, which were 97% identical at the DNA level, and encoded proteins which differed in 22 out of 553 amino acids (95% identity).



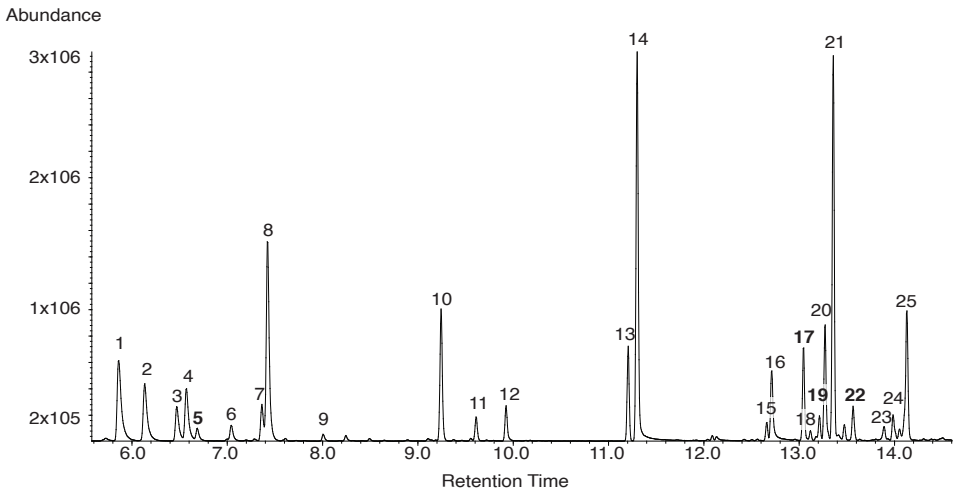



**Figure 3.1** GC-MS analysis of pentane extract of *Cinnamomum camphora* root tissue. The compounds were identified by their mass spectra using the NIST library and confirmed comparing their retention indexes with the reference list provided by Adams (1995). The y-axis reports the GC-MS response units, while the x-axis reports the retention times. The compounds identified for each peak are listed. 1. α-pinene; 2. camphene; 3. sabinene; 4. β-pinene; **5. myrcene**; 6. α-pinene; 7. limonene; 8. cineole; 9. cis-β-terpineol; 10. camphor; 11. borneol; 12. α-terpineol; 13. bornyl acetate; 14. safrene; 5. α-cubebene; 16. β-elemene; **17. α-santalene**; 18. trans-β-caryophyllene; **19. trans-α-bergamotene**; 20. α-guaiene; 21. guaia-6,9-diene ; **22. β-santalene**; 23. germacrene D; 24. β-selinene; 25. α-bulnesene.

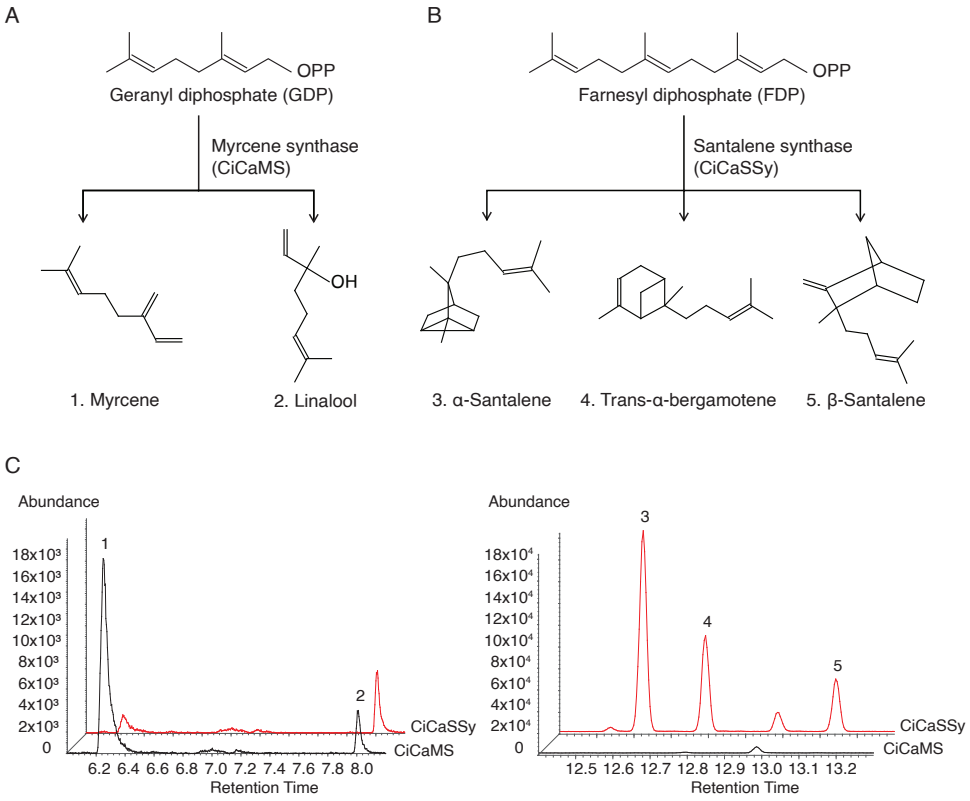



**Figure 3.2** (A) Conversion of GDP into 1. myrcene and 2. linalool by CiCaMS and (B) conversion of FDP into 3. α-santalene, 4. trans-α-bergamotene and 5. β-santalene by CiCaSSy. (C) Chromatograms of *in vitro* enzyme assay GC-MS analysis. The graphic on the left shows the production using GDP as substrate for CiCaMS and CiCaSSy. Products were identified as 1. myrcene (K.I. 991) and 2. linalool (K.I. 1098). The graphic on the right shows the production using FDP as substrate for CiCaMS and CiCaSSy. No sesquiterpene activity was detected for CiCaMS, while products of CiCaSSy were identified as 3. α-santalene (K.I. 1420); 4. trans-α-bergamotene (K.I. 1436) and 5. β-santalene (K.I. 1462). The chromatograms are obtained by extracting ion 93 from the total ion count.

The enzyme activity of both variants was investigated *in vitro*, by testing product formation using GDP or FDP as substrate. One variant converted FDP to santalenes and was referred to as *Cinnamomum camphora* Santalene Synthase (CiCaSSy). Products were identified as α-santalene, trans-α-bergamotene, and β-santalene, by their mass spectra and by comparison to a sandalwood oil standard. The other variant produced a monoterpene when GDP was used as a substrate and was referred to as CiCaMS. The product of CiCaMS was identified as myrcene, by comparison to a myrcene standard and using the retention time and the retention index[44] (**Figure 3.2**). The product profile of the two synthases was confirmed by overexpression in a plant system. For expression in *Nicotiana benthamiana*, both CiCaSSy and CiCaMS full length coding regions were cloned into binary vectors and the effect of their transient expression on the headspace of *N. benthamiana* was investigated (**Figure S3.1**). Expression of CiCaSSy led to the presence of α-santalene, trans-α-bergamotene, and β-santalene in the headspace, while no monoterpene formation was observed (**Figure S3.2E**). Transient expression of CiCaMS led to emission of myrcene (**Figure S3.2C**), while no sesquiterpene formation could be observed. Standards were used for identification of the compounds (**Figure S3.2B** and **D**) Therefore, CiCaSSy was identified as a santalene synthase, while CiCaMS was identified as a myrcene synthase.



A BLAST analysis in the non-redundant database[45] revealed that CiCaSSy was very closely related (92% identical) to a predicted (-)-α-terpineol synthase from *Cinnamomum micranthum*[46]. The closest characterized synthases include an (-)-α-terpineol synthase from *Magnolia grandiflora*[47], with 46% identity, and an (-)-α-terpineol synthases from *Vitis viniferae* [48] with 43%. Sequence identity to STSs such as *Santalum album* and *Santalum spicatum* santalene synthase (SaSSy and SspSSy) was low (37.8% and 38%; [20]) (**Figure S3.3**). Thus, it appeared that CiCaSSy was most closely related to MTSs, rather than STSs. An alignment of 175 synthases was generated to construct a phylogenetic tree (**Figure S3.4**). By following the clades proposed by Chen[39] we identified CiCaMS and CiCaSSy to belong to the TS-g clade.

### 3.3.2 Activity of CiCaMS/CiCaSSy hybrids for production of sesquiterpenes in *E. coli*

Mono- and STSs in plants have a common structural fold. They consist of two domains, an N-terminal domain, which is not part of the catalytic site, and a C-terminal domain forming the hydrophobic active site cavity[12].

The C-terminal domain region contains two $Mg^{2+}$ coordination sites at its opening (labelled DDxxD and NSE/DTE in **Figure 3.3B**) and a loop containing the diphosphate binding site for the substrate (labelled RxR in **Figure 3.3B**). A structure model for CiCaSSy was generated using existing TS crystal structures as templates. **Figure 3.3** highlights the regions which display variation between CiCaSSy and CiCaMS. The genes encoding CiCaMS and CiCaSSy were divided into 6 regions based on the position of the substitutions in the linear protein sequence. Regions R1, R2 and R3 are located in the N-terminal domain of the protein. In our 3D model of CiCaSSy (**Figure 3.3B**), these three regions are coloured green, purple, and light blue respectively. The residue substitutions in these regions are distant from the substrate binding cavity, and are therefore unlikely to affect product formation. Regions R4 and R5, and to a lesser extent R6 are close to the active site.



This information was used to investigate which regions of CiCaSSy and CiCaMS play a role in substrate- and/or product specificity. Hybrids exchanging one or two regions between both enzymes were generated (**Figure 3.3A**). Initially, hybrids were tested by *in vitro* enzyme assays, using cell-free extracts, and FDP or GDP as a substrate (**Figure S3.6A-B, Figure 3.2A-B**). Wild type CiCaMS hardly produces sesquiterpenes when FDP is supplied as a substrate. CiCaSSy produces some linalool when GDP is supplied as a substrate. Interestingly, we did not observe any linalool formation by CiCaSSy when expressed in *N. benthamiana* (**Figure S3.1 B-C**).

As a first step, substrate specificity of hybrids was investigated, by determining the ratio between monoterpene and sesquiterpene products (**Figure 3.4**). While exchanging regions R1, R2, R3 or R6 did not have a strong effect on the ratio between monoterpenes and sesquiterpenes produced, exchange of region R4 or R5 appeared to have a profound effect on both sesquiterpene and monoterpene synthase, as it completely reversed the substrate specificity of the enzyme: MS_R4 and MS_R5 produce sesquiterpenes, while SSy_R4 and SSy_R5 do not. Apart from having a strong effect on substrate preference (as observed for R4 and R5 in **Figure 3.4**), exchange of these regions also had an effect on the product profile. Monoterpene profiles of hybrid enzymes were derived from in vitro assays supplying GDP as a substrate (**Figure 3.5A-B**). All CiCaSSy-derived hybrids behaved like CiCaSSy, and showed linalool formation, with traces of myrcene.

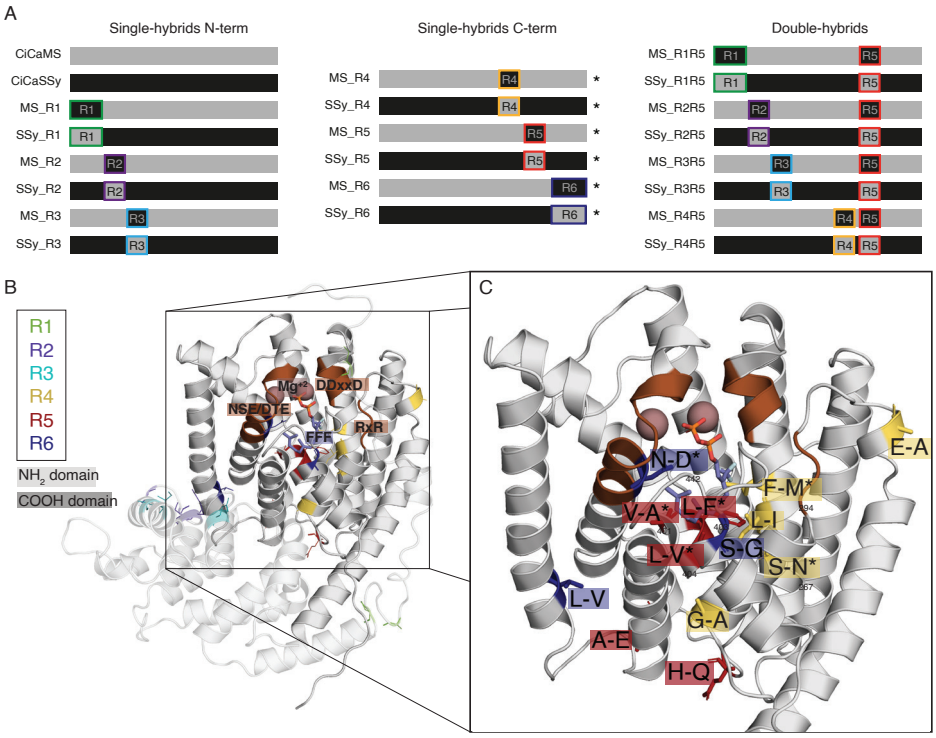



**Figure 3.3** (A) Schematic representation of hybrid design. The MTS (CiCaMS) is represented in grey, while the STS is represented in black. The region swaps are highlighted and depicted as for the parental enzyme (from R1 to R6, respectively: green, purple, light blue, orange, red, ocean blue). The asterisks indicate the regions where single residue exchanges were made. (B) 3D model of CiCaSSy, including Mg2+ ions [pink] FDP analogue, trifluorofarnesyl diphosphate (FFF) [sticks]. The model was generated with Modeller [29], using three templates, of which: one STS (5JO7) and two MTSs (2ONH and 5UV2). The different regions are highlighted in different colours as described in (A). (C) Zoom in into the active site of CiCaSSy. R4, R5 and R6 are visible. In the picture, all amino acid substitutions of the three regions are named, from CiCaMS to CiCaSSy. The asterisks point out the residues essential for the evolution of CiCaMS into a santalene synthase.

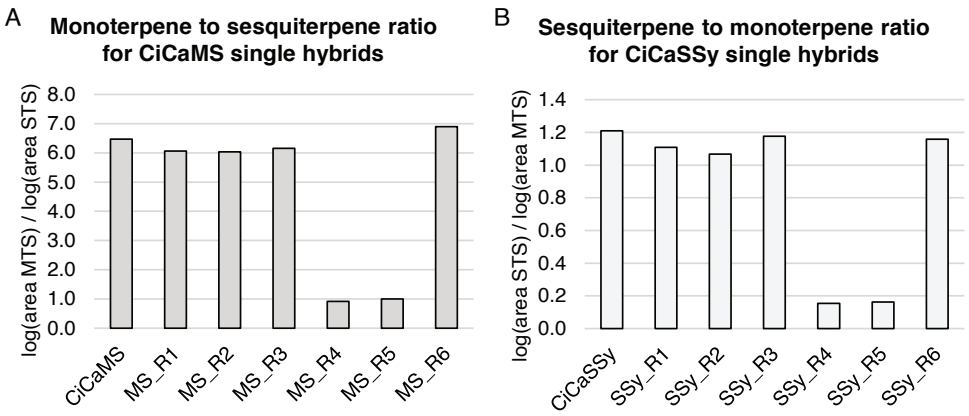



**Figure 3.4** Calculated ratio for monoterpene vs. sesquiterpene production in CiCaMS and CiCaMS-derived single hybrids (A) and sesquiterpene vs. monoterpene production in CiCaSSy and CiCaSSy-derived single-hybrids (B). Data from *in vitro* enzyme assays were used. Each enzyme was tested in duplicate. For the ratios, the sum of the areas of the principal compounds was calculated and converted in logarithmic scale. Note that, as all enzymes showed monoterpene activity in vitro, graphic (A) displays an overall higher ratio compared to graphic (B). The trend shows consistency between the two graphics.



While quantification based on enzyme assays was difficult, in general linalool peaks in CiCaSSy-derived hybrids were very low, compared to myrcene production in CiCaMS-derived hybrids (see chromatogram in **Figure 3.2C**). All CiCaMS-derived hybrids showed a predominance of myrcene production, as reported for CiCaMS. An exception was observed when both region R4 and R5 were exchanged: MS_R4R5 and SSy_R4R5 showed product profiles more similar to CiCaSSy and CiCaMS respectively (**Figure 3.5A-B**). Apparently, these regions define the identity of the formed monoterpene.

To exclude that the absence of STS activity of some of the hybrids (e.g. SSy_R5) was related to poor solubility of the hybrid proteins, an SDS PAGE protein gel analysis was performed of the cell free extracts. This did not reveal obvious differences in the amount of soluble TS protein between active and inactive synthases (**Figure S3.8**). This supports the hypothesis that hybrids were not compromised in their protein folding, and that the lack of sesquiterpene production in some of the hybrids is due to the changes in the structure of the active site of the enzyme.

The sesquiterpene profile of different hybrids was more diverse, and was eventually addressed using an *in vivo* production method. To this end, the WT parent enzymes and the hybrids were expressed in *E. coli*, in

combination with a plasmid which supplies FDP, and their performance was tested in a flask fermentation, using a dodecane overlay for products collection. After fermentation, the dodecane layer was analysed by GC-MS, and the sesquiterpene profile was extracted. In the *in vivo* system, all hybrids displayed sesquiterpene product profiles similar to those observed in the *in vitro* experiments (**Figures 5-7; Figure S3.6A-B**). No change in sesquiterpene profile was observed for hybrids covering regions R1, R2 and R3. Hybrids covering regions R4, R5 and R6 displayed marked changes in sesquiterpene profiles.

Region R4 localizes in the C-terminal domain (yellow in **Figure 3.3C**). It contains three substitutions located in the active site and two in its proximity, suggesting that its exchange could have an impact on the resultant terpene profiles. When the R4 region of CiCaSSy was introduced in MTS CiCaMS, the hybrid protein (MS_R4) displayed STS activity, producing predominantly trans-$\alpha$-bergamotene, with some β-santalene (**Figure 3.5C**). Conversely, when the R4 region of CiCaMS was introduced into CiCaSSy (SSy_R4), all sesquiterpene production was lost (**Figure 3.5D**).



Region R5 (represented in red in **Figure 3.3C**) is only 12 amino acids long but contains five substitutions, three of which are located very close to the substrate in the 3D model. Again, the exchange of R5 caused the complete loss of sesquiterpene synthase activity in CiCaSSy (SSy_R5, **Figure 3.5D**). On the other hand, introducing region R5 from CiCaSSy into CiCaMS (MS_R5), resulted in production of trans-$\alpha$-bergamotene, as was also observed for MS_R4 (**Figure 3.5C**).

Region R6 (represented in blue in **Figure 3.3C**) is located at the C-terminus of the protein. It comprises three substitutions between CiCaMS and CiCaSSy. Replacing region R6 from CiCaSSy by R6 from CiCaMS (SSy_R6) resulted in the production of all wild type sesquiterpene products, but with a relatively higher production of trans-$\alpha$-bergamotene compared to CiCaSSy (**Figure 3.5D**). Introduction of R6 from CiCaSSy into CiCaMS did not result in any sesquiterpene production (MS_R6) (**Figure 3.5C**).

A set of double hybrids was generated for CiCaMS and CiCaSSy (**Figure 3.3A**), in which regions R1 to R4 were simultaneously exchanged in combination with R5. All double hybrids carrying R1-R3 regions in combination to R5 showed the same product profile as R5 single hybrids (**Figure S3.6C**), indicating that none of these regions contribute to product

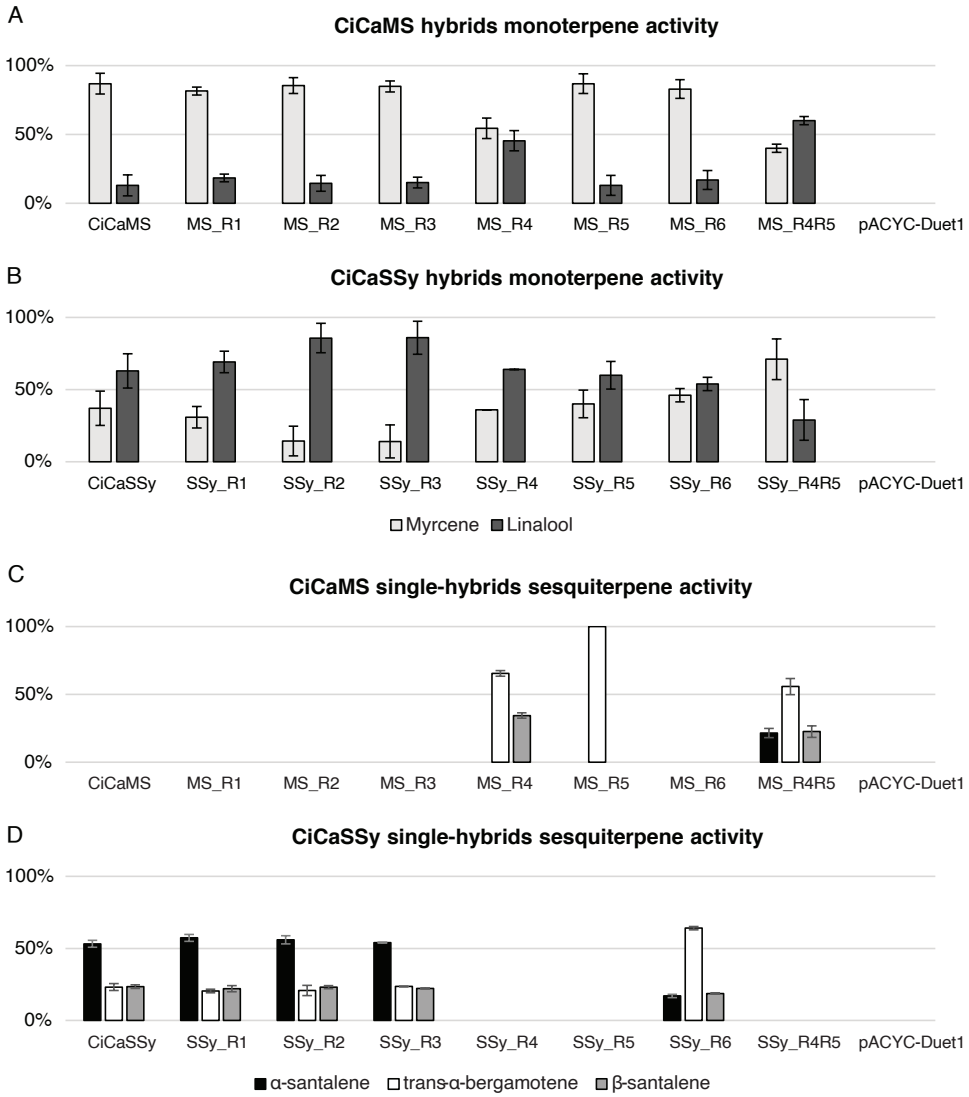



**Figure 3.5** (A-B) Monoterpene activity of relevant hybrids derived from CiCaMS (A) and CiCaSSy (B). Activity was tested in an *in vitro* system using cell-free extracts and GDP as substrate. Production of myrcene and linalool was observed for all enzymes. (C-D) Profile of produced sesquiterpenes by enzyme hybrids in an *E. co*li fermentation system. (C) Sesquiterpene production profiles mediated by CaCaMS and its derived hybrids; (D) sesquiterpene production profiles mediated by CiCaSSy and its derived hybrids. (A-D) Values on the Y axis express the relative ratio of each compound, relative to the total sesquiterpenes. Each variant was tested in three independent experiments. Error bars indicate the standard deviation.

specificity. However, the double hybrid carrying R4 and R5 from CiCaSSy in the mainframe of CiCaMS (MS_R4R5) restored the production of all three main products of CiCaSSy. The relative peak ratio of the products of MS_R4R5 was comparable to the profile produced by SSy_R6, confirming that the residues essential for restoring CiCaSSy product profiles in CiCaMS are located in R4, R5 and R6 (**Figure 3.5C-D**)

### 3.3.3 Single-residue mutants identify critical residues for sesquiterpene synthase activity

As a next step, the roles of 10 individual positions in the amino acid sequence for producing santalenes were investigated by exchanging them between CiCaSSy and CiCaMS (**TableS3.3**). In addition to the profile, total sesquiterpene production in dodecane was analysed for the most relevant mutants.



The residues were grouped, based on their position in respect to the active site cavity. From R4, residues 267, 291 and 294 appear to belong to the active site (**Figure 3.3C**). These residues were substituted in CiCaMS and CiCaSSy, obtaining MS_S267N, MS_L291I, MS_F294M from CiCaMS and SSy_N267S, SSy_I291L and SSy_M294F from CiCaSSy. When testing the product profile of these mutant enzymes only one residue appeared to be responsible for STS activity. Mutant MS_F294M showed low but well detectable production of trans-$\alpha$-bergamotene. Conversely, complementary mutant SSy_M294F had lost the ability to produce sesquiterpenes in this system. Mutation of residues N267 and I291 each resulted in a change of the product ratio in CiCaSSy (**Figure 3.6 A-B**), but no major change in total sesquiterpene production was observed. The more distant residues in R4, 273 and 308, were also probed for their role in terpene synthesis, by testing mutants MS_G273A and MS_E308D for CiCaMS and SSy_A273G and SSy_D308E for CiCaSSy. Among these, mutant SSy_A273G displayed an altered sesquiterpene profile and lower productivity, compared to CiCaSSy. SSy_D308E showed a product profile which was comparable to the wild type CiCaSSy.

The same approach was used for region R5, where positions 401, 403 and 404 participate in the active site, and positions 415 and 419 appear to be located further away, near the bottom of the cavity (**Figure 3.3C**). Among the CiCaMS mutants, production of trans-$\alpha$-bergamotene was observed for both MS_L403F and MS_L404V (**Figure 3.6C**). Apparently, either of these mutations in CiCaMS is sufficient to confer STS activity, albeit that the

product accumulation was lower than observed for CiCaSSy and hybrid MS_R5. Conversely, substitution V404L in CiCaSSy resulted in a complete loss of STS activity. CiCaSSy mutants A401V and F403L displayed altered product ratios compared to wildtype CiCaSSy, with trans-α-bergamotene being the dominant product (**Figure 3.6D**). Mutations in the two residues further away from the active site did not alter the product spectrum of CiCaSSy (SSy_Q415H, SSy_E419A), nor did they confer STS activity on CiCaMS (MS_H415Q, MS_A419E). **Figure 3.6** reports the quantification data calculated for MS_R4, MS_R5 and the most relevant single mutants, providing an overview of the efficiency of the different variants. Thus, the substitution analysis of region R4 and R5 indicates that three single amino acid positions are crucial to introduce STS activity in CiCaMS: substitutions F294M, L403F and L404V each result in STS activity of the MTS. In CiCaSSy, substitutions M294F and V404L result in a complete loss of STS activity in CiCaSSy.



Although R6 was not observed to be crucial for sesquiterpene production, its exchange resulted in a significant variation in product ratio (**Figure 3.7**). Among the R6 variant residues, residue D442 maps close to the active site of CiCaSSy and is part of its NSE/DTE motif (Figure 3.3B-C). Two mutations, MS_N442D and SSy_D442N, were tested in variant MS_R4R5 and in CiCaSSy, respectively. As shown in Figure 3.7A, hybrid MS_R4R5 with the additional single mutation N442D recovers the product ratio and the activity of the wild type CiCaSSy. Conversely, SSy_D442N results in the same product ratio as SSy_R6 and MS_R4R5 (**Figure 3.7B**). This confirms residue 442 as a major determinant of product ratio in CiCaSSy. With these results, we hypothesised that the combination of the six residues identified above (267, 294, 401, 403, 404, 442) would be sufficient to effectively establish the CiCaSSy STS profile in CiCaMS.

To prove this, we generated a CiCaMS-derived mutant carrying all these six amino acid substitutions. The obtained variant, referred to as MS_6S, was at least as active as CiCaSSy in producing sesquiterpenes and showed a product profile very similar to wild type CiCaSSy (**Figure S3.9**), with the presence of all three compounds and α-santalene as major peaks (**Figure 3.7B**).

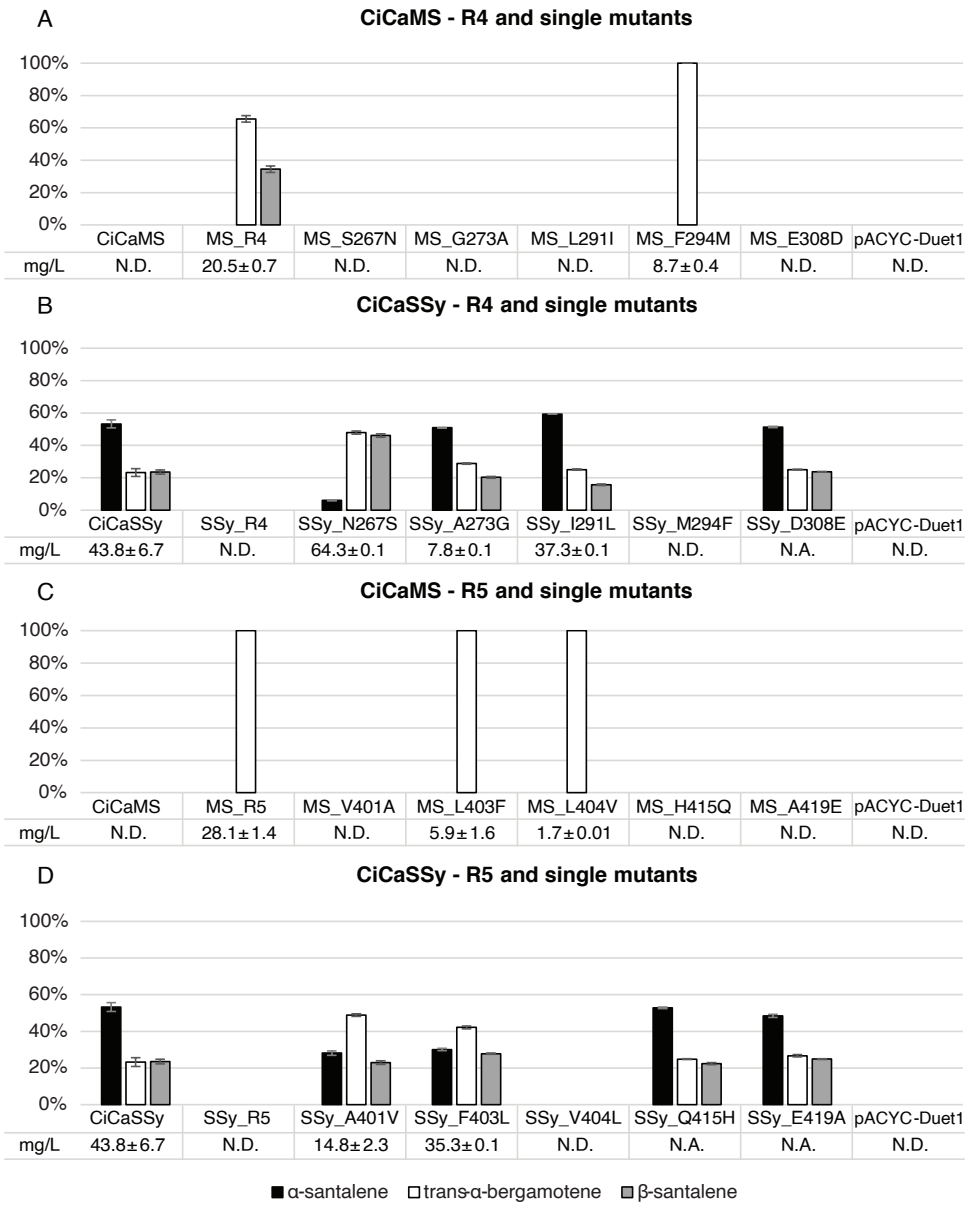



**Figure 3.6** Profile of produced sesquiterpenes by amino-acid mutants in an *E. coli* fermentation system. (A) sesquiterpene production profiles mediated by CiCaMS R4 hybrid and single mutants; (B) sesquiterpene production profiles mediated by CiCaSSy R4 hybrid and single mutants. (C) sesquiterpene production profiles mediated by CiCaMS R5 hybrid and single mutants; (D) sesquiterpene production profiles mediated by CiCaSSy R5 hybrid single mutants. Activity of WT enzymes has been included for comparison. Values on the Y axis express the relative ratio of each compound, relative to the total sesquiterpenes produced. Each variant was tested in three independent experiments. Error bars indicate the standard deviation, tables report the calculated mg/L of total sesquiterpenes produced. In the tables, N.D. stands for "not detected" while N.A. for "not analysed".

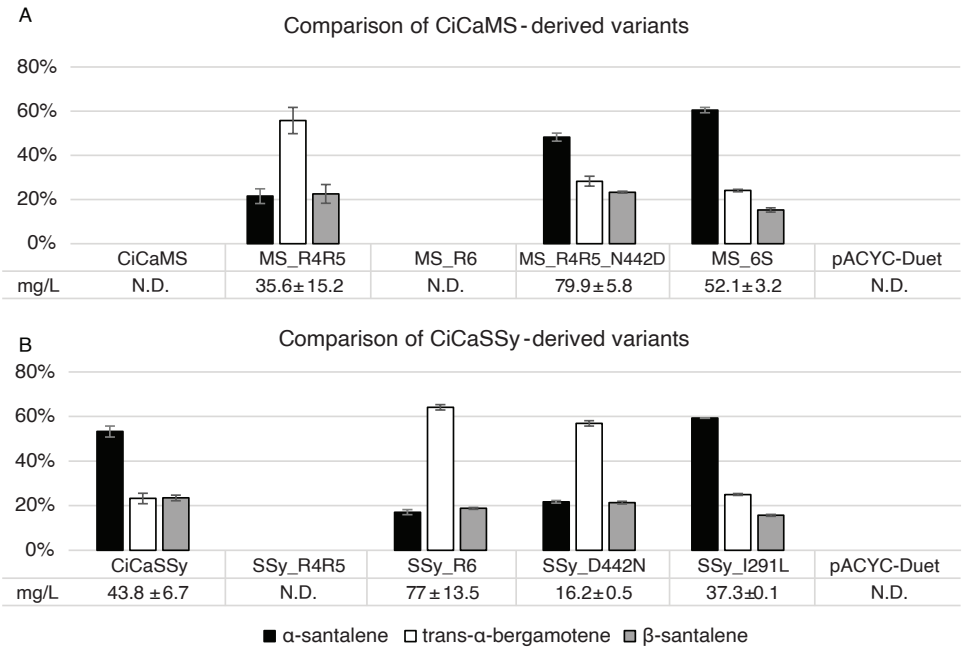



**Figure 3.7** Profile of produced sesquiterpenes by CiCaMS (A) and CiCaSSy (B) variants in an E. coli fermentation system. (A) sesquiterpene production profiles mediated by CiCaMS double hybrid MS_R4R5 in comparison with the single mutant MS_R4R5_N44D. (B) sesquiterpene production profiles mediated by SSy_R6 hybrid and R6 single mutant (SSy_D442N). In graphic (A) is also reported MS_6S, in comparison with SSy_I291L in (B). Error bars indicate standard deviation, tables report the calculated mg/L of total sesquiterpene produced. In the tables, N.D. stands for “not detected” while N.A. for “not analysed”.

## 3.4 Discussion

In our study we characterized a novel santalene synthase from *C. camphora* (CiCaSSy) which shows low similarity with the previously characterized santalene synthases from *Santalum* spp. (~38% identity [20]). We also identified a closely related MTS (CiCaMS) which does not show any sesquiterpene activity despite differing from CiCaSSy in only 22 out of 553 (95%) amino acids. Among these, three residue changes (M294F, L403F and V404L) were each able to convert the MTS into a STS. Three more substitutions (S267N, V401A and N442D) appear to be involved in defining the product profile of CiCaSSy. Thus, six residues define the specific product properties of CiCaSSy, relative to its MTS counterpart. To examine the amino acid differences between these two enzymes in a functional context, we used a structural model of CiCaSSy, depicted in **Figure 3.3B**.

Two of the most important residues addressed in this study are at position 294 and 404. Both in CiCaSSy and in CiCaMS, the identity of the residues in these positions determine STS activity. In the structure model, the sidechains of these residues point into the active site cavity, although the precise topology of the side chain cannot be accurately inferred from the structure model. For residue 294, the size properties of the Phe side chain as found in CiCaMS may hinder entry of FDP into the active site pocket, thus preventing STS activity. The role of this region in TS function has been earlier addressed by Kampranis *et al.*, (2007)[49], who showed that in the 1,8-cineole MTS from *Salvia fruticosa*, a mutation of an Asn to Ala, in a position corresponding to 291 in CiCaSSy, allowed for the enlargement of the active site cavity to accommodate the bulkier FDP substrate and induce sesquiterpene production. Both position 291 and 294 have differing amino acids in the two *Cinnamonum* synthases, but only 294 appears to affect sesquiterpene production. An alignment of relevant synthases, highlighting the residues of interest, is presented in **Figure S3.7**.

Residue 404, together with residues 401 and 403, lies around the kink in the G2 helix, which has been studied in many different contexts as being crucial for product specificity in TSs[49–52]. These previous studies reveal that mutations in these positions can lead to changes in the product specificity, consistent with our observations that substitutions in positions 401 and 403 have a strong impact on the product profile of CiCaSSy. However, a stronger effect is observed for position 404, which induces STS activity in CiCaMS (MS_L404V) and disrupts completely the STS activity in CiCaSSy (SSy_V404L). The results presented here indicate that conversion

of CiCaMS into a bergamotene synthase can be mediated by a variety of mutations which affect the shape of the active site cavity.

More subtle changes result from mutations in positions 267, 291 and 442, which predominantly have an effect on the product profile of the STS, leading to altered ratios of trans-$\alpha$-bergamotene, $\alpha$- and β-santalene. Position 267 (Ser in CiCaMS and Asn in CiCaSSy; **Figure 3.6B**) has been implicated in the second cyclization required to produce bicyclic monoterpenes such as $\alpha$-pinene[53], corresponding to the third cyclization for the production of tricyclic sesquiterpenes. Thus, this functional activity may explain the difference in product ratio of the bicyclic sesquiterpenes (trans-α-bergamotene and β-santalene) compared to the tricyclic α-santalene seen in SSy_N267S (**Figure 3.6B**). Residue 442 in R6 forms part of the catalytic NSE/DTE motif, which acts as the second $Mg^{2+}$ binding motif in TSs[54]. Our previous research[13] showed that this position is predominantly (65%) an Asp among 250 characterized plant STSs. Possibly, the prominent role of this residue in determining the product profile of CiCaSSy is related to its involvement in orienting the magnesium ion.



The fact that CiCaSSy and CiCaMS are so closely related raises the question whether the common ancestor enzyme was a STS or a MTS. Our results do not provide clear answers to this question. The highest sequence similarity of both enzymes was found with a terpineol synthase, suggesting that the santalene synthase CiCaSSy could have evolved from a more common MTS present in *Cinnamomum* spp. On the other hand, the ability of CiCaSSy to produce a suite of bicyclic and tricyclic sesquiterpenes, which require a complex cascade of proton migrations in the sesquiterpene, as opposed to the presumed more simple linear monoterpene myrcene, which requires hardly any proton transfer, could lead to consider CiCaMS as a sort of loss-of-function mutant of CiCaSSy[39,55,56]. This view would also be in agreement with the absence of a clear plastid transit peptide in both CiCaMS and CiCaSSy. However, it should be noted that CiCaMS is a functional monoterpene synthase, since it does mediate myrcene formation when expressed in *N. benthamiana.* Moreover, the chromatogram of the root tissue (**Figure 3.1**) shows a myrcene peak consistent with the activity observed for CiCaMS. Thus, different hypotheses into the evolutionary order of myrcene and santalene evolution find support.

TSs with a mixed mono/sesquiterpene product profile have been described before. Several studies describe the ability of few STS to also behave as

MTSs in the presence of GDP[57]. The santalene synthases and the bisabolene synthases from *Santalum* spp. have also been observed to produce linalool, geraniol and terpineol when supplied with GDP[20]. Other examples are the trans-$\alpha$-bergamotene synthase from *Lavandula angustifolia*[58] and the $\alpha$-bisabolene synthase from *Abies grandis*[59]. Interestingly, all these enzymes result in products which are part of a specific subclass of sesquiterpenes, which derive from the bisabolyl cation[60]. One could hypothesize that sesquiterpenes derived from the bisabolyl cation, which present a cyclised "head" and a uncyclized "tail", can be produced by synthases which are closely related to MTSs, and may have evolved from them. From this perspective, santalenes, bergamotenes and bisabolenes can also be seen as cyclized monoterpenes with an isopentenyl extension. As mentioned above, a change in residue 294, 403, or 404 seems to be sufficient to change the shape of the active site pocket of CiCaMS and to allow the accommodation of the larger FDP substrate. Thus, this hypothesis is sustained both by the ability of CiCaSSy to produce monoterpenes in vitro and, more importantly, by the demonstration that several single residue substitutions in the active site of MTS CiCaMS each are sufficient to trigger the production of sesquiterpene trans-$\alpha$-bergamotene.



## 3.5 Conclusions

With this study we characterized two novel TSs, one MTSs and one STS from *C. camphora*. Residues essential for the conversion of the MTS into a STS were identified and we effectively succeeded into converting CiCaMS in a santalene synthase showing the same product profile as CiCaSSy, by substituting only six residues. This work provides new insights into the function of specific residues and their role in the catalytic site of TSs, contributing to a better understanding of this class of enzymes.

### Acknowledgements

This work is part of the research programme Novel Enzymes for Flavour and Fragrance with project number TTW 15043 which is financed by the Netherlands Organisation for Scientific Research (NWO). Celina Vossen is acknowledged for her support with quantitative analysis of sesquiterpenes.

### Appendix A

Supplementary data to this article can be found online at:

https://doi.org/10.1016/j.abb.2020.108647

## References


1. Balandrin, M. F., Klocke, J. A., Wurtele, E. S. & Bollinger, W. H. Natural plant chemicals: Sources of industrial and medicinal materials. *Science (80-. ).* **228**, 1154–1160 (1985).
2. Alves Gomes Albertti, L. *et al.* Identification of the bisabolol synthase in the endangered candeia tree (eremanthus erythropappus (dc) mcleisch). *Front. Plant Sci.* **9**, (2018).
3. Dhanya, B., Viswanath, S. & Purushothman, S. Sandal (santalum album L.) conservation in Southern India: A review of policies and their impacts. *J. Trop. Agric.* **48**, 1–10 (2010).
4. Butaud, J.-F. *et al. Sandalwood, current state of knowledge and implications for conservation and enhancement*. https://www.researchgate.net/publication/282507135_Sandalwood_current_state_of_knowledge_and_implications_for_conservation_and_enhancement (2014).
5. Smit, A. & Mushegian, A. Biosynthesis of isoprenoids via mevalonate in archaea: The lost pathway. *Genome Res.* **10**, 1468–1484 (2000).
6. Degenhardt, J., Köllner, T. G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70**, 1621–1637 (2009).
7. Aharoni, A., Jongsma, M. A. & Bouwmeester, H. J. Volatile science? Metabolic engineering of terpenoids in plants. *Trends Plant Sci.* **10**, 594–602 (2005).
8. Liang, P. H., Ko, T. P. & Wang, A. H. J. Structure, mechanism and function of prenyltransferases. *Eur. J. Biochem.* **269**, 3339–3354 (2002).
9. Lange, B. M., Rujan, T., Martin, W. & Croteau, R. Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13172–13177 (2000).
10. Martin, V. J. J., Piteral, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.* **21**, 796–802 (2003).
11. Pateraki, I., Heskes, A. M. & Hamberger, B. Cytochromes p450 for terpene functionalisation and metabolic engineering. *Adv. Biochem. Eng. Biotechnol.* **148**, 107–139 (2015).
12. Christianson, D. W. Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).
13. 13. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).
14. Beekwilder, J. *et al.* Valencene synthase from the heartwood of Nootka cypress ( Callitropsis nootkatensis ) for biotechnological production of valencene. *Plant Biotechnol. J.* **12**, 174–182 (2014).
15. Teixeira da Silva, J. A. *et al.* Sandalwood: basic biology, tissue culture, and genetic transformation. *Planta* vol. 243 847–887 (2016).
16. Arunkumar, A. N., Dhyani, A. & Joshy, G. *Santalum album IUCN Red List Assessment. The IUCN Red List of Threatened Species* (2019).
17. Celedon, J. M. *et al.* Heartwood-specific transcriptome and metabolite signatures of tropical sandalwood (Santalum album) reveal the final step of (Z)-santalol fragrance biosynthesis. *Plant J.* **86**, 289–299 (2016).
18. Diaz-Chavez, M. L. *et al.* Biosynthesis of Sandalwood Oil: Santalum album CYP76F Cytochromes P450 Produce Santalols and Bergamotol. *PLoS One* **8**, e75053 (2013).
19. Birkbeck, A. A. The synthesis of fragrant natural products from Santalum album L.: (+)-(Z)-α-Santalol and (-)-(Z)-β-Santalol. *Chimia (Aarau).* **71**, 823–835 (2017).
20. Jones, C. G. *et al.* Sandalwood fragrance biosynthesis involves sesquiterpene synthases of both the terpene synthase (TPS)-a and TPS-b subfamilies, including santalene synthases. *J. Biol. Chem.* **286**, 17445–17454 (2011).
21. Frizzo, C. D. *et al.* Essential oils of camphor tree (cinnamomum camphora nees & eberm) cultivated in Southern Brazil. *Brazilian Arch. Biol. Technol.* **43**, 313–316 (2000).
22. Jongedijk, E. *et al.* Capturing of the monoterpene olefin limonene produced in Saccharomyces cerevisiae. *Yeast* **32**, 159–171 (2015).

23. Cankar, K. *et al.* (+)-Valencene production in Nicotiana benthamiana is increased by down-regulation of competing pathways. *Biotechnol. J.* **10**, 180–189 (2015).
24. Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS One* **4**, (2009).
25. Xia, Y., Chu, W., Qi, Q. & Xun, L. New insights into the QuikChangeTM process guide the use of Phusion DNA polymerase for site-directed mutagenesis. *Nucleic Acids Res.* **43**, e12 (2015).
26. Peralta-Yahya, P. P. *et al.* Identification and microbial production of a terpene-based advanced biofuel. *Nat. Commun.* **2**, 483 (2011).
27. Reece-Hoyes, J. S. & Walhout, A. J. M. Gateway recombinational cloning. *Cold Spring Harb. Protoc.* **2018**, 1–6 (2018).
28. van Engelen, F. A. *et al.* pBINPLUS: An improved plant transformation vector based on pBIN19. *Transgenic Res.* **4**, 288–290 (1995).
29. Sambrook, J. & Russell, D. W. SDS-Polyacrylamide Gel Electrophoresis of Proteins. *Cold Spring Harb. Protoc.* **2006**, pdb.prot4540 (2006).
30. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinforma.* **15**, 5.6.1-5.6.30 (2006).
31. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* vol. 31 365–370 (2003).
32. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
33. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, (2011).
34. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Research* vol. 42 (2014).
35. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
36. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
37. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
38. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
39. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011).
40. Stubbs, B. J., Specht, A. & Brushett, D. The essential oil of cinnamomum camphora (L.) Nees and eberm.—Variation in oil composition throughout the tree in two chemotypes from eastern Australia. *J. Essent. Oil Res.* **16**, 200–205 (2004).
41. Pélissier, Y., Marion, C., Prunac, S. & Bessière, J. M. Volatile components of leaves, stems and bark of Cinnamonum camphora Nees et Ebermaier. *J. Essent. Oil Res.* **7**, 313–315 (1995).
42. Pino, J. A. & Fuentes, V. Leaf Oil of cinnamomum camphora (L.) J. Presl. From Cuba. *J. Essent. Oil Res.* **10**, 531–532 (1998).
43. Chen, C. *et al.* Transcriptome analysis and identification of genes related to terpenoid biosynthesis in Cinnamomum camphora. *BMC Genomics* **19**, 550 (2018).
44. Adams, R. P. *Identification of Essential Oil Components by Gas Chromatography/Quadrupole Mass Spectroscopy*. (Allured Pub Corp, 1995).
45. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **47**, D23–D28 (2019).
46. Chaw, S. M. *et al.* Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* **5**, 63–73 (2019).
47. Lee, S. & Chappell, J. Biochemical and genomic characterization of terpene synthases in Magnolia grandiflora. *Plant Physiol.* **147**, 1017–1033 (2008).
48. Martin, D. M. & Bohlmann, J. Identification of Vitis vinifera (-)-α-terpineol synthase by in silico

screening of full-length cDNA ESTs and functional characterization of recombinant terpene synthase. *Phytochemistry* **65**, 1223–1229 (2004).
49. Kampranis, S. C. *et al.* Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: Structural insights into the evolution of terpene synthase function. *Plant Cell* **19**, 1994–2005 (2007).
50. Loizzi, M., González, V., Miller, D. J. & Allemann, R. K. Nucleophilic Water Capture or Proton Loss: Single Amino Acid Switch Converts δ-Cadinene Synthase into Germacradien-4-ol Synthase. *ChemBioChem* **19**, 100–105 (2018).
51. Köllner, T. G., Schnee, C., Gershenzon, J. & Degenhardt, J. The variability of sesquiterpenes emitted from two Zea mays cultivars is controlled by allelic variation of two terpene synthase genes encoding stereoselective multiple product enzymes. *Plant Cell* **16**, 1115–1131 (2004).
52. Garms, S., Chen, F., Boland, W., Gershenzon, J. & Köllner, T. G. A single amino acid determines the site of deprotonation in the active center of sesquiterpene synthases SbTPS1 and SbTPS2 from Sorghum bicolor. *Phytochemistry* **75**, 6–13 (2012).
53. Hsu, L. J. & Chu, F. H. Plasticity residues involved in secondary cyclization of terpene synthesis in Taiwania cryptomerioides. *Tree Genet. Genomes* **11**, 796 (2015).
54. Christianson, D. W. Structural Biology and Chemistry of the Terpenoid Cyclases. *Chem. Rev.* **106**, 3412–3442 (2006).
55. Bohlmann, J., Meyer-Gauen, G. & Croteau, R. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4126–4133 (1998).
56. Karunanithi, P. S. & Zerbe, P. Terpene Synthases as Metabolic Gatekeepers in the Evolution of Plant Terpenoid Chemical Diversity. *Frontiers in Plant Science* vol. 10 (2019).
57. Pazouki, L. & Niinemetst, U. Multi-substrate terpene synthases: Their occurrence and physiological significance. *Front. Plant Sci.* **7**, (2016).
58. Landmann, C. *et al.* Cloning and functional characterization of three terpene synthases from lavender (Lavandula angustifolia). *Arch. Biochem. Biophys.* **465**, 417–429 (2007).
59. Bohlmann, J., Crock, J., Jetter, R. & Croteau, R. Terpenoid-based defenses in conifers: cDNA cloning, characterization, and functional expression of wound-inducible (E)-α-bisabolene synthase from grand fir (Abies grandis). *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6756–6761 (1998).
60. IUBMB. Bisabolene Biosynthesis. https://www.qmul.ac.uk/sbcs/iubmb/enzyme/reaction/terp/bisabol.html (2011).



## Database

Sequences of the parental enzymes are available in GenBank with the IDs:

CiCaMS: MN756611

CiCaSSy: MN756612

## Enzymes

CiCaSSy, santalene synthase (EC4.2.3.81; EC4.2.3.82; EC4.2.3.83) from *Cinnamomum camphora;* CiCaMS, myrcene synthase (EC4.2.3.15) from *Cinnamomum camphora*.

# 4

# The use of Proton Transfer Reaction Mass Spectrometry for high throughput screening of terpene synthases

Alice Di Girolamo[1], Michele Pedrotti[2], Francel Verstappen[1], Alex Koot[3], Adèle van Houwelingen[4],Celina Vossen[5], Harro J. Bouwmeester[6], Dick de Ridder[7], Jules Beekwilder[5]

[1] Laboratory of Plant Physiology, Department of Plant Sciences, Wageningen University, Netherlands
[2] Post-harvest technology group, Wageningen Food & Biobased Research, Wageningen Research, Netherlands
[3] Authenticity & Nutrients group, Wageningen Food Safety Research, Wageningen Research, Netherlands
[4] Bioscience, Wageningen Plant Research, Wageningen University, Netherlands
[5] Isobionics BV, Geleen, The Netherlands
[6] Swammerdam Institute for Life Sciences, University of Amsterdam, Netherlands
[7] Bioinformatics Group, Department of Plant Sciences, Wageningen University, Netherlands

# abstract

In this work, we introduce the novel application of Proton Transfer-Reaction Mass Spectrometry (PTR-MS) for the selection of improved terpene synthase mutants as an alternative to the use of Gas Chromatography-Mass Spectrometry (GC-MS). The advantages of using PTR-MS include reduction of sample preparation steps and analysis time, providing a promising platform for the high throughput screening (HTS) of large enzyme mutant libraries. To investigate the feasibility of the method, we selected a small library of *Callitropsis nootkatensis* valencene synthase (CnVS) mutants. We investigated the robustness of the method by comparing data obtained through PTR-MS analysis and data obtained with GC-MS analysis. Here, we propose the use of 96 deep well plates for bacterial growth as a miniaturized set-up for real time, *in vivo* HTS. We hereby demonstrate the feasibility for future application of PTR-MS as *in vivo* HTS method for the rapid and reliable selection of highly productive terpene synthases.

## 4.1 Introduction

Many Volatile Organic Compounds (VOCs) produced in plants are valuable compounds for the flavour and fragrance industry. In the last decade, great effort has been invested in the development of microbial platforms able to fermentatively produce plant-derived VOCs, resulting in the successful implementation of terpene synthases for the commercial production of terpenes[1,2]. To improve such microbial-based production systems and to select the best performing enzymes, reliable screening methods are needed. Currently, the most widely used method is the Gas Chromatography Mass Spectrometry (GC-MS). This technique is considered reliable and allows both qualitative and quantitative analysis of VOCs. A disadvantage is that this method is time consuming as it requires sample preparation and long analysis times, since the chromatography takes at least a few minutes per sample[3]. Therefore, when screening for improved enzyme mutants, GC-MS can only be employed as a medium throughput screening method.

Proton-Transfer Reaction Mass Spectrometry (PTR-MS) is a widely used technique for the analysis of VOCs[4]. This technique has been developed to have high sensitivity, being able to detect VOCs in mixing ratios in the range of parts per trillion. The PTR-MS uses positively ionized small molecules, normally $H_3O^+$, as proton donors for VOC ionization. This proton transfer takes place inside a drift tube with a fixed length, which provides a fixed reaction time that allows the determination of the ion residence time. Thus, a kinetic analysis of the (proton donor)/(proton acceptor) ion signal ratio allows the calculation of the absolute concentration of the acceptor molecule[5]. Several types of mass spectrometers can be employed for the detection of the ionized VOCs, as mass-to-charge ratio ($m/z$)[6]. A recent development in PTR-MS technology is the time-of-flight mass spectrometer (ToF-MS)[7], which has an increased dynamic range and sensitivity, in comparison to other PTR-MS formats and reaches higher performance when coupled with a Quadrupole ion guide (Qi)[8,9].

The PTR-MS technique has found application in a variety of fields, from air quality control monitoring[10], food technology[11] and medical research[12]. The most compelling feature of the PTR-MS technique is the possibility to detect VOCs in real time, allowing a significant reduction in sample preparation time when compared to other analytical techniques, thereby introducing a potential for high throughput screening (HTS) of a large number of samples. In food research, several studies report the application of PTR-MS for monitoring the shelf life and storage effects on milk products[13,14]

and their volatile profiling for quality control[15], as well as monitoring the effects of different bacterial contaminants on the VOCs profile[16].

In plant metabolomic analysis, PTR-Qi-ToF-MS has been recently employed to analyse emission of isoprenoids, with a throughput of over 700 samples per day[17]. This work revealed that PTR-MS analysis allows a much faster throughput of samples when compared to different GC-MS-based techniques[18,19]. Among its advantages, the PTR-MS method does not need sample preparation, nor a chromatography step, and is only limited by the MS analysis time. Sesquiterpenes belong to the isoprenoid family and are synthesized by plants as part of their essential oils, playing an important role in plant communication[20]. Being VOCs, sesquiterpenes can be detected by PTR-MS in different types of samples. A reliable quantification of total sesquiterpene signal can be obtained by monitoring emissions of *m*/*z* 205 from freshly harvested plant samples[21]. Real time measurements permit to monitor changes in VOCs emission over a period of time, for example in changing physiological conditions, which has been employed for non-invasive analysis of plant metabolites[22]. Valencene is a sesquiterpene that has been of interest for the flavour and fragrance industry. It is a component of citrus essential oils and the precursor of nootkatone, the main grapefruit flavor[23].



In this work, we investigate the potential application of PTR-Qi-ToF-MS for screening of sesquiterpene synthase mutants, using a previously tested valencene synthase mutant library, including variants showing different efficiency. By comparing these results with those obtained by using GC-MS analysis, we were able to demonstrate that PTR-Qi-ToF-MS offers a promising alternative as a HTS method to identify improved sesquiterpene synthase mutants.

## 4.2 Materials and methods

### 4.2.1 Bacterial strains

An *Escherichia coli* strain derived from BL21DE3, harbouring a plasmid expressing all genes necessary for the synthesis of farnesyl diphosphate (FDP) (pBbE5k-MevT(CO)-MBIS (CO))[24] was used for the production of valencene. *Callitropsis nootkatensis* valencene synthase (CnVS)[25] was chosen as target for this experiment, together with a small library of single mutants (plasmids were kindly provided by Isobionics BV) (**Table 4.1**). Synthases were expressed using the inducible plasmid pACYC-Duet-1, as previously described[24].

### 4.2.2 *In vitro* enzyme assay of CnVS mutants

Single colonies of each individual transformant were inoculated in 5 mL of LB medium (10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract) with 1% glucose and 50 μg/mL chloramphenicol as antibiotic and grown overnight at 37°C. 200 μL of each culture were then diluted in 20 mL 2xYT medium (16 g/L tryptone, 10 g/L NaCl, 10 g/L yeast extract) with 50 μg/mL chloramphenicol in a 100 mL Erlenmeyer flask closed with RotiLabo Cultuur Stoppen (Größe 29; Brunschwig) and incubated at 37°C and 250 rpm until an OD600 of 0.6 to 0.8 was reached. Then, 1 mM IPTG was added for plasmid induction, and incubation was continued overnight at 250 rpm and 18°C. The next day, cultures were harvested by centrifugation in 50 mL tubes at 3400 rpm for 15 minutes. Pellets were resuspended in 1 mL 50 mM Tris-HCl pH 7.5. About 0.2 g zirconium/silica beads 0.1 mm (Biospec products) were added. Then lysis was done by shaking for 10 seconds in a FastPrep (FP120 Bio101 Savant) at speed 6.5, then transfer of the tubes to ice for 2 minutes, and another round of shaking 10 seconds speed 6.5. Subsequently, lysates were centrifuged for 10 minutes at 13000 rpm and 4°C, and supernatants were immediately used for enzyme assays.

Enzyme assays were performed by mixing 65 μL Tris-HCl (50 mM, pH=7.5), 800 μL Assay buffer (15 mM MOPSO (pH=7.0), 12.5% glycerol (v/v), 1 mM ascorbic acid, 0.1% Tween 20, 1 mM $MgCl_2$ and 2 mM Dithiothreitol), 35 μL crude enzyme extract, 5 μL FDP (10 mM) and 20 μL sodium orthovanadate (250 mM). The glass tubes were incubated at 30°C with mild agitation for 2 hours. The sesquiterpenes produced in the enzyme assay were extracted with 2 mL ethyl acetate ≥99.5% (Alfa Aesar). The ethyl acetate was centrifuged for 10 min at 1200 x g, dried over a sodium sulphate column and used for GC-MS analysis.



### 4.2.3 Cultivation conditions

For the experimental set up, different conditions were chosen, consistently with the analysis technique to apply.

#### 4.2.3.1 For GC-MS analysis

Fermentations were performed as previously described[24]. In short: pre-cultures were grown in 5 mL LB medium supplemented with 50 μg/mL kanamycin and 50 μg/mL chloramphenicol in 50 mL Greiner tubes. Subsequently, the pre-cultures were diluted to an OD600 of 0.150 in 20 mL 2xYT medium with antibiotics in 100 mL glass flasks and incubated at

37°C 250 rpm until an OD600 of 0.4–0.5 was reached. Then, IPTG 1 mM and 10% v/v n-dodecane 99% (Alfa Aesar) were added to the fermentations, following a 24 h incubation at 28°C 250 rpm. The 2 mL dodecane was added as organic phase to trap the produced sesquiterpenes, then recovered for GC-MS analysis by centrifugation at 4600 rpm for 15 min. All selected CnVS mutants were grown in duplicates for GC-MS analysis.

#### 4.2.3.2 For PTR-Qi-ToF-MS analysis

Fermentations were performed using alternatively 30mL of 2xYT medium with antibiotics in 250 mL bluecap glass bottles, or 20 mL of 2xYT medium with antibiotics in 100 mL bluecap glass bottles (Greiner). The same growth conditions were applied as described for the GC-MS analysis. As no extraction of the sesquiterpenes in organic phase is necessary for PTR-Qi-ToF-MS analysis, no dodecane layer was added to the cultures.

In a smaller scale set up, reusable 24-square deep-well polypropylene (17x17 mm, depth 40 mm, total volume 10 mL, Applikon Biotechnology, Netherlands) and single-use 96-square deep-well polypropylene (8x8 mm, depth 40 mm, total volume 2 mL, VWR Netherlands)) plates were used. 800 µL and 200 µL of 2xYT medium with antibiotics were inoculated for the 24-well plates respectively and the 96-well-plates respectively. At an OD600 of 0.1, 1 mM IPTG was immediately added. 96-well plates were covered by silicon Axygen AxyMat 96 Square Well Sealing Mat for Deep Well Plates (VWR, Netherlands), while for 24 wells plates, Air-O-seal hydrophobic gas permeable adhesive seals (Bioké, Netherlands) were used. The plates were incubated using the MicroFlask shaker clamp system (Applikon Biotechnology, Netherlands) at 28°C and 300 rpm for 24 hours.

After 24 hours growth, the OD600 was measured using Ultrospec 10 Cell Density Meter portable Spectrophotometer (Amersham Biosciences, UK) for each sample to determine the growth. The growth in flasks was used as a benchmark and performed in duplicate for each experiment. The same pre-culture of each strain was used for all cultivation conditions. Two plates per type were measured as biological replicates. 96 wells plates set up was designed to contain 5 biological replicates of each strain, while 24 wells plates set up was designed to contain 3 biological replicates of each strain.

### 4.2.4 GC-MS measurements and data analysis

The GC-MS analysis was performed as described before[24]. In brief, a 7980A GC system combined with a 597C inert MSD detector (70eV) (Agilent Technologies, Netherlands) was used. The system is equipped with a 7683 auto-sampler and injector and a Phenomenex Zebron ZB-5ms column of 30 m length x 0.25 mm internal diameter and 0.25 μm stationary phase, with a Guardian precolumn (5 m). For each sample, 1 μL of sample was injected. The injection chamber was at 250°C, the injection was splitless, and the ZB5 column was maintained at 55°C for 2 min after which a gradient of 10°C per minute was started, until 300°C. Peaks were detected in chromatograms of the total ion count. Valencene was identified and quantified by comparing the sample peak with a valencene standard (80% purity, Isobionics, Netherlands). The obtained data was analysed as corrected area/units. Average and standard deviation of replicates was calculated using Microsoft Excel. GC-MS results were used as benchmark for the validation of PTR-Qi-ToF-MS measurements.

### 4.2.5 PTR-Qi-ToF-MS measurements

For the measurements of bottles and plates, a PTR-Qi-ToF-MS (Ionicon Analytik GmbH, Innsbruck, Austria) operating in V mode (standard configuration) was used to measure the headspace of the samples. The following ionization conditions were set in the drift tube [900 V, 60°C, and 3.80 mBar corresponding to an E/N value of 134 Townsend (1 Td = $10^{-21}$Vm$^2$)]. The mass resolution (m/Δm) was higher than 3800. A Polyetheretherketone (PEEK) capillary tube (inner diameter, 0.50 mm), heated at 60 °C was used as inlet. The inlet tube was connected to a standard ISO polypropylene screw cap GL45 (max temperature 140°C), supplemented with two fitting valves. The second valve being necessary for the air flow in entrance. The flow rate was of 52 sccm, with an acquisition rate of one spectrum per second (*m*/*z* range: 20-510). The reagent ion used in this system is $H_3O^+$. Two sampling approaches were taken for the different sample typology. For bottles, discrete 60 seconds measurements were performed. Before each measurement, the samples were heated at 30°C for 30 min to allow gases equilibration. After 10 seconds of sampling laboratory air, each sample was measured for 40 seconds. For the measurements of 96 and 24 wells plates, the probe was attached to a sterile Henke-Ject® 0.90x25 mm needle (Henke Sass Wolf, Germany). Another needle of the same type was used to guarantee the airflow in the wells during the measurements. In this case, a continuous measurement was performed, and samples in different wells were measured without interruption. Each sample was measured for

10 seconds, with a recovery time (airflow) of 40 to 60 seconds between each sample to avoid memory effects. All measurements were performed manually. All *m/z* values in this study are reported as integers mass accuracy approximated for consistency. The parent ion for sesquiterpenes $(C_{15}H_{24})^+$ is *m/z* 205.195[26] and is indicated in this work as *m/z* 205.

### 4.2.6 PTR-Qi-ToF-MS data processing and analysis

All the PTR-MS data were processed through the PTR-MS Viewer software v.3.4.2.1 (Ionicon Analytik GmbH, Austria). Firstly, an internal mass axis calibration was performed by using the following calibration mass peaks: 21.0226 ($H_3O^+$), 29.9974 ($NO^+$), 59.0497 ($C_3H_6OH^+$) and 203.9430 (fragment of 1,3-Diodobenzene used as internal gas standard). The final mass accuracy of the processing was sufficient for determining the sum formula of volatile compounds (≈0.001) and therefore the experimental *m/z* values are reported up to the third decimal digits. Then a mass peaks integration was conducted to extract peak intensities in counts per seconds (CPS) and in normalized CPS (NCPS) using the formula:

$$NCPS = 10^6 \times ([MH^+] / (([H_3O^+]/0.2) + [cluster\ H_2O + H^+])$$



where $[MH^+]$: *m/z* 205 cps; $[H_3O^+]$: *m/z* 21 cps; [cluster $H_2O$ $H^+$]: *m/z* 37 cps.

For bottles, an average of the NCPS from second 15 to second 45 after insertion of the probe was calculated, while for plates an average of the NCPS from the moment of insertion of the probe start until 30 seconds after insertion of the probe was used. The baseline was calculated as an average of all the blank samples (culture medium) and subtracted from the average NCPS. For statistical analysis, IBM SPSS was used. To assess homogeneity of variance, Levene's Test for Equality of Variances of the one-way ANOVA analysis was performed. For p-values higher than 0.05, the assumption of homogeneity of variance is met and the one-way ANOVA was performed, using Tuckey and Scheffe post-hoc tests. For p-values lower than 0.05, the assumption of homogeneity of variance is violated and a non-parametric Kruskal-Wallis test was used to conduct the analysis.

| CnVS library | Enzyme Assay | In vivo Fermentations | Quantification (mg valencene/L)* | Reference |
|---|---|---|---|---|
| CnVS | Wild type | Wild Type | 69.4 ± 5.2 | Beekwilder *et al.* 2014[27] |
| F300Y | Higher production | Lower production | 57.5 ± 0.4 | Isobionics BV |
| L444F | Loss of function | Loss of function | N.D. | Isobionics BV |
| L566S | Higher production | Higher production | 90.9 ± 2.1 | Isobionics BV |
| pACYC Duet1 | Empty plasmid: negative control | Empty plasmid | N.D. | Laboratory stock |

*Quantification performed on n-dodecane extracted from *in vivo* fermentations using GC-MS.

**Table 4.1** Library of selected sesquiterpene synthases, comparative observations of *in vitro* enzyme assay and *in vivo* fermentations. The column on the right reports the quantification data obtained for the *in vivo* fermentation of CnVS mutants and their standard deviation.

## 4.3 Results

### 4.3.1 Mutants of CnVS show different enzyme activity *in vivo* and *in vitro*

*C. nootkatensis* valencene synthase (CnVS) has previously been shown to produce mainly valencene, with around 10% of germacrene A[27] (**Figure 4.1A**). A small library, consisting of the empty vector (pACYC-Duet-1), CnVS and three CnVS mutants were tested, where amino-acid positions 300 (F300Y), 444 (L444F) and 566 (L566F) were altered (**Table 4.1**). These enzymes were produced in *E. coli*, and cell free extracts were prepared for testing with FDP as a substrate. A comparison of the amount of valencene and β-elemene produced, indicated the L444F mutant as a loss-of-function variant, while L566S and F300Y displayed a higher activity in the enzyme assay compared to CnVS wild type (**Figure 4.1B**).

To determine sesquiterpene production by the CnVS mutants *in vivo*, and to create a benchmark, the same strains were grown in flask in 20 mL cultures, with an organic phase of n-dodecane, to trap the produced sesquiterpenes. We observed that valencene production in mutant L566S *in vivo* was higher than that detected in CnVS, while as expected in L444F the valencene was not detected (**Figure 4.1C)**. Interestingly, variant F300Y performed less well than CnVS in the *in vivo* assay, compared to the *in vitro* enzyme assay (**Figure 4.1B vs C**).

### 4.3.2 PTR-Qi-ToF-MS analysis of cultures in bottles

The CnVS mutants described above were grown in glass bottles of 250 mL (with 30 mL of culture) and 100 mL (with 20 mL of culture). In the

bottles, no n-dodecane was added since no extraction is necessary for the analysis of the produced sesquiterpene. Moreover, we observed that the presence of 10% n-dodecane interferes with detection of sesquiterpenes as the high concentration saturates the detector causing ion suppression. In parallel, starting from the same pre-cultures, flasks with 20 mL culture and n-dodecane layer were used as benchmark for growth and production of sesquiterpenes. The growth in flasks was comparable with the growth in bottles, with no major difference for different bottle sizes, as shown in **Figure 4.2A**. After 24 hours fermentation, the headspace of the bottles was analysed with PTR-Qi-ToF-MS using 60 seconds discrete measurements (**Figure 4.2 B**).

For both bottle sizes, the PTR-Qi-ToF-MS signals of the different CnVS mutants were consistent with the GS-MS based quantifications of the flask cultivations. In all three set ups, no signal was detected for sesquiterpenes in L444F, as expected, while F330Y, CnVS and L566S, showed increasing sesquiterpenes signals, respectively (**Figure 4.3A, B and C).** To further corroborate the similarity of the analytical methods, a correlation analysis was performed (**Figure 4.3D, E**), showing a good correlation between the flasks ($R^2 > 0.95$) both bottle types. These experiments indicate the applicability of PTR-Qi-ToF-MS as a screening method for differences in terpene synthase efficiency, showing a good correlation to the accepted GC-MS methodology, but with advantage of a shorter measurement time. A larger difference between the ratio of L566S/F300Y was observed in bottles compared to flasks, especially for the 100 mL bottles set up (**Table 4.2**), which may indicate a different sensitivity of the methods at lower concentrations.



### 4.3.3 PTR-Qi-ToF-MS analysis of cultures in deep-well plates

To test the suitability of the PTR-Qi-ToF-MS as a HTS method for larger mutant libraries, a smaller scale cultivation set-up was tested. The same strains used in the bottles were grown in 24 deep-well plates, using 800 μL of medium, leaving 9.2 mL of headspace in every well. For each CnVS mutant, three replicates were grown in parallel, and the experiment was repeated twice. For this set-up, a continuous measurement approach was used, in which each well was measured for 10 seconds and a time varying between 30 and 90 seconds was taken between measurements as recovery time, to eliminate any possible memory effect (**Figure 4.4A**).

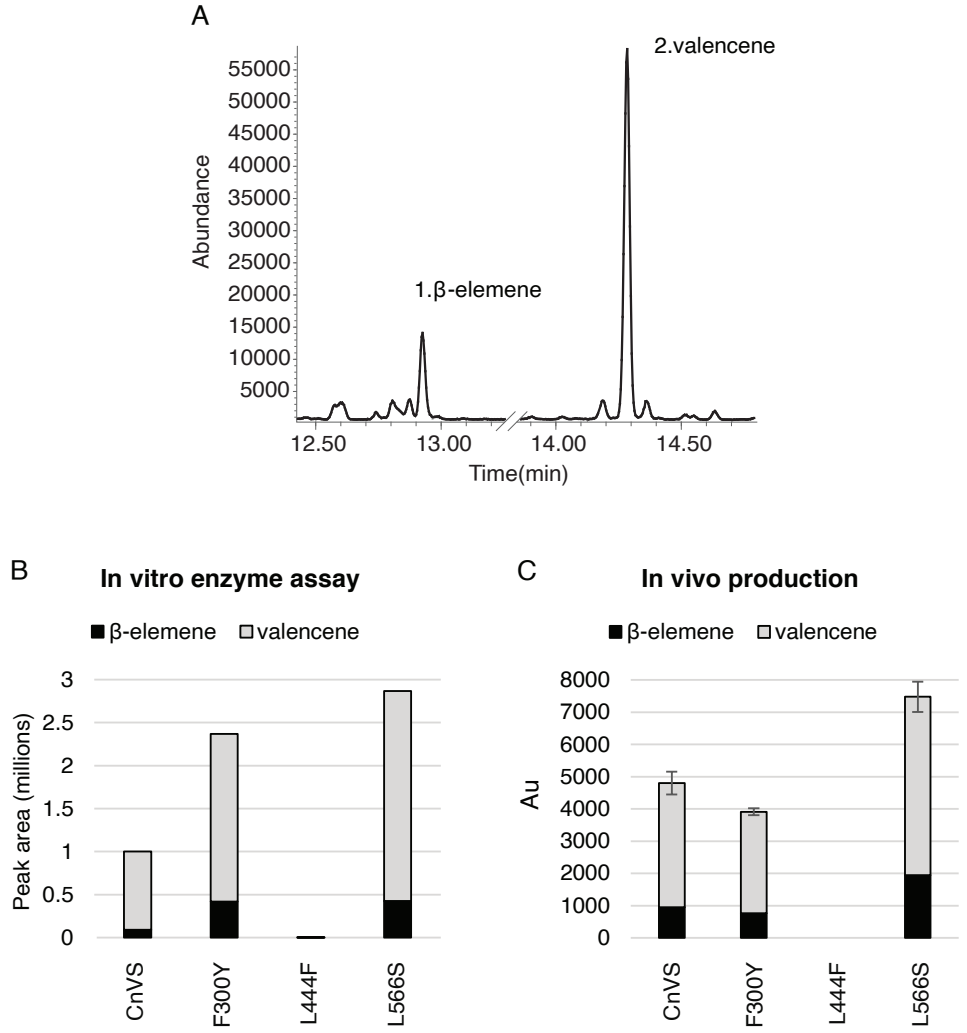



**Figure 4.1** Characterization of valencene synthase from *Callitropsis nootkatensis* (CnVS) mutants. (A) GC-MS chromatogram of wild type CnVS, peaks for 1. β-elemene and 2. valencene are reported. (B) *In vitro* enzyme assay and (C) *in vivo* production of sesquiterpenes by CnVS mutants. The histograms report the peak area for the *in vitro* enzyme assay (B) of the sesquiterpene produced, and the corrected area units (Au) for the *in vivo* fermentation (C). In both cases, the sesquiterpenes produced are analysed using GC-MS, showing valencene as the main product detected, with traces of β-elemene (germacrene A). Standard deviation is calculated for *in vivo* fermentations.

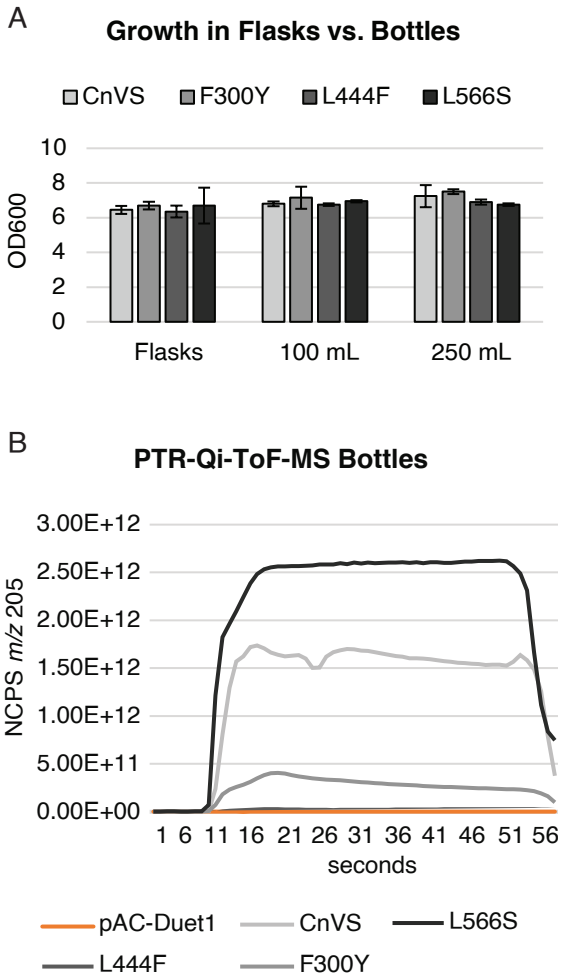



**Figure 4.2** (A) Growth comparison of selected CnVS mutants between cultures in flasks and in bottles. Cultures were grown in triplicates for 24 hours and OD600 was measured. Growth in flasks and in bottles showed no significant difference. (B) Summary graph of PTR-Qi-ToF-MS analysis of cultures grown in bottles. The graph reports the measurement of the headspace of cultures grown in 30 mL medium. L566S shows the highest signal.

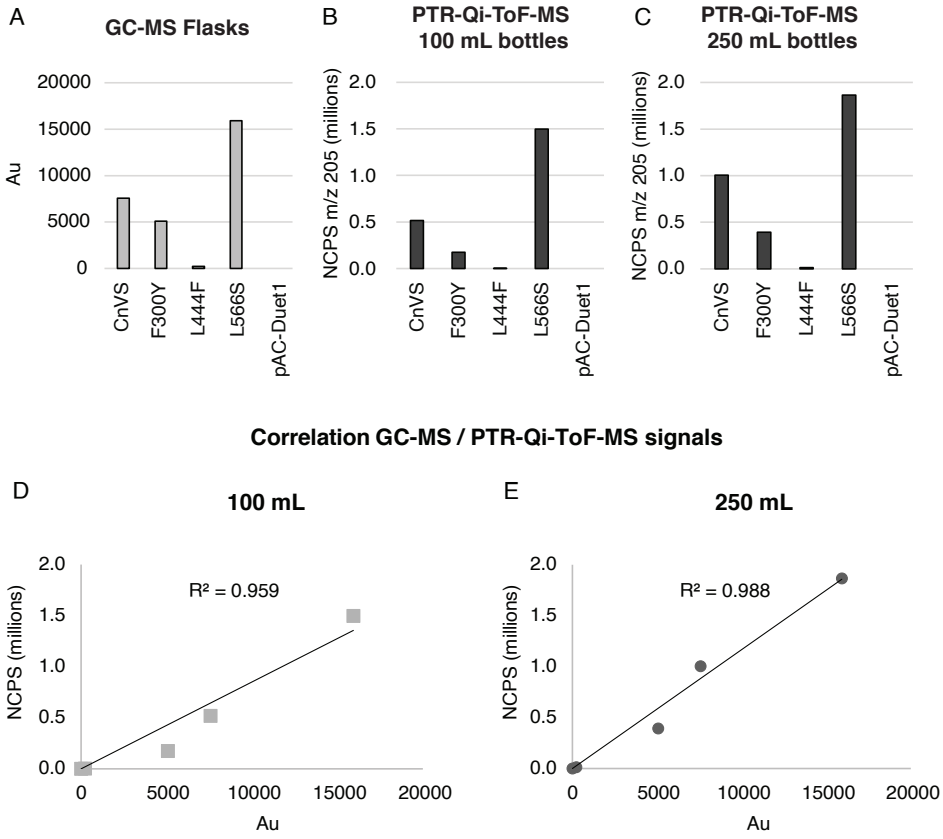



**Figure 4.3** Comparison of GC-MS and PTR-Qi-ToF-MS signals. (A) GC-MS analysis of fermentations in flasks using n-dodecane as overlay. The graph reports the total sesquiterpenes calculated as corrected area units (Au). (B-C) PTR-Qi-ToF-MS analysis of fermentation in (B) 100mL and (C) 250mL glass bottles. (D-E) Correlation between corrected area units (Au, x-axis) and NCPS in 100mL bottles (D) and 250mL bottles (E).

| Ratios | GC-MS | PTR-Qi-ToF-MS | |
|---|---|---|---|
| | Flasks | Bottles | |
| | | 100mL | 250mL |
| L566S / CnVS | 2.1 | 2.9 | 1.9 |
| L566S / F300Y | 3.1 | 8.6 | 4.7 |
| CnVS / F300Y | 1.5 | 3.0 | 2.6 |

**Table 4.2** Calculated ratio for valencene production in flasks (measured with GC-MS), and bottles (measured with PTR-Qi-ToF-MS).

Cultures were also grown in flasks with n-dodecane overlay as benchmark, the valencene production was analysed via GC-MS and quantified using a valencene standard (**Figure 4.4B**). With the 24 deep-well plates set-up, clear differences were observed between the cultures measured with PTR-Qi-ToF-MS (**Figure 4.4C, Table S4.1**). Consistently with the previous larger volume experiments, when comparing cultures grown in flasks and in bottles (**Figure 4.3A, B, C**), the strain expressing L566S showed on average higher valencene production than the CnVS strain. Similarly, the signal for valencene in strain F300Y was consistently lower than CnVS, and valencene production by L444F was hardly detectable (**Figure 4.4C**). The correlation of the signal obtained from PTR-Qi-ToF-MS analysis with the GC-MS quantification of the parallel flask experiment, showed an $R^2$ of 0.863 (**Figure 4.4D**). After measuring the headspace of the 24 deep-well plates, bacterial growth was measured as OD600, showing no major differences from the cultures grown in flasks (**Figure 4.4E**).

A further miniaturization of the cultivation was introduced by growing strains in a 96 deep-well plate format. In this case, the well volume is 2 mL, and the cultures were grown in 200 μL medium, leaving 1.8 mL of headspace volume. The same PTR-Qi-ToF-MS analysis time and set up as for the 24 deep-well plates were used (**Figure 4.5A**), and experiments were repeated twice, on different days. Each strain was cultivated in five replicates per plate. In this experiment, the difference between strain L566S and CnVS was less pronounced (**Figure 4.5B**) and did not qualify as significant in the ANOVA analysis, although the mean value for the L566S strain was higher than the value for CnVS. It must be noted that the GC-MS results for the flask cultivations ran in parallel also showed a less pronounced difference between the production of valencene in CnVS and variant L566S (**Figure 4.4B**). The correlation between the PTR-Qi-ToF-MS values for the 96-well plates and the GC-MS values for the parallel flask experiments gave an $R^2$ value of 0.894 (**Figure 4.5C**). Mean, standard deviation, and standard error for 24-well and 96-well plates were calculated (**Table S4.1**). Finally, the OD600 was measured to assess the bacteria growth, showing a consistent growth for all mutants in the 96 deep-well plates, comparable to that of the cultures in flasks. The ratio of sesquiterpene produced between the different cultures showed again, for both 24 and 96 deep-well plates set-ups, a larger difference between the ratio L566S/F300Y (**Table 4.3**).

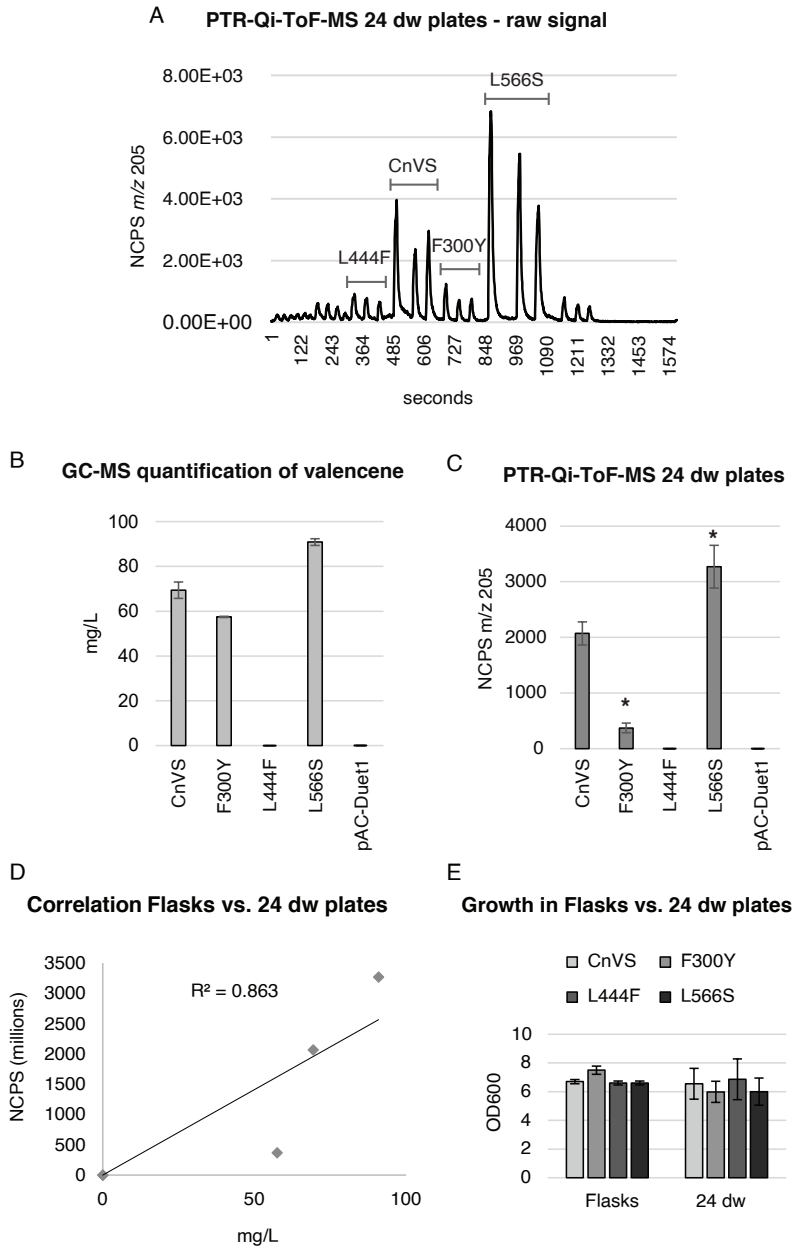



**Figure 4.4** (A) Raw signal obtained for continuous measurements of 24 deep-well (dw) plates measured with PTR-Qi-ToF-MS. Triplicates can be recognized in the varying peak heights. (B) Quantification of valencene produced in flasks by CnVS mutants using GC-MS. (C) Calculated average of triplicates in 24 dw plates. The asterisk indicates a significant difference compared with the wild type CnVS ($p < 0.05$). F300Y and L566S show significant difference compared to CnVS. (D) Correlation between valencene produced, quantified by GC-MS (mg/L of dodecane, x-axis) and signal obtained with PTR-Qi-ToF-MS. (E) Comparison of bacterial growth in flasks and in 24 dw plates. The growth was comparable among the two different set-ups.

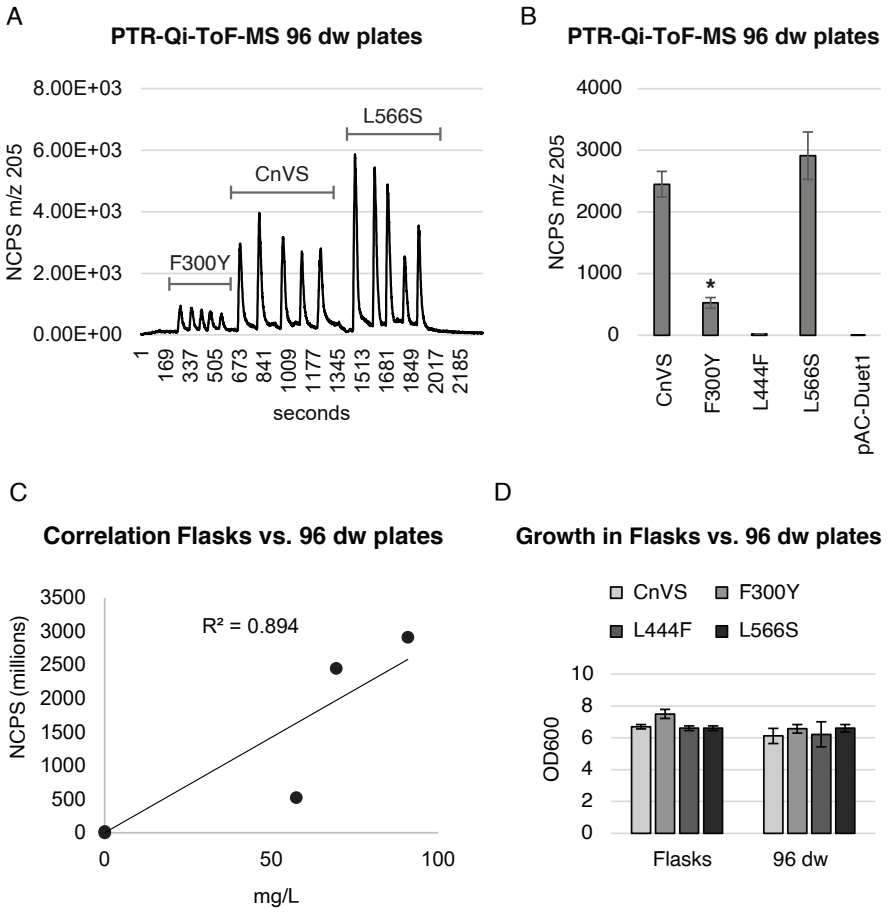
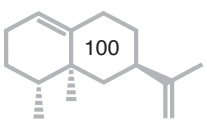



**Figure 4.5** (A) Raw signal obtained for continuous measurements of 96 deep-well (dw) plates measured with PTR-Qi-ToF-MS. The 5 replicates can be recognized in the varying peak heights. (B) Calculated average of replicates in 96 dw plates. The asterisk indicates a significant difference compared with the wild type CnVS ($p < 0.05$). Only F300Y shows significant difference compared to CnVS, however L566S shows the highest signal on average (Table S4.1). (C) Correlation between valencene produced, quantified by GC-MS (mg/L of dodecane, x-axis) and signal obtained with PTR-Qi-ToF-MS. (D) Comparison of bacterial growth in flasks and in 96 dw plates. The growth was comparable among the two different set-ups.

| | GC-MS | PTR-Qi-ToF-MS | |
|---|---|---|---|
| | Flasks | Plates | |
| Ratios | | 24 dw | 96 dw |
| L566S / CnVS | 1.3 | 1.6 | 1.2 |
| L566S / F300Y | 1.6 | 6.2 | 5.5 |
| CnVS / F300Y | 1.2 | 4.6 | 5.0 |

**Table 4.3** Calculated ratio for valencene production in flasks (measured with GC-MS) and deep-well (dw) plates (measured with PTR-Qi-ToF-MS).



## 4.4 Discussion

So far, HTS of terpene-producing microbial strains has mainly been developed using colorimetric methods. Reactive dyes such as Purpald[28,29] or malachite green[30,31] have been employed for screening of terpene synthase activity *in vitro*[28,30]. However, the activity of synthases *in vitro* cannot be directly translated to *in vivo* performance of strains expressing variant synthases (**Figure 4.1**). A clear example of this can be observed with the activity of variant F300Y in an *in vitro* setting compared with the more reliable *in vivo* results. Apparently, the *in vitro* conditions may not be representative for *in vivo* conditions, as different physical and chemical conditions, such as pH, salinity and availability of the substrate, might hinder or enhance the activity of the enzyme without a correspondence in the *in vivo* system.

Other colorimetric approaches for HTS for *in vivo* activity have been developed, such as the use of Nile Red to select *E. coli* strains that accumulated higher concentrations of terpenes in their intracellular space[32]. Another qualitative method reported for the screening of monoterpene synthases (limonene) relied on the use of 2,2-diphenyl-1-picrylhydrazyl (DPPH) and its chromatic toning in organic phase due to the mild antioxidant activity of some monoterpenes[33]. These methods, while being undoubtedly faster than GC-MS analysis, present a number of limitations related to the selectivity of the staining agent and the subsequent detection method. For instance, the detection of Nile Red is subject to variation due to background noise related to the staining of lipids. DPPH is known to react with other antioxidant compounds, and has been widely applied as an *in vitro* assay for antioxidant activity[34]. This lack of specificity might hinder the accuracy of the results as small variations in the presence of other, more strongly antioxidant compounds could interfere with the detection of terpenes. Furthermore, both these methods are indirect and require an extra step of sample preparation that reduces the throughput of the screening.

In the current work, we evaluate real time detection of sesquiterpenes in the headspace of the cultures as a tool for HTS. In different formats for screening a mini library of valencene synthase mutants, the PTR-Qi-ToF-MS methodology provided results which sowed high correlation with the classical analytical method based on GC-MS. Indeed, measurements performed by comparing cultures grown in bottles and flasks, show that sesquiterpenes present in the headspace measured by PTR-Qi-ToF-MS correlate well ($R^2 > 0.95$) with sesquiterpenes present in the dodecane

layer measured by GC-MS (**Figure 4.4C, D**). This indicates that PTR-Qi-ToF-MS analysis has potential to be an accurate replacement of the time and labour-intensive GC-MS detection method in HTS of large terpene synthase libraries.

To further address the robustness of the PTR-Qi-ToF-MS method, we tested the same library in different cultivation set ups (**Figures S4.1**). The biological variation between replicates was observed when analysing deep-well plates, with a higher difference for the 96-well plate set-up (**Table S4.1**). This variation might be caused by the smaller culture volume and the potentially more dramatic impact of the biomass growth in correlation with the production. Nevertheless, under each cultivation condition tested, the ranking of the mutants was the same, indicating a high sensitivity of the PTR-Qi-ToF-MS for detecting changes in terpene production, even in the smallest scale tested. We observed that the mutant L566S consistently produced more valencene when compared to wild type CnVS (**Figures 4.3B, C; 4.4C** and **4.5B**), while mutant L444F was the least efficient. In all *in vivo* cultivation conditions tested, F300Y was less productive than CnVS, in contrast to the values observed in the *in vitro* enzyme assays (**Figure 4.1B, C**). However, it should be noted that this mutant showed differences in performance on different repeats of the flask cultivation with dodecane, suggesting that it might be more sensitive to small changes in the cultivation conditions. In both 24 and 96 deep-well plates, the production of valencene of F300Y was significantly lower than that of CnVS. Possibly, this variation is related to the differences existing in the precultures used for inoculating the flasks and multi-well plates. However, when applying PTR-Qi-ToF-MS as HTS method, the focus would uniquely be on the enzymes showing higher productivity compared to the wild type.



Although the results obtained by using the multi well plates are consistent with the GC-MS results obtained in the flasks from the same pre-culture, a smaller cultivation volume could influence the difference in terpene accumulated within the 24 hours of cultivation. To our knowledge, the impact of a relatively small headspace on terpene production and accumulation has not yet been investigated, and this might be an important factor to consider in the choice of the timing for the measurements. Thus, for the development of a robust HTS culture screening method this aspect would need further investigation.

PTR-Qi-ToF-MS offers notable advantages in analysis time, which could make the technique attractive for screening of larger terpene synthase libraries. When considering the 96-well plate set up, 10-15 seconds per measurement appears to be sufficient to determine the abundance of valencene. Calculating an average of 50 seconds of recovery time between samples, we can estimate that only 1 minute is required per each sample. This translates to the possibility of measuring one 96 well plate in less than 2 hours, which, in turn, opens the possibility to measure about 12 plates in 24 hours, for a total of 1152 samples per day. Considering that the current throughput for GC-MS allows the measurement of less than 60 samples per day, we propose that the employment of PTR-Qi-ToF-MS as a HTS could bring considerable benefits (**Figure 4.6**). For this to be feasible, it is essential to implement the PTR-Qi-ToF-MS with an *ad-hoc* autosampler to reduce human error and optimize the timing for the measurements.

Despite the advantages presented by the use of PTR-Qi-ToF-MS, it is important to make some remarks. First, as mentioned above, it is crucial to establish an automated system for the successful application of this technique as a HTS for large mutant libraries, as it has been proposed and applied in food research[35]. The system should be able to optimize both the measurement and the recovery time and guarantee consistency in the measuring process. Ideally, an automated liquid handler should also be implemented for the preparation of the deep-well plates, as described by Leferink *et al.* (2019)[3]. Second, the variation observed between the replicates must be kept into consideration when designing the sampling scheme. As shown in **Figure S4.1**, the design includes 5 replicates per sample. To make the design efficient for HTS applications the number of replicates must be lowered. While an autosampler might help minimize this variation, our observations suggest that, to have confidence over the results, the design should include at least three replicates per mutant, as tested for the 24 deep-well plates. Furthermore, to reduce the memory effect and keep the contamination under control, a blank sample should be included between each set of replicates. A 96 deep-well plate design appears to have good potential for HTS application. In fact, no major reduction in growth was detected in comparison to the other conditions (**Figure 4.4E, 4.5D**) and no issues were encountered with the measurements of the different wells in comparison with the 24 deep-well plates.

Finally, we must consider that as per most of the HTS methods, the PTR-Qi-ToF-MS technique can only be used to identify preliminary candidates. In fact, while GC-MS analysis allows the identification of different isoforms

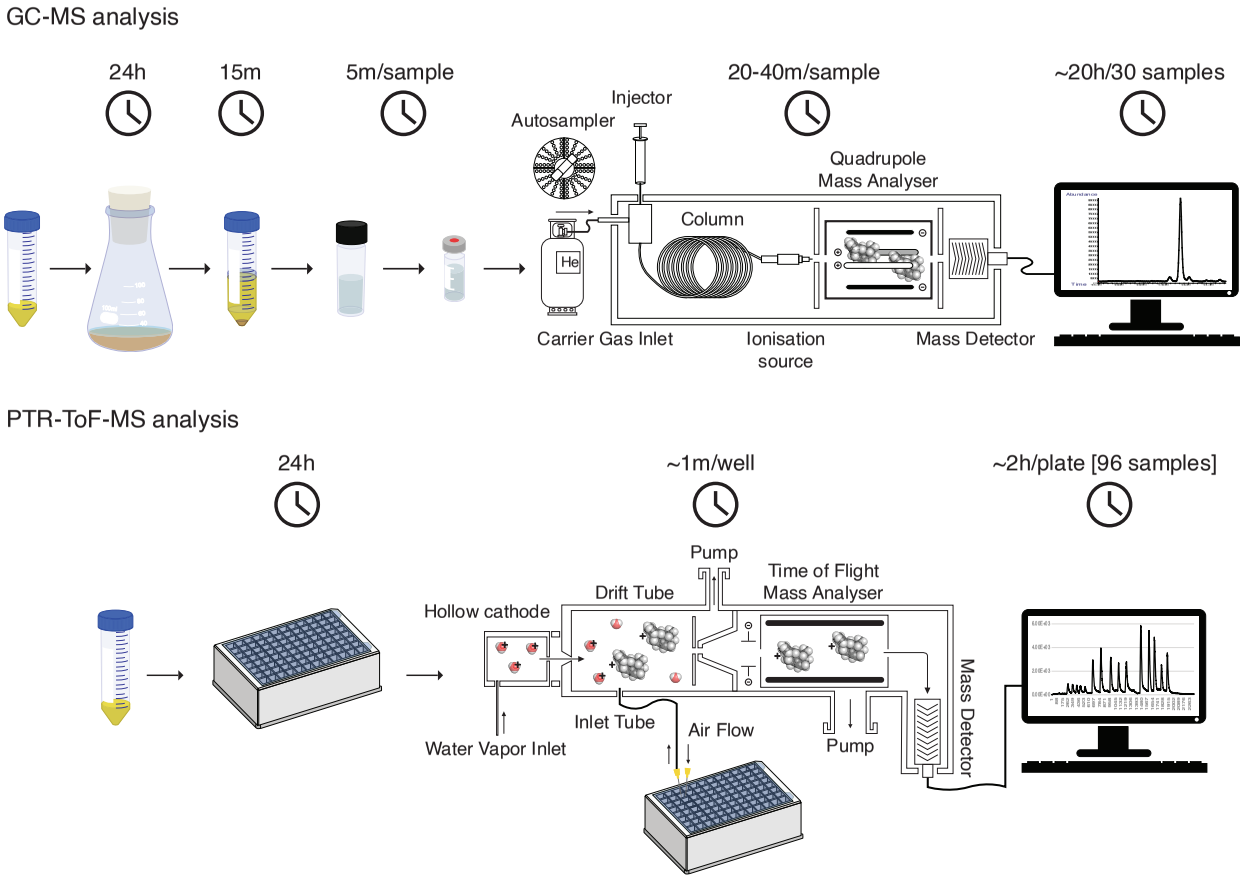



**Figure 4.6** Schematic representation of steps and terms involved in GC-MS vs PTR-Qi-ToF-MS analysis.

of ions with the same mass by detecting their fragmentation pattern, PTR-Qi-ToF-MS only provides a quantification of the total signal for each specific mass. While in literature it is shown that differentiation between compounds with the same mass as isoforms is possible[36,37], it is difficult to obtain a reliable quantification for each compound. Therefore, the analysis can only be quantitative but not accurately qualitative. Following the complementarity principle introduced by Majchrzak and colleagues[38], where it is stated that "*for analytical techniques to be complementary, they should meet at least one condition concerning their complementarity, namely that their application enables the qualitative and quantitative analysis to be carried out in a much less time-consuming way than in the case of using one analytical technique*", a GC-MS analysis of selected mutant strains will be necessary to obtain the full characterization of the mutants of interest, however saving time over the full characterization of those mutants that did not pass the initial selection.

## 4.5 Conclusions

In this study, we propose the application of the PTR-Qi-ToF-MS as a HTS method for the identification of improved sesquiterpene synthases. The successful miniaturization to a 96-well plate set up for bacterial cultures provides a promising starting point for the implementation of this technique for the screening of large libraries of mutants. While optimization is still necessary to employ PTR-Qi-ToF-MS for the screening of large mutant libraries, and HTS application depends on efficient automation, the versatility of the method clearly allows for further application to other volatile compounds produced in microbial hosts and could be extended not only for enzyme but also for strain engineering.

# Appendix

**Figure S4.1** Design for 24 deep-well plates and 96 deep-well plates.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 2xYT | L444F | L444F | 2xYT | L444F | L444F |
| B | F300Y | F300Y | F300Y | F300Y | F300Y | F300Y |
| C | CnVS | CnVS | CnVS | CnVS | CnVS | CnVS |
| D | L566S | L566S | L566S | L566S | L566S | L566S |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2xYT | pAC-Duet1 | 2xYT | L444F | L444F | 2xYT | 2xYT | L444F | L444F | L444F | 2xYT | 2xYT |
| B | 2xYT | pAC-Duet1 | 2xYT | F300Y | F300Y | 2xYT | 2xYT | F300Y | F300Y | F300Y | 2xYT | 2xYT |
| C | 2xYT | pAC-Duet1 | 2xYT | CnVS | CnVS | 2xYT | 2xYT | CnVS | CnVS | CnVS | 2xYT | 2xYT |
| D | 2xYT | pAC-Duet1 | 2xYT | L566S | L566S | 2xYT | 2xYT | L566S | L566S | L566S | 2xYT | 2xYT |
| E | 2xYT | pAC-Duet1 | 2xYT | L444F | L444F | 2xYT | 2xYT | L444F | L444F | L444F | 2xYT | 2xYT |
| F | 2xYT | pAC-Duet1 | 2xYT | F300Y | F300Y | 2xYT | 2xYT | F300Y | F300Y | F300Y | 2xYT | 2xYT |
| G | 2xYT | pAC-Duet1 | 2xYT | CnVS | CnVS | 2xYT | 2xYT | CnVS | CnVS | CnVS | 2xYT | 2xYT |
| H | 2xYT | pAC-Duet1 | 2xYT | L566S | L566S | 2xYT | 2xYT | L566S | L566S | L566S | 2xYT | 2xYT |

| | 24 deep-well* | | | 96 deep-well** | | |
|---|---|---|---|---|---|---|
| | Mean NCPS | Std.dev | Std. Error | Mean NCPS | Std.dev | Std. Error |
| CnVS | 2070.8 | 510.4 | 208.4 | 2449.6 | 494.0 | 137.0 |
| F300Y | 372.1 | 214.9 | 87.7 | 525.5 | 65.7 | 19.0 |
| L444F | -12.9 | 15.4 | 7.7 | 22.2 | 43.9 | 12.7 |
| L566S | 3270.4 | 940.7 | 384.1 | 2913.9 | 830.9 | 230.4 |

*Calculated baseline = 195 **Calculated baseline = 137

**Table S4.1** Calculated mean, standard deviation and standard error for 24 and 96-well plates. The NCPS were calculated using formula (1) and the calculated baseline was subtracted from the average NCPS for each sample.

## References


1. Lim, H. S. *et al.* (-)-α-Bisabolol Production in Engineered Escherichia coli Expressing a Novel (-)-α-Bisabolol Synthase from the Globe Artichoke Cynara cardunculus var. Scolymus. *J. Agric. Food Chem.* **69**, 8492–8503 (2021).
2. Niu, F. X., He, X., Wu, Y. Q. & Liu, J. Z. Enhancing Production of Pinene in Escherichia coli by Using a Combination of Tolerance, Evolution, and Modular Co-culture Engineering. *Front. Microbiol.* **9**, 1623 (2018).
3. Leferink, N. G. H. *et al.* An automated pipeline for the screening of diverse monoterpene synthase libraries. *Sci. Rep.* **9**, 1–12 (2019).
4. Ellis, A. M. & Mayhew, C. A. *Proton Transfer Reaction Mass Spectrometry. Proton Transfer Reaction Mass Spectrometry: Principles and Applications* (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118682883.
5. Blake, R. S., Monks, P. S. & Ellis, A. M. Proton-transfer reaction mass spectrometry. *Chemical Reviews* vol. 109 861–896 (2009).
6. Cappellin, L. *et al.* On quantitative determination of volatile organic compound concentrations using proton transfer reaction time-of-flight mass spectrometry. *Environ. Sci. Technol.* **46**, 2283–2290 (2012).
7. Graus, M., Müller, M. & Hansel, A. High resolution PTR-TOF: Quantification and Formula Confirmation of VOC in Real Time. *J. Am. Soc. Mass Spectrom.* **21**, 1037–1044 (2010).
8. Cappellin, L. *et al.* Extending the dynamic range of proton transfer reaction time-of-flight mass spectrometers by a novel dead time correction. *Rapid Commun. Mass Spectrom.* **25**, 179–183 (2011).
9. Sulzer, P. *et al.* A proton transfer reaction-quadrupole interface time-of-flight mass spectrometer (PTR-QiTOF): High speed due to extreme sensitivity. *Int. J. Mass Spectrom.* **368**, 1–5 (2014).
10. Xia, S. Y. *et al.* Long-term observations of oxygenated volatile organic compounds (OVOCs) in an urban atmosphere in southern China, 2014–2019. *Environ. Pollut.* **270**, (2021).
11. Majchrzak, T., Wojnowski, W. & Wasik, A. Revealing dynamic changes of the volatile profile of food samples using PTR–MS. *Food Chemistry* vol. 364 130404 (2021).
12. Löser, B. *et al.* Changes of exhaled volatile organic compounds in postoperative patients undergoing analgesic treatment: A prospective observational study. *Metabolites* **10**, 1–13 (2020).
13. Pedrotti, M. *et al.* Rapid and noninvasive quality control of anhydrous milk fat by PTR-MS: The effect of storage time and packaging. *J. Mass Spectrom.* **53**, 753–762 (2018).
14. Bottiroli, R. *et al.* Application of PTR-TOF-MS for the quality assessment of lactose-free milk: Effect of storage time and employment of different lactase preparations. *J. Mass Spectrom.* **55**, (2020).
15. Liu, N., Koot, A., Hettinga, K., de Jong, J. & van Ruth, S. M. Portraying and tracing the impact of different production systems on the volatile organic compound composition of milk by PTR-(Quad)MS and PTR-(ToF)MS. *Food Chem.* **239**, 201–207 (2018).
16. Alothman, M., Lusk, K. A., Silcock, P. & Bremer, P. J. Comparing PTR-MS profile of milk inoculated with pure or mixed cultures of spoilage bacteria. *Food Microbiol.* **64**, 155–163 (2017).
17. Li, M., Cappellin, L., Xu, J., Biasioli, F. & Varotto, C. High-throughput screening for in planta characterization of VOC biosynthetic genes by PTR-ToF-MS. *J. Plant Res.* **133**, 123–131 (2020).
18. Aprea, E. *et al.* Rapid white truffle headspace analysis by proton transfer reaction mass spectrometry and comparison with solid-phase microextraction coupled with gas chromatography/mass spectrometry. *Rapid Commun. Mass Spectrom.* **21**, 2564–2572 (2007).
19. Mallette, N. D., Berk Knighton, W., Strobel, G. A., Carlson, R. P. & Peyton, B. M. Resolution of volatile fuel compound profiles from Ascocoryne sarcoides: A comparison by proton transfer reaction-mass spectrometry and solid phase microextraction gas chromatographymass spectrometry. *AMB Express* **2**, 1–13 (2012).
20. Kappers, I. F., Verstappen, F. W. A., Luckerhoff, L. L. P., Bouwmeester, H. J. & Dicke, M. Genetic variation in jasmonic acid- and spider mite-induced plant volatile emission of cucumber

accessions and attraction of the predator Phytoseiulus persimilis. *J. Chem. Ecol.* **36**, 500–512 (2010).
21. Bouvier-Brown, N. C., Holzinger, R., Palitzsch, K. & Goldstein, A. H. Quantifying sesquiterpene and oxygenated terpene emissions from live vegetation using solid-phase microextraction fibers. *J. Chromatogr. A* **1161**, 113–120 (2007).
22. Živković, S. *et al.* Rehydration process in rustyback fern (Aspleniumceterach l.): Profiling of volatile organic compounds. *Biology (Basel).* **10**, (2021).
23. Furusawa, M., Hashimoto, T., Noma, Y. & Asakawa, Y. Highly efficient production of nootkatone, the grapefruit aroma from valencene, by biotransformation. *Chem. Pharm. Bull.* **53**, 1513–1514 (2005).
24. Di Girolamo, A. *et al.* The santalene synthase from Cinnamomum camphora: Reconstruction of a sesquiterpene synthase from a monoterpene synthase. *Arch. Biochem. Biophys.* **695**, (2020).
25. Beekwilder, J. *et al.* Valencene synthase from the heartwood of Nootka cypress ( Callitropsis nootkatensis ) for biotechnological production of valencene. *Plant Biotechnol. J.* **12**, 174–182 (2014).
26. Kim, S. *et al.* Measurement of atmospheric sesquiterpenes by proton transfer reaction-mass spectrometry (PTR-MS). *Atmos. Meas. Tech.* **2**, 99–112 (2009).
27. Beekwilder, J. *et al.* Valencene synthase from the heartwood of Nootka cypress (Callitropsis nootkatensis) for biotechnological production of valencene. *Plant Biotechnol. J.* **12**, 174–182 (2014).
28. Lauchli, R. *et al.* High-throughput screening and directed evolution of terpene synthase-catalyzed cylization. *Angew Chem Int Ed Engl* **52**, 1–12 (2013).
29. Lauchli, R., Pitzer, J., Kitto, R. Z., Kalbarczyk, K. Z. & Rabe, K. S. Improved selectivity of an engineered multi-product terpene synthase. *Org. Biomol. Chem.* **12**, 4013–4020 (2014).
30. Vardakou, M., Salmon, M., Faraldos, J. A. & O'Maille, P. E. Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* **1**, e187–e196 (2014).
31. Pegan, S., Tian, Y., Sershon, V. & Mesecar, A. A Universal, Fully Automated High Throughput Screening Assay for Pyrophosphate and Phosphate Release from Enzymatic Reactions. *Comb. Chem. High Throughput Screen.* **13**, 27–38 (2009).
32. Siltanen, C. A. *et al.* An Oil-Free Picodrop Bioassay Platform for Synthetic Biology. *Sci. Rep.* **8**, (2018).
33. Behrendorff, J. B. Y. H., Vickers, C. E., Chrysanthopoulos, P. & Nielsen, L. K. 2,2-Diphenyl-1-picrylhydrazyl as a screening tool for recombinant monoterpene biosynthesis. *Microb. Cell Fact.* **12**, 76 (2013).
34. Nwachukwu, I. D., Sarteshnizi, R. A., Udenigwe, C. C. & Aluko, R. E. A concise review of current in vitro chemical and cell-based antioxidant assay methods. *Molecules* vol. 26 (2021).
35. Capozzi, V. *et al.* PTR-ToF-MS coupled with an automated sampling system and tailored data analysis for food studies: Bioprocess monitoring, screening and nose-space analysis. *J. Vis. Exp.* **2017**, 1–11 (2017).
36. Tani, A. Fragmentation and reaction rate constants of terpenoids determined by proton transfer reaction-mass spectrometry. *Environ. Control Biol.* **51**, 23–29 (2013).
37. Misztal, P. K., Heal, M. R., Nemitz, E. & Cape, J. N. Development of PTR-MS selectivity for structural isomers: Monoterpenes as a case study. *Int. J. Mass Spectrom.* **310**, 10–19 (2012).
38. Majchrzak, T. *et al.* PTR-MS and GC-MS as complementary techniques for analysis of volatiles: A tutorial review. *Analytica Chimica Acta* vol. 1035 1–13 (2018).

# 5

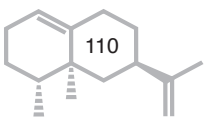
# Identification of transcriptional regulators affecting gene expression in a valencene-producing *Rhodobacter sphaeroides*

Alice Di Girolamo[1], Janani Durairaj[2], Thamara Hesselink[3], Aurin Vos[3], Adèle van Houwelingen[3], Harro J. Bouwmeester[4], Aalt DJ. van Dijk[2,5], Dick de Ridder[2], Jules Beekwilder[6]

[1] Laboratory of Plant Physiology, Department of Plant Sciences, Wageningen University, Netherlands
[2] Bioinformatics Group, Department of Plant Sciences, Wageningen University, Netherlands
[3] Bioscience, Wageningen Plant Research, Wageningen University, Netherlands
[4] Swammerdam Institute for Life Sciences, University of Amsterdam, Netherlands
[5] Biometris, Department of Plant Sciences, Wageningen University, Netherlands
[6] Isobionics BV, Geleen, The Netherlands

# abstract

*Rhodobacter sphaeroides* is an alpha proteobacterium well-known for its versatile metabolism. While some of its regulatory circuits have been widely studied, it is still unclear if the production of exogenous sesquiterpenes is affecting gene expression regulation. In this study, we analysed RNA-seq measurements on a valencene-producing *R. sphaeroides* and a wild type *R. sphaeroides* harbouring an empty vector, aiming to identify transcriptional regulators sensitive to the presence of this sesquiterpene. Five different conditions were employed, including the presence or absence of an organic phase in the medium. We identified 98 differentially expressed genes throughout the different growth conditions, of which 7 transcriptional regulators. Two of the transcriptional regulators, IlvR and PstR, were chosen for further characterization and knockouts were generated and tested for valencene production. We observed an increased production of valencene in *ΔIlvR* and a reduced production of carotenoids and bacteriochlorophylls, while *ΔPstR* shows a reduced production of valencene. We provide an initial characterization of the identified regulators, concluding that further experiments are necessary to identify the ligands of these regulators and their DNA binding sites.



## 5.1 Introduction

*Rhodobacter sphaeroides* is a purple non-sulphur alpha proteobacterium studied for its ability to grow at very high densities, in both phototrophic and chemotrophic conditions, aerobically or anaerobically[1]. Being a facultative photosynthetic organism able to naturally produce a variety of isoprenoids, such as bacteriochlorophyll and carotenoids[2,3], *R. sphaeroides* is a promising candidate for biotechnological production of terpenoids[4]. One of the first applications of *R. sphaeroides* as a cell factory aimed at the production of native Coenzyme Q10 (CoQ10)[5]. It is currently also applied for production of sesquiterpenes, such as valencene. By learning how to improve the production of naturally occurring compounds, new insights on the regulation of supply pathways have been obtained and have been applied for heterologous expression of different terpenoid synthases[6]. In recent years, the use of CRISPR-Cas9 technology combined with homologous recombination has been optimized for the engineering of *R. sphaeroides*, yielding an efficient and reliable genome editing tool for molecular research[7].

A number of studies have been performed on *R. sphaeroides* and its metabolic pathways involved in chlorophyll[8–10] and carotenoid production[11–13], and on central carbon metabolism[6] and changes in gene expression in the transition from exponential to stationary phase[14]. The regulation of photosynthetic (PS) genes in *R. sphaeroides* has been elucidated, and the transcriptional regulators (TRs) involved have been characterized. PpsR and AppA are the two main players in the regulation of the expression of the light harvesting complex II (LHII)[15]. AppA is inactive in the presence of oxygen and blue light[16,17], PpsR binds to the promoter regions of the PS operon *(crt, puc, bch, puf)* acting as a repressor[18]. Under anaerobic conditions, AppA is active and binds PpsR, releasing the promoter region and activating transcription. In turn, the expression of *appA* is regulated by the PrrBA (known also as RegBA) complex, which is sensitive to redox balance[19,20]. It has been shown that the deletion of *ppsR* enables the transcription of the PS operon in aerobic conditions and thus improves the production of carotenoids under low oxygen conditions[21].

However, not much is known about the response of *R. sphaeroides* to production of non-native terpenes, such as valencene, or the regulatory mechanisms behind such responses[22]. Considering the prevalence of terpene synthases in bacteria[23] and the well-known interactions between plant-derived secondary metabolites and bacterial regulators[24], it is logical

to hypothesise that regulatory circuits must be in place for the recognition of endogenous and exogenous terpenes. Valencene has generally low toxicity to microorganisms, but it yields differential expression of specific genes in different bacteria[25,26]. An example is provided by the overexpression of an mmpL2-type efflux pump in *Corynebacterium glutamicum* ATCC 13032 in the presence of valencene in the culture medium[27]. Transcriptional analysis of the response of bacteria to monoterpenes has revealed that monoterpene tolerance is related to the differential expression of membrane and flagellar proteins, as well as transcriptional regulators (TRs) belonging to the MarR and FabR families. These TRs are likely involved in the differential expression of gene clusters observed in the presence of terpenes[28,29]. In *R. sphaeroides*, transcriptome analyses have aimed mostly at the different growth phases[14,30] and the PS regulatory network[31], and as part of a comparative analysis of phototrophic growth between different species[3]. However, no studies have been conducted with the aim of identifying TRs that respond to the presence of terpenes, modulating gene expression. The identification of TRs sensitive to valencene, or to other sesquiterpenes, would open the possibility to implement these native regulatory circuits into biosensors able to recognize the product of interest[32].

In this work, we investigated regulatory circuits of *R. sphaeroides* by RNA sequencing (RNA-seq) analysis of a valencene-producing strain compared with a non-producing strain. With the analysis of the differentially expressed genes (DEGs) we identified candidates TRs potentially involved in the recognition of valencene. We then proceeded to generate knock out (KO) strains of some of the identified TRs and characterized these strains in comparison with the wild type. One of the KO obtained, *ΔIlvR*, behaved very similar to wild type in the absence of a terpene synthase. However, in combination with valencene synthase expression, the *ΔIlvR* showed decreased growth and native isoprenoids produced, in combination with an increase in valencene produced, compared to the wild type strain. This suggests a correlation between the TR and the metabolism of isoprenoid precursors.



## 5.2 Materials and Methods

### 5.2.1 Bacterial strains and culture media

The strains used in this study are reported in **Table 5.1**. In short, *E. coli* DH5α was used for cloning; *E. coli* S17-1 served as vector donor for *R. sphaeroides* in conjugation; wild type *R. sphaeroides 241 (Rs241)* was used as

starting strain for all experiments.

*E. coli* strains were grown on LB (Luria-Bertani) medium (10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract) with 50 μg/mL kanamycin at 37°C, 200 rpm. LB supplemented with 10% w/v agar was used for plates.

*Rs241* was grown on complex medium RS102 for valencene production as previously described[33], in short per 1L of medium: 20g yeast extract, 0.5 g NaCl, 0.5 g $MgSO_4 \cdot 7H_2O$, 33 g D-glucose monohydrate, 2 mL microelements solution (80 g/L $(NH_4)_2Fe(SO_4)_2 \cdot 6H_2O$, 6 g/L $ZnSO_4 \cdot 7H_2O$, 2 g/L $MnSO_4 \cdot H_2O$, 0.2 g/L $NiSO_4 \cdot 6H_2O$, 2 g/L vitamin C), 2mL CaFe solution (75g/L $CaCl_2 \cdot 2H_2O$, 5g $FeCl_3 \cdot 6H_2O$, 3.75mL/L HCl (37%)). 15% w/v of Micro Agar (Duchefa Biochemie, Netherlands) was added for plates. Kanamycin was used as antibiotic at a final concentration of 100μg/mL.

*Rs241* used for conjugation was grown on RÄ minimal medium, as previously described[34]. In short: 3 g/L malic acid (as only carbon source), 0.2 g/L $MgSO_4 \cdot 7H_2O$, 1.2 g/L $(NH_4)_2SO_4$, 0.07 g/L $CaCl_2 \cdot 2H_2O$, 1.5 mL of microelements stock solution (0.5 g/L Fe(II)-Citrate, 0.02 g/L $MnCl_2 \cdot 4H_2O$, 0.005 g/L $ZnCl_2$, 0.0025 g/L KBr, 0.0025 g/L KI, 0.0023 g/L $CuSO_4 \cdot 5H_2O$, 0.041g/L $Na_2MoO$, 0.005 g/L $CoCl_2 \cdot 6H_2O$, 0.0005 g/L $SnCl_2 \cdot 2H_2O$, 0.0006 g/L $BaCl_2 \cdot 2H_2O$, 0.031 g/L AlCl, 0.41 g/L $H_3BO$, 0.02 g/L EDTA), 2 mL of vitamin stock solution (0.2 g/L nicotinic acid, 0.4 g/L thiamine HCl, 0.008 g/L biotin,0.2g/L nicotinamide) and 5 mL of phosphate buffer (0.6 g/L $KH_2PO_4$and 0.9 g/L $K_2HPO_4$). Micro agar and antibiotics were added as per RS102 medium.

PY plates used for conjugation contain 1% w/v Bacto-Peptone, 0.05% w/v yeast extract, 2mM $MgCl_2$, 2mM $CaCl_2$, 0.04mM $FeSO_4$, 15% w/v Micro Agar.

## 5.2.2 Constructs

Plasmids used in this study are listed in **Table 5.2**. Two constructs were used for the RNA-seq analysis of differential expression: pBBR-K-PcrtE-trxValCopt, containing kanamycin resistance and *Callitropsis nootkatensis* valencene synthase (CnVS) and pBBR1MCS-2 (Addgene #85168) as control[33].

| Strain | Description | Reference |
|---|---|---|
| *Escherichia coli* | | |
| DH5α | Used for assembly and plasmid amplification. fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17 | Laboratory stock |
| S17-1 | Host strain for transconjugation. *thi pro recA hsdR* [RP4-2Tc::Mu-Km::Tn7] $Tp^r$ $Sm^r$ | Laboratory stock |
| *Rhodobacter sphaeroides* | | |
| 241 | Wild type | Laboratory stock |
| *ΔIlvR* | 241 *ΔIlvR* (RSP_1297) | This study |
| Δ*PstR* | 241 ΔPstR (RSP_3339) | This study |

**Table 5.1** Strains used in this study

## 5.2.3 Generation of plasmids for gene knock out

Constructs to knock out selected TRs in *Rs241* were designed to include homologous regions for the target genes and one single guide, using as template vector the previously described pBBR_Cas9_NT plasmid[7]. Two TRs were identified for knockout, a LysR-family TR (RSP_1297, *IlvR*) and a GntR-family (RSP_3339, *PstR*). One construct was generated for each target gene and the spacer was selected by using the Benchling CRISPR tool[35]. Primers for plasmid assembly were designed containing at least 20 bases overhang. Primers LYS294 and LYS293 contained the spacer targeting RSP_1297, while GNT302 and GNT301 the spacer for RSP_3339. The homologous regions for the recombination were designed to be at least 1 kb. Alongside the TRs, we attempted to generate a KO of RSP_1559, annotated as NADP-dependent isocitrate dehydrogenase (*icd*). The assembly scheme is reported in **Table 5.2**. All primers used were purchased from Integrated DNA Technologies (IDT) and are listed in **Table S5.1**. Q5 High-Fidelity DNA polymerase (NEB) was used for cloning, while DreamTaq Green PCR Master Mix (2x) (Thermo Fisher) was used for colony PCR, following manufacturer protocols. All PCR reactions performed on genomic DNA from *R. sphaeroides* as template were supplemented with 3% v/v DMSO due to the high GC content. PCR fragments amplified for plasmid assembly were purified from gel using the Zymoclean Gel DNA recovery kit (Zymo Research) and assembled using the HiFi Assembly kit (NEB). Plasmids were isolated using QIAprep Spin Miniprep Kit (QIAgen) and sequenced by Macrogen Europe.

### 5.2.4 Growth conditions for *Rs241*

Glycerol stocks of *Rs241* were streaked on RS102 agar plates with 100 μg/mL kanamycin and incubated at 30°C for 24-48 hours. Single colonies were picked and grown as precultures in 10 mL RS102 with 100 μg/mL kanamycin in 100 ml flasks. Subsequently, cultures were diluted to an OD600 of 0.05 in 20 mL of the same medium. 10% v/v n-dodecane (≥99% purity, Alfa Aesar, Fisher scientific, Benelux) or 10% v/v n-dodecane with valencene 0.4mM (80% purity, Isobionics BV, Netherlands) was added to specific cultures, while other production cultures were grown without second phase (**Table 5.3**). Triplicate experiments were performed for all conditions. Growth was followed by measuring OD600 until 72 hours. Cultures were sampled after 48 hours (and 72 hours for HPLC analysis) of growth at 30°C, 180 rpm shaking. The dodecane layer was recovered via centrifugation.

### 5.2.5 RNA isolation

Samples were taken at 30 hours and OD600 was measured to determine the cell count of 0.5 mL cultures. 20-30mg of dry weight pellet was measured for extraction. Samples were centrifuged at full speed for 30 seconds to remove the culture medium, then an equal volume of cold MeOH (-80°C) was added to the pellets, re-suspended and centrifuged for 1 min at full speed. The supernatant was removed, and the pellets frozen in liquid nitrogen. 1.5 mL of RLT buffer (QIAgen) with 15μL of β-mercaptoethanol was added to each sample and the mix was sonicated 4 times (10s followed by 10s break; amplitude 12). For purification, RNeasy mini kit (QIAgen) was used, following manufacturer protocol, including a second purification step. RNA concentration was measured with NanoDrop One (Thermo Scientific). Finally, a DNase treatment was performed using Invitrogen DNase I, amplification grade, using 1μL DNase I (Invitrogen) per 1μg RNA, following the manufacturers protocol. Finally, prior to sequencing, the NEBNext® rRNA Depletion Kit (Bacteria) (NEB) was used for rRNA depletion, following the manufacturer protocol.

| Plasmid | Description | Reference |
|---|---|---|
| pBBRMCS-2 | Empty vector, used as control. Kan resistance | Laboratory stock |
| pBBR-K-PcrtE-trxValCopt | Expression vector: *CnVS.* Kan resistance | Isobionics BV |
| pBBR_Cas9_NT | Empty vector: contains cas9 gene codon harmonized for R. sphaeroides. Kan resistance | Mougiakos *et al.* 2019 |

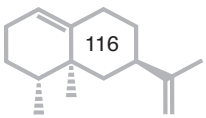
Assembly scheme plasmids for KO of Rs241, based on pBBR_Cas9_NT

| Plasmid | Fragment ID | Primer ID | Fragment size (bp) | Spacer | Reference |
|---|---|---|---|---|---|
| pBBR_Cas9_*ΔIlvR* | Cas9_L | LYS294 + CAS007 | 4622 | tgacgatcaa tctcgcgacg | This study |
| | Kan_L | CAS006 + LYS293 | 4875 | | |
| | HR_UP_L | LYS272 + LYS273 | 1160 | | |
| | HR_DW_L | LYS274 + LYS275 | 1157 | | |
| pBBR_Cas9_*ΔPstR* | Cas9_G | GNT302 + CAS007 | 4622 | gagctgggc cacgaagtcg | This study |
| | Kan_G | CAS006 + GNT301 | 4875 | | |
| | HR_UP_G | GNT276 + GNT277 | 1092 | | |
| | HR_DW_G | GNT278 + GNT279 | 1140 | | |
| pBBR_Cas9_*Δicd* | Cas9_D | ICD282 + CAS007 | 4622 | cctctcgaga tagcccatcg | This study |
| | Kan_D | CAS006 + ICD281 | 4875 | | |
| | HR_UP_D | ICD268 + ICD269 | 1100 | | |
| | HR_DW_D | ICD270 + ICD271 | 1076 | | |

**Table 5.2** Plasmids used in this study

| Strain | Construct | Valencene | Dodecane | Abbreviation |
|---|---|---|---|---|
| *Rs241* | pBBR-K-PcrtE-trxValCopt | Producing | Yes | CnVS(D) |
| *Rs241* | pBBR-K-PcrtE-trxValCopt | Producing | No | CnVS |
| *Rs241* | pBBR1MCS-2 | No | No | MCS |
| *Rs241* | pBBR1MCS-2 | No | Yes | MCS(D) |
| *Rs241* | pBBR1MCS-2 | 0.4mM valencene in dodecane | Yes | MCS(val) |

**Table 5.3** Growth conditions for RNA-seq experiment



## 5.2.6 RNA-seq analysis and data processing

Isolated RNA was sequenced on an Illumina NextSeq 550. The reference genome used for the RNA-seq mapping is *Cereibacter sphaeroides* 2.4.1 (previously annotated as *Rhodobacter sphaeroides* 2.4.1), assembly GCA_000012905.2, comprising of 7 replicons (**Table 5.4**). Kallisto[36] was used to map reads to the genome and transcript levels were quantified as TPM values (transcript per million) based on Kallisto pseudoalignments. Around 100 million reads per sample from RNA sequencing were mapped[36]. Only 5-10% of the reads in each sample mapped to genes, while the remaining reads mapped to ribosomal RNA, indicating that the ribosomal depletion step was not successful. Only reads which did not map to ribosomal RNA were considered in further analysis steps, resulting in 3-10 million reads per sample. Counts were processed in the SLEUTH[37] RNA-seq differential analysis framework to determine differential transcript expression between samples. Principal component analysis (PCA) was performed using SLEUTH. The Galaxy web tool was used for the generation of heatmaps and data visualization[38,39]. Data was plotted using the Z-score calculated on the average TPM values per gene for each condition using the formula:

$$Z\ score(g) = \left( \frac{TPM(g) - average_{\ TPM\,(all\ samples)}}{st.dev_{TPM\ (all\ samples)}} \right)$$

Here, "*g*" represents the specific gene in analysis for each experimental condition. STRING[40] was used to predict functional interaction networks for genes of interest; these were visualized using Cytoscape[41].



## 5.2.7 RT-qPCR analysis of DEGs

To verify results obtained with RNA-seq analysis, primers were designed to amplify a selected subset of genes (**Table S5.2**). 500ng RNA was used for cDNA synthesis. iScript cDNA synthesis kit (Biorad), was used, following the manufacturer protocol. For qPCR reaction, qPCRBIO SyGreen Blue Mix Lo-Rox (Sopachem) was used. Forward and reverse primers were used at a final concentration of 400nM, while cDNA was diluted at 1:4 or 1:8, for a final reaction volume of 10µL. Primers were tested using a cDNA pool in dilutions 1:4, 1:8, 1:16, 1:32 and 1:64. The relative expression level was calculated using the $2^{-\Delta\Delta Ct}$ method[42] and *rpoD* (RSP_0395) was used as reference gene. The log10 of the $2^{-\Delta\Delta Ct}$ was calculated for a clearer graphical representation.

## 5.2.8 Conjugation and plasmid curation

Wild type *Rs241* was inoculated from glycerol stocks or from RS102 agar plates in 10 mL RÄ medium and grown for 48 hours at 30°C and 200 rpm. OD600 was measured and a volume corresponding to OD600 0.1 was inoculated in 10 mL fresh RÄ medium. Cells were harvested after 24 hours, at an OD600 of 3. In parallel, *E. coli* S17-1 harbouring the plasmid of interest was grown overnight at 37°C, 200 rpm in LB medium supplemented with 50µg/mL kanamycin. *E. coli* S17-1 was then inoculated in fresh LB medium with kanamycin and grown until the OD600 reached 1. Cultures were washed twice with 1 mL RÄ medium, mixed with *Rs241* at a ratio of 1:1 and centrifuged for 1 minute at maximum speed. 100µL RÄ was added to the pellet and the suspension was transferred to a 0.22 µm 47 mm diameter nitrocellulose filter disc (Sigma-Aldrich) on a PY plate and incubated at 28°C for 6 hours. The conjugation mixture was resuspended in 2 mL RÄ, then 100µL were plated on RÄ plates supplemented with kanamycin and incubated at 30°C for 3 days or until colonies appeared. Colonies were picked and plated on a new RÄ plates with kanamycin for strains harbouring pBBRMCS-2 and pBBR-K-PcrtE-trxValCopt, or on RÄ plates without antibiotics for knockout constructs. Colonies were further plated on RS102 medium until no *E. coli* was present. Colony PCR was performed to confirm the absence of *E. coli*.



KO colonies were verified via PCR and inoculated in parallel in RS102 medium with and without antibiotics. Colonies that were only able to grow without antibiotics were considered cured of the plasmid. A second round of conjugation was performed on the obtained KO to introduce pBBRMCS-2 and pBBR-K-PcrtE-trxValCopt.

## 5.2.9 GC-MS analysis of valencene produced

The valencene in the dodecane layer was quantified by GC-MS analysis, as previously described[43]. Briefly, a 7980A GC system combined with a 597C inert MSD detector (70eV) (Agilent Technologies, The Netherlands) was used. The system is equipped with a 7683 auto-sampler and injector and a Phenomenex Zebron ZB-5ms column of 30 m length x 0.25 mm internal diameter and 0.25 µm stationary phase, with a Guardian precolumn (5 m). For each sample, 1µL of sample was injected. The injection chamber was set at 250 °C, injection was splitless, and the ZB5 column was held at 55 °C for 2 minutes. A gradient of 10 °C per minute to 300 °C was then started. Peaks were detected in the chromatograms of total ion count. Valencene was identified and quantified by comparing the sample peak with a

valencene standard (80% purity, Isobionics, The Netherlands). The data obtained were expressed as corrected area units. The average and standard deviation of the replicates were calculated using Microsoft Excel functions.

### 5.2.10 HPLC analysis of isoprenoids

*Rs241* cultures for HPLC analysis were grown following the same growth conditions described above. Samples were taken at 48 and 72 hours. Cultures were centrifuged in 50 mL Greiner tubes and washed twice with milliQ water to remove traces of culture medium. 12 mL disposable glass tubes with screwcap (DWK life science) were weighted and used for the extraction. The biomass was suspended in 500 µL milliQ and transferred to the glass tubes, frozen in liquid nitrogen and lyophilized for 48 to 72 hours. The biomass was weighted before starting the extraction.

Extraction and analysis of non-polar metabolites was performed as described by Fraser *et al.* (2000)[44] and Bino *et al.* (2005)[45] with modifications. 4.5 mL extraction solvent, methanol/chloroform 5:4 containing 0.1% butylated hydroxytoluene (BHT), containing 0.3 µg/mL Sudan 1 as internal standard was supplement to the freeze-dried biomass in glass vials. After 10 min incubation on ice and 10 min sonication, 2.5 ml Tris-HCl pH7.4 was added and incubated on ice for 10 min after vortexing. Phases were separated by centrifugation at 1800 rcf for 10 min and 1.8 ml of the chloroform phase transferred to a new 12 mL glass vial. The extraction was repeated twice with 1 ml of chloroform containing 0.1% BHT and the 3.8 ml collected chloroform evaporated in a heating block at 35 °C under a nitrogen stream. Dried pellets were solubilized in 500 µl ethyl acetate containing 0.1% BHT and centrifuged at max speed for 10 minutes. 2 mL amber crimp vials were used for the HPLC analysis and 10 µl injected for HPLC-PDA analysis.

Relative and absolute quantification of isoprenoids was performed using a E2695 Separations module (Waters) with a YMC Carotenoid column (4.6 mm ID, S-5µm, 250 x 4.6 mm) and pre column and Waters 2996 PDA module. The column and autosampler were maintained at 35 °C and 15 °C, respectively. A flowrate of 1 mL/min was maintained using mobile phases consisting of methanol (A), t-butyl-methyl-ether (B), and methanol:water 4:1 with 0.2 % ammonium acetate (C). Isocratic elution with 95% A and 5% B was maintained for 12 min where after a linear gradient was used increasing from 80% A, 15% B, and 5% C to 30% A, 65% B, and 5% over 18 min whereafter it was maintained for 10 min and returned to 95% B and 5% C in 1 min and maintained for 4 min. Data analysis was performed using

Empower 3 software (Waters). Area units were corrected for biomass dry weight used for extraction.

Corrected area units were calculated as

$$y = \frac{A \times d}{z}$$

where $y$ = corrected area units; $A$ = peak area; $d$ = dilution; $z$ = mg dry weight (biomass). When possible, μg of compound per mg dry weight were calculated as:

$$x = \left( \frac{\left(\frac{y}{m}\right)}{2} \right) \times 1000$$

where $x$ = μg compound/mg dry weight and $m$ = slope of standard curve. The factor ($y/m$) is divided by 2 as the standards were diluted in 1mL whereas the samples were diluted in 0.5 mL. Standard curves for the different compounds were obtained using:

- Coenzyme Q10 (CoQ10) (≥98%, Acros Organics), retention time ~23.4 minutes and a maximum absorbance at 270 nm;
- β-Carotene (≥95%, Sigma Aldrich), retention time of ~28.3 minutes and maximum absorbance between 466nm and 497 nm;
- Bacteriochlorophyll a from *Rhodopseudomonas sphaeroides* (Partially purified, Sigma Aldrich), retention time ~13.03 minutes, absorbance peaks at 395nm, 441nm, 685 nm.

Different dilutions were used for the different compounds (**Table S5.3**). Average and standard deviation were calculated for replicates. Compounds produced in *Rs241* were identified based on the standards available. Calculation of mass produced per dry weight of biomass (μg / mg dw) was obtained using the standard curves reported in **Table S5.3** for CoQ10 and chlorophylls. The standard for carotenoids was used for identification but not for quantification, as the standard curve was not linear. Sphaeroidene and sphareoidenone were identified by comparing with the compounds described by Ashikhmin *et al.* (2017)[46].

## 5.3 Results

In this study we identified differentially expressed genes (DEGs) in wild type *Rs241* between different growth conditions. We studied variation in gene expression between *Rs241* producing valencene and *Rs241* harbouring an empty vector. Five conditions were tested: *Rs241* + pBBR-K-PcrtE-trxValCopt with and without n-dodecane as second phase, indicated as CnVS(D) and CnVS respectively; *Rs241* + pBBR1MCS-2 with and without n-dodecane, indicated as MSC(D) and MCS; and *Rs241* + pBBR1MCS-2 with 0.4mM valencene in n-dodecane, MCS(val) (**Table 5.3**). MCS, MCS(D) and MCS(val) were considered control conditions and the differential expression is always intended between CnVS, CnVS(D) and all the other conditions.

In order to assess the consistency of the growth of *Rs241* in the different conditions, the optical density at 600nm absorbance (OD600) was measured through 72h growth. No major difference was observed between the growth curves of strains (**Figure 5.1A**). The production of valencene in the dodecane fractions was measured with GC-MS. At 48 hours growth, the production of valencene appears to reach its plateau (**Figure 5.1B**), although the OD600 measured showed biomass still increasing at 72 hours. With these data, we decided to use the 30 hours timepoint for the sampling of the RNA for expression analysis.

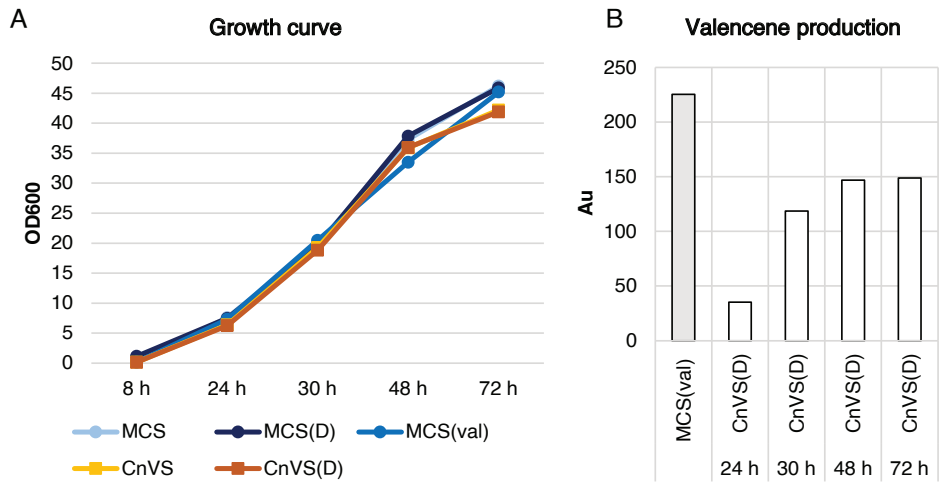



**Figure 5.1** (A) Growth curve expressed in optical density absorbance at 600nm (OD600) and (B) valencene production of *Rs241* in different growth conditions, expressed in corrected area units (Au). The exponential growth phase is still active at 30h, while the production of valencene reaches its plateau at 48h.

## 5.3.1 Gene expression analysis

RNA-seq data were mapped to the *Cereibacter sphaeroides* 2.4.1 reference genome. 3-10 million reads per sample mapped to genes (**Table 5.4**). Expression was quantified and differential expression between samples was assessed[37] (**Tables S5.4, S5.5**). By comparing gene expression data between CnVS samples and MCS samples, we observed that, in CnVS(D) and CnVS, 86 genes of the 113 on plasmid D do not appear to be expressed in the CnVS strain, as no transcript mapped to those genes (**Table S5.6**). To avoid these genes dominating the Principal Component Analysis (PCA), masking the relevant differences in other genes, we decided to not include them in the analysis. The PCA showed that sample replicates generally had less variation in expression compared to different sample groups, providing confidence in the results (**Figure 5.2**). The main direction of variation (PC1) related to production of valencene, as samples CnVS and CnVS(D) can be found on the left side of the graph, separated from MCS, MCS(D) and MCS(val). PC2 appeared to relate to the presence of dodecane, as a separation can be observed between samples CnVS(D), MCS(D) and MCS(val), positioning in the top half of the graph, while CnVS and MCS positioned in the bottom half.

| ***R. sphaeroides* genomic elements** | **RefSeq** | **GenBank** | **Length (bp)** | **Genes** | **DEGs** |
|---|---|---|---|---|---|
| Chromosome 1 | NC_007493.2 | CP000143.2 | **3188524** | 3052 | 54 |
| Chromosome 2 | NC_007494.2 | CP000144.2 | **943018** | 849 | 47 |
| Plasmid A | NC_009007.1 | DQ232586.1 | **114045** | 103 | 4 |
| Plasmid B | NC_007488.2 | CP000145.2 | **114179** | 104 | 7 |
| Plasmid C | NC_007489.1 | CP000146.1 | **105284** | 107 | 2 |
| Plasmid D* | NC_007490.2 | CP000147.2 | **100827** | 113 | 2 |
| Plasmid E | NC_009008.1 | DQ232587.1 | **37100** | 41 | 2 |
| | | | Total | 4369 | 98 |

***86 not expressed in CnVS and CnVS(D)**

**Table 5.4** Genomic organization of *R. sphaeroides*, including RefSeq and GenBank numbers, genomic length, number of genes and identified DEGs.

Gene expression was compared between all conditions and the log fold change was calculated (**Table S5.4**). Genes were considered differentially expressed between any two experimental conditions when the calculated log fold change was higher than 1.2 or lower than -1.2, with a false discovery rate (FDR) lower than 0.05. With this threshold, we identified 98 DEGs (**Table S5.5**).

We observed that 9 genes were differentially expressed between MCS and MCS(D), while only 6 genes were differentially expressed between CnVS and CnVS(D), indicating that the presence of dodecane does not strongly impact the gene expression in *Rs241* (**Table S5.4**). We identified four patterns of differential expression through the different conditions: genes upregulated in both CnVS(D) and CnVS (**Figure 5.3A**); downregulated in both CnVS(D) and CnVS (**Figure 5.3B**); upregulated only in CnVS (**Figure 5.3C**); and upregulated only in CnVS(D) (**Figure 5.3D**). Among the genes upregulated in CnVS(D) and CnVS (**Figure 5.3A**), thioredoxin A (trxA, RSP_1529) and the NADP-dependent isocitrate dehydrogenase (icd, RSP_1559, EC 1.1.1.42) displayed strong regulation (**Figure S5.1**). The upregulation of RSP_1442, a carbohydrate ABC transporter substrate-binding protein (CUT1 family), might have a role in improving glucose supply[47].

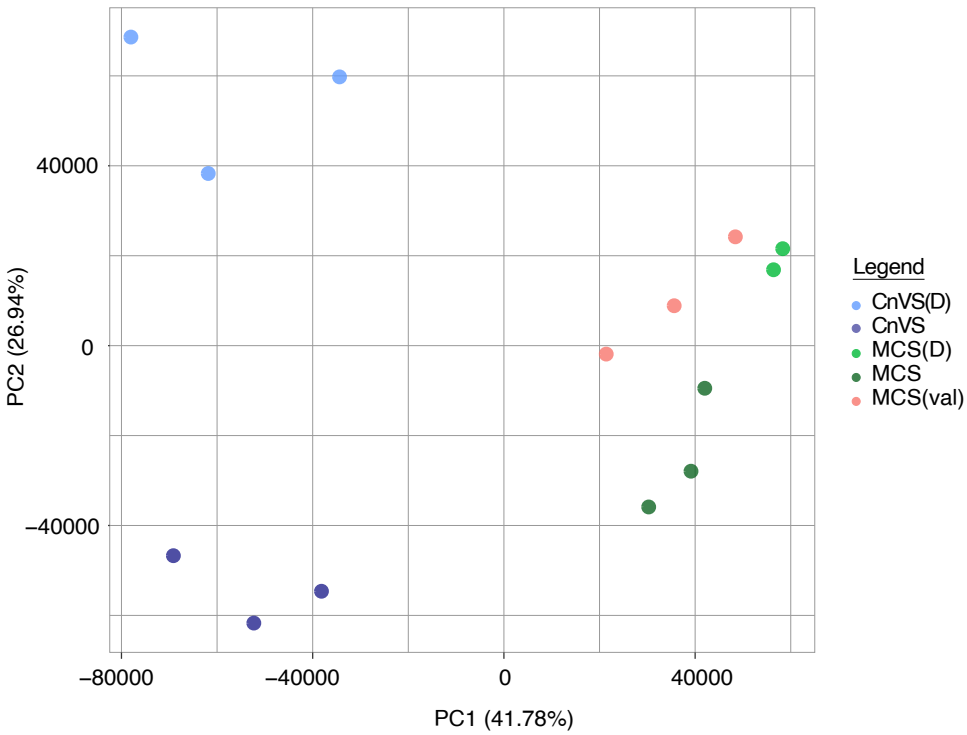



**Figure 5.2** PCA plot of RNA-seq results. PC1 represents the presence of the valencene synthase (CnVS), showing a clear separation of the stains producing valencene (CnVS and CnVS(D)) and the non-producing controls. PC2 represents the presence of dodecane, showing separation between the samples grown with dodecane (CnVS(D), MCS(D) and MCS(val)) and the samples grown without dodecane (CnVS and MCS). The percentage of variance is reported for PC1 and PC2.

## 5.3.2 Identification of transcriptional regulators associated with DEGs

The analysis of the DEGs was mainly focused on the identification of TRs that may be involved in the sensing of valencene. We hypothesized that TRs might be involved in the regulation of the neighbouring operons containing the DEGs. Therefore, our attention was directed towards the TRs localised in the neighbourhood of the identified DEGs We identified 7 TRs neighbouring clusters of DEGs (**Table S5.7**). One of the identified transcriptional regulators, RSP_1297, is associated with an operon downregulated in CnVS(D) and CnVS compared to all other conditions (**Figure 5.3B**). RSP_1297 is annotated as a LysR-family TR, but it has not been yet characterized. It co-localizes with four genes that are potentially involved in the degradation of amino acids such as valine, leucine and isoleucine and fatty acid degradation, among other metabolic pathways (**Table 5.5**). RSP_1297 is localized upstream of the operon, on the complementary strand (**Figure 5.4A**). Functional interactions of RSP_1297 were predicted with STRING[40] (**Figure 5.4B**), and we observed that RSP_1297 is predicted to interact with another LysR family TR, RSP_2965. Any co-expression of RSP_1297 and RSP_2965 could not be observed in our analysis: RSP_2965 was not observed to follow the same expression pattern as RSP_1297 in CnVS and CnVS(D) (**Figure S5.2**). We propose for RSP_1297 the name of *IlvR* for its association with genes involved in the degradation of valine, leucine and isoleucine[48].



A second transcriptional regulator, RSP_3339, co-localizes and co-expresses with an operon showing upregulation uniquely in CnVS (Figure 5.3C). This operon is predicted to encode for the ABC spermidine/putrescine transporter (Figure 5.4C), while RSP_3339 is annotated as a GntR-family TR. The interactions of RSP_3339 were predicted with STRING, which yielded 10 genes, including the 4 in its genomic neighbourhood (Figure 5.4D). Four more GntR family TRs are indicated as co-occurring based on the STRING network. However, in the RNA-seq analysis, we did not observe any differential expression for any of these TRs; we therefore focused our attention on RSP_3339. Because of its proximity to the spermidine/putrescine ABC transporter genes, we decided to name it "putrescine/spermidine transporter regulator", PstR.

To confirm the expression data, we chose two genes adjacent to each of the TRs and tested their expression by RT-qPCR. For IlvR (RSP_1297), RSP_1294, RSP_1295 were tested, and for PstR (RSP_3339), RSP_3336 and

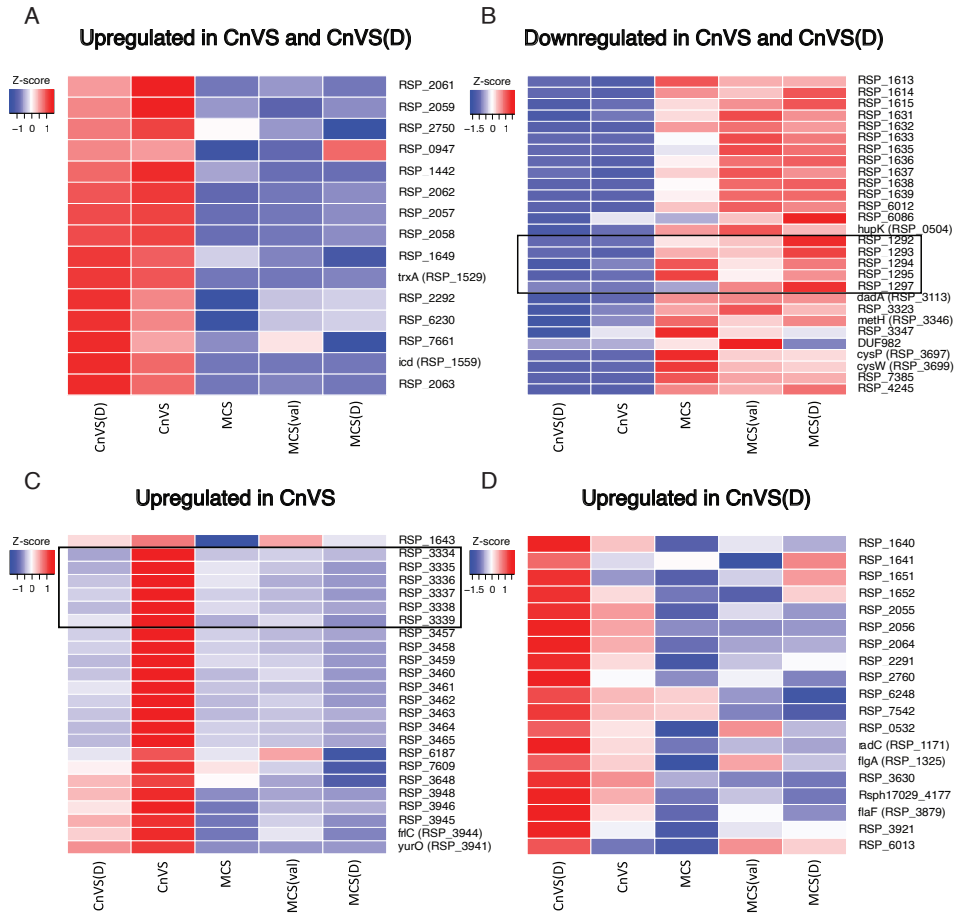



**Figure 5.3** Heatmaps of DEGs. The heatmaps are obtained using the Z-scores calculated for each condition compared to the others. (A) Genes upregulated in CnVS and CnVS(D). *trxA* and *icd* can be found in this heatmap. (B) Genes downregulated in CnVS and CnVS(D). The operon containing RSP_1297 (LysR-family TR) is found in this heatmap (black box). (C) Genes upregulated only in CnVS. The operon containing RSP_3339 (GntR-family TR) is found here (black box). (D) Genes upregulated only in CnVS(D).

RSP_3337. The qPCR results confirmed that RSP_1294 and RSP_1295 were downregulated in CnVS and CnVS(D) (Figure 5.5A-B, D-E), while RSP_3336 and RSP_3337 were selectively upregulated only in CnVS (Figure 5.6A-B, D-E). As these transcriptional regulators appeared to be involved in the different gene expression when valencene was produced, we decided to investigate their role in the regulation by generating KO of RSP_1297 and RSP_3339 in Rs241.

| Gene ID | Gene name | Symbol | EC number | Metabolic pathways |
|---|---|---|---|---|
| RSP_1292 | Putative Aldehyde dehydrogenase | ALDH | EC 1.2.1.3 | Glycolysis/Gluconeogenesis, Fatty acid degradation, Amino acid metabolism, Limonene and pinene degradation |
| RSP_1293 | ACP S-malonyltransferase | APC | EC 2.3.1.39 | Fatty acid biosynthesis |
| RSP_1294 | branched-chain alpha-keto acid dehydrogenase E1 component | bkdA | EC1.2.4.4 | Val, leu and ile degradation, Propanoate metabolism |
| RSP_1295 | Putative acyl-CoA dehydrogenase: glutaryl-CoA dehydrogenase | gcdH | EC 1.3.8.6 | Fatty acid degradation, Lys degradation, Trp metabolism, Benzoate degradation |
| **RSP_1297** | **LysR family transcriptional regulator** | **IlvR** | | |
| RSP_3334 | aspartate/glutamate racemase family protein | hpxA | EC 5.1.99.3 | Purine metabolism |
| RSP_3335 | ABC transporter permease | ABC.SP.P | | |
| RSP_3336 | ABC spermidine/putrescine transporter, inner membrane subunit | ABC. SP.P1 | | |
| RSP_3337 | extracellular solute-binding protein | ABC.SP.S | | |
| RSP_3338 | ABC transporter ATP-binding protein | ABC.SP.A | | |
| **RSP_3339** | **GntR family transcriptional regulator** | **PstR** | | |
| RSP_3341 | Rrf2 family transcriptional regulator, nitric oxide-sensitive transcriptional repressor | nsrR | | |

**Table 5.5** Identified DEGs associated with TRs. Gene name, symbol, EC number (when available) and metabolic pathways are reported.

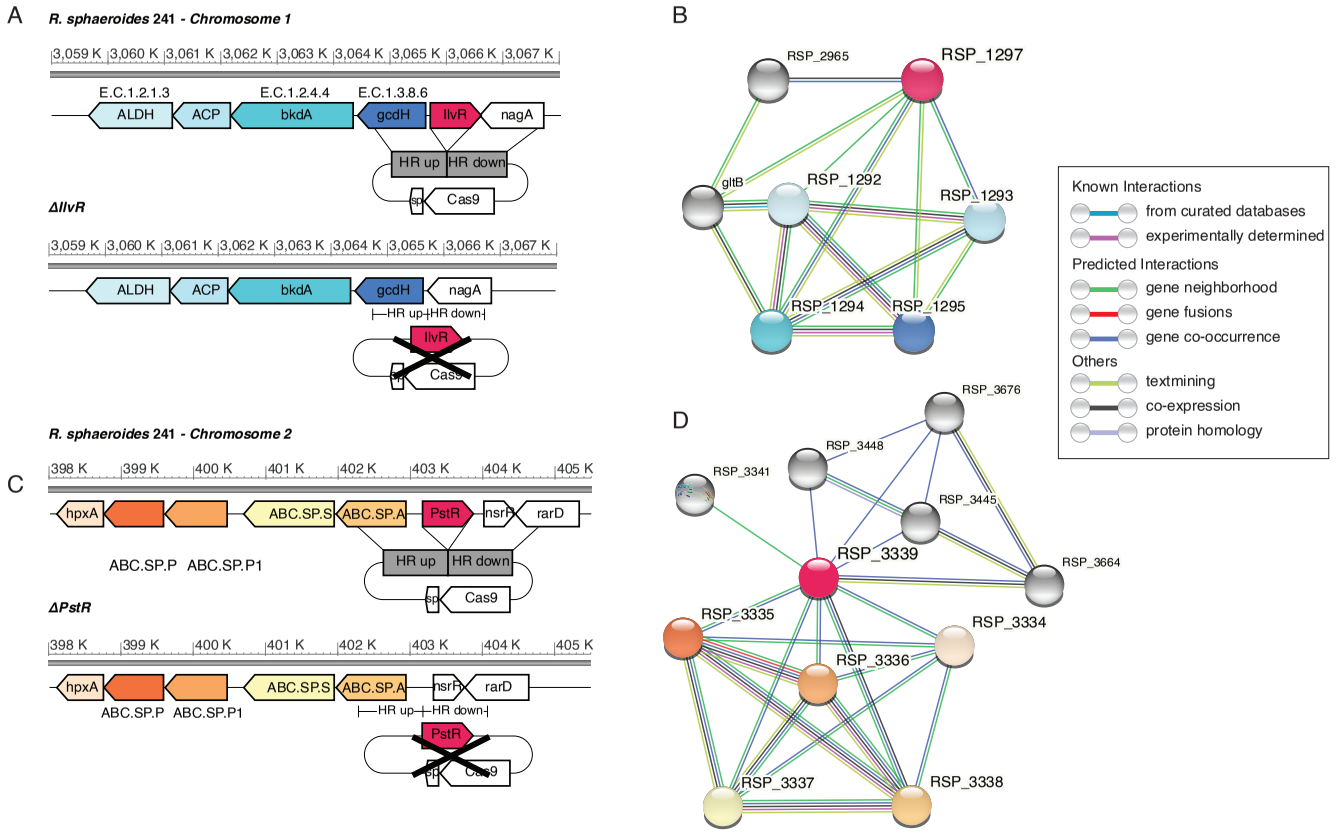



**Figure 5.4 (A) Genomic organization of** *R. sphaeroides 241* RSP_1297 *(IlvR)* operon in comparison with the KO *ΔIlvR* obtained using the CRISPR-Cas9 counter selection method; and (B) its STRING network. (C) Genomic organization of RSP_3339 *(PstR)* operon in comparison with the KO *ΔPstR* obtained using the CRISPR-Cas9 counter selection method; and (D) its STRING network. Colours used for the genomic organization are maintained in the STRING networks.

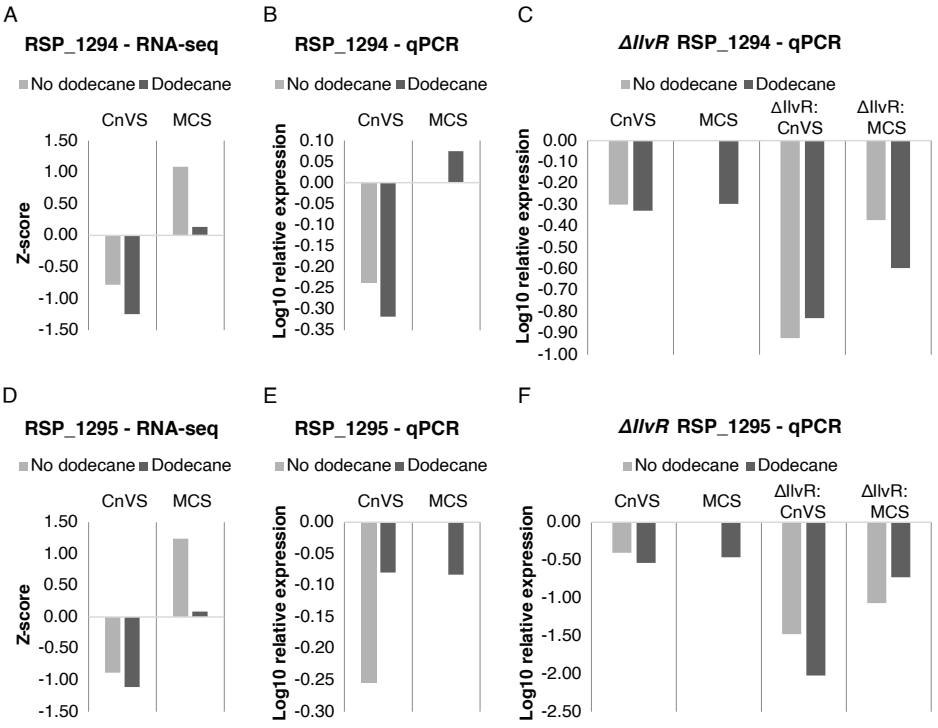



**Figure 5.5** RNA-seq and RT-qPCR analysis of IlvR-regulated genes (A-C) RSP_1294 and (D-F) RSP_1295. (A;D) RNA-seq results for wild type *Rs241* harbouring a valencene synthase (CnVS) or an empty plasmid (MCS). (B; E) RT-qPCR results. (C; F) gene expression in wild type *Rs241* harbouring CnVS or MCS in comparison with KO strain *ΔIlvR* harbouring CnVS or MCS. RNA-seq analysis revealed downregulation of (A) RSP_1294 and (D) RSP_1295 in wild type harbouring CnVS, both in presence and absence of dodecane. RT-qPCR results (B; E) confirmed the result. For both (C) RSP_1294 and (F) RSP_1295 we observe a further downregulation in the knockout strains compared to the wild type of both strains harbouring CnVS and MCS. In all qPCR graphs, MCS (no dodecane) is used as control for the calculation of the ΔΔCt, rendering its relative equal to 0.

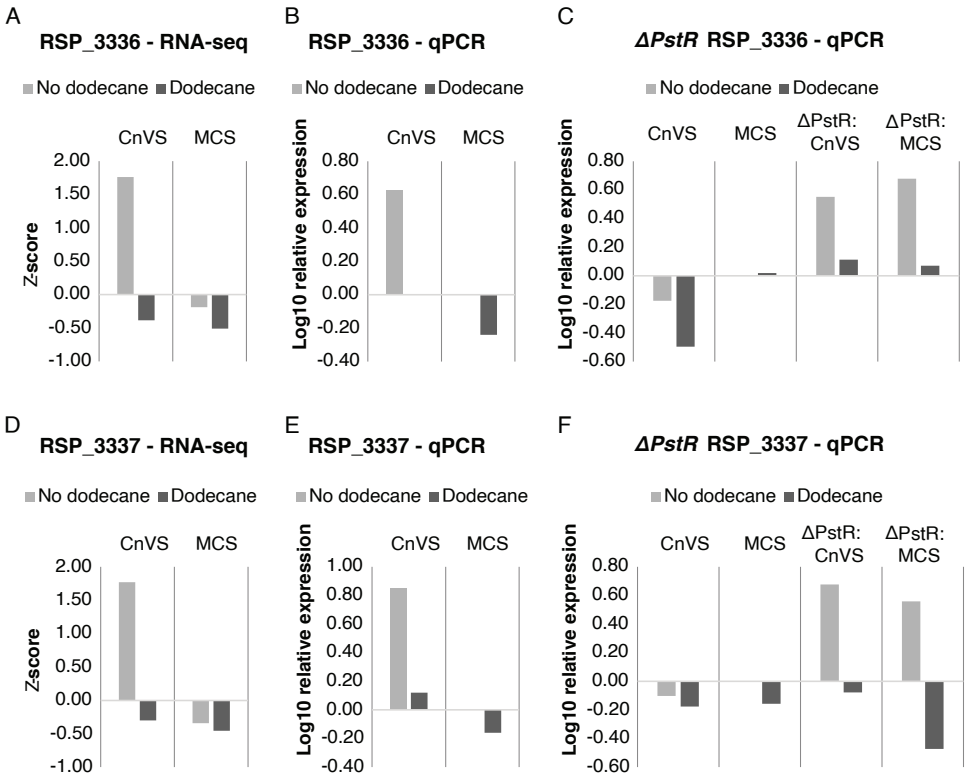



**Figure 5.6** RNA-seq and RT-qPCR analysis of PstR-regulated genes (A-C) RSP_3336 and (D-F) RSP_3337. (A;D) RNA-seq results for wild type *Rs241* harbouring a valencene synthase (CnVS) or an empty plasmid (MCS). (B; E) RT-qPCR results. (C; F) gene expression in wild type *Rs241* harbouring CnVS or MCS in comparison with KO strain *ΔPstR* harbouring CnVS or MCS. RNA-seq analysis revealed upregulation of (A) RSP_3336 and (D) RSP_3337 in wild type harbouring CnVS only in the absence of dodecane. RT-qPCR results (B; E) confirmed the result. For both (C) RSP_3336 and (F) RSP_3337 we observed upregulation in the knockout strains compared to the wild type of both strains harbouring CnVS and MCS, in the absence of dodecane. The previously observed upregulation in wild type CnVS was not observed in the second experiment. In all qPCR graphs, MCS (no dodecane) is used as control for the calculation of the ΔΔCt, rendering its relative equal to 0.

### 5.3.3 Generation of *Rs241* knock out strains

To knock out genes RSP_1297, RSP_3339 and RSP_1559 from *Rs241*, and obtain KO strains *ΔIlvR, ΔPstR* and *Δicd* respectively, we employed the *Rhodobacter* optimized Cas9 counter selection method[7] (**Figure 5.4**). We aimed to remove the entire open reading frame of each gene. The KO strains were verified by PCR (**Figure 5.7**) and sequencing, confirming the deletion of the genes as described in the Materials and Methods. While we could obtain *ΔIlvR* and *ΔPstR,* we could not obtain a viable KO strain for *Δicd*.

### 5.3.4 Gene expression in *ΔIlvR* and *ΔPstR*

Both *ΔIlvR* and *ΔPstR* strains were transformed with the pBBR-K-PcrtE-trxValCopt plasmid and with the pBBR1MCS-2 plasmid. With these strains, the effect of valencene synthase expression on operon gene expression was tested (**Table 5.6**). Strains were cultivated under the same conditions previously applied for the RNA-seq analysis, and all six strains were cultivated with and without n-dodecane. Samples were harvested at 30 hours of growth, for comparison with the data obtained through the RNA-seq analysis. For *ΔIlvR,* we tested expression of RSP_1294 and RSP_1295; for *ΔPstR,* we tested RSP_3336 and RSP_3337, as previously done to confirm the RNA-seq data. For the operon associated with IlvR, we observed for both *ΔIlvR*:CnVS and *ΔIlvR*:MCS a lower expression of genes RSP_1294 and RSP_1295, compared to the wild type strains harbouring CnVS or MCS (**Figure 5.6C**). This result suggests that IlvR acts as a transcriptional activator of these genes. For the gene associated with *PstR*, we observed an upregulation in both *ΔPstR*:MCS and *ΔPstR:*CnVS (**Figure 5.6F**) in the absence of dodecane, indicating that this TR might act as a repressor. However, we could not observe selective upregulation of both RSP_3336 and RSP_3337 in wild type harbouring CnVS. Thus, these newly grown cultures did not confirm the result obtained with the RNA-seq. Therefore, the characterization of RSP_3339, and its role in the regulation of the expression of the neighbouring operon requires further experiments.



### 5.3.5 Bacterial growth and valencene production

To characterize the KO strains and the effect of the deletions on the bacterial growth, we performed a 72-hours fermentation experiment, measuring the OD600 at 24, 30, 48 and 72 hours. We observed that, at 72 hours, the OD600 of *ΔIlvR*:CnVS was about 40% of the OD600 of Wt:MCS and Wt:CnVS. Interestingly, *ΔIlvR*:MCS did not show such difference in OD600, showing

a growth curve comparable to those of Wt:MCS and Wt:CnVS (**Figure 5.8A**).

The amount of valencene present in the dodecane was investigated by GC-MS analysis. Despite the lower OD600, the production of valencene was higher in *ΔIlvR* CnVS compared to Wt:CnVS (**Figure 5.8B**). By normalizing the valencene production for the OD600, we found that production of valencene in *ΔIlvR:*CnVS is about 3.4 times higher than the production in Wt:CnVS (**Figure 5.8C**). In contrast, for *ΔPstR:*CnVS we observed no significant differences in growth compared to Wt:CnVS, while valencene production was about 30% lower compared to Wt:CnVS.

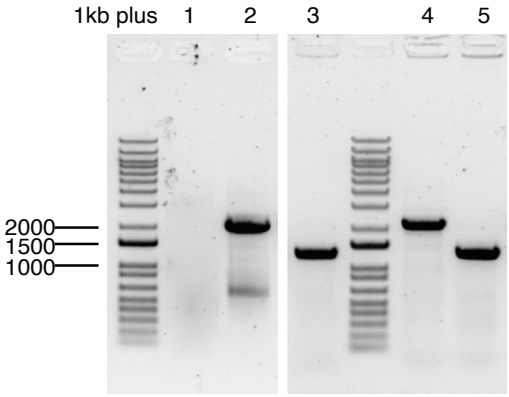



**Figure 5.7** PCR verification of KO strains. From left to right: (1) negative control; (2) wild type *Rs241* using primer pair LYS310 and LYS311 (Table S5.1); (3) KO *ΔIlvR* using same primers as (2). (4) wild type *Rs241* using primer pair GNT308 and GNT309 (Table S5.1); (5) *ΔPstR* using same primers as (4). The smaller bands confirm the deletion of the genes of interest (3 and 5).



| Strain | Construct | Valencene | Abbreviation |
|---|---|---|---|
| *Rs241* | pBBR1MCS-2 | No | Wt:MCS |
| *Rs241* | pBBR-K-PcrtE-trxValCopt | Producing | Wt:CnVS |
| *ΔIlvR* | pBBR1MCS-2 | No | *ΔIlvR*:MCS |
| *ΔIlvR* | pBBR-K-PcrtE-trxValCopt | Producing | *ΔIlvR*:CnVS |
| *ΔPstR* | pBBR1MCS-2 | No | *ΔPstR*:MCS |
| *ΔPstR* | pBBR-K-PcrtE-trxValCopt | Producing | *ΔPstR*:CnVS |

**Table 5.6** Sample table for characterization of KO strains *ΔIlvR* and *ΔPstR.*

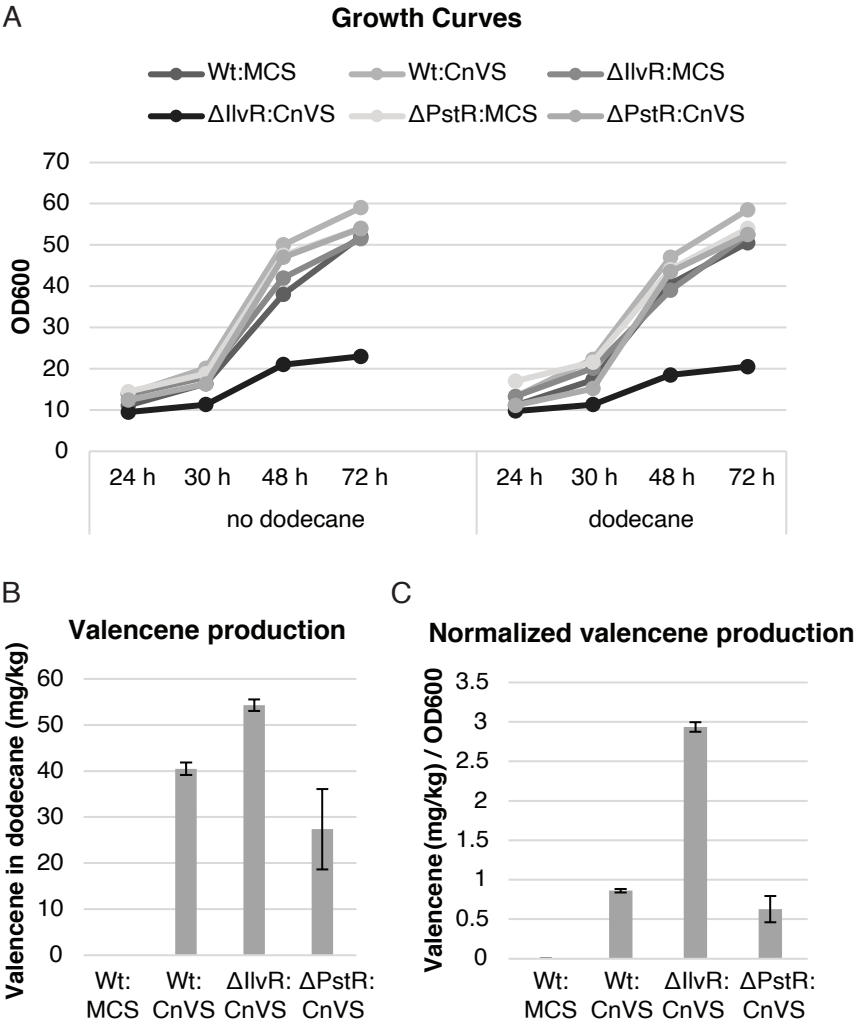



**Figure 5.8** (A) Growth curves over 72 hours of wild type *Rs241* and KO strain *ΔIlvR* and *ΔPstR. ΔIlvR*:CnVS shows decreased growth compared to all other conditions. (B) Valencene production in wild type *Rs241* compared to *ΔIlvR* and *ΔPstR,* all harbouring CnVS. *ΔIlvR*:CnVS shows increased production of valencene compared to both Wt:CnVS and *ΔPstR*:CnVS, while *ΔPstR*:CnVS shows decreased production compared to Wt:CnVS. (C) Normalized valencene production for the measured OD600.

### 5.3.6 Isoprenoids production in *ΔIlvR*

A difference in colour between the pellets of Wt:CnVS, *ΔIlvR*:CnVS and *ΔIlvR*:MCS was observed, possibly related to carotenoid and/or chlorophyll content of the cells (**Figure 5.9**). To assess the production of other isoprenoids in the different strains, we performed HPLC analysis of samples at 48 and 72 hours (**Figure 5.10**). Production of CoQ10 was stable in all samples (Figure 5.10A). Bacteriochlorophyll (BChl; retention time ~7.4 and 13.1 minutes) and bacteriopheophytin (Bphe; retention time ~16.0 and 18.4 minutes) (**Figure 5.10B**) showed higher abundance in Wt:MCS, comparable abundance in Wt:CnVS and *ΔIlvR*:MCS, and a strongly decreased abundance in *ΔIlvR*:CnVS. Carotenoids, sphaeroidenone a and sphaeroidene b (Sph a, Sph b; retention time between 24.2 and 25.9) (**Figure 5.10C**) showed comparable abundance in Wt:CnVS and *ΔIlvR*:MCS, but were significantly decreased in *ΔIlvR*:CnVS.

## 5.4 Discussion

*R. sphaeroides* is a widely studied microorganism, but the architecture of its regulatory circuits is still largely unknown, especially with regard to the effects of heterologous production of terpenes. In this study, we investigated the effects of expression of a valencene synthase (*CnVS)*[33] on transcriptional networks with the aim of identifying TRs with potential sensitivity to this sesquiterpene. N-dodecane is a rather commonly used organic solvent for the extraction of hydrophobic compounds in biphasic fermentations[27,33]. Nonetheless, gene expression was expected to be altered by the presence of n-dodecane in the culture medium, and in the presence of both n-dodecane and valencene. However, no obvious changes were observed between wild type strains growing with or without n-dodecane, nor between stains grown with n-dodecane and valencene.

In contrast, strong changes in transcription were observed when *CnVS* was expressed in *R. sphaeroides*. In strains carrying *CnVS*, 86 genes belonging to plasmid D appeared to be not expressed (**Table S5.6**). Possibly this lack of expression is due to absence of part of this endogenous plasmid, but this would need to be further confirmed. It is important to keep in mind that, if pD is in fact largely absent, it might affect the transcription or the regulation of other genes. A number of operons located on the genome showed a different transcription pattern in the presence of *CnVS*. Notably, *trxA* and *icd* were upregulated in the presence of *CnVS*. TrxA is

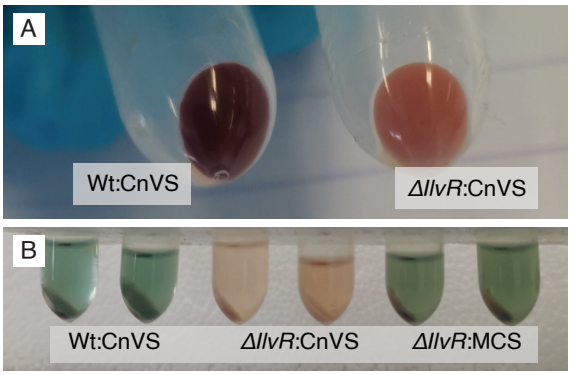



**Figure 5.9** (A) Bacterial pellets of Wt:CnVS compared to *ΔIlvR*:CnVS. Equal biomass was used for the comparison. *ΔIlvR*:CnVS shows decreased pigmentation. (B) Isoprenoid extraction using 100% methanol of Wt:CnVS compared to *ΔIlvR*:CnVS and *ΔIlvR*:MCS.

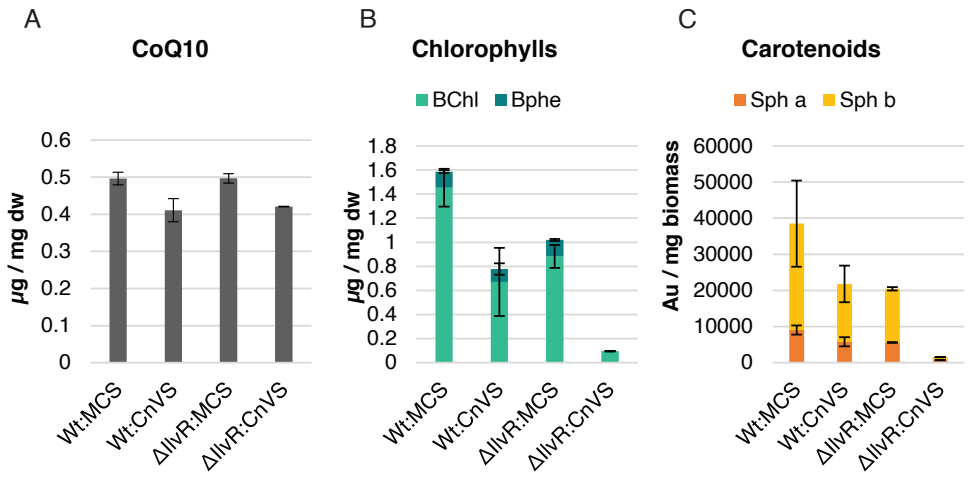



**Figure 5.10**: HPLC analysis of isoprenoids in *Rs241* strains. (A) Coenzyme Q10 (CoQ10) was produced without major difference in both Wt and *ΔIlvR strains.* (B) Bacteriochlorophyll (BChl) and Bacteriopheophytin (Bphe) were identified for all samples. Both Wt:CnVS and *ΔIlvR*:MCS show a reduction in chlorophylls produced of about 50% of the Wt:MCS. *ΔIlvR*:CnVS shows a production of chlorophylls reduced to about 10% of the Wt:MCS, with only traces of Bphe. (C) Sphaeroidenone a (Sph a) and Sphaeroidene b (Sph b) were identified for all samples. Product quantification was calculated using the standard curves (Table S5.3) for CoQ10 and chlorophylls. Graphs report the isoprenoid produced after 72 hours fermentation.

an oxidoreductase and has an essential role in redox balance, causing a lethal phenotype when deleted[49,50]. By regulating oxygen availability, it is also responsible for the expression of the PS genes[50]. The upregulation of this redox active protein in the presence of valencene suggests that the synthesis of this sesquiterpene causes oxidative stress within the cell. Icd is a well-known enzyme responsible for the decarboxylation of D-isocitrate to 2-oxoglutarate[51]. The upregulation of this enzyme might indicate a higher metabolic pressure on the TCA cycle in the presence of *CnVS.* Furthermore, upregulation of RSP_1442 supports a potentially higher intake of glucose in the *CnVS* strain[47]. To verify this, growth and glucose consumption should be measured over time for each strain. Previously, unsuccessful attempts have been made to knock out *trxA*[49], and in the current work we did not succeed in knocking out *icd*. Possibly this could indicate that these genes are essential for *R . sphaeroides*. However, Burger *et al.* (2017)[52] did observe transposon insertions in these genes, which could indicate that the conditions they tested (aerobic growth in LB or SMM media) have a different effect than the conditions applied by us for generating deletion mutants by CRISPR-Cas9 technology.

In the genome of *Rs241,* 231 genes are annotated as TRs, of which 182 do not belong to the sigma factor superfamily and 117 are predicted to have a helix-turn-helix (HTH) DNA binding domain (DBD)[53]. Typically, these TRs are constituted by a ligand binding domain (LBD) and the HTH DBD, which make them good candidates for the development of biosensors, using both the native TR[32,54,55] or a chimeric TR[56,57]. In the current work, 5 TRs associated with differentially expressed operons were identified (**Table S5.7**). Two of these TRs are annotated as LysR-family and one as GntR-family.



*IlvR* (RSP_1297), which seems to be regulated by the presence of valencene synthase, is a member of the LysR transcriptional regulatory (LTTR) family. Members of this family have been described to act as both transcriptional activators and repressors, and they often appear to rely on co-inducers acting in a feedback loop for their regulatory activity[58]. Our observations suggest that in *Rs241,* IlvR acts as a transcriptional activator of the neighbouring operon, containing predicted ALDH (RSP_1292), ACP (RSP_1293), bkdA (RSP_1294) and gdcH (RSP_1295) (**Figure 5.11A**).

Interestingly, gene annotation indicates that two out of the four genes (*bkdA* and *ALDH*) in this cluster encode enzyme activities involved in

degradation of isoleucine, leucine, and valine; while the other two genes (*gcdH* and *ACP*) potentially play a role in fatty acid metabolism. In *Pseudomonas putida,* the operon containing the *bkd* genes is regulated by the TR bkdR[59]. Similar genomic organization can be found in *Pseudomonas aeruginosa* PAO1[60] (Figure 5.11C), though bkdR is not member of the LysR family. Phylogenetic studies of these gene operons throughout different bacterial species would provide further insight into their roles in the regulatory networks.

There is no obvious relationship between degradation of amino acids and the production of terpenes, but downregulation of the putative *Ilv* operon was observed when *CnVS* was present, suggesting that valencene or its precursors might play a role in the regulation. We observed via qPCR that both *bkdA* and *gdcH* appear to be downregulated in *ΔIlvR* strains compared to the control, regardless of the presence of valencene (**Figure 5.5**). Furthermore, we observed that the production of valencene was increased in strain *ΔIlvR* CnVS, while production of biomass was reduced (**Figure 5.8**), and formation of carotenoids and chlorophylls was strongly suppressed (**Figure 5.10**). These results suggest that in the absence of a functional regulation of *bkdA* and *gdcH*, a stronger substrate competition for the production of isoprenoids might be present. The next step to characterize the possible interaction of IlvR with valencene would be to express IlvR in a heterologous system in association with a reporter gene and test its response to different ligands, as described by De Paepe *et al.* (2018)[54].



Since valencene is not normally produced in *R. sphaeroides*, the physiological role of the operon regulated by IlvR remains elusive. Some connection between the metabolism of terpenes and branched chain amino acids has been observed in *Pseudomonas* species, which are known to utilize monoterpenes as carbon source. Gene clusters involved in the degradation of valine, leucine and isoleucine, such as LiuABCDE in *Pseudomonas aerigunosa* PAO1[61], are homologous to clusters involved in acyclic monoterpene degradation, such as AtuABCDEF[62,63] (**Figure 5.11D-E**, **Table S5.8**). These operons are regulated by TetR/AcrR-family TRs, LiuR and AtuR respectively. Mutagenesis studies of the genes belonging to these operons showed that both *atu* and *liu* are involved in the catabolism of acyclic monoterpenes and leucine. The presence of all genes in these operons seem to be necessary for the growth on citronellol and leucine, indicating a concerted mechanism of degradation[62]. This suggests that the

TRs may recognize citronellol or one of the degradation products as ligand, although possibly not exclusively.

The *liuE* gene encodes the 3-hydroxy-3- methylglutaryl-coenzyme A lyase (HMG-CoA lyase), which catalyses the cleavage of HMG-CoA to acetyl-CoA and acetoacetate and is one of the key enzymes for the utilization of leucine and acyclic monoterpenes as carbon sources[61]. In *R. sphaeroides* the HMG-CoA lyase *hmgE* gene (RSP_2510) is also present, and the genomic organization is similar to the one of *P. aeruginosa* LiuABCDE. The operon containing the LiuABCDE homologous genes (RSP_2506, RSP_2508-2511, **Figure 5.11B**) in *R. sphaeroides* is associated with a CreA-family protein (RSP_2505), which may be involved in the regulation of this operon. The function of CreA-family proteins is still relatively unknown, but it has been shown that they act as TRs in fungi such as *Aspergillus nidulans*[64,65]. However, in our analysis the genes belonging to this operon did not show differential expression. Therefore, a clear role of the identified IlvR-associated operon in biochemical pathways associated to terpene metabolism cannot yet be inferred from the identity of the genes.

One could still speculate on a role for IlvR gene regulation for terpene metabolism. No genes encoding the enzymes for the degradation of acyclic monoterpenes are found in *R. sphaeroides.* Nonetheless, the bacterium is able to switch quickly between heterotrophic metabolism, and photosynthesis, the latter requiring isoprenoids for chlorophyll and carotenoid biosynthesis[20]. Possibly, IlvR plays a role in this switch, by a mechanism similar to the mechanisms needed for monoterpene breakdown in *P. aeruginosa*. There are substantial differences between the LiuR/AtuR regulators and IlvR. While LiuR and AtuR are transcriptional repressors, IlvR seems to act as a transcriptional activator. Hence, a deletion in LiuR/AtuR would be expected to result in an increase of the expression of their associated operons, increasing the degradation pathway. Conversely, a deletion of *IlvR* would potentially result in a lack of leucine degradation, and/or lack of degradation of the terpene biosynthesis intermediates. If indeed the degradation of the terpene biosynthetic intermediates would be hampered by deletion of IlvR, this could provide an explanation for the observation that more valencene is produced by the *ΔIlvR*:CnVS strain. It could also be the reason for this strain to stop growing in an earlier stage, which could be the result of toxification of the culture by terpenes, or by side products. Many sesquiterpene synthases are known to produce monoterpenes and farnesol as side products[5,43], which are known to be

toxic, and could be substrates for this pathway. The role of this pathway in the biosynthesis of carotenoids and chlorophylls is still unclear, since it seems that the DXP pathway, delivering carbon for biosynthesis of these molecules, is hardly affected, at least judged from the valencene titers. More experiments need to be performed to elucidate the physiological roles of these enzymes and their regulatory network.

The GntR family comprises TRs that control the expression of genes involved in different metabolic processes, from amino acid metabolism to fatty acid metabolism[66–68]. This TRs family takes the name from the first characterized TR of this family, the repressor of the gluconate operon in *Bacillus subtilis*[69]. While all TRs of this family show the characteristic N-term HTH domain, the different sub-families have C-term LBD domains of different length, depending on their ligand[68]. In this study, we identified a predicted GntR-family transcriptional regulator (*PstR,* RSP_3339) associated with a gene cluster responsible for the production of the putrescine/spermidine ABC transporter (RSP_3335, RSP_3336, RSP_3337, RSP_3338). Polyamines such putrescine and spermidine are essential for cell proliferation[70] and have been described to have a role in quorum sensing[71] and biofilm formation[72].

The RNA-seq results showed a selective upregulation of this gene cluster in wild type CnVS. We hypothesized an involvement of valencene in the regulation of this gene cluster as the selective upregulation in the absence of n-dodecane might represent a longer dwelling of the produced valencene in the culture environment. To test the role of RSP_3339 we generated a KO strain, *ΔPstR*, and tested the expression of the gene cluster via qPCR measurement of RSP_3336 and RSP_3337 transcripts. The results of this experiment showed selective upregulation of both genes in the absence of n-dodecane regardless of the presence of the valencene synthase, while in the presence of n-dodecane we could not observe a definite pattern **(Figure 5.6**). However, in this experiment, we could not reproduce the results obtained in the RNA-seq analysis for Wt:CnVS, as we observed no selective upregulation of either gene in Wt:CnVS. Therefore, while the results obtained for *ΔPstR* strains suggest that PstR might act as a repressor, the experiment would need to be repeated to gain confidence in these results. Given the different behaviour of this operon in the presence and absence of n-dodecane, we can hypothesize that PstR might not be the only TR acting on it. The obtained results are therefore not sufficient for a full characterization of this KO strain. We also observed that the growth of *ΔPstR* strains was comparable to the Wt strains, while the production of

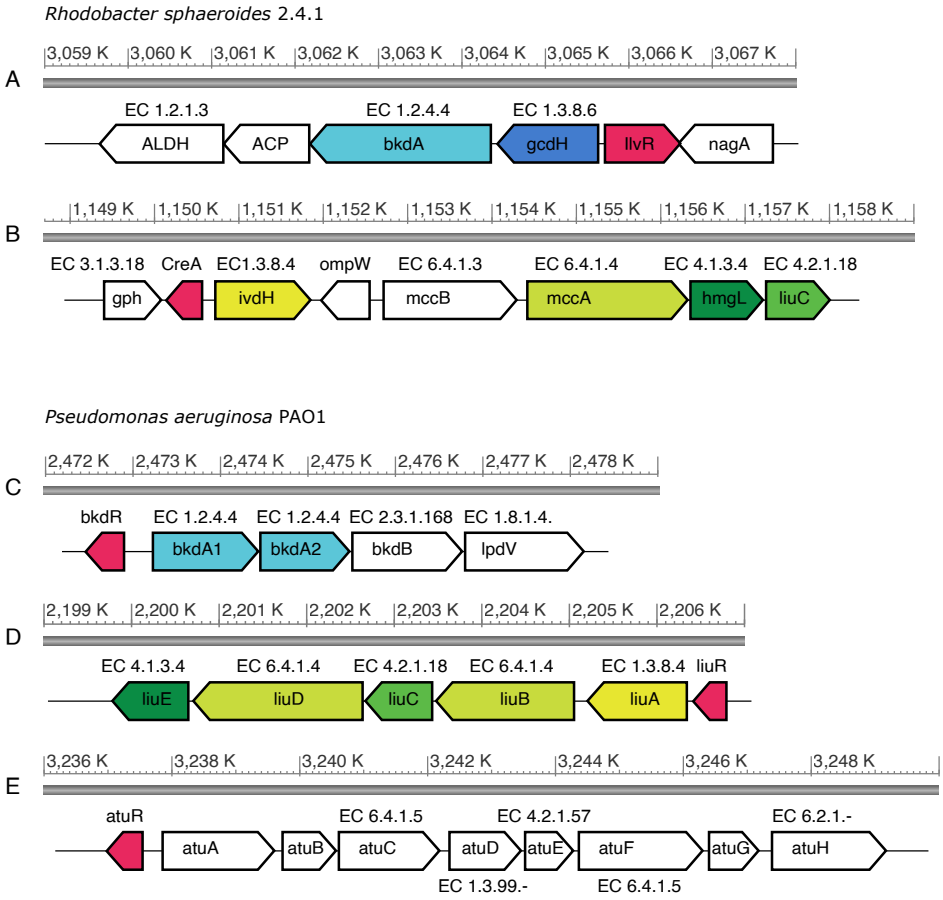



**Figure 5.11** Genomic organization of (A) *R. sphaeroides* 241 *llv* operon and (B) *CreA* operon. The (C) *bkdR* operon from *P. aeruginosa* PAO1 contains homologous genes to the *bkdA* in *R. sphaeroides*, regulated by the bkdR TR. The *CreA* operon in *R. sphaeroides* (B) is homologous to the *liu* operon in *P. aeruginosa* PAO1 (D), which shows similarity to the *atu* operon and contribute to the metabolism of acyclic monoterpenes and leucine.

valencene was lower in *ΔPstR*:CnVS compared to Wt:CnVS (**Figure 5.8**). As for IlvR, the heterologous expression of PstR is one of the next steps necessary for the characterization of the TR and the identification of its ligand. PstR might belong to the YtrA subfamily, as these TRs have been described to be associated with ABC transporters[73,74].

Since the DNA binding domain (DBD) of both TRs is only predicted, the characterization of the TRs could be performed in two steps, by generating chimeric proteins between a well characterized TR belonging to the same family, and the TRs of interest, as proposed by De Paepe *et al.*[57,75]. With this approach, it would be possible to separately the study of LBD independently from its DBD and vice versa, tackling one problem at a time. The identification of ligands for these TRs would open the possibility to employ them as genetic biosensors for the compound of interest.

## 5.5 Conclusions

In this study we identified two TRs that are potentially involved in the sensing of valencene. We generated KO strains of *Rs241* and we characterized them by following their growth and isoprene production in the presence and absence of *CnVS*. Our results show an interesting change in isoprenoids produced in the KO strain *ΔIlvR* that might represent a potentially beneficial improvement for a sesquiterpene-producing strain. It also shows how the prospects for the engineering of microbial platforms might be hidden in unexplored regulatory networks. Understanding the mechanisms of these networks is therefore an extremely promising approach to follow for strain engineering. Furthermore, the identification of TRs and their regulatory mechanisms can be the starting point for the development of genetic biosensors based on the interaction of the TR of choice and a reporter gene expressing an easily detectable protein, such as the fluorescent proteins GFP or RFP. The challenge of discovering TRs that bind the product of interest could also be bypassed by the directed engineering of already characterized TRs and their LBD.



### Acknowledgements

We thank Anne-Catrin Hailer for sharing rpoD primers for RT-qPCR analysis. This work is part of the TTW 15054 project financed by the Netherlands Organisation for Scientific Research.

# Appendix

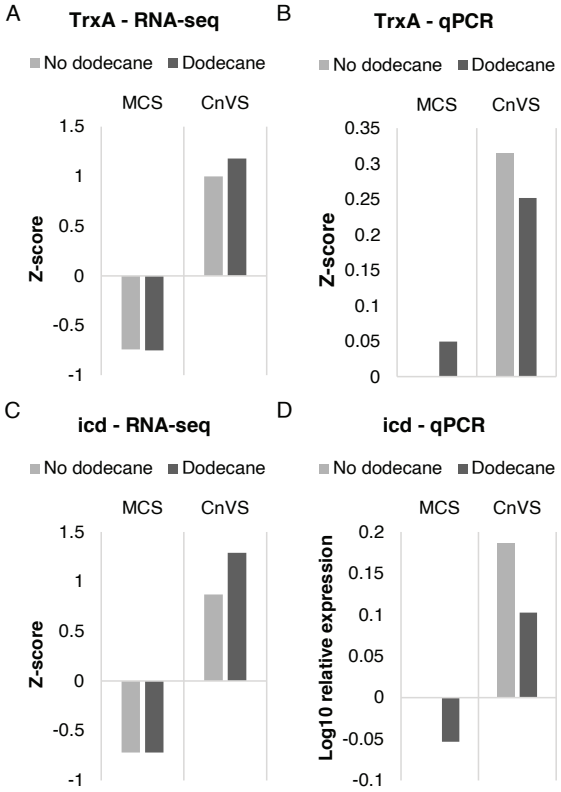



**Figure S5.1** RNA-seq results for *trxA* (A) and *icd* (D), confirmed by RT-qPCR analysis. Both genes are upregulated in the presence of CnVS.



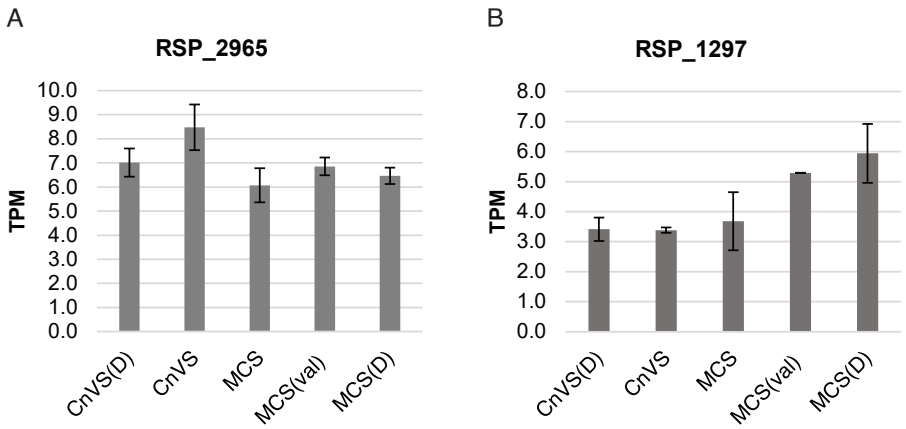



**Figure S5.2** RNA-seq results for (A) RSP_2965 and (B) RSP_1297. The histograms report the TPMs for each gene. RSP_2965 does not show a significant expression difference between different samples, while RSP_1297 shows downregulation in CnVS(D) and CnVS compared to the other conditions.

**Table S5.1** Primers used in this study for plasmid assembly, colony PCR and sequencing. Caps indicate binding sequence; low case indicate overhang and non-binding sequence.

| Primer ID | Sequence | Primer Length (bp) | Fw/Re | Template | Description |
|---|---|---|---|---|---|
| **Primers used for plasmid assembly** | | | | | |
| CAS006 | CGGCATGATGAACCTGAATCGCCAGCG | 27 | Fw | pBBR_Cas9_NT | Backbone of pBBR_Cas9_NT, 5' Cas9 |
| CAS007 | TTTGTGATGGCTTCCATGTC | 20 | Re | pBBR_Cas9_NT | Backbone of pBBR_Cas9_NT, 3' KanR cassette |
| LYS294 | ATCGCTGGTTTgacgatcaatctcgcgacg GTTTTAGAGCTAGAAATAGCAAG | 53 | Fw | pBBR_Cas9_NT | Contains spacer for IlvR |
| LYS293 | GCTCTAAAACcgtcgcgagattgatcgtca AACCAGCGATCCCGTCCGCC | 50 | Re | pBBR_Cas9_NT | Contains spacer for IlvR |
| GNT302 | ATCGCTGGTTcgagctgggccacgaagtcg GTTTTAGAGCTAGAAATAGCAAG | 53 | Fw | pBBR_Cas9_NT | Contains spacer for PstR |
| GNT301 | GCTCTAAAACcgacttcgtggcccagctcg AACCAGCGATCCCGTCCGCC | 50 | Re | pBBR_Cas9_NT | Contains spacer for PstR |
| ICD282 | ATCGCTGGTTcctctcgagatagcccatcg GTTTTAGAGCTAGAAATAGCAAG | 53 | Fw | pBBR_Cas9_NT | Contains spacer for icd |
| ICD281 | GCTCTAAAACcgatgggctatctcgagagg AACCAGCGATCCCGTCCGCC | 50 | Re | pBBR_Cas9_NT | Contains spacer for icd |
| LYS272 | gcattctgccgacatggaagccatcacaaa TGCGCTTGACGATCGAGATC | 50 | Fw | *Rs241* | Amplify homologous region UP for IlvR |
| LYS273 | gcgtgccgccaccctctgactgtctcggag GAACTCCGGCGAAATTAGCGTTATC | 55 | Re | *Rs241* | Amplify homologous region UP for IlvR |
| LYS274 | gccggagttcctccgagacagtcagagggt GGCGGCACGCCTATTGGCTG | 50 | Fw | *Rs241* | Amplify homologous region DOWN for IlvR |
| LYS275 | tgccgctggcgattcaggttcatcatgccg CCGGACAACGGGTGGAGCAG | 50 | Re | *Rs241* | Amplify homologous region DOWN for IlvR |
| GNT276 | gcattctgccgacatggaagccatcacaaa ATTCCACCGCCTGCACCACG | 50 | Fw | *Rs241* | Amplify homologous region UP for PstR |

| | | | | | |
|---|---|---|---|---|---|
| GNT277 | gcaggggcgaaccctctgactgtctcggag TGGCCGATTCCTCGCCTCAG | 50 | Re | *Rs241* | Amplify homologous region UP for PstR |
| GNT278 | gaatcggccactccgagacagtcagagggt TCGCCCCTGCGGCACATGTG | 50 | Fw | *Rs241* | Amplify homologous region DOWN for PstR |
| GNT279 | tgccgctggcgattcaggttcatcatgccg GCAGCTGGTGGCCATCGTGC | 50 | Re | *Rs241* | Amplify homologous region DOWN for PstR |
| ICD268 | gcattctgccgacatggaagccatcacaaa AAGCCGCTGCTCCGGCTGAC | 50 | Fw | *Rs241* | Amplify homologous region UP for icd |
| ICD269 | gcgccgagcgaccctctgactgtctcggag GGAGAGGCTCCCGAATGGTTGG | 52 | Re | *Rs241* | Amplify homologous region UP for icd |
| ICD270 | gagcctctccctccgagacagtcagagggt GCGCTCGGCGCCTGAGCGAAG | 51 | Fw | *Rs241* | Amplify homologous region DOWN for icd |
| ICD271 | tgccgctggcgattcaggttcatcatgccg CTCGGGCGTGAACGGCCTGC | 50 | Re | *Rs241* | Amplify homologous region DOWN for icd |
| **Primers for colony PCR, plasmid confirmation and knock out sequencing** | | | | | |
| CAS004 | CATTAAGCATTCTGCCGACATG | 22 | Fw | pBBR_Cas9 assembled constructs | general FW primer upstream of HR.up regions |
| CAS005 | TCTCATGCTGGAGTTCTTCGC | 21 | Re | pBBR_Cas9 assembled constructs | general primer for pBBR sequencing - RE - downstream HR regions - can be used with all the FW seq primers |
| GNT308 | CTGAGGCGAGGAATCGGCCA | 20 | Fw | pBBR_Cas9_ ΔPstR | colony PCR of pBBR_Cas9_ΔPstR constructs - 3' HR_UP_G |
| GNT309 | GCACATGTGCCGCAGGGGCG | 20 | Re | pBBR_Cas9_ ΔPstR | colony PCR of pBBR_Cas9_ΔPstR constructs 5' - HR_DW_G |
| LYS310 | ATGCCGGCCAAGGATAACGC | 20 | Fw | pBBR_Cas9_ΔIlvR | colony PCR of pBBR_Cas9_ΔIlvR constructs - 3' HR_UP_L |
| LYS311 | CAGCCAATAGGCGTGCCGC | 19 | Re | pBBR_Cas9_ΔIlvR | colony PCR of pBBR_Cas9_ΔIlvR constructs - 5' HR_DW_L |

| | | | | | |
|---|---|---|---|---|---|
| ICD306 | CTTCGCTCAGGCGCCGAGCG | 20 | Fw | pBBR_Cas9_Δicd | colony PCR of pBBR_Cas9_Δicd constructs - 3' HR_UP_D |
| ICD307 | CAACCATTCGGGAGCCTCTC | 20 | Re | pBBR_Cas9_Δicd | colony PCR of pBBR_Cas9_ΔicdR constructs - 5' HR_DW_D |
| ICD312 | CCGTCGCGTCGCTAGTGGTC | 20 | Fw | Δicd | FW genomic primer icd - 5' of HR_UP_D |
| ICD313 | TCCGCGCCTTCTGTGCCGAC | 20 | Re | Δicd | RE genomic primer icd - 3' of HR.DW_D |
| GNT314 | GCGAGGTGCACGGACTCGC | 19 | Fw | ΔPstR | genomic primer ΔPstR - 5' of HR_UP_G |
| GNT315 | GCGCTCATCTCGGTCAACTG | 20 | Re | ΔPstR | genomic primer ΔPstR - 3' of HR_DW_G |
| LYS316 | GTGCCCTCGTAGGTGTTCAC | 20 | Fw | ΔIlvR | genomic primer ΔIlvR - 5' of HR_UP_L |
| LYS317 | AGGAGATCCGATGTCTCTCG | 20 | Re | ΔIlvR | genomic primer ΔIlvR - 3' of HR_DW_L |



**Table S5.2** Primer pairs used for RT-qPCR, include gene name and regulation observed in RNA-seq.

| RSP number | Primer Fw ID | Sequence | Primer Re ID | Sequence | Regulation |
|---|---|---|---|---|---|
| RSP_0395 | FW348 | CGGTGGACAAGTTCGAATACC | RE349 | GATCCACCAGGTCGCATAGG | reference gene |
| RSP_1294 | FW325 | CGGCGATGGAATTGTCGTTG | RE326 | CCCTGCCCTTCTTCTCGAAC | downregulated in CnVS and CnVS(D) |
| RSP_1295 | FW340 | TGAGGCCGTAGGAGACGTAG | RE341 | CCGCGAGGAACAGACCGATC | downregulated in CnVS and CnVS(D) |
| RSP_3337 | FW337 | AGCTATCGCGTGGCGCTCGA | RE338 | CGCCGCATCGACCGTCATCT | upregulated in CnVS |
| RSP_3338 | FW333 | CTCGGTGCTTGCCGTGATGA | RE334 | CGCCGTCCATGAACACGACC | upregulated in CnVS |
| RSP_1559 | FW319 | TGCGATCAAGCAGTATGGCG | FW320 | CGTTCGGGCTCTTCCACATG | upregulated in CnVS and CnVS(D) |
| RSP_1529 | FW365 | GATTTCTGGGCCGAATGGTG | RE366 | GTTCTCGTCGACATTGACCTTC | upregulated in CnVS and CnVS(D) |

**Table S5.3** Standard curves used in HPLC product quantification. B-carotene slope was not used for quantification as it results to be non linear.

| | mg/mL | Area units | | mg/mL | Area units | | mg/mL | Area units |
|---|---|---|---|---|---|---|---|---|
| Bacteriochlorophyll a | 1 | 9570035 | Coenzyme Q10 | 1 | 9624446 | β-Carotene | 50 | 903001 |
| (BChl a) | 0.8 | 7868603 | (CoQ10) | 0.8 | 8156247 | (Ctr) | 10 | 498986 |
| | 0.6 | 6624109 | | 0.6 | 6356651 | | 8 | 436392 |
| | 0.4 | 3116832 | | 0.4 | 4310039 | | 6 | 373735 |
| | 0.24 | 1737040 | | 0.2 | 2293901 | | 4 | 275274 |
| | 0.2 | 1537763 | | 0.1 | 1225763 | | 2 | 162458 |
| | 0.103 | 630760 | | 0.04 | 424219 | | 1 | 89467 |
| | 0.04 | 213256 | | 0.008 | 88325 | | 0.4 | 36408.5 |
| | 0.008 | 36019 | | 0.0016 | 18329 | | 0.08 | 8751 |
| slope | | 10157260.5 | | | 9837052.0 | | | 61291.4* |

**Table S5.7** Transcriptional regulators associated with the DEGs. For each transcriptional regulator, the neighbouring operon is reported.

| RSP number | Function | Associated genes | Function |
|---|---|---|---|
| RSP_1953 | Transcriptional regulator, CRP/FNR family | RSP_1952 | Cold-shock DNA-binding protein family |
| RSP_2853 | Transcriptional regulator, TetR family | RSP_2851 | Multidrug (Bicyclomycin) efflux pump |
| | | RSP_2852 | Putative polyketide biosynthesis associated protein |
| | | RSP_2854 | Cation/multidrug efflux pump |

| | | | |
|---|---|---|---|
| | | RSP_2855 | Cation/multidrug efflux pump |
| RSP_1297 | Transcriptional regulator, LysR family | RSP_1292 | Putative Aldehyde dehydrogenase E.C. 1.2.1.3) |
| | | RSP_1293 | Putative acyl-carrier protein |
| | | RSP_1294 | branched-chain alpha-keto acid dehydrogenase E1 component (E.C.1.2.4.4) |
| | | RSP_1295 | Putative acyl-CoA dehydrogenase E.C. 1.3.8.6) |
| RSP_3339 | GntR family transcriptional regulator | RSP_3334 | aspartate/glutamate racemase family protein |
| | | RSP_3335 | ABC transporter permease |
| | | RSP_3336 | ABC transporter permease Binding-protein-dependent transport systems inner membrane component ABC spermidine/putrescine transporter, inner membrane subunit |
| | | RSP_3337 | extracellular solute-binding protein |
| | | RSP_3338 | ABC transporter ATP-binding protein |
| RSP_3464 | LysR family transcriptional regulator | RSP_3457 | Amino acid ABC transporter substrate-binding protein, PAAT family |
| | | RSP_3458 | ABC transporter related |
| | | RSP_3459 | Polar amino acid ABC transporter, inner membrane subunit ABC transporter permease subunit Amino acid ABC transporter membrane protein 1, PAAT family |
| | | RSP_3460 | Amino acid ABC transporter membrane protein 2, PAAT family |

| | | |
|---|---|---|
| RSP_3461 | Predicted amidohydrolase Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase N-carbamyl-D-amino acid amidohydrolase N-carbamoyl-D-amino-acid hydrolase |
| RSP_3462 | Dihydropyrimidinase E.C. 3.5.2.2) |
| RSP_3463 | Asp/Glu/hydantoin racemase Putative hydantoin racemase |
| RSP_3465 | Flavin_Reduct domain-containing protein |

**Table S5.8** Comparison of homologous operons in *R. sphaeroides* 2.4.1 and *P. aeruginosa* PAO1

| Gene ID | Gene name | Symbol | EC number | Metabolic pathway (KEGG) |
|---|---|---|---|---|
| IlvR operon *Rhodobacter sphaeroides* 2.4.1 | | | | |
| RSP_1292 | Putative Aldehyde dehydrogenase | ALDH | EC:1.2.1.3 | Glycolysis/Gluconeogenesis, Fatty acid degradation, Amino acid metabolism, Limonene and pinene degradation |
| RSP_1293 | ACP S-malonyltransferase | ACP | EC:2.3.1.39 | fatty acid biosynthesis |
| RSP_1294 | branched-chain alpha-keto acid dehydrogenase E1 component | bkdA | EC:1.2.4.4 | Val, leu and ile degradation, Propanoate metabolism |
| RSP_1295 | Putative acyl-CoA dehydrogenase: glutaryl-CoA dehydrogenase | gcdH | EC:1.3.8.6 | Fatty acid degradation, Lys degradation, Trp metabolism, Benzoate degradation |
| RSP_1297 | Transcriptional regulator, LysR family | IlvR | | |
| CreA operon *Rhodobacter sphaeroides* 2.4.1 | | | | |
| RSP_2504 | Phosphoglycolate phosphatase | gph | EC:3.1.3.18 | Glyoxylate and dicarboxylate metabolism |
| RSP_2505 | Putative CreA family protein | CreA | | |
| RSP_2506 | Isovaleryl-CoA dehydrogenase | ivdH | EC:1.3.8.4 | Val, leu and ile degradation, Propanoate metabolism |

| | | | | |
|---|---|---|---|---|
| RSP_2507 | Putative outer membrane protein | OmpW | | |
| RSP_2508 | Methylcrotonyl-CoA carboxylase beta chain | mccB | EC:6.4.1.3 / EC:6.4.1.4 | Val, leu and ile degradation, Propanoate metabolism |
| RSP_2509 | 3-methylcrotonoyl-CoA carboxylase, alpha subunit | mccA | EC:6.4.1.4 | Val, leu and ile degradation, Propanoate metabolism |
| RSP_2510 | Hydroxymethylglutaryl-CoA lyase | hmgL | EC:4.1.3.4 | Val, leu and ile degradation; geraniol degradation; butanoate metabolism |
| RSP_2511 | Methylglutaconyl-CoA hydratase | liuC | EC:4.2.1.17 / EC:4.2.1.18 | Val, leu and ile degradation, Propanoate metabolism |
| AtuR operon *Pseudomonas putida* PAO1 | | | | |
| PA2885 | AtuR, TetR-family transcriptional regulator | atuR | | Geraniol degradation |
| PA2886 | hypothetical protein | atuA | | Geraniol degradation |
| PA2887 | Citronellol/citronellal dehydrogenase | atuB | | Geraniol degradation |
| PA2888 | geranyl-CoA carboxylase subunit beta | atuC | EC:6.4.1.5 | Geraniol degradation |
| PA2889 | citronellyl-CoA dehydrogenase | atuD | EC:1.3.99.- | Geraniol degradation |
| PA2890 | isohexenylglutaconyl-CoA hydratase | atuE | EC:4.2.1.57 | Geraniol degradation |
| PA2891 | geranyl-CoA carboxylase subunit alpha | atuF | EC:6.4.1.5 | Geraniol degradation |
| PA2892 | Short chain citronellol/citronellal dehydrogenase | atuG | | Geraniol degradation |
| PA2893 | citronellyl-CoA synthetase | atuH | EC:6.2.1.- | Geraniol degradation |
| LiuR operon *Pseudomonas putida* PAO1 | | | | |
| PA2011 | hydroxymethylglutaryl-CoA lyase | liuE | EC:4.1.3.4 | Valine, leucine and isoleucine degradation; geraniol degradation; butanoate metabolism |
| PA2012 | 3-methylcrotonyl-CoA carboxylase alpha subunit | liuD | EC:6.4.1.4 | Val, leu and ile degradation |
| PA2013 | methylglutaconyl-CoA hydratase | liuC | EC:4.2.1.18 | Val, leu and ile degradation |
| PA2014 | 3-methylcrotonyl-CoA carboxylase beta subunit | liuB | EC:6.4.1.4 | Val, leu and ile degradation |
| PA2015 | isovaleryl-CoA dehydrogenase | liuA | EC:1.3.8.4 | Val, leu and ile degradation |
| PA2016 | LiuR transcriptional regulator | liuR | | |

| BkdR operon in *Pseudomonas putida* PAO1 | | | | |
|---|---|---|---|---|
| PA2246 | Lrp/AsnC family transcriptional regulator, leucine-responsive regulatory protein | bkdR | | |
| PA2247 | 2-oxoisovalerate dehydrogenase E1 component alpha subunit | bkdA1 | EC:1.2.4.4 | Val, leu and ile degradation, Propanoate metabolism |
| PA2248 | 2-oxoisovalerate dehydrogenase E1 component beta subunit | bkdA2 | EC:1.2.4.4 | Val, leu and ile degradation, Propanoate metabolism |
| PA2249 | 2-oxoisovalerate dehydrogenase E2 component (dihydrolipoyl transacylase) | bkdB | EC:2.3.1.168 | Val, leu and ile degradation, Propanoate metabolism |
| PA2250 | dihydrolipoamide dehydrogenase | lpdV | EC:1.8.1.4 | Glycolysis / Gluconeogenesis; Citrate cycle (TCA cycle); Glycine, serine and threonine metabolism; Valine, leucine and isoleucine degradation; Lysine degradation; Tryptophan metabolism; Pyruvate metabolism; Glyoxylate and dicarboxylate metabolism; Propanoate metabolism |

## References


1. Mackenzie, C. *et al.* Postgenomic adventures with Rhodobacter sphaeroides. *Annual Review of Microbiology* vol. 61 283–307 (2007).
2. Niedzwiedzki, D. M., Swainsbury, D. J. K., Canniffe, D. P., Neil Hunter, C. & Hitchcock, A. A photosynthetic antenna complex foregoes unity carotenoid-to-bacteriochlorophyll energy transfer efficiency to ensure photoprotection. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 6502–6508 (2020).
3. Licht, M. K. *et al.* Adaptation to photooxidative stress: Common and special strategies of the alphaproteobacteria Rhodobacter sphaeroides and Rhodobacter capsulatus. *Microorganisms* **8**, (2020).
4. Orsi, E., Beekwilder, J., Eggink, G., Kengen, S. W. M. & Weusthuis, R. A. The transition of Rhodobacter sphaeroides into a microbial cell factory. *Biotechnology and Bioengineering* vol. 118 531–541 (2021).
5. Chen, X. *et al.* Co-production of farnesol and coenzyme Q10 from metabolically engineered Rhodobacter sphaeroides. *Microb. Cell Fact.* **18**, (2019).
6. Orsi, E. *et al.* Metabolic flux ratio analysis by parallel 13C labeling of isoprenoid biosynthesis in Rhodobacter sphaeroides. *Metabolic Engineering* vol. 57 228–238 (2020).
7. Mougiakos, I. *et al.* Efficient Cas9-based genome editing of Rhodobacter sphaeroides for metabolic engineering. *Microb. Cell Fact.* **18**, 204 (2019).
8. Ortega-Ramos, M. *et al.* Engineered biosynthesis of bacteriochlorophyll gF in Rhodobacter sphaeroides. *Biochim. Biophys. Acta - Bioenerg.* **1859**, 501–509 (2018).
9. Nagashima, K. V. P. *et al.* Probing structure–function relationships in early events in photosynthesis using a chimeric photocomplex. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10906–10911 (2017).
10. Anthony, J. R., Warczak, K. L. & Donohue, T. J. A transcriptional response to singlet oxygen, a toxic byproduct of photosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6502–6507 (2005).
11. Liu, S., Li, X., Zhang, G. & Zhang, J. Effect of magnesium ion on crt gene expression in improving carotenoid yield of Rhodobacter sphaeroides. *Arch. Microbiol.* **197**, 1101–1108 (2015).
12. O'Gara, J. P. & Kaplan, S. Evidence for the role of redox carriers in photosynthesis gene expression and carotenoid biosynthesis in Rhodobacter sphaeroides 2.4.1. *J. Bacteriol.* **179**, 1951–1961 (1997).
13. Steiger, S., Takaichi, S. & Sandmann, G. Heterologous production of two unusual acyclic carotenoids , the crtC and crtD genes from Rhodobacter and Rubrivivax. *J. Biotechnol.* **97**, 51–58 (2002).
14. McIntosh, M., Eisenhardt, K., Remes, B., Konzer, A. & Klug, G. Adaptation of the Alphaproteobacterium Rhodobacter sphaeroides to stationary phase. *Environ. Microbiol.* **21**, 4425–4445 (2019).
15. Gomelsky, M. & Kaplan, S. Molecular genetic analysis suggesting interactions between AppA and PpsR in regulation of photosynthesis gene expression in Rhodobacter sphaeroides 2.4.1. *J. Bacteriol.* **179**, 128–134 (1997).
16. Gomelsky, M. & Kaplan, S. appA, A novel gene encoding a trans-acting factor involved in the regulation of photosynthesis gene expression in Rhodobacter sphaeroides 2.4.1. *J. Bacteriol.* **177**, 4609–4618 (1995).
17. Masuda, S. & Bauer, C. E. AppA is a blue light photoreceptor that antirepresses photosynthesis gene expression in Rhodobacter sphaeroides. *Cell* **110**, 613–623 (2002).
18. Pandey, R., Flockerzi, D., Hauser, M. J. B. & Straube, R. An extended model for the repression of photosynthesis genes by the AppA/PpsR system in Rhodobacter sphaeroides. in *FEBS Journal* vol. 279 3449–3461 (2012).
19. Elsen, S., Swem, L. R., Swem, D. L. & Bauer, C. E. RegB/RegA, a Highly Conserved Redox-Responding Global Two-Component Regulatory System. *Microbiol. Mol. Biol. Rev.* **68**, 263–279 (2004).
20. Imam, S., Noguera, D. R. & Donohue, T. J. Global insights into energetic and metabolic networks

in Rhodobacter sphaeroides. *BMC Syst. Biol.* **7**, 89 (2013).
21. Qu, Y., Su, A., Li, Y., Meng, Y. & Chen, Z. Manipulation of the Regulatory Genes ppsR and prrA in Rhodobacter sphaeroides Enhances Lycopene Production. *J. Agric. Food Chem.* **69**, 4134–4143 (2021).
22. Grützner, J. *et al.* The small DUF1127 protein CcaF1 from Rhodobacter sphaeroides is an RNA-binding protein involved in sRNA maturation and RNA turnover. *Nucleic Acids Res.* **49**, 3003–3019 (2021).
23. Yamada, Y. *et al.* Terpene synthases are widely distributed in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 857–862 (2015).
24. Peck, M. C., Fisher, R. F. & Long, S. R. Diverse flavonoids stimulate NodD1 binding to nod gene promoters in Sinorhizobium meliloti. *J. Bacteriol.* **188**, 5417–5427 (2006).
25. Vickers, C. E., Bongers, M., Liu, Q., Delatte, T. & Bouwmeester, H. Metabolic engineering of volatile isoprenoids in plants and microbes. *Plant. Cell Environ.* **37**, 1753–75 (2014).
26. Liu, K., Chen, Q., Liu, Y., Zhou, X. & Wang, X. Isolation and Biological Activities of Decanal, Linalool, Valencene, and Octanal from Sweet Orange Oil. *J. Food Sci.* **77**, C1156–C1161 (2012).
27. Frohwitter, J., Heider, S. A. E., Peters-Wendisch, P., Beekwilder, J. & Wendisch, V. F. Production of the sesquiterpene (+)-valencene by metabolically engineered Corynebacterium glutamicum. *J. Biotechnol.* **191**, 205–213 (2014).
28. Niu, F. X., Huang, Y. Bin, Ji, L. N. & Liu, J. Z. Genomic and transcriptional changes in response to pinene tolerance and overproduction in evolved Escherichia coli. *Synth. Syst. Biotechnol.* **4**, 113–119 (2019).
29. Niu, F. X., He, X., Wu, Y. Q. & Liu, J. Z. Enhancing Production of Pinene in Escherichia coli by Using a Combination of Tolerance, Evolution, and Modular Co-culture Engineering. *Front. Microbiol.* **9**, 1623 (2018).
30. Bathke, J., Konzer, A., Remes, B., McIntosh, M. & Klug, G. Comparative analyses of the variation of the transcriptome and proteome of Rhodobacter sphaeroides throughout growth. *BMC Genomics* **20**, (2019).
31. Imam, S., Noguera, D. R. & Donohue, T. J. Global Analysis of Photosynthesis Transcriptional Regulatory Networks. *PLoS Genet.* **10**, 1004837 (2014).
32. Siedler, S., Stahlhut, S. G., Malla, S., Maury, J. & Neves, A. R. Novel biosensors based on flavonoid-responsive transcriptional regulators introduced into Escherichia coli. *Metab. Eng.* **21**, 2–8 (2014).
33. Beekwilder, J. *et al.* Valencene synthase from the heartwood of Nootka cypress ( Callitropsis nootkatensis ) for biotechnological production of valencene. *Plant Biotechnol. J.* **12**, 174–182 (2014).
34. Orsi, E. *et al.* Functional replacement of isoprenoid pathways in Rhodobacter sphaeroides. *Microb. Biotechnol.* **13**, 1082–1093 (2020).
35. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
36. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
37. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).
38. Doyle, M. Visualization of RNA-Seq results with heatmap2 (Galaxy Training Materials). https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-heatmap2/tutorial.html (2021).
39. Batut, B. *et al.* Community-Driven Data Analysis Training for Biology. *Cell Syst.* **6**, 752-758.e1 (2018).
40. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
41. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular

interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
42. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods* **25**, 402–408 (2001).
43. Di Girolamo, A. *et al.* The santalene synthase from Cinnamomum camphora: Reconstruction of a sesquiterpene synthase from a monoterpene synthase. *Arch. Biochem. Biophys.* **695**, (2020).
44. Fraser, P. D., Elisabete S Pinto, M., Holloway, D. E. & Bramley, P. M. Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. *Plant J.* **24**, 551–558 (2000).
45. Bino, R. J. *et al.* The light-hyperresponsive high pigment-2dg mutation of tomato: Alterations in the fruit metabolome. *New Phytol.* **166**, 427–438 (2005).
46. Ashikhmin, A., Makhneva, Z., Bolshakov, M. & Moskalenko, A. Incorporation of spheroidene and spheroidenone into light-harvesting complexes from purple sulfur bacteria. *J. Photochem. Photobiol. B Biol.* **170**, 99–107 (2017).
47. Schneider, E. ABC transporters catalyzing carbohydrate uptake. *Res. Microbiol.* **152**, 303–310 (2001).
48. Malakooti, J. & Ely, B. Identification and characterization of the ilvR gene encoding a LysR-type regulator of Caulobacter crescentus. *J. Bacteriol.* **176**, 1275–1281 (1994).
49. Pasternak, C., Assemat, K., Clément-Métral, J. D. & Klug, G. Thioredoxin is essential for Rhodobacter sphaeroides growth by aerobic and anaerobic respiration. *Microbiology* **143**, 83–91 (1997).
50. Pasternak, C., Haberzettl, K. & Klug, G. Thioredoxin is involved in oxygen-regulated formation of the photosynthetic apparatus of Rhodobacter sphaeroides. *J. Bacteriol.* **181**, 100–106 (1999).
51. Spaans, S. K., Weusthuis, R. A., van der Oost, J. & Kengen, S. W. M. NADPH-generating systems in bacteria and archaea. *Frontiers in Microbiology* vol. 6 (2015).
52. Burger, B. T., Imam, S., Scarborough, M. J., Noguera, D. R. & Donohue, T. J. Combining Genome-Scale Experimental and Computational Methods To Identify Essential Genes in Rhodobacter sphaeroides. *mSystems* **2**, (2017).
53. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiology Reviews* vol. 29 231–262 (2005).
54. Peters, G. *et al.* Development of N-acetylneuraminic acid responsive biosensors based on the transcriptional regulator NanR. *Biotechnol. Bioeng.* **115**, 1855–1865 (2018).
55. Wang, R. *et al.* Design and Characterization of Biosensors for the Screening of Modular Assembled Naringenin Biosynthetic Library in Saccharomyces cerevisiae. *ACS Synth. Biol.* **8**, 2121–2130 (2019).
56. Dimas, R. P., Jiang, X. L., De La Paz, J. A., Morcos, F. & Chan, C. T. Y. Engineering repressors with coevolutionary cues facilitates toggle switches with a master reset. *Nucleic Acids Res.* **47**, 5449–5463 (2019).
57. De Paepe, B., Maertens, J., Vanholme, B. & De Mey, M. Chimeric LysR-Type Transcriptional Biosensors for Customizing Ligand Specificity Profiles toward Flavonoids. *ACS Synth. Biol.* **8**, 318–331 (2019).
58. Fu, Y. *et al.* Four LysR-type transcriptional regulator family proteins (LTTRs) involved in antibiotic resistance in Aeromonas hydrophila. *World J. Microbiol. Biotechnol.* **35**, 127 (2019).
59. Hester, K., Luo, J., Burns, G., Braswell, E. H. & Sokatch, J. R. Purification of Active E1alpha2beta2 of Pseudomonas Putida Branched-Chain-Oxoacid Dehydrogenase. *Eur. J. Biochem.* **233**, 828–836 (1995).
60. Stover, C. K. *et al.* Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature* **406**, 959–964 (2000).
61. Chávez-Avilés, M., Díaz-Pérez, A. L., Reyes-De La Cruz, H. & Campos-García, J. The Pseudomonas aeruginosa liuE gene encodes the 3-hydroxy-3-methylglutaryl coenzyme A lyase, involved in leucine and acyclic terpene catabolism. *FEMS Microbiol. Lett.* **296**, 117–123 (2009).
62. Aguilar, J. A. *et al.* The atu and liu clusters are involved in the catabolic pathways for acyclic

monoterpenes and leucine in Pseudomonas aeruginosa. *Appl. Environ. Microbiol.* **72**, 2070–2079 (2006).
63. Förster-Fromme, K. & Jendrossek, D. AtuR is a repressor of acyclic terpene utilization (Atu) gene cluster expression and specifically binds to two 13 bp inverted repeat sequences of the atuA-atuR intergenic region. *FEMS Microbiol. Lett.* **308**, 166–174 (2010).
64. de Assis, L. J. *et al.* Regulation of Aspergillus nidulans CreA-mediated catabolite repression by the F-Box proteins Fbx23 and Fbx47. *MBio* **9**, (2018).
65. Tamayo, E. N. *et al.* CreA mediates repression of the regulatory gene xlnR which controls the production of xylanolytic enzymes in Aspergillus nidulans. *Fungal Genet. Biol.* **45**, 984–993 (2008).
66. Tramonti, A., Nardella, C., di Salvo, M. L., Pascarella, S. & Contestabile, R. The MocR-like transcription factors: pyridoxal 5′-phosphate-dependent regulators of bacterial metabolism. *FEBS Journal* vol. 285 3925–3944 (2018).
67. Rigali, S., Derouaux, A., Giannotta, F. & Dusart, J. Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J. Biol. Chem.* **277**, 12507–12515 (2002).
68. Suvorova, I. A., Korostelev, Y. D. & Gelfand, M. S. GntR Family of Bacterial Transcription Factors and Their DNA Binding Motifs: Structure, Positioning and Co-Evolution. *PLoS One* **10**, e0132618 (2015).
69. Fujita, Y., Fujita, T., Miwa, Y., Nihashi, J. & Aratani, Y. Organization of transcription of the gluconate operon, gnt, of Bacillus subtilis. *J. Biol. Chem.* **261**, 13744–13753 (1986).
70. Igarashi, K. & Kashiwagi, K. Polyamine transport in bacteria and yeast. *Biochemical Journal* vol. 344 633–642 (1999).
71. Chan, Y. Y. & Chua, K. L. Growth-related changes in intracellular spermidine and its effect on efflux pump expression and quorum sensing in Burkholderia pseudomallei. *Microbiology* **156**, 1144–1154 (2010).
72. Karatan, E., Duncan, T. R. & Watnick, P. I. NspS, a predicted polyamine sensor, mediates activation of Vibrio cholerae biofilm formation by norspermidine. *J. Bacteriol.* **187**, 7434–7443 (2005).
73. Salzberg, L. I., Luo, Y., Hachmann, A. B., Mascher, T. & Helmann, J. D. The Bacillus subtilis GntR family repressor YtrA responds to cell wall antibiotics. *J. Bacteriol.* **193**, 5793–5801 (2011).
74. Tsypik, O. *et al.* Transcriptional regulators of GntR family in Streptomyces coelicolor A3(2): analysis in silico and in vivo of YtrA subfamily. *Folia Microbiol. (Praha).* **61**, 209–220 (2016).
75. De Paepe, B., Peters, G., Coussement, P., Maertens, J. & De Mey, M. Tailor-made transcriptional biosensors for optimizing microbial cell factories. *J. Ind. Microbiol. Biotechnol.* **44**, 623–645 (2017).

# 6

# General Discussion

Alice Di Girolamo[1]

[1] Laboratory of Plant Physiology, Department of Plant Sciences, Wageningen University, Netherlands

I opened this thesis presenting two major challenges in the biotechnological development of novel terpene synthases (TSs) for flavours and fragrances. Efforts to characterize the enzyme responsible for terpene production date back to the mid-1980s. In 1984, an extract of *Salvia officinalis* was obtained for the first time that could convert farnesyl diphosphate (FDP) to the sesquiterpenes humulene and caryophyllene[1]. This study led, in 1988, to the purification, albeit partial, of two of the first plant sesquiterpene synthases (STSs)[2]. In 1989, a trichodiene synthase from the pathogenic fungus *Fusarium sporotrichioides* was characterized[3] and expressed in *E. coli*[4], yielding the first sesquiterpene produced in a heterologous system. Production of terpenes using microbial platforms provides an alternative to the exploitation of natural resources to obtain plant specialized metabolites. Native plant synthases often need optimization before being successfully implemented in the microbial system and might require several rounds of mutagenesis before reaching competitive amounts of compounds produced.

If on the one hand the need for highly efficient synthases requires the generation of large libraries of mutant enzymes, on the other hand it is crucial to implement reliable high-throughput screening (HTS) methods. Due to a lack of efficient screening methods, the generation of smart libraries by directed mutagenesis has often been preferred over large libraries by random mutagenesis. The various approaches that have been undertaken to tackle these challenges have been described in detail, and it is clear that several methods have contributed extensively to progress in the field. In the following sections I will discuss the results of this thesis and how they contribute to current and future developments in the field.



## 6.1 Improving characterized enzymes by direct mutagenesis

The first step to approach the challenge presented by the vast number of uncharacterized sesquiterpene synthases (STSs) was to gather in a public database all the relevant information available on the characterized plant STSs. Although in the last 35 years numerous studies focused on the characterization of natural STSs, there is still a large gap in the knowledge of these enzymes. In **Chapter 2** we describe the generation of a manually curated database of characterized plant STS[5] and report an initial number of 262 characterized STSs, producing 117 different compounds in total. By gathering these STSs, we observed that the sequence similarity is, in most cases, driven by the phylogenetic relationship rather than the

compound specificity. Considering the relatively small number of characterized synthases available, we suggest that finding common structural characteristics within the C-term domain is a promising starting point for the prediction of the uncharacterized synthases. To date, the database counts 307 enzymes covering 113 species and 117 sesquiterpene produced[6]. The database presented in **Chapter 2** can be easily extended to other TS families or used as scaffold for other classes of enzymes producing specialized metabolites.

Public databases of natural compounds such as COCONUT[7,8] report over 350 compounds having the general formula for sesquiterpenes $C_{15}H_{24}$. This infers that a large number of STSs remain uncharacterized, and their products can only be obtained by extraction from the natural source or via chemical synthesis, when possible. Characterizing such STSs is therefore still relevant to expand the number of compounds available for microbial production. Next to the inclusion of newly characterized synthases, a useful addition to the database could be data on the mutagenesis studies, including the change in product specificity and product profile obtained in the mutants. Evidently, the characterization of natural synthases represents the prelude for improving our knowledge of these enzymes, offering opportunities for specific case studies. The case studies are essential fragments of knowledge necessary to build accurate prediction tools[9]. Several studies aim to contribute to the development of accurate predictive tools based on machine learning[9,10] and feature extraction[11,12]. These tools, once trained on a reliable dataset, offer an initial screening approach for the uncharacterized TSs and allow a preliminary selection of native enzymes that might be of interest for characterization. The scope of this thesis was however not to develop such tools, but to contribute to build the fragments of knowledge.



As mentioned, a general study of structure-function relationships is bound by the amount of information available on the characterized synthases. Given that sequence similarity is high for synthases with different product specificity in the same species, it is crucial to investigate the effect of mutations in such synthases. The information obtained by these mutagenesis studies not only provides insights on the function of the specific residues in the cyclization cascade but can represent a starting point for evolutionary studies as well. In **Chapter 3** we performed mutagenesis studies of two TSs from *Cinnamomum camphora*[13]. The characterization of these TSs showed different product specificity despite the high sequence

similarity. In particular, CiCaMS showed monoterpene synthase (MTS) activity, producing myrcene, while CiCaSSy showed sesquiterpene synthase (STS) activity, producing a blend of santalenes and bergamotenes. We first addressed the 22 amino acid differences between these two TSs by grouping them in discrete regions. Three of the regions were located in the N-terminal end of the protein, while three more were in the C-terminal end. After determining the regions crucial for sesquiterpene synthase activity by generating hybrids between the two enzymes, we substituted each residue in the region to assess its function. We showed that only six substitutions were necessary to change the product profile and the synthase activity of CiCaMS from monoterpene to sesquiterpene. No major changes were observed when substituting the residues in the N-terminus, but profound effects were obtained by substituting the residues in the C-terminus.

While the work performed elucidated the residues crucial for the function, questions remain about the evolution of the synthases. It is known that, in plants, MTSs are localized in plastids[14,15] while STSs are normally localized in the cytoplasm[16]. This different localization is related to two aspects: the substrate production[14,17] and the presence of a transit peptide[18]. Although examples are known of STSs that also localize in the plastids[19] and vice versa[20], it is still widely recognized that the presence of a transit peptide in the amino acid sequence is necessary for the plastid localization. Furthermore, cleavage of the transit peptide is necessary for the correct folding and functionality of the mature protein[18,21,22]. In fact, due to the presence of the transit peptide, MTS genes encode usually for 600 to 650 amino acids, while STS genes encode for 530 to 600 amino acids[23]. Interestingly, both CiCaMS and CiCaSSy encode for 553 amino acids, suggesting that no transit peptide is present in these enzymes. The fact that both enzymes were functional with no need to remove the first 50-70 amino acid corroborates this hypothesis. In our work, we performed transient expression of both CiCaMS and CiCaSSy in *Nicotiana benthamiana* to characterize the production in plant. To assess the cellular localization of the synthases, fusion enzymes containing GFP could be cloned and transformed in *N. benthamiana,* and the transformed leaves analysed for fluorescence by confocal microscopy.

We observed that CiCaMS and CiCaSSy have a higher sequence similarity to other MTSs, such as the (-)-$\alpha$-terpineol synthases from *Cinnamomum micranthum*[24] and *Magnolia grandiflora*[25], than to other santalene synthases[26]. This suggests that these enzymes might be phylogenetically more closely

related to MTSs than STSs. These enzymes could in fact be MTSs that permanently lost their transit peptide[27]. While retaining MTS activity, the new cytosolic environment might have provided the conditions for the establishment of STS activity. Understanding how these enzymes evolve in nature contributes to direct enzyme engineering towards more strategic approaches. In this framework, it would be valuable to perform directed mutagenesis studies using the closely related (-)-α-terpineol synthases, aiming to change substrate specificity by few directed substitutions, as obtained for CiCaMS. A similar approach has been taken in the past to investigate the phylogenetic relationships between synthases, yielding relevant information on the structure-function relationships between the catalytic site and the substrate[28–30]. Our results could provide a starting point for determining the residues to address for mutagenesis in the (-)-α-terpineol synthases. If the substrate specificity would indeed change from GDP to FDP by few substitutions in the key residues identified in the catalytic site, we would have further indication of the phylogenetic relation of our enzymes to MTSs. Moreover, such results would contribute to strengthen the knowledge on the residues crucial for STS activity and would represent valuable information to integrate in a prediction tool. When studying the properties of catalytic pockets, computational tools such as Caretta[11], Geometricus[12] and CASTp[31] can support protein engineering by providing structural information.

By performing *in vitro* enzyme assays, we observed that both parental enzymes and all hybrids were able to produce MTs (myrcene and linalool) when supplied with GDP as substrate. Thus, unlike the STS activity, no substitutions were detrimental for MTS activity. Several promiscuous synthases have been identified to produce both mono- and sesquiterpenes[32–35]. CiCaSSy could then be considered a promiscuous synthase, although no linalool was detected when assayed in *N. benthamiana.* This might be due to the overall higher abundance of FDP in the cytosol or to the higher affinity of CiCaSSy for FDP. An analysis of the enzyme kinetics in the presence of the two different substrates could help elucidate this aspect[28]. The obtained hybrid would also provide useful insights on substrate affinity when transformed in a plant system. The *in vitro* system presents some advantages, such as being able to test the same samples with different substrates and accurate quantification of the substrate used in the reaction. However, it is important to remember that discrepancies could be observed with the *in vivo* system due to the artificial conditions of the assay. In these cases, the results validated *in vivo* should rather be preferred

over the ones obtained *in vitro*, as they will be more representative of the production system.

## 6.2 High throughput screening methods

If on the one hand directed mutagenesis is often a preferred approach for changes in substrate and product specificity, the improvement of enzyme activity is still bound to the generation of larger libraries of mutants. By employing approaches such as site saturation mutagenesis[36,37], the numbers of enzyme mutants obtained often exceed the throughput capacity of the methods available.

### 6.2.1 The challenge of implementing analytical techniques

While gas chromatography mass spectrometry (GC-MS) is still the preferred technique for the characterization of the product profile of synthases, the laborious sample preparation and the time required for analysis renders the method low-to-medium throughput. Recent work by Leferink *et al.* (2019)[38] describes the development of an automated pipeline for GC-MS-based analysis of MTS libraries. With this method, fully automated screening was achieved by combining a miniaturized system using 96 deep-well plates for biphasic growth handled by a robotic platform, a GC-MS compatible autosampler and automated data extraction scripts for data analysis. The complexity of this system makes this a very specialized approach, combining high-end analytical hardware with ad-hoc analytical scripts. In **Chapter 1** I provided an overview of the numerous studies aimed to develop less complex, generalised HTS methods. These studies present different approaches, from chemical-based enzyme assays using malachite green (MG)[39] or 2,2-diphenyl-1-picrylhydrazyl (DPPH)[40] to synthetic substrates such as vinyl methyl ether in combination with Purpald[41]. While these methods have relatively broad applicability, they have limited throughput because crude extracts or purified enzymes must be obtained for the assays. As mentioned earlier, methods that rely heavily on *in vitro* assays may produce results that are only partially representative of enzyme performance *in vivo*. Therefore, it is important to develop downscaling methods that are representative of industrial conditions in order to assess the performance of synthases *in vivo*.

In **Chapter 4** we explored the applicability of Proton Transfer Reaction Mass Spectrometry (PTR-MS) as a HTS method for sesquiterpene synthase mutants. PTR-MS has been widely employed for the detection of volatile

organic compounds (VOCs) in very diverse fields, from air quality control[42–44] to food technology[45–47] and medical diagnosis[48–50]. Another crucial advantage of PTR-MS systems is the possibility to measure headspace in real time, without disturbing the sample analysed[51] and without the need for an organic phase for extraction. This entails that the production of VOCs can be tracked over a longer period of time for the same samples[52], allowing to collect time series data. Real-time analysis of the VOCs by PTR-MS can be a powerful tool to understand enzyme kinetics and potentially optimize the production conditions and the harvesting time.

In our work, we aimed to test the reliability of PTR-MS measurements in a miniaturized fermentation system, employing 96 deep-well plates and a small library of *Callitropsis nootkatensis* valencene synthase[53] (CnVS) mutants. We could demonstrate a good correlation between the valencene produced by the control cultures grown in 20mL medium in flasks, measured with GC-MS, and the samples grown in 96 deep-well plates, measured with PTR-MS. The use of PTR-MS has proven valuable for the detection of sesquiterpenes[54] and it has been shown that it is possible to identify different isomers based on different fragmentation patterns[55,56], as is usually the case for GC-MS analysis. However, the PTR-MS system presents different challenges in this respect. First, because PTR-MS measures in real time, and all compounds in the samples are measured at the same time, it is not possible to determine a specific retention time for the different compounds in a mixture. The principal component analysis in PTR-MS is in fact based on the detection of compounds with different masses[45,57]. Therefore, the presence of isomers will make identification and accurate quantification in the presence of complex mixtures, such as bacterial cultures, very challenging. Although extraction scripts could be employed to attempt isomer quantification[55], for the application of PTR-MS as HTS method for STSs it is sufficient to obtain the total sesquiterpene produced to identify the enzyme(s) with the highest activity, and further proceed with characterization and validation using GC-MS as complementary method[58]. A preliminary experimental validation of the method with a second small library was performed using two mutants of CiCaSSy described in **Chapter 3**. From our previous characterization, we chose SSy_R2 as a higher performing enzyme and SSy_N267S for its altered product profile. In **Figure 6.1** we show a comparison of the measurements of total sesquiterpenes obtained with GC-MS (**Figure 6.1A**) and with PTR-MS (**Figure 6.1B**). While the correlation between GC-MS and PTR-MS results was not as high as observed for CnVS mutants, it was

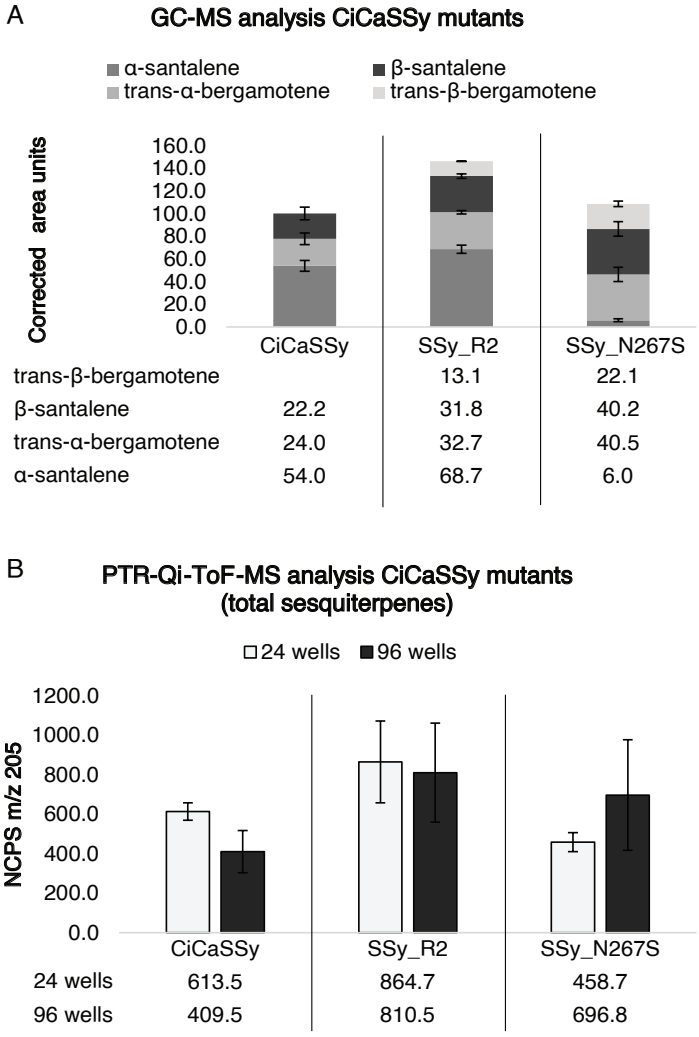



**Figure 6.1** Preliminary analysis using PTR-MS of a small mutant library derived from CiCaSSy compared with GC-MS analysis. (A) The GC-MS analysis reveals the different product profiles of mutant SSy_N267S compared to wild type CiCaSSy, while SSy_R2 shows overall higher activity. (B) The PTR-MS analysis shows similar profiles for the different enzymes. Compared to GC-MS, a higher variation can be observed resulting in a discrepancy between the production of sesquiterpenes detected in the 24-well plate compared to the 96-well plate, especially for mutant SSy_N267S. However, PTR-MS is also in this case useful to determine the higher performing enzyme.

still possible to identify SSy_R2 as the best performing enzyme. Based on the observations, SSy_N267S does not outperform CiCaSSy, although its altered product profile might be of interest. This example is relevant as it shows that it is possible to miss unexpected but desirable mutants. In such cases, it could be advisable to include enzymes showing similar activity to the wild type.

As pointed out above, using PTR-MS as a HTS method is preferable when selecting for higher production, rather than for a different product profile. This shortcoming can be overcome when purified standards of the single compounds are available. As described by Misztal *et al.* (2012)[55] for monoterpenes, it is then possible to have a precise ratio of each protonated ion associated with sesquiterpenes for each compound of interest. Having these ratios, an extraction script could be employed to determine the relative abundance of each different sesquiterpene in the samples. However, as we suggest using GC-MS as complementary method after the initial screening with PTR-MS, the assessment of the product profile might be delayed to this second characterization. What is nevertheless essential for the application of PTR-MS as a real HTS method is the implementation of a dedicated autosampler and, to improve on precision and reduce the chance of cross-contamination, the availability of a robotic liquid handling platform to prepare the plates for screening, as described by Leferink *et al.* (2019)[38]. Hence, our study provides a feasibility account and a starting point for the implementation of PTR-MS as a novel method for HTS of STS mutant libraries.

The development of HTS methods is a multidisciplinary endeavour. The need to innovate and adapt available systems to answer new questions requires open communication and collaboration between experts in biology, analytical chemistry, and mechanical engineering. Once such technologies are established, it would be beneficial to promote access for external researchers to avoid wasting resources.



### 6.2.2 The study of regulatory circuits for high throughput screening application

While the adaptation of available analytical techniques represents a valuable resource for the development of novel generalized HTS methods, the possibility to use genetic regulatory circuits to build compound-specific biosensors is just as attractive. Unlike the analytical techniques described above, that require a complex infrastructure of high-tech hardware and ad-

hoc software, transcriptional biosensors can be developed and employed with far less equipment. While final characterization of the identified synthase would necessarily be performed with GC-MS, the selection would require only the time necessary for the organism of choice to grow, allowing a much higher throughput compared to any analytical technique. In **Chapter 4** we suggest that with an automated, fully optimized system, the throughput of PTR-MS analysis could reach about 1,000 samples per day. A biosensor based on individual colonies on agar plates could allow for higher throughput, considering that each plate could contain 100 to 300 colonies and the plates could be measured for the fluorescence[59]. The only limitation would then be the time needed to analyse the images obtained with an on-plate reader. Systems have also been developed to select colonies for fluorescence directly on-plate[60], so that the most promising ones can directly be isolated and characterization can continue. Libraries of mutants transformed in biosensor cells can also be screened using fluorescence activated cell sorting (FACS)[61,62]. This method virtually allows the sorting of libraries containing millions of different clones per day, based on the fluorescence emitted and has been widely applied for enzyme development[63–66]. The development of a biosensor for sesquiterpenes would therefore open the possibility to screen large libraries of mutants using FACS.

The development of biosensors has however different complexities compared to analytical methods. One of the main challenges is to obtain a circuit that is specific and sensitive enough to concentration variations of the ligand. This often requires several steps of fine-tuning the system itself, followed by validation tests for the specific enzyme in analysis[67]. An elegant example of development of a transcriptional biosensor, for naringenin, is described by De Paepe *et al.* (2018)[68]. Naringenin is a trihydroxy flavanone belonging to the flavonoid family and is found in a wide variety of plants. The key element to develop such a biosensor is evidently the identification of a native regulatory circuit containing the transcriptional regulator (TR) sensitive to the compound of interest. To our knowledge, no TRs of the helix-turn-helix (HTH) type have been identified in bacteria for being sensitive to sesquiterpenes. A promising candidate was identified in *Corynebacterium glutamicum* in a TetR-like TR associated with a putative mmpL2 multidrug efflux pump[69].

We investigated the function of this TR in a heterologous system, modelled on the design proposed by De Paepe *et al.* (2018)[68]. Two complementary

plasmids were generated following the detector/effector module system (**Figure 6.2A**). By transforming the effector module in *Escherichia coli* alone, we first tested the functionality of the Transcriptional Regulator Binding Site (TRBS) in the absence of the regulator. Second, by combining the two modules we attempted to test the interaction of the TRBS and the TR (**Figure 6.2C**). With the native *Corynebacterium glutamicum* predicted TRBS (**Figure 6.2B**) we could not obtain a functional construct in the *E. coli* system, as no RFP was produced in the absence of the TR. A possible explanation might be the lack of recognition of the exogenous TRBS by *E. coli*. Recognition might be facilitated by adding extra ribosomal binding sites in the proximity of the TRBS. Another approach could be to generate chimeric TRs. Several studies have shown the possibility of engineering the DNA-binding domains (DBDs) from a well characterized TR and ligand binding domains (LBDs) from an uncharacterized TR[70,71], with the aim of identifying the ligand specificity of the uncharacterized TR[72]. This approach may be supported by dedicated algorithms trained to predict the most promising residues to use for the fusion of the two domains[71]. We designed a chimeric TR using the DBDs from TetR (PDB ID: 3ZQI) and CprB[73] (PDB ID: 4PXI) and the LBD of Cg1053 in the attempt of identifying the Cg1053 ligand.

In a preliminary experiment, the chimeric TR CprB-Cg1053 appeared to be able to bind CprB TRBS, but we could not identify a ligand able to release the TR and activate the transcription of RFP. Similarly, we tested the design using the repressor of acyclic terpene utilization, AtuR from *Pseudomonas aeruginosa* described by Förster-Fromme and Jendrossek (2010)[74]. The expression of the *atu* gene cluster in *P. aeruginosa* is dependent on the presence of acyclic monoterpenes in the medium, suggesting that AtuR plays a role in the sensing of these compounds. When AtuR and its TRBS were implemented in our biosensor constructs, we could observe repressed expression of RFP. However, we could not induce the TR by adding different acyclic monoterpenes to the medium. One explanation could be the inability of the *E. coli* heterologous system to take up the acyclic monoterpenes. Evidently, if the ligand is not transported inside of the bacterial cell, the TR will not be able to bind it and release the TRBS. To overcome this, it would be useful to test the biosensor construct in an *E. coli* system producing acyclic monoterpenes.

These experiments represented a first approach to the development of a biosensor for sesquiterpenes and require further work to be conclusive.

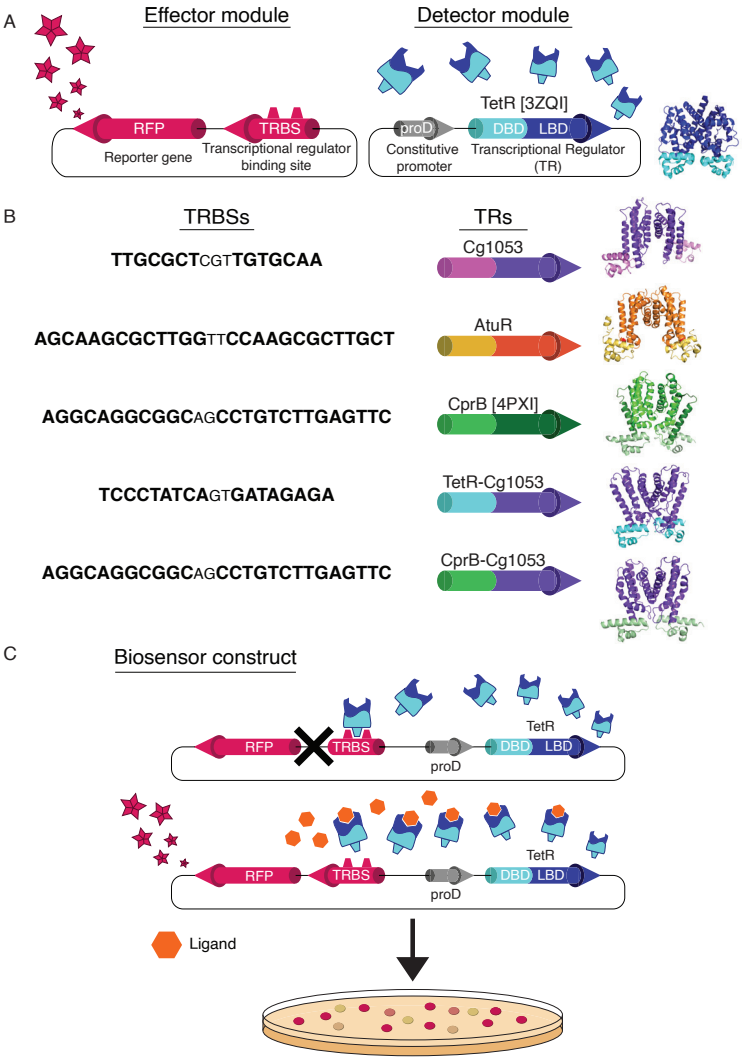



**Figure 6.2** Schematic overview of a genetic biosensor. (A) The general biosensor design consists of two separate modules: the effector module, carrying the TRBS and the reporter gene (RFP), and the detector module, carrying the TR under a constitutive promoter. To test the functionality of the designed system, the *E. coli* TetR was used[70] (PDB: 3ZQI). (B) The biosensor design was used to test two candidates from literature: the Cg1053 TetR-like regulator from *C. glutamicum*[69] with its predicted TRBS and the AtuR monoterpene utilization regulator from *P. putida*[74] with its TRBS. TetR and CprB (PDB: 4PXI) from *S. coelicolor*[73] were used for the generation of chimeric TRs, using their DNA binding domains (DBD) in combination with the ligand binding domain (LDB) of Cg1053, to identify the ligand for Cg1053. Models of the TRs structure were obtained using SWISS-MODEL[108]. (C) Graphic representation of the functional mechanism of the genetic biosensor. In the studied cases, the TRs act as repressors, and no reporter gene is expressed in the absence of ligand. Conversely, in the presence of ligand, the reporter gene is expressed. When the ligand is produced endogenously, the response of the biosensor is expected to be concentration-dependent, resulting in colonies of different colours depending on the efficiency of the enzyme producing the ligand.

However, both Cg1053 and AtuR do not seem to be able to bind sesquiterpenes as ligands. We therefore approached this challenge from a different perspective, aiming to identify sesquiterpene-specific TRs, we explored the differential expression of genes in *Rhodobacter sphaeroides* in the presence of endogenous valencene.

Investigating the regulatory activity of TRs in their native microorganism under different conditions is one of the best approaches to identify candidate TRs to be implemented in a biosensor construct. Since the effects observed in regulation could be due to secondary effects such as solvent sensitivity, ligand modifications or ligand availability, it is crucial to consider these aspects when design the experiment. In **Chapter 5** we report the differentially expressed genes (DEGs) in a *R. sphaeroides* strain supplemented with CnVS compared to the same strain harbouring an empty vector as control. With our RNA-seq analysis we identified two TRs, associated with DEGs. As the identified TRs were not previously characterized, we proposed two new names, based on the genomic cluster they are associated with. RSP_1297 was annotated as a LysR family TR and is localized in the neighbourhood of genes involved in the metabolism of valine, leucine and isoleucine. For this reason, we named it IlvR[75]. The second TR, RSP_3339, was annotated as a GntR family TR. For its association with the spermidine/putrescine ABC transporter gene cluster, we proposed the name "putrescine/spermidine transporter regulator", PstR. To characterize these TRs and study their function in *R. sphaeroides* we generated knock-out (KO) strains for each TR.

By obtaining *ΔIlvR* and *ΔPstR* strains, we were able to compare the production of valencene and the production of other isoprenoids naturally produced by *R. sphaeroides*, namely bacteriochlorophylls and carotenoids. While no major differences from the wild type were observed for *ΔPstR*, of particular interest were the results obtained for *ΔIlvR*. This knockout showed a higher production of valencene compared to the wild type and a lower production of other isoprenoids, although bacterial growth was consistently lower than the wild type. The regulation of *R. sphaeroides* has been studied using different integrated approaches[76–78]. However, the role of many TRs remains to be elucidated. Metabolic plasticity has been studied for model microorganisms such as *E. coli*[79], and it is clear that altering regulatory circuits by deleting TRs causes a general rearrangement of gene expression.

Our results suggest that IlvR acts as a transcriptional activator of the adjacent operon. This is, however, only a partial representation of the role of IlvR and does not yet answer the question whether this TR uses valencene as a ligand. Further experiments are still necessary to characterize the binding properties of IlvR and PstR. First, the TRBS should be identified for both TRs. By performing electromobility shift assays (EMSA) using the predicted binding motifs, as described for AtuR[74], it will be possible to identify the minimum sequence necessary for the TRs to bind and activate or repress transcription. Computational methods have also been developed to predict TRBS[80,81] and regulatory motifs[82,83] and could aid the identification in the *R. sphaeroides* genome. Second, the TRs must be tested for their affinity for the ligand of interest. There are several experimental approaches to investigate the binding properties of TRs, from *in vitro* ligand binding assays[84–86] to heterologous *in vivo* expression of the native TR[87,88] or using chimeric TRs[70,72,89]. Finally, to elucidate the global effect of the absence of the identified TRs in *R. sphaeroides*, RNA-seq should be performed on the KOs, in comparison with the wild type.

Ultimately, the identified TRs could be implemented in a biosensor system designed as previously described (**Figure 6.2**) to test the ability of the TRs to bind to the ligands of interest in the system of choice. Gaining insights into the regulatory circuits of microbial production platforms such as *R. sphaeroides* is a valuable exercise in itself, as it is still possible to improve production without mutating the synthase. *R. sphaeroides* has a very versatile metabolism[90,91] and its ecological niche makes it a plausible candidate for natural regulatory circuits able to sense sesquiterpenes. However, the mechanisms for sensing these compounds might reside in structures other than HTH-type TRs, such as two component systems[92,93] or chemosensory pathways[94,95]. To increase the chances of finding sesquiterpene-specific TRs, it would be useful to investigate the genomes of sesquiterpene-producing bacteria[96,97], such as cyanobacteria *Nostoc punctiforme*[98]and *Anabaena variabilis*[96], or actinobacteria *Streptomyces citricolor*[99] and *Streptomyces viridochromogenes*[100]. Next to soil ecosystems, other ecological niches that might favour the presence of regulatory circuits specific for sesquiterpenes are the microbial ecosystems of flowers[101–105] and fruits[106,107]. A thorough investigation of the regulatory networks of these microbial communities could provide the basis for identifying candidate TRs. Once again, a concerted approach based on genome mining and protein predictions, together with experimental validation, would be ideal for a timely translation of native circuits into biotechnological tools.

## 6.3 Closing remarks

Reliable databases are essential for the development of novel bioinformatics tools and constitute the backbone of current knowledge. Protein engineering is nowadays increasingly bound to the predictive power of available tools but still depends on biotechnologist's ability to access these tools. As protein engineering is largely facilitated by computational tools, the prediction tools are only as good as the experimental data they contain. To increase the efficiency of research in this field, it seems almost inevitable to establish closer collaboration between bioinformaticians and biotechnologists to facilitate exchange of information between the different disciplines.

In this thesis, the focus was mainly on sesquiterpenes for flavours and fragrances and their synthases. However, the proposed research approaches extend beyond this niche. Analytical technologies such PTR-MS have widely proven their flexibility in different applications and still provide the possibility for novel approaches to HTS. The availability of miniaturized fermentation setups opens the possibility for fast and reliable analysis of any library of VOCs-producing enzyme engineered in a microbial host.

Finally, the synergistic study of metabolic regulatory circuits of known microorganisms combined with genome mining of newly discovered species offers a remarkable opportunity for the identification of TRs to implement in biosensor systems. Considering the enormous number of microbial species identified in the different ecological niches, very few regulatory networks have been elucidated and experimentally validated. While it may be initially overwhelming to deal with such enormous task, the prospect of discovering TRs to engineer for novel biotechnological applications is extremely enthusing. Once again, curated databases are essential to make information accessible and reliable.



This thesis explored the challenges of developing novel synthases for flavours and fragrances using a multidisciplinary approach that combines elements of bioinformatics with protein engineering and applied analytical chemistry. I conclude this discussion by pointing to the future development and broad applicability of the approaches described, with particular emphasis on the growing need for close collaboration between different disciplines. Only with concerted multidisciplinary efforts such as these can we fruitfully develop new biotechnological solutions to use the wealth of natural products in a sustainable manner.

## References


1. Croteau, R. & Gundy, A. Cyclization of farnesyl pyrophosphate to the sesquiterpene olefins humulene and caryophyllene by an enzyme system from sage (Salvia officinalis). *Arch. Biochem. Biophys.* **233**, 838–841 (1984).
2. Dehal, S. S. & Croteau, R. Partial purification and characterization of two sesquiterpene cyclases from sage (Salvia officinalis) which catalyze the respective conversion of farnesyl pyrophosphate to humulene and caryophyllene. *Arch. Biochem. Biophys.* **261**, 346–356 (1988).
3. Hohn, T. M. & Beremand, P. D. Isolation and nucleotide sequence of a sesquiterpene cyclase gene from the trichothecene-producing fungus Fusarium sporotrichioides. *Gene* **79**, 131–138 (1989).
4. Hohn, T. M. & Plattner, R. D. Expression of the trichodiene synthase gene of Fusarium sporotrichioides in Escherichia coli results in sesquiterpene production. *Arch. Biochem. Biophys.* **275**, 92–97 (1989).
5. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).
6. Characterized Plant Sesquiterpene Synthases. https://www.bioinformatics.nl/sesquiterpene/synthasedb/.
7. COCONUT: Natural Products Online. https://coconut.naturalproducts.net/.
8. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J. Cheminform.* **13**, 1–13 (2021).
9. Durairaj, J. *et al.* Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comput. Biol.* **17**, (2021).
10. Priya, P., Yadav, A., Chand, J. & Yadav, G. Terzyme: A tool for identification and analysis of the plant terpenome. *Plant Methods* **14**, 1–18 (2018).
11. Akdel, M., Durairaj, J., de Ridder, D. & van Dijk, A. D. J. Caretta – A multiple protein structure alignment and feature extraction suite. *Comput. Struct. Biotechnol. J.* **18**, 981–992 (2020).
12. Durairaj, J., Akdel, M., de Ridder, D. & van Dijk, A. D. J. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics* **36**, I718–I725 (2020).
13. Di Girolamo, A. *et al.* The santalene synthase from Cinnamomum camphora: Reconstruction of a sesquiterpene synthase from a monoterpene synthase. *Arch. Biochem. Biophys.* **695**, (2020).
14. Dong, L., Jongedijk, E., Bouwmeester, H. & Van Der Krol, A. Monoterpene biosynthesis potential of plant subcellular compartments. *New Phytol.* **209**, 679–690 (2016).
15. Turner, G., Gershenzon, J., Nielson, E. E., Froehlich, J. E. & Croteau, R. Limonene synthase, the enzyme responsible for monoterpene biosynthesis in peppermint, is localized to leucoplasts of oil gland secretory cells. *Plant Physiol.* **120**, 879–886 (1999).
16. Nagegowda, D. A. Plant volatile terpenoid metabolism: Biosynthetic genes, transcriptional regulation and subcellular compartmentation. *FEBS Letters* vol. 584 2965–2973 (2010).
17. Vranová, E., Coman, D. & Gruissem, W. Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annual Review of Plant Biology* vol. 64 665–700 (2013).
18. Williams, D. C., McGarvey, D. J., Katahira, E. J. & Croteau, R. Truncation of limonene synthase preprotein provides a fully active 'pseudomature' form of this monoterpene cyclase and reveals the function of the amino-terminal arginine pair. *Biochemistry* **37**, 12213–12220 (1998).
19. Zhu, B. Q. *et al.* Identification of a plastid-localized bifunctional nerolidol/linalool synthase in relation to linalool biosynthesis in young grape berries. *Int. J. Mol. Sci.* **15**, 21992–22010 (2014).
20. Gutensohn, M. *et al.* Cytosolic monoterpene biosynthesis is supported by plastid-generated geranyl diphosphate substrate in transgenic tomato fruits. *Plant J.* **75**, 351–363 (2013).
21. Bohlmann, J., Steele, C. L. & Croteau, R. Monoterpene synthases from grand fir (Abies grandis): cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)-limonene synthase, and (-)-(1S,5S)-pinene synthase. *J. Biol. Chem.* **272**, 21784–21792 (1997).
22. Reichardt, S. *et al.* The carrot monoterpene synthase gene cluster on chromosome 4 harbours genes encoding flavour-associated sabinene synthases. *Hortic. Res.* **7**, 190 (2020).

23. Bohlmann, J., Meyer-Gauen, G. & Croteau, R. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4126–4133 (1998).
24. Chaw, S. M. *et al.* Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* **5**, 63–73 (2019).
25. Lee, S. & Chappell, J. Biochemical and genomic characterization of terpene synthases in Magnolia grandiflora. *Plant Physiol.* **147**, 1017–1033 (2008).
26. Jones, C. G. *et al.* Sandalwood fragrance biosynthesis involves sesquiterpene synthases of both the terpene synthase (TPS)-a and TPS-b subfamilies, including santalene synthases. *J. Biol. Chem.* **286**, 17445–17454 (2011).
27. Hillwig, M. L. *et al.* Domain loss has independently occurred multiple times in plant terpene synthase evolution. *Plant J.* **68**, 1051–1060 (2011).
28. Kampranis, S. C. *et al.* Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: Structural insights into the evolution of terpene synthase function. *Plant Cell* **19**, 1994–2005 (2007).
29. Li, J. X. *et al.* Rational engineering of plasticity residues of sesquiterpene synthases from Artemisia annua: Product specificity and catalytic efficiency. *Biochem. J.* **451**, 417–426 (2013).
30. Lei, D., Qiu, Z., Qiao, J. & Zhao, G. R. Plasticity engineering of plant monoterpene synthases and application for microbial production of monoterpenoids. *Biotechnology for Biofuels* vol. 14 1–15 (2021).
31. Tian, W., Chen, C., Lei, X., Zhao, J. & Liang, J. CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.* **46**, W363–W367 (2018).
32. Johnson, S. R. *et al.* Promiscuous terpene synthases from Prunella vulgaris highlight the importance of substrate and compartment switching in terpene synthase evolution. *New Phytol.* **223**, 323–335 (2019).
33. Pazouki, L. & Niinemetst, U. Multi-substrate terpene synthases: Their occurrence and physiological significance. *Front. Plant Sci.* **7**, (2016).
34. Pazouki, L., Memari, H. R., Kännaste, A., Bichele, R. & Niinemets, Ü. Germacrene A synthase in yarrow (Achillea millefolium) is an enzyme with mixed substrate specificity: Gene cloning, functional characterization and expression analysis. *Front. Plant Sci.* **6**, (2015).
35. Ashaari, N. S. *et al.* Functional characterization of a new terpene synthase from Plectranthus amboinicus. *PLoS One* **15**, (2020).
36. Szymczyk, P., Szymańska, G., Lipert, A., Weremczuk-Jeżyna, I. & Kochan, E. Computer-Aided Saturation Mutagenesis of Arabidopsis thaliana Ent-Copalyl Diphosphate Synthase. *Interdiscip. Sci. Comput. Life Sci.* **12**, 32–43 (2020).
37. Labrou, N. Random Mutagenesis Methods for In Vitro Directed Enzyme Evolution. *Curr. Protein Pept. Sci.* **11**, 91–100 (2010).
38. Leferink, N. G. H. *et al.* An automated pipeline for the screening of diverse monoterpene synthase libraries. *Sci. Rep.* **9**, 1–12 (2019).
39. Vardakou, M., Salmon, M., Faraldos, J. A. & O'Maille, P. E. Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* **1**, e187–e196 (2014).
40. Behrendorff, J. B. Y. H., Vickers, C. E., Chrysanthopoulos, P. & Nielsen, L. K. 2,2-Diphenyl-1-picrylhydrazyl as a screening tool for recombinant monoterpene biosynthesis. *Microb. Cell Fact.* **12**, 76 (2013).
41. Lauchli, R. *et al.* High-throughput screening and directed evolution of terpene synthase-catalyzed cylization. *Angew Chem Int Ed Engl* **52**, 1–12 (2013).
42. Xia, S. Y. *et al.* Long-term observations of oxygenated volatile organic compounds (OVOCs) in an urban atmosphere in southern China, 2014–2019. *Environ. Pollut.* **270**, (2021).
43. Schripp, T. *et al.* Application of proton-transfer-reaction-mass-spectrometry for Indoor air quality research. *Indoor Air* **24**, 178–189 (2014).
44. Maji, S., Beig, G. & Yadav, R. Winter VOCs and OVOCs measured with PTR-MS at an urban site of India: Role of emissions, meteorology and photochemical sources. *Environ. Pollut.* **258**,

(2020).
45. Liu, N., Koot, A., Hettinga, K., de Jong, J. & van Ruth, S. M. Portraying and tracing the impact of different production systems on the volatile organic compound composition of milk by PTR-(Quad)MS and PTR-(ToF)MS. *Food Chem.* **239**, 201–207 (2018).
46. Bottiroli, R. *et al.* Application of PTR-TOF-MS for the quality assessment of lactose-free milk: Effect of storage time and employment of different lactase preparations. *J. Mass Spectrom.* **55**, (2020).
47. Pedrotti, M. *et al.* Rapid and noninvasive quality control of anhydrous milk fat by PTR-MS: The effect of storage time and packaging. *J. Mass Spectrom.* **53**, 753–762 (2018).
48. Löser, B. *et al.* Changes of exhaled volatile organic compounds in postoperative patients undergoing analgesic treatment: A prospective observational study. *Metabolites* **10**, 1–13 (2020).
49. Cassagnes, L. E. *et al.* Online monitoring of volatile organic compounds emitted from human bronchial epithelial cells as markers for oxidative stress. *J. Breath Res.* **15**, (2020).
50. Li, Q., Li, J., Wei, X., Li, Y. & Sun, M. An exploratory study on online quantification of isoprene in human breath using cavity ringdown spectroscopy in the ultraviolet. *Anal. Chim. Acta* **1131**, 18–24 (2020).
51. Capozzi, V. *et al.* PTR-ToF-MS coupled with an automated sampling system and tailored data analysis for food studies: Bioprocess monitoring, screening and nose-space analysis. *J. Vis. Exp.* **2017**, 1–11 (2017).
52. Tejero Rioseras, A. *et al.* Comprehensive Real-Time Analysis of the Yeast Volatilome. *Sci. Rep.* **7**, 1–9 (2017).
53. Beekwilder, J. *et al.* Valencene synthase from the heartwood of Nootka cypress (Callitropsis nootkatensis) for biotechnological production of valencene. *Plant Biotechnol. J.* **12**, 174–182 (2014).
54. Kim, S. *et al.* Measurement of atmospheric sesquiterpenes by proton transfer reaction-mass spectrometry (PTR-MS). *Atmos. Meas. Tech.* **2**, 99–112 (2009).
55. Misztal, P. K., Heal, M. R., Nemitz, E. & Cape, J. N. Development of PTR-MS selectivity for structural isomers: Monoterpenes as a case study. *Int. J. Mass Spectrom.* **310**, 10–19 (2012).
56. Tani, A. Fragmentation and reaction rate constants of terpenoids determined by proton transfer reaction-mass spectrometry. *Environ. Control Biol.* **51**, 23–29 (2013).
57. Ellis, A. M. & Mayhew, C. A. *Proton Transfer Reaction Mass Spectrometry*. *Proton Transfer Reaction Mass Spectrometry: Principles and Applications* (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118682883.
58. Majchrzak, T. *et al.* PTR-MS and GC-MS as complementary techniques for analysis of volatiles: A tutorial review. *Analytica Chimica Acta* vol. 1035 1–13 (2018).
59. Margolin, W. Green fluorescent protein as a reporter for macromolecular localization in bacterial cells. *Methods* **20**, 62–72 (2000).
60. Környei, Z. *et al.* Cell sorting in a Petri dish controlled by computer vision. *Sci. Rep.* **3**, (2013).
61. Michener, J. K. & Smolke, C. D. High-throughput enzyme evolution in Saccharomyces cerevisiae using a synthetic RNA switch. *Metab. Eng.* **14**, 306–316 (2012).
62. Bott, M. Need for speed - finding productive mutations using transcription factor-based biosensors, fluorescence-activated cell sorting and recombineering. *Microbial Biotechnology* vol. 8 8–10 (2015).
63. Binder, S. *et al.* A high-throughput approach to identify genomic variants of bacterial metabolite producers at the single-cell level. *Genome Biol.* **13**, 1–12 (2012).
64. Schallmey, M., Frunzke, J., Eggeling, L. & Marienhagen, J. Looking for the pick of the bunch: high-throughput screening of producing microorganisms with biosensors. *Curr. Opin. Biotechnol.* **26**, 148–154 (2014).
65. Jang, S., Kim, M., Hwang, J. & Jung, G. Y. Tools and systems for evolutionary engineering of biomolecules and microorganisms. *J. Ind. Microbiol. Biotechnol.* **46**, 1313–1326 (2019).
66. Townshend, B., Xiang, J. S., Manzanarez, G., Hayden, E. J. & Smolke, C. D. A multiplexed, automated evolution pipeline enables scalable discovery and characterization of biosensors.

*Nat. Commun.* **12**, (2021).
67. De Paepe, B., Peters, G., Coussement, P., Maertens, J. & De Mey, M. Tailor-made transcriptional biosensors for optimizing microbial cell factories. *J. Ind. Microbiol. Biotechnol.* **44**, 623–645 (2017).
68. De Paepe, B., Maertens, J., Vanholme, B. & De Mey, M. Modularization and Response Curve Engineering of a Naringenin-Responsive Transcriptional Biosensor. *ACS Synth. Biol.* **7**, 1303–1314 (2018).
69. Frohwitter, J., Heider, S. A. E., Peters-Wendisch, P., Beekwilder, J. & Wendisch, V. F. Production of the sesquiterpene (+)-valencene by metabolically engineered Corynebacterium glutamicum. *J. Biotechnol.* **191**, 205–213 (2014).
70. Dimas, R. P. *et al.* Engineering DNA recognition and allosteric response properties of TetR family proteins by using a module-swapping strategy. *Nucleic Acids Res.* **47**, 8913–8925 (2019).
71. Dimas, R. P., Jiang, X. L., De La Paz, J. A., Morcos, F. & Chan, C. T. Y. Engineering repressors with coevolutionary cues facilitates toggle switches with a master reset. *Nucleic Acids Res.* **47**, 5449–5463 (2019).
72. De Paepe, B., Maertens, J., Vanholme, B. & De Mey, M. Chimeric LysR-Type Transcriptional Biosensors for Customizing Ligand Specificity Profiles toward Flavonoids. *ACS Synth. Biol.* **8**, 318–331 (2019).
73. Bhukya, H., Bhujbalrao, R., Bitra, A. & Anand, R. Structural and functional basis of transcriptional regulation by TetR family protein CprB from S. coelicolorA3(2). *Nucleic Acids Res.* **42**, 10122–10133 (2014).
74. Förster-Fromme, K. & Jendrossek, D. AtuR is a repressor of acyclic terpene utilization (Atu) gene cluster expression and specifically binds to two 13 bp inverted repeat sequences of the atuA-atuR intergenic region. *FEMS Microbiol. Lett.* **308**, 166–174 (2010).
75. Malakooti, J. & Ely, B. Identification and characterization of the ilvR gene encoding a LysR-type regulator of Caulobacter crescentus. *J. Bacteriol.* **176**, 1275–1281 (1994).
76. Imam, S., Noguera, D. R. & Donohue, T. J. An Integrated Approach to Reconstructing Genome-Scale Transcriptional Regulatory Networks. *PLoS Comput. Biol.* **11**, 1004103 (2015).
77. Imam, S., Noguera, D. R. & Donohue, T. J. Global Analysis of Photosynthesis Transcriptional Regulatory Networks. *PLoS Genet.* **10**, 1004837 (2014).
78. Burger, B. T., Imam, S., Scarborough, M. J., Noguera, D. R. & Donohue, T. J. Combining Genome-Scale Experimental and Computational Methods To Identify Essential Genes in Rhodobacter sphaeroides. *mSystems* **2**, (2017).
79. McCloskey, D. *et al.* Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism. *Nat. Commun.* **9**, (2018).
80. Jayaram, N., Usvyat, D. & Martin, A. C. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* **17**, (2016).
81. Mercier, E. *et al.* An Integrated Pipeline for the Genome-Wide Analysis of Transcription Factor Binding Sites from ChIP-Seq. *PLoS One* **6**, e16432 (2011).
82. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
83. Bailey, T. L. & Noble, W. S. Searching for statistically significant regulatory modules. *Bioinformatics* **19**, ii16–ii25 (2003).
84. Pollard, T. D. MBOC technical perspective: A guide to simple and informative binding assays. *Molecular Biology of the Cell* vol. 21 4061–4067 (2010).
85. Hirooka, K. & Fujita, Y. Identification of Aromatic Residues Critical to the DNA Binding and Ligand Response of the Bacillus subtilis QdoR (YxaF) Repressor Antagonized by Flavonoids. *Biosci. Biotechnol. Biochem.* **75**, 1325–1334 (2011).
86. Hirooka, K. *et al.* Dual Regulation of the Bacillus subtilis Regulon Comprising the lmrAB and yxaGH Operons and yxaF Gene by Two Transcriptional Repressors, LmrA and YxaF, in Response to Flavonoids. *J. Bacteriol.* **189**, 5170–5182 (2007).
87. Sang, K. A., Tahlan, K., Yu, Z. & Nodwell, J. Investigation of transcription repression and small-

molecule responsiveness by TetR-Like transcription factors using a heterologous Escherichia coli-based assay. *J. Bacteriol.* **189**, 6655–6664 (2007).
88. Siedler, S., Stahlhut, S. G., Malla, S., Maury, J. & Neves, A. R. Novel biosensors based on flavonoid-responsive transcriptional regulators introduced into Escherichia coli. *Metab. Eng.* **21**, 2–8 (2014).
89. Mahfouz, M. M. *et al.* Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant Mol. Biol.* **78**, 311–321 (2012).
90. Bathke, J., Konzer, A., Remes, B., McIntosh, M. & Klug, G. Comparative analyses of the variation of the transcriptome and proteome of Rhodobacter sphaeroides throughout growth. *BMC Genomics* **20**, (2019).
91. Mackenzie, C. *et al.* Postgenomic adventures with Rhodobacter sphaeroides. *Annual Review of Microbiology* vol. 61 283–307 (2007).
92. Schmidt, R. *et al.* Fungal volatile compounds induce production of the secondary metabolite Sodorifen in Serratia plymuthica PRI-2C. *Sci. Rep.* **7**, 862 (2017).
93. Amaya, S. *et al.* Inhibition of quorum sensing in Pseudomonas aeruginosa by sesquiterpene lactones. *Phytomedicine* **19**, 1173–1177 (2012).
94. Martín-Mora, D. *et al.* Functional Annotation of Bacterial Signal Transduction Systems: Progress and Challenges. *Int. J. Mol. Sci.* **19**, 3755 (2018).
95. Wuichet, K. & Zhulin, I. B. Origins and Diversification of a Complex Signal Transduction System in Prokaryotes. *Sci. Signal.* **3**, (2010).
96. Yamada, Y. *et al.* Terpene synthases are widely distributed in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 857–862 (2015).
97. Dickschat, J. S. Bacterial terpene cyclases. *Nat. Prod. Rep.* **33**, 87–110 (2016).
98. Agger, S. A., Lopez-Gallego, F., Hoye, T. R. & Schmidt-Dannert, C. Identification of sesquiterpene synthases from Nostoc punctiforme PCC 73102 and Nostoc sp. strain PCC 7120. *J. Bacteriol.* **190**, 6084–6096 (2008).
99. Nakano, C., Kudo, F., Eguchi, T. & Ohnishi, Y. Genome mining reveals two novel bacterial sesquiterpene cyclases: (-)-germacradien-4-ol and (-)-epi-α-bisabolol synthases from Streptomyces citricolor. *ChemBioChem* **12**, 2271–2275 (2011).
100. Rabe, P., Schmitz, T. & Dickschat, J. S. Mechanistic investigations on six bacterial terpene cyclases. *Beilstein J. Org. Chem.* **12**, 1839–1850 (2016).
101. Aleklett, K., Hart, M. & Shade, A. The microbial ecology of flowers: An emerging frontier in phyllosphere research1. *Botany* **92**, 253–266 (2014).
102. Massoni, J., Bortfeld-Miller, M., Widmer, A. & Vorholt, J. A. Capacity of soil bacteria to reach the phyllosphere and convergence of floral communities despite soil microbiota variation. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
103. Vannette, R. L. The Floral Microbiome: Plant, Pollinator, and Microbial Perspectives. *Annual Review of Ecology, Evolution, and Systematics* vol. 51 363–386 (2020).
104. Rebolleda-Gómez, M. *et al.* Gazing into the anthosphere: considering how microbes influence floral evolution. *New Phytologist* vol. 224 1012–1020 (2019).
105. Alvarez-Perez, S. *et al.* Acinetobacter pollinis sp. nov., Acinetobacter baretiae sp. nov. and Acinetobacter rathckeae sp. nov., isolated from floral nectar and honey bees. *Int. J. Syst. Evol. Microbiol.* **71**, (2021).
106. Droby, S. & Wisniewski, M. The fruit microbiome: A new frontier for postharvest biocontrol and postharvest biology. *Postharvest Biol. Technol.* **140**, 107–112 (2018).
107. Zhang, H. *et al.* Unravelling the fruit microbiome: The key for developing effective biological control strategies for postharvest diseases. *Comprehensive Reviews in Food Science and Food Safety* vol. 20 4906–4930 (2021).
108. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

## Summary

The terpenoids form one of the largest classes of natural compounds. Terpenoids are widely produced by plants and microorganisms and play an important role in various fields, from food and cosmetics to healthcare and agriculture. The production of these compounds is mediated by terpene synthases, enzymes with a high diversity in their sequences but with a remarkably similar general structure. The identification of terpene synthases in plants made it possible to shift the extraction of terpenes from natural resources to production in microbial cell factories. This thesis expands on the current knowledge on sesquiterpene synthases by addressing two key challenges: the generation of smart libraries for terpene synthases to study functional residues and the development of novel approaches for high throughput screening methods, providing relevant examples.

In **Chapter 1**, I introduce the terpene synthases and the metabolic pathways responsible for the production of the substrates. By gathering the known methods of generating enzyme mutant libraries and the methods used for screening, I build the framework for this thesis. A thorough analysis of the strength and weaknesses of the different methods allows for a clearer starting point for the development of novel strategies. With the aim of providing a collection of characterized sesquiterpene synthases, in **Chapter 2** we collected all the described enzymes in a manually curated database. We provide an analysis based on protein sequence, and we observe that phylogenetic relations affect sequence similarities more than product specificity. We suggest that structural analysis could represent a more suitable approach to study the synthases and predict their products. Mutagenesis studies of closely related enzymes provide important insights on the residues critical for cyclization and product specificity. An example is explored in **Chapter 3**, where a case study of two synthases from *Cinnamomum camphora* having high sequence similarity but different product specificity is presented. By employing both domain swapping and site-directed mutagenesis, we identify the residues crucial for the sesquiterpene synthase activity, determining both product profile and product specificity.

The attention is then moved from the study of the synthases' biochemistry to the development of screening methods. In **Chapter 4** we propose the use of Proton Transfer Reaction Mass Spectrometry (PTR-MS) as a high throughput screening method for valencene mutant libraries. We show that PTR-MS can provide a sensitive and reliable method to measure the

valencene produced in a small-scale fermentation system using 96 deep-well plates. We convey that the use of PTR-MS offers a promising time efficient method for high throughput screening of terpene synthase mutant libraries, provided the implementation of a dedicated auto-sampler. Finally, **Chapter 5** aims to identify a candidate valencene-sensitive transcriptional regulator, through a comparative expression analysis of *Rhodobacter sphaeroides*. RNA-seq analysis was performed on a wild type *R. sphaeroides* supplemented with a valencene synthase against a wild type *R. sphaeroides* harbouring an empty vector. We identify two transcriptional regulators associated with differentially expressed genes in the presence of the valencene synthase. By generating knockout mutants of these regulators in *R. sphaeroides*, we attempt to characterize their function. We suggest that the heterologous expression of these transcriptional regulators is necessary for further characterization and for their potential application as biosensor elements.

I close this thesis with **Chapter 6**, in which I discuss the progress achieved with the work presented and address the challenges and opportunities that remain to be explored. I emphasise that special attention needs to be paid to the crucial role of reliable high throughput screening methods in enzyme development. I conclude by suggesting that close cooperation between bioinformatics and biotechnology is key to future advances in protein engineering.

# Acknowledgments

The last 5 years have been a constant challenge, with a spectacular series of events and opportunities for growth. I always tell everyone that, despite all the struggles, doing this PhD has been one of my best choices and I would do it all over again. And it would not have been the same without all the people that populated my life in the last 4+ years. So, I will take the chance to thank at least some of them.

First and foremost, I want to thank **Jules** for giving me the opportunity to join this project. You have been a wonderful supervisor and my main reference point throughout these years, and I know that without you I would not have been able to achieve such a rewarding result. You have always been understanding and open minded, and your humour have always made our meetings the most enjoyable I have ever had. I want to then thank **Harro**; you have always been kind and helpful when I reached out and you provided very useful input throughout the project. A special thanks also to **Dick** and **Aalt-Jan**, you helped me a lot through the last stages of the thesis drafting and have always been available and on spot with your suggestions. I have really appreciated your contribution. **Janani**, you are one of the most brilliant people I have ever met. Since day one I have admired your resourcefulness and your incredible knowledge. Our trip to Halle was one of a kind and I will always remember it. It has been an honour and a pleasure to work with you in this project, and I hope we will collaborate again in the future.

**Adèle**, you were the first person I was introduced to when I started in the lab. I have always looked up to you, and you have been the most amazing colleague one can hope to have. Your kindness and your patience made me feel welcome from the very beginning and I knew I could always rely on you. Thank you. So much. **Thamara**, you are one of the nicest people I met throughout these years. You have the most contagious smile and a remarkable sense of humour, and I am so grateful I had the chance to work with you. You have been incredibly helpful, especially at the end of my project. Without you, Chapter 5 would have been only another “boring” RNA-seq thing. **Francel**, you helped me a lot with the analytical chemistry work and your knowledge has been a spark for me to get into the subject. I will never forget our “autosampler” experience with the PTR-MS. High throughput, yes, maybe, potentially, but not quite yet! **Michele**, thank you for helping me with the PTR-MS data and to make sense out of it.

**Francesca** (**Ncsk,** per gli amici). Maronn, non so neanche da dove cominciare. We have the happiest housemate relationship in the history of housemates and I think that moving in together was one of the best decisions of these last years. I love you like a sister, our breakfast/dinner/snackini time have given me so many times the strength I

was missing to move forward. Thank you so much for being such a wonderful friend (and paranymph). **Eirini,** I remember our first AMS outing together, how immediately we started laughing and I knew we were going to get along. Doing everything with you was always fun, my only regret is not having had more opportunities to spend time together! Thank you for being my amazing paranymph.

**Katja** and **Jeroen**, for the fun times in the lab and outside; **Melanie** for our coffee and lunch-breaks that always turned into lengthy conversations; **Aurin**, your energy has always been incredibly motivating and I enjoyed a lot our science talks. Thank you for helping me with the HPLC in the very last stretch, allowing me to add some extra spark to the chapter. To all the AMS cluster, **Ingrid**, **Bart**, **Margriet**, **Maurice**, **Hanny**, **Hetty**, **Ric**, **Maarten**, **Dirk**, **Twan**, **Henriëtte** and **Roland**, thank you so much for the beautiful outings and the kindness, while spending my PhD between departments, it was really nice to be able to count you as my colleagues as well. To the Italian girls, always staying around way too shortly, but having so much impact. **Giusi** e **Carmen** per il tempo passato insieme, sia qui che in Sicilia. **Cristina**, per tutte le avventure e le chiacchierate e quella tua naturale predisposizione a prenderti cura degli altri.

**Renake**, there are so many things I could say, but most of them I do not need to, because you know them all. So, I will just say thank you, for everything. Your light really is exaordinary.

**Iris**, thank you for being my "COVID supervisor", it meant a lot to me to have someone to refer to inside PPH when all the things felt like they were crumbling apart, one direction or another. You have always been a wonderful listener and I always enjoyed our chats. **Nora** ci siamo trovate subito, ci siamo perse per un po'. Grazie per aver sempre trovato il tempo per me, anche in mezzo a tutte le tue corse.

To my pre-COVID office-mates; **Mariëlle, Eva, Mara, Nikita** and all PPH group**, Christa, Leònie, Robert, Rina, Leo, Lot, Rumyana, Scott, Eva, Jasper, Mariana, Tom, Kris, Mara, Kees, Hongfei, Jeroen, Thijs, Mic, Jessica, Annika, Patricija, Roland, Damian, Joram, Annabel, Jesse, Eliza, Silvia,** for the amazing time spent together, on the outings and on the spontaneous PPH trips. You are a beautiful group of people and I am grateful I got to spend these years with you.

**Sara** e **Francesco** (e cane) thank you for all the BBQs and the dinners and all the neighbouring chats, you have been a fundamental part of the "Lombardi family", together with **Jenny** and **Ben.** Thank you especially for taking care of Demetra. You are amazing people, and I am very happy I got to meet you.

**Bianca**, ci conosciamo ormai da una vita e ad ogni traguardo, in un modo o nell'altro,

ci sei sempre stata. Anche se diventa sempre più difficile fare parte della vita l'una dell'altra, tu sei per me sempre una certezza.

**Ireneo**, bro, have you seen this? I did it!! You are one of the few people I still consider a friend from the times of high school and knowing you has been a privilege. We have had wonderful adventures and we still have the best conversations (and I hope it will never change!).

**Nicoletta**, amica, che parto questo dottorato, alla fine ce l'abbiamo fatta! Sì, ti aggiungo al successo perché le abbiamo attraversate tutte insieme, anche se a distanza. Perché so che la nostra è una di quelle poche amicizie che sono fatte per restare.

**Rosalba**, sei stata la spinta che mi serviva per muovermi e decidere di partire per venire qui (nella spensierata Wageningen), e nonostante la distanza e il tempo che passa, penso sempre a te come ad un punto di riferimento. **Cecilia** grazie per avermi sempre permesso di fare il parassita di Netflix, per tutti i messaggi di gati e per essere ancora quella che "se non lo sai, chiedi alla Ceci".

**Kylie**, sestra, you have been one of my bestest friends since the time of the Master and we shared so many things I can't even begin to. You are the most loyal person I know and I owe you so much. I am glad we finally found a good place for the horses, a little certainty for me in the midst of a new beginning. And I am grateful to be able to always count on you.

**Lena**, you have been the biggest part of these last five years, and I have learned so much from being with you. I would not be the person I am now if I did not have you in my life. Through all the difficulties, I think we did a great job.

**Mamma**, grazie per avermi sempre supportata (e molto spesso anche sopportata) in questi anni. Se sono riuscita a fare tutto questo è, in larga parte, grazie a te. Sei sempre la mia roccia, la terraferma nel mare in tempesta.

## About the author

Alice Di Girolamo was born on the 23rd of June in Bologna, Italy. Since childhood, she was passionate about animals and nature, having an unquenchable thirst for learning new things. She started horse riding at the age of 8 and keeps enjoying the company of horses. After obtaining her high school diploma in Arts & Design at the "Liceo Artistico Francesco Arcangeli" in Bologna, she decided to make a change, and started a Bachelor's degree in Biotechnology at the University of Bologna. She moved to Wageningen in 2015 to continue her studies, starting a Master's degree in Cellular/Molecular Biotechnology.

In Wageningen, she had the chance to focus on different aspects of Biotechnology applied to the engineering of microorganisms, first with her MSc thesis at the Microbiology department, and the with her internship in Corbion. In February 2017 she graduated from her Master and decided to look for a PhD that would strengthen her knowledge in Biotechnology applied to industrial microorganisms.

The perfect opportunity presented itself shortly after the graduation, and in August 2017 she started her PhD in Wageningen University and Research, under the supervision of Jules Beekwilder. The project allowed her to strengthen her skills in Molecular Biology and allowed her to discover the wonders of protein engineering. Now, she is starting her next chapter in the Biocatalysis group at the University of Amsterdam, following her passion for protein engineering and enzyme development.

# List of publications

**An analysis of characterized plant sesquiterpene synthases**. Durairaj *et al*. Phytochemistry, Volume 158, February 2019

https://doi.org/10.1016/j.phytochem.2018.10.020

**The santalene synthase from *Cinnamomum camphora*: Reconstruction of a sesquiterpene synthase from a monoterpene synthase.**

Di Girolamo *et al*. Archives of Biochemistry and Biophysics, Volume 695, November 2020

https://doi.org/10.1016/j.abb.2020.108647

## Education Statement of the Graduate School
## Experimental Plant Sciences

**Issued to:** Alice Di Girolamo
**Date:** 21 October 2022
**Group:** Plant Physiology
**University:** Wageningen University and Research

| 1) Start-Up Phase | date | cp |
|---|---|---|
| ► **First presentation of your project** | | |
| 'Novel enzymes for terpene production' | 02 Nov 2017 | 1.5 |
| ► **Writing or rewriting a project proposal** | | |
| ► **MSc courses** | | |
| *Subtotal Start-Up Phase* | | 1.5 |

| 2) Scientific Exposure | date | cp |
|---|---|---|
| ► **EPS PhD student days** | | |
| EPS PhD student days 'Get2Gether', Soest, NL | 15-16 Feb 2018 | 0.6 |
| EPS PhD student days 'Get2Gether', Soest, NL | 11-12 Feb 2019 | 0.6 |
| EPS PhD student days 'Get2Gether', Soest, NL | 10-11 Feb 2020 | 0.6 |
| EPS PhD student days 'Get2Gether', Gathertown, online | 01-02 Feb 2021 | 0.4 |
| ► **EPS theme symposia** | | |
| EPS theme 3 symposium 'Metabolism and Adaptation', Wageningen, NL | 13 Mar 2018 | 0.3 |
| EPS theme 3 symposium 'Metabolism and Adaptation', Nijmegen, NL | 14 Oct 2019 | 0.3 |
| EPS theme 4 symposium 'Genome Biology', online | 11 Dec 2020 | 0.2 |
| ► **Lunteren Days and other national platforms** | | |
| Annual Meeting Experimental Plant Sciences, Lunteren, NL | 09-10 Apr 2018 | 0.6 |
| Annual Meeting Experimental Plant Sciences, Lunteren, NL | 08-09 Apr 2019 | 0.6 |
| Annual Meeting Experimental Plant Sciences, Gathertown, online | 12-13 Apr 2021 | 0.5 |
| ► **Seminars (series), workshops and symposia** | | |
| Seminar: B-Wise Seminar, Anton Feenstra and Ehsan Motazedi | 09 Jan 2018 | 0.2 |
| Seminar: Science Week - 'What is life?', Wageningen, NL | 14 Mar 2018 | 0.1 |
| Seminar: B-Wise Seminar, Gurnoor Singh and Janani Durairaj - Predicting specificity and interactions from sequence | 04 Sep 2018 | 0.2 |
| Seminar: Prof. Yongfeng Guo - Peptide signalling in plant senescence - CLE14 and more. | 05 Apr 2019 | 0.1 |
| Seminar: Peter Openshaw - Coronavirus: what do we know? - Episode 5 | 30 Apr 2020 | 0.1 |
| Seminar: Shamma Khan and Joseph Yeoh - A beginner's guide to RNA-seq | 04 Feb 2021 | 0.2 |
| Seminar: Prof. Liesje Mommer - Plant the Future together | 20 Mar 2021 | 0.1 |
| Symposium: NBV annual symposium, 'Biotechnology in harmony', Ede, NL | 22 May 2018 | 0.3 |
| ► **Seminar plus** | | |
| ► **International symposia and congresses** | | |
| Congress: Terpnet 2019, Halle, DE | 26-30 Aug 2019 | 1.5 |
| Congress: AIChE 2021 - Metabolic Engineering 14, online | 11-15 Jul 2021 | 1.5 |
| ► **Presentations** | | |
| Poster: From plant to bacteria: understanding and improving fragrance enzymes, Annual EPS meeting 2019 | 08 Apr 2019 | 1.0 |
| Poster: Hybrid terpene synthases for biotechnological production of fragrance, Terpnet 2019 | 28 Aug 2019 | 1.0 |
| Poster: The santalene synthase from Cinnamomum camphora, AIChE 2021 congress | 13 Jul 2021 | 1.0 |
| Talk: From Plants to Bacteria: Understanding and Improving Fragrance Enzymes, EPS Get2Gether 2019 | 12 Feb 2019 | 1.0 |
| Talk: The santalene synthase from Cinnamomum camphora, Annual EPS meeting 2021 | 12-13 Apr 2021 | 1.0 |
| ► **3rd year interview** | | |
| ► **Excursions** | | |
| Company visit at KeyGene, organized by EPS PhD Council | 12 Oct 2017 | 0.2 |
| *Subtotal Scientific Exposure* | | 14.2 |

| 3) In-Depth Studies | date | cp |
|---|---|---|
| ► **Advanced scientific courses & workshops** | | |
| EdX Course: Using Python for Research, HarvardX, online | 29 Apr - 29 Jun 2020 | 1.7 |
| Course: Advanced course in Biocatalysis, TU Delft, online | 30 Nov - 04 Dec 2020 | 2.0 |
| ► **Journal club** | | |
| Participated in literature discussion group at Plant Physiology | 2017-2021 | 3.0 |
| ► **Individual research training** | | |
| *Subtotal In-Depth Studies* | | 6.7 |

| 4) Personal Development | date | cp |
|---|---|---|
| ► **General skill training courses** | | |
| EPS Course: EPS Introduction Course, Wageningen, NL | 26 Sep 2017 | 0.3 |
| WGS Course: Brain Training, Wageningen, NL | 08 Nov 2017 | 0.3 |
| WGS PhD Competence Assessment, Wageningen, NL | 12 Jun 2018 | 0.3 |
| WGS Course: Research and Data Management, Wageningen, NL | 13, 20, 27 Sep 2018 | 0.5 |
| EPS Workshop: Entrepreneurship clinic, StartHub, Wageningen, NL | 03 Oct 2019 | 0.2 |
| EPS Workshop: Scientific paper writing, Wageningen, NL | 24 Oct 2019 | 0.1 |
| Workshop: Data Visualization - the beautiful way, online | 24 Nov 2020 | 0.3 |
| EdX Course: Data Science Ethics, MichiganX, online | 05 Jun - 07 Jul 2022 | 0.6 |
| DARE Workshop: Theory of Microaggressions & What is an Ally, Wageningen, NL | 22 Jun 2022 | 0.2 |
| EPS Workshop: Applying for Marie Skłodowska-Curie Fellowship: from proposal to project, online | 22 Jul 2022 | 0.2 |

| | | |
|---|---|---|
| ► **Organisation of meetings, PhD courses or outreach activities** | | |
| ► **Membership of EPS PhD Council** | | |
| *Subtotal Personal Development* | | 3.0 |

| **5) Teaching & Supervision Duties** | *date* | *cp* |
|---|---|---|
| ► **Courses** | | |
| PPH-10806 Structure and Function of Plants | 16 Mar - 24 Apr 2020 | 1.5 |
| NEM-20306 Research Methodology in Plant Sciences | 9 May - 1 Jul 2020 | 1.5 |
| ► **Supervision of BSc/MSc students** | | |
| Run Qi (MSc.) - Analysis of Mutant Libraries of Cyclic and Acyclic Sesquiterpene Synthases | May - Nov 2019 | 3.0 |
| *Subtotal Teaching & Supervision Duties* | | 6.0 |

| **TOTAL NUMBER OF CREDIT POINTS*** | **31.4** |
|---|---|
| Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits. | |
| ** A credit represents a normative study load of 28 hours of study.* | |

This research was financially supported by The Dutch Research Council (NWO-TTW), Isobionics BV(BASF), PFW Aroma Chemicals BV.
Project number: TTW 15043
Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design and thesis layout by the author
**Printed by**: Ridderprint | Ridderprint.nl


CH3
HO
CH3
O
O
H
H
H
O
O
O
O
O
O
H