Technische Universität München
TUM School of Natural Sciences

The potential of a comparative genome and transcriptome approach for the discovery of valuable terpenoid molecules in *Caryopteris x clandonensis*

Manfred J. Ritz

Vollständiger Abdruck der von der TUM School of Natural Sciences der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Tom Nilges

Prüfer*innen der Dissertation:

1. Priv.-Doz. Dr. Norbert Mehlmer
2. Prof. Dr. Matthias Feige
3. Priv.-Doz. Dr. Markus Teige

Die Dissertation wurde am 26.07.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Natural Sciences am 29.11.2023 angenommen.

If you are not willing to risk the unusual, you will have to settle for the ordinary.

| Jim Rohn

I Abstracts

Chapter I: Comparative genome-wide analysis of two *Caryopteris x clandonensis* cultivars: Insights on the biosynthesis of volatile terpenoids

In the first chapter, the biosynthetic potential of two *Caryopteris x clandonensis* cultivars, Dark Knight and Pink Perfection, is examined using comparative genomics. A principal component analysis was employed to determine out of four cultivars the selected, representative plants for genomic evaluation. GC-MS data yielded a divergent synthesis of monoterpenes, regarding limonene. Here, different limonene derived molecules were discovered. Through functional analysis and comparative genomics of the long-read data, cultivar-specific terpene synthases and cytochrome p450 enzymes were identified. This allows the use of biotechnological methods for the development of terpene production systems for industrial applications such as flavorings, food preservatives or pharmaceutical formulations. The completeness of the presented genomes was determined to be 96.8% using BUSCO and an approximate size of 355 Mb was identified. Gene models were generated and their annotation yielded 52,090 potential genes. Within these, 42 genes were related to terpene biosynthesis. Furthermore, 1340 models related to cytochrome p450 enzymes were discovered.

Chapter II: Differential RNA-Seq Analysis Reveals Genes Related to Terpenoid Tailoring in *Caryopteris x clandonensis*

The second chapter reveals the terpenoid tailoring mechanisms in *Caryopteris x Clandonensis*. The focus lies on cytochrome p450 enzymes, which are able to tailor monoterpenes. In the first section limonene derived molecules were discovered. The origin and molecular mechanisms in their synthesis were further elucidated in this part. The published reference genome was further refined and scaled down with a factor of 14. This refinement was necessary to efficiently map short read sequencing data on the genome and detect differential expressed genes. 3305 genes were evaluated for further identification of underlying synthesis mechanisms of above mentioned molecules. 61 enzymes related to terpene tailoring using cytochrome p450 proteins were observed. Further 23 were upregulated in plants considered limonene derived molecules positive. This section provides excellent data for further functional assays to validate this putative synthesis.

II Zusammenfassungen

Kapitel 1: Vergleichende, genomweite Analyse von zwei *Caryopteris x clandonensis* Kultivaren: Einblicke in die Biosynthese von flüchtigen Terpenen

Im ersten Kapitel wird mittels vergleichender Genomik die Biosynthese von zwei *Caryopteris x clandonensis* Kultivaren, Dark Knight und Pink Perfection, näher betrachtet. Ausgangspunkt ist eine Hauptkomponentenanalyse flüchtiger terpenoider Verbindungen, welche mittels GC-MS analysiert wurde, so dass zwei repräsentative Pflanzen ausgewählt werden konnten. Durch diese Daten konnte bereits eine Abweichung der Terpenbiosynthese hinsichtlich Terpenen mit Limonen-Grundgerüst aufgezeigt werden. Über funktionelle Analyse und vergleichende Genomik der long-read Daten konnten kultivar-spezifische Terpensynthasen und Cytochrom p450 Enzyme identifiziert werden. Dies ermöglicht den Einsatz biotechnologischer Methoden zur Entwicklung von Terpenproduktionssystemen für industrielle Applikationen wie beispielsweise Aromastoffe, Konservierungsstoffe oder pharmazeutische Formulierungen. Die Vollständigkeit der vorgestellten Genome wurde mit BUSCO auf 96,8 % bestimmt und eine ungefähre Größe von 355 Mb ermittelt. Die Erstellung und Annotation von Genmodellen ergab 52 090 potenzielle Gene. Innerhalb dieser konnten für die Terpenbiosynthese 42 Gene identifiziert werden. Des Weiteren wurden 1340 Modelle entdeckt welche im Zusammenhang mit Cytochrom p450 Enzyme stehen.

Kapitel 2: Eine differenzielle RNA-Seq Analyse deckt Terpen-modifizierende Gene in *Caryopteris x clandonensis* auf

Im zweiten Kapitel wird das analysierte Genom des *Caryopteris x clandonensis* Kultivars Pink Perfektion als Referenz verwendet und genauer untersucht. Ein weiterer, zuvor durchgeführter, Reinigungsschritt konnte die genomische Referenz weiter verbessern. Hier wurde die Anzahl der Contigs um einen Faktor von 14 reduziert. Auf Grundlage dieses Genoms wurde eine Mapping von short Reads durchgeführt um eine differenzielle Genexpression analysieren zu können. 3305 Gene wurden ermittelt, welche im Hinblick auf ihre Terpenmodifikationsmöglichkeiten näher betrachtet wurden. Von diesen konnten 61 den Cyotchrome p450 Enzymen zugeordnet werden. Diese sind dafür bekannt Terpen-Grundstrukturen weiter modifizieren zu können. Im ersten Abschnitt konnten auf Limonen basierende Moleküle (LDM) identifiziert werden, welche als Grundlage für eine weitere Einengung der Sequenzauswahl herangezogen wurde. Pflanzen, welche eine erhöhte Menge an LDM produzieren wurden mit Pflanzen mit geringen Mengen an LDM auf Transkriptebene verglichen. Durch diese differenzielle Analyse konnten 23 Enzyme identifiziert werden, welche mit der Biosynthese dieser Moleküle in Verbindung stehen.

III Acknowledgements

First and foremost, I want to thank **Prof. Dr. Thomas Brück** for giving me the opportunity to conduct research at the Werner Siemens Chair for Synthetic Biotechnology (WSSB). This thesis and related manuscripts are the result of the supervision, resources and support I got during the last years at his chair. I am grateful for all scientific advices, however also for remarks on how to set a path for the right future.

In addition to that, I want to say thank you to **Priv.-Doz. Dr. Norbert Mehlmer** who supported me in all scientific related questions. Nucleic acids, proteins, bioinformatics? No problem. Especially the setup of all necessary computational resources gave me the opportunity to dive into bioinformatics, which I wished to do from the beginning of my thesis. Thanks for the help and evaluation of different scripts as well as your knowledge in Linux.

For the evaluation and examination of this thesis I want to express my outmost gratitude to **Prof. Dr. Tom Nilges**, **Prof. Dr. Matthias Feige** and **Priv.-Doz. Dr. Markus Teige**.

What would a doctorate be without colleagues to talk to, drink some Spezi, share the lunch break and wander off for some after work activities. Sadly, during these challenging times I was not able to get to know all of you better. However, I still am grateful to your contribution during my thesis no matter which nature - scientific, emotional, mind challenging - I will carry these memories with me and they will always remind me of the time at the **WSSB. My appreciation to all of you**.

Selina, Nadim, Nate, **Kevin** and **Zora,** a special thanks to you, for all the time we spent off-work playing board games, drinking, laughing, sailing and having a good time. I hope we will have further time off-work and continue our endeavors and will find more time to tell sailor's yarn.

Und zum Abschluss ein herzliches Vergelts' Gott an die mit wichtigsten Unterstützer dieser aufregenden Zeit, die ich auf dem Weg zum Doktor verbringen durfte - Meine Freunde, meine Familie und all diejenigen die mich inspiriert haben, mir in manch harten Zeiten beigestanden und abgelenkt haben. Vor allem die Ablenkung hat sehr gut getan. **Mama**, **Papa**, **Martin** und **Melanie,** danke, dass ihr für mich da wart, auch wenn ihr manchmal nicht nachvollziehen konntet über was ich rede.

Danke, **Daniel** für deine unendliche Geduld, Liebe und Unterstützung in allen Hochs, Tiefs und weiteren Tiefs die es während den letzten Jahren zu durchstehen galt. Ich bin froh dich an meiner Seite zu haben und bin dankbar für all die Zeit die wir zusammen verbringen dürfen. Ohne dich hätte ich es nicht so gut meistern können.

Oma, ja I glab des woars mid da Schaei, aeitz gaeids mim Oarban laous (ł 2020).

IV Table of Contents

V List of Abbreviations

°C	degree Celsius
BLAST	Basic Local Alignment Search Tool
BUSCO	Benchmarking Universal Single-Copy Orthologs
CDP-ME	4-diphosphocytidyl-2-C-methyl-D-erythritol
CoA	coenzyme A
COG	Cluster of Orthologous Groups
CTAB	cetyltrimethylammonium bromide
CYP	cytochrome p450
DK	Dark Knight
DMAPP	dimethylallyl diphosphate
DNA	deoxyribonucleic acid
DXP	1-deoxy-d-xylulose 5-phosphate
EDTA	Ethylene diamine tetraacetic acid
FPP	farnesyl pyrophosphate
FID	flame ionization detector
GPP	geranyl pyrophosphate
GC	gas chromatography
gDNA	genomic deoxy nucleic acid
GO	Gene Ontology
HGAP	Hierarchical Genome Assembly Process
HMBPP	hydroxy-3-methylbut-2-enyl diphosphate
IPA	Improved Phased Assembler
IPP	isopentenyl diphosphate
KEGG	Kyoto Encyclopedia of Genes and Genomes
MEcPP	2-C-methyl- D-erythritol-2,4-cyclodi-phosphate
MEP	methyl erythritol
mM	millimolar
mRNA	messenger ribonucleic acid
MVA	mevalonic acid
NGS	next generation sequencing
ONT	Oxford Nanopore Technologies
PCA	principal component analysis

PEG	polyethylene glycol
PP	pyrophosphate entity
PP	Pink Perfection
PVP	Polyvinylpyrrolidone
RNA	ribonucleic acid
RT	room temperature
SBH	sequencing by hybridization
SBL	sequencing by ligation
SMRT	single molecule real-time
TS	terpene synthase

1 Introduction

1. 1 *Origin and importance of Caryopteris x Clandonensis*

The kingdom of *Viridiplantae* consists of three phyla: *Chlorophyta*, *Prasinodermophyta*, and *Streptophyta*. Latter is divided into four classes: *Chlorokybophyceae*, *Klebsormidiophyceae*, *Mesostigmatophyceae*, and *Streptophytina*. Furthermore, the subphylum *Streptophytina* harbors the clade *Embryophyta*. A closer look at this clade reveals the different plants of the earth flora and, in a detailed differentiation in clades and orders, the family of flowering plants, *Lamiaceae* [1], [2]. This family harbors the genus *Mentha*, including the species *Mentha spicata* (Spearmint) and *Mentha x piperita* (Peppermint), which is the origin of the widely spread term - the mint family. *Lamiaceae* is a diverse and large family with around 7,000 species distributed worldwide [3]. One of the characteristic features of the *Lamiaceae* family is the presence of glandular hairs, trichomes. These produce volatile and essential oils and give the plants their characteristic scent and flavor [4], as well as their herbaceous, aromatic values, which can be exploited for culinary, ornamental, and medicinal purposes [5], [6]. The oils are a complex mixture of different compounds including terpenes, phenols, and flavonoids [7], [8] and are known to display a variety of biological activities such as antimicrobial, antioxidant, and anti-inflammatory properties [9]. Especially the family of terpenes is of great interest in industrial applications.

In this study, selected cultivars of *Caryopteris x clandonensis* were investigated in detail. This species is located in the *Lamiaceae* family, more specifically within the subfamily *Ajugoideae* and the corresponding tribe *Ajugeae.* Originating from a hybrid of *Caryopteris incana* and *Caryopteris mongholica*, it was subjected to further breeding [10]. These plants are mostly known for their ornamental values and vast variety of different cultivars, which besides different essential oil constituents also harbor different compositions in their volatile compound profile [10], [11]. Already the parental generation *C. incana* and *C. mongholica* where known as a source of new glycosides [12] and new alkaloids [13], respectively. Further molecules were found to be constituents of the resulting hybrids essential oil, as is the pyrano-juglon derivate α-caryopteron [14], and the keto-glycosides harpagides and clandonosides [15].

Research into the biological activity of these oils revealed insecticidal activities for *C. clandonensis*, as shown for *C. incana* [16]. Additionally, anti-cancer activity was discovered for ethanolic stem extracts however, a single molecule responsible for these mode of actions was not yet determined [10], [17]. Compared to the biological activities discovered in the respective *Lamiaceae* family, which range from anti-viral [18], anti-inflammatory [19], immunomodulatory [20] and many more [21], there is a plethora of hidden possibilities in the investigated species.

The first chapter focuses on the synthesis of these compounds inside plants. It describes why these metabolites are produced, how they can be differentiated and how they can be exploited for biotechnological use.

1. 2 Plant metabolites

Molecules which are produced by the plant to maintain their metabolism are called metabolites. Within each plant this is a diverse group of organic compounds that are synthesized to perform essential functions such as growth, development, reproduction, and defense against biotic and abiotic stresses [22], [23]. Biotic stresses include plant damage by herbivores or pest, whereas abiotic stresses are mostly physical circumstances such as radiation, chemicals or temperature [24]. To sustain potential damage and maintain a regular metabolism, different types and concentrations of plant metabolites are synthesized and vary among different species, tissues, and developmental stages, and are influenced, amongst others, by above mentioned environmental factors such as light, temperature, soil nutrients, and herbivory [25].

The main types of plant metabolites are: primary and secondary metabolites.

1. 2. 1 Primary metabolites

These essential compounds are required for the basic metabolic processes of the plant. They are synthesized through central metabolic pathways like the Calvin cycle citrate cycle, glycolysis or shikimate pathway and are involved in various processes such as respiration, photosynthesis and protein synthesis. Some of the most important primary metabolites include carbohydrates, lipids, nucleic acids and proteins [25], [26].

Carbohydrates, one major source of energy for plants, are synthesized through the process of photosynthesis. The most common carbohydrate in plants is glucose which is used to further build up other sugars such as sucrose, fructose and starch. These sugars are used to fuel cellular processes and provide structural support for the plant cell walls [27].

Lipids are another important primary metabolite in plants and are involved in energy storage, membrane structure and signaling pathways. Plant lipids include triglycerides, phospholipids, and sterols. These lipids are stored in lipid droplets and serve the plant as energy source during periods of low photosynthetic activity [28]. In addition, an important primary metabolite in plants are nucleic acids which are involved in the storage and transmission of genetic information. Plant nucleic acids include DNA and RNA which are synthesized from nucleotides consisting of purine or pyrimidine bases and needed in the process of replication and transcription. These bases are either produced *de novo* or *via* salvage pathways. Nucleic acids are used to code for the synthesis of proteins and enzymes, and are involved in cellular processes like cell division and differentiation [29]–[31]. A last example are

proteins, which are essential for plant development and growth, and are synthesized for a range of processes, such as enzyme catalysis, transport, and signaling. Plant proteins are synthesized through the process of translation, which involves the conversion of mRNA into amino acid sequences. Examples of plant proteins include enzymes such as Rubisco, which is involved in the process of carbon fixation during photosynthesis [32], and transport proteins such as aquaporins, which regulate the movement of water and nutrients across the cell membrane [33].

Understanding the role of primary metabolites in plant growth and development is important for improving agricultural practices and developing new plant-based products. An even more elaborated use in medicine, aroma and herb production is the case for the more specialized groups of metabolites.

1. 2. 2 **Secondary metabolites**

These compounds are synthesized by plants for specialized functions, such as defense against herbivores, pathogens, environmental stresses or communication. Examples include alkaloids, flavonoids, phenolics, and terpenoids. Secondary metabolites are synthesized through different pathways that branch off from the central metabolic pathways. As mentioned above, these molecules are not essential for survival of the plant however, plants producing secondary metabolites have a higher rate of interaction with their surroundings, thus a better exchange in symbionts or attractors. Furthermore, in case of defense mechanisms against herbivores and pest, plants have a higher likelihood for propagation [34].

Some of the most important secondary metabolites in plants include alkaloids, phenolics, flavonoids and terpenoids. Alkaloids are a diverse group of secondary metabolites, that contain nitrogen and are often bitter-tasting and toxic to herbivores and other pathogens [35]. Examples of alkaloids in plants include purine alkaloids like caffeine, tropane alkaloids, such as scopolamine and morphine as an example of a benzylisoquinoline alkaloid. These molecules are synthesized dependent on their corresponding group from purines, monoterpenoid-indole structures or from amino acids originating from the primary plant metabolism [36]. Phenolics are synthesized from the shikimic acid pathway [37] and are also involved in defense against pathogens, protection from UV radiation, and herbivory. Examples of phenolics in plants include lignin, tannins, and flavonoids. Flavonoids are synthesized from the phenylpropanoid pathway and originate from phenolic compounds [38]. Examples of flavonoids in plants include anthocyanins, isoflavones, and flavones. Flavonoids are involved in various processes such as pollination, attraction of pollinators, and defense against herbivores and pathogens [39], [40]. Terpenoids are synthesized from the five-carbon isoprene unit and are involved in plant defense, attraction of pollinators, and production of essential oils. Examples of terpenoids in plants include limonene, menthol, carotenoids, and essential oils such as spearmint

and lavender. The biosynthesis is facilitated in two different pathways, the mevalonic acid (MVA) in the cytosol, and the methyl-erythritol-phosphate (MEP) respectively, which produces the backbones used for further tailoring and diversification [41]. In the next chapter a more detailed look into these pathways as well as this huge compound family in general shall be provided.

In summary, plant metabolites can be divided in essential primary metabolites, which are necessary for sustaining the metabolism as well as energy production and propagation. In contrast to primary metabolites, secondary metabolites are synthesized by plants for specialized functions such as defense against herbivores, pathogens, and environmental stresses. Understanding the role and function of metabolites in plant growth and development is important for improving agricultural practices and developing new plant-based products, for example medicines, food additives and cosmetics.

1. 3 Terpenes and terpenoids - biosynthesis and differentiation

In the past decades, around 70,000 terpenes and their derivates were characterized and still are of interest for many researchers worldwide [42]. Precursor biosynthesis for terpene production is maintained through two pathways, the MVA in the cytosol, and the MEP pathway in plastids [43]. In Figure 1 are both illustrated.

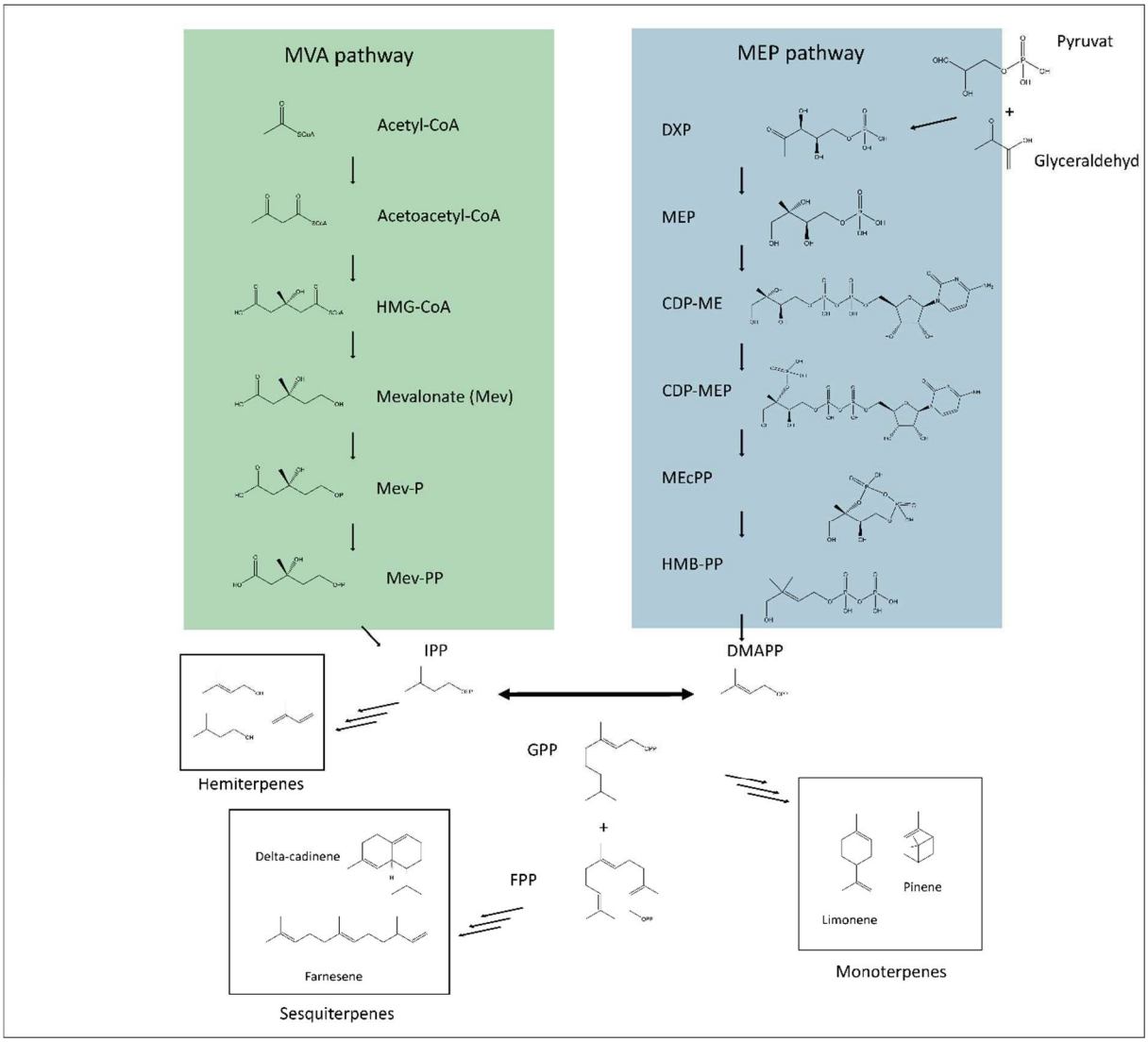


Figure 1. The mevalonate (MVA, left side) and methyl-eryhtritol pathway (MEP, right side) for the synthesis of terpene structures. DXP: 1-deoxy-d-xylulose 5-phosphate, CDP-ME: 4-diphosphocytidyl-2-C-methyl-D-erythritol, MEcPP: 2-C-methyl- D-erythritol-2,4-cyclodi-phosphate, HMB-PP: hydroxy-3-methylbut-2-enyl diphosphate. An isomerase converts IPP (isopentenyl diphosphate) to DMAPP (dimethylallyl diphosphate). Prenyltransferases build up geranylpyrophosphate (GPP) and yield in monoterpenes. A further fusion with farnesylpyrophosphate (FPP) results in the synthesis of sesquiterpenes. The scheme is adapted from [44].

The first is the MVA pathway. It was described earlier and is present in almost all living organisms [45]. The starting molecules are two acetyl-CoA, which are fused in a condensation reaction to acetoacetyl-CoA. A second aldol condensation forms S-3-hydroxy-3-methylglutaryl-CoA (HMG-CoA),

which is reduced to the name-giving mevalonate. Two subsequent phosphorylation steps result in MVA-P and MVA-PP. In a last step MVA-PP is decarboxylated and yields IPP [46].

The 1-deoxy-d-xylulose 5-phosphate (DXP)-pathway occurs as a seven step pathway in all plastid-bearing organism such as plants and further bacteria [46]. The starting molecules are pyruvate and glyceraldehyde, which are fused through a transketolase-like condensation, and form DXP, which is catalyzed by a DXP synthase. In a second step, a reductoisomerase converts DXP into 2-C-methyl-D-erythritol 4-phosphate (MEP). Two further steps convert MEP into 4-diphosphocytidyl-2-C-methyl-D-erythritol (CDP-ME) and after phosphorylation, into CDP-MEP, and subsequent cyclization and loss of CMP it results in 2-C-methyl- D-erythritol-2,4-cyclodi-phosphate (MEcPP). The fifth step of the MEP pathway reduces MEcPP into hydroxy-3-methylbut-2-enyl diphosphate (HMBPP). The last reduction results in isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) in a ratio of about 5:1 [46], [47]. These two building blocks are the basis for the diversity of terpenes and terpenoids in plants.

An isomerase is able to convert IPP into DMAPP and *vice versa* resulting in the first isoprene units for the terpene structures. Starting from the building block IPP and DMAPP terpene structures, which increase by five C-atoms (one isoprene unit), are built. Mediated by prenyltransferases, prenyl diphosphates are formed through head-to tail condensation yielding, dependent on the supply of IPP and DMAPP, in monoterpenes (C10), sesquiterpenes (C15), diterpenes (C20) and higher structures. Here, it is to mention that higher structures such as triterpenes are formed through tail-to-tail condensation. These backbones are the basis to form biologically more active terpenes and terpenoids [48]. There are two main classes of enzymes which are tailoring these backbones: Terpene synthases (TS) and cytochrome p450 enzymes (CYPs).

Among others, TS belong to an enzyme class, that is responsible for the vast diversification of the basic carbon backbones into the 70,000 to 80,000 different structures known and characterized so far [42], [49]. TS catalyzing the cyclization of above mentioned carbon backbones are also known as terpene cyclases (TC). Generally, there are two main types: Type I and Type II. The differentiation depends on the mechanism of carbocation ionization. For Type I, a trinuclear metal-dependent cluster can be elucidated in the active site. For this enzyme class two conserved sites are known, the highly conserved motif ‘DDXXD’ and a less conserved ‘NSE/DTE’ motif. For type II, an acidic environment is responsible for carbocation protonation with a ‘DXDD’ and EDXXD like motif. A further notable motif discovered in many TC is a ‘RR(x)8W’-motif [49], [50]. These domains are important for functional annotation, which will be discussed in another chapter. To differentiate TS, their function, and occurrence within the kingdom, they are further clustered in eight TS-families (TS-a to TS-h).

Another enzyme class responsible for the diversification of terpene backbones are CYPs. This enzyme class catalyzes a plethora of activities such as hydroxylation, ether forming activity,

carboxylation or acetylation [51] resulting in a variety of functionalized molecules. Medicinally important and well-studied molecules include taxol [52] and artemisinin [53], which are also biotechnologically produced terpenoids. CYPs are present in all living organisms, therefore, for an easy access and differentiation, a classification system was proposed [54]. Plant CYP families are categorized as CYP71-99, CYP701-999, and in a four-digit format as CYP7001-9999. The classification into these groups is based on sequence similarity. A minimum of 40% matching amino acids is required within the same family (represented by an Arabic number), while a minimum of 55% matching amino acids is required within the subfamily (represented by an Arabic letter) [55], [56]. Consequently, the enzyme CYP76S40 [57] belongs to the CYP76S subfamily and the CYP76 family as the 40th individual enzyme. As described for TS, CYPs also exhibit conserved domains as an oxygen binding and activation motif, 'A/G-G-X-E/D-T-T/S', or the 'C-X-G'-motif for heme-binding essential for the redox reaction of CYPs [58], [59]. Both, TC and CYPs, are playing part in the molecular mechanism in producing plant terpenes and their essential oils. These can further be used in a variety of applications.

1.4 Industrial applications

Since that plant secondary metabolites show various biological effects these substances were already used during the beginning of mankind. In the early days wounds were treated with plant extracts, tea was made out of leaves and distinct flowers were planted to attract pollinators to make sure food plants are pollinated [60]. These methods were discovered either empirically or by accident. To this day, those applications were investigated and in some cases a single active agents was elucidated and tested in clinical trials to ensure their safety for distribution [5], [61]. This way, one can find food preservatives, aromas or even food packaging from plant derived molecules [62], [63].

Terpenes play an important part in the plants interaction with its environment and the metabolites can be exploited for industrial application. The diverse molecule structures and the promiscuous catalyzation possibilities of TS and CYPs as well as the regio-specificity open up a plethora of possibilities to exploit the protein diversity found in nature. One option to mine these treasures lies within bioinformatics methods. In the next chapters, the tools to discover enzymes, which are putatively functionally active are described. The identified sequences can then be further used in a biotechnological application.

1.5 Next Generation Sequencing

The basis for sequencing was set in 1977 through the incorporation of di-deoxynucleosides as DNA polymerase inhibitors. This led to the basis of Sanger sequencing, which made it possible to decipher DNA information faster and with more accuracy [64]. However, for plant genomes and larger sequences in general, this method was not feasible regarding time and costs. The sequencing of the

human genome took 13 years using sanger sequencing and even ten years after finishing the human genome a conifer genome (e.g. the coast redwood with 26.5 Gb [65]) was still out of reach due to economic and technological hindrances [66]. The advancement of Sanger sequencing led to new technologies called next generation sequencing (NGS).

These new methods have revolutionized the field of genomics research by enabling the rapid and cost-effective analysis of large and complex genomes. NGS techniques involve sequencing millions of DNA fragments in parallel and assembling the yielded bases to reconstruct the entire genome sequence. In general, there are two main approaches to NGS: short-read sequencing and long-read sequencing, whereas short-read sequencing is the most widely used technology for genome sequencing today. The name results from its up to 300 and 500 base fragments sequenced in one run. There are different approaches, sequencing by ligation (SBL) sequencing by hybridization (SBH) and sequencing by synthesis (SBS) [67]. SBL employs labelled sequences (probe and anchor), which bind complementarily on the template. The probe provides one or two known bases fused to an universal base fragment to identify the unknown bases on the complementary strand through anchor initiated ligation. A new anchor with a known offset starts the reaction from the beginning and yields up to 100 bp paired-end reads. SBS can be split in two categories. The first method uses single nucleotide addition. Here, either the pyrophosphate or the proton is used as a base signal which is released after incorporation. In an iterative process, one of each single deoxynucleotide is added, detected, washed from the beads and another nucleotide added. During the second approach, also called Illumina sequencing, DNA fragments are first amplified and attached to a solid surface, and then fluorescently labeled nucleotides are added one at a time. The sequence of the DNA fragment is determined by detecting the color of the fluorescent label as each nucleotide is added. After base identification, remaining nucleotides are washed from the surface and the fluorophore, which prevents further elongation, is cleaved and a new set of labelled nucleotides added [67], [68]. Illumina sequencing can generate up to 2x300 base pair reads per run, which is ideal for sequencing small genomes and detecting single nucleotide variations or sequence the transcriptome of an organism [69]. In general, short read sequencing yields small DNA fragments in high coverage with a relatively low error rate [70].

Long-read sequencing technologies, on the other hand, can sequence much longer DNA fragments, up to tens of thousands of base pairs, resulting in more contiguous genome assemblies even spanning over highly repetitive sequences, such as telomere and centromere regions, which are commonly known to produce high error rates using short read sequencer [71]. One of the uprising long-read sequencing platforms is provided by Oxford Nanopore Technologies. It involves threading a single nucleic acid strand through a small pore and measuring the changes in electrical current as the

nucleotides pass. One drawback of this method are the small distances between each nucleic acid (3.4 Ångstrom), which results in low accuracy and a higher error rate as only oligomer changes are measured rather than single nucleotides changes [67]. This method can generate reads ranging from a few kilobases to more than 100 kilobases (N50), which is useful for *de novo* genome assembly and identifying structural variations.

A further commercially available and most widely used method is single molecule real-time (SMRT) sequencing provided by Pacific Biosciences (PacBio). A scheme of the prepared library and the generation of HiFi reads is depicted in Figure 2. During library preparation the double strand is closed to a hair-pin like structure with adapter and primer sequences, and a polymerase is added to the provided primer. Within the flow cells of the instrument, containing zero-mode waveguides (ZMW), the polymerase is immobilized [72]. Fluorescently labelled nucleotides are used to detect the incorporation of a new nucleotide during strand synthesis. The small diameter of the ZMW in combination with the wavelength allows the accurate detection of single molecule reactions within the cell. The primer and adapter fused to the nucleic acid is in further downstream analysis used to count passes of polymerase reads and also to differentiate samples during a multiplex setup. During sequencing, the polymerase is able to read the sequence of nucleic acids without amplification. Due to the circular structure of the DNA-molecule, the sequencing process allows sequencing the bases of both the forward and the reverse strand. During elongation of the DNA strand and completing the second strand errors can occur, which are reduced with the numbers of passes during the run.

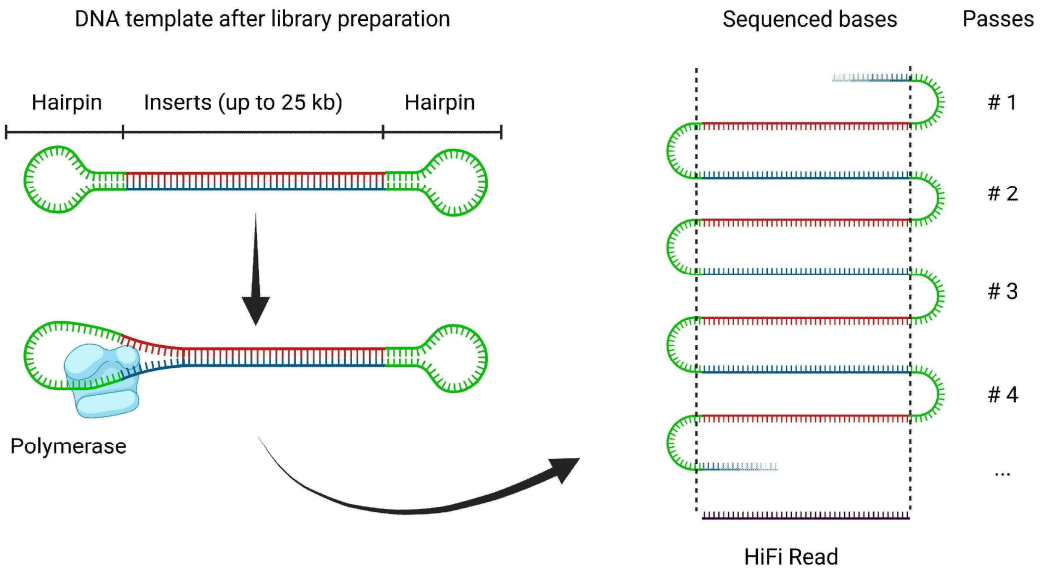


Figure 2. Scheme of the library preparation for long-read sequencing using PacBio. DNA fragments with a length up to 25 kb are ligated using a hairpin loop to generate circular DNA. A polymerase uses a primer to replicate the nucleic acid. Due to the circular structure of the prepared template HiFi-Reads can be generated after a certain amount of passes ensuring a High-Quality long-read with a low error rate.

During base calling, quality scores are measure for each base detected. These scores represent the probability of a base call being correct. In a logarithmic phred scale, a quality score of 20 corresponds to a probability of an incorrect base of less than 1 %. Thus, a score of 40 represents an error probability of 1 in 10,000. Along with this quality score, the length of the read and number of passes, HiFi reads are generated and separated from other reads. Compared to ONT a true single nucleotide calling is possible, which ensures the low error rates in this sequencing method [73].

The short and long-read sequencing methods are only one part of deciphering the nucleotide sequences behind an organisms' genomic make-up. To identify the genomes and their organization, the detected reads need to be assembled to contigs and scaffolds. In a last step, they can be mapped into chromosome like structures including centromere and telomere regions.

1. 6 **Bioinformatic analysis**

There are different possibilities to analyze huge amounts of sequencing data. Free software tools to organize, manipulate or visualize the results are available via platforms such as Bioconductor or github [74]. A comprehensive and easy to use graphical user interface is provided by the galaxy project [75]–[77]. Semi-automated analysis is commercialized by providers like BioBam, Qiagen or Illumina and full analysis can be provided from companies such as Eurofins Genomics and Twist Biosciences. In the next chapters a typical bioinformatics workflow is provided.

We assume having a high-quality long-read whole genome sequencing data set. First, the data has to undergo a comprehensive quality step to ensure correct downstream analysis. Further steps involve genome assembly, phasing and polishing. Between these steps further quality checkpoints can be installed, especially regarding contaminated sequences. This steps ensure a high-quality genome at the end of this process. A last control needs to be performed before secondary analysis. As presented later in this thesis, this might involve: gene prediction and functional annotation, the collection of RNA-Seq data and mapping on the prepared reference genome, structural variant calling, and further analysis depending on the research question.

1. 6. 1 Quality steps for sequencing data

After sequencing, the reads still contain low quality reads, adapter sequences, barcodes and contaminating sequences. During sequencing, each nucleotide emits a signal, which is detected and scored. Is the signal low or if a different wavelength is interfering, the base gets a low quality score. In a FASTQ file the sequence including the quality score is stored and provided at the end of each sequencing run. At this point, reads with overall bad read quality can be discarded. However, in some cases only the first few bases reach low quality scores and those need to be trimmed. Trimming parameters can be adjusted according to the severity of low quality reads or a distinct number of bases in the beginning can be cropped. Adapter sequences are trimmed using the known adapter or a pool of usually used sequences. The barcode is necessary to ensure the differentiation between barcoded samples during multiplexing. However, after separation the barcodes need to be removed. The quality refining yields a basis for genome or transcriptome assembly.

1. 6. 2 Genome assembly

The cleaned reads can be assembled depending on their origin and available references. Genomes can be created using a *de novo* approach or can be mapped on a reference. Both methods are used with short- and long-read data. The latter is preferred for *de novo*, whereas short-read libraries are preferably used for reference mapping however, both can be used for *de novo* assembly. The longer the read, the easier is the assembly and more comprehensive the resulting genome. A first step can be an assembly of the long reads. ONT and SMRT sequencing result in high quality long reads, which allow a fast and accurate assembly of the species genetic makeup. To achieve this, numerous tools have been published. To get the best results out of each tool, it is necessary to take the organisms and the species in consideration. Examples for genome assemblers are flye [78], HGAP [79] or HiCanu [80]. They are able to produce contigs, which are built from overlapping fragments of raw reads, which can be further assembled to scaffolds which represent overlapping contigs. This way the reads can be concatenated over repetitive sequences like centromere and telomere regions building up whole

chromosomes. With plants such as *Lamiaceae*, further refinement steps are necessary to check for haplotypes, categorize each sequence reads in either primary or haplotypic sequences and build the contigs, scaffolds and chromosomes from this step onwards. Pacific Biosciences developed a software tool called Improved Phased Assembly. This assembly method allows to first differentiate into primary contigs and haplotigs. A further refinement step with polishing and the deletion of duplicates finalizes the steps of assembly.

1. 6. 3 **Gene and genome annotation**

Prior to the annotation step, gene models need to be generated. Different approaches are able to forecast these models. Hybrid approaches encompass the use of genomic and transcriptomic reads to identify open reading frames. This method starts with a mapping of transcribed genes on a reference genome and the generation of putative gene models starting with a start codon (in most cases methionine) and ending with a stop codon (typically for plants are: UAG, UGA or UAA [81]).

Using these gene cassettes, distinct algorithms are trained to detect genes abundant in transcriptomic data, however visible on the genome. This can also be supplemented with proteomic data to evaluate the translated transcripts of the sample. Another method is the *de novo* method, which uses statistical models and can be used if no expression (transcriptomic or proteomic) data is available. For prokaryotic sequences the prediction of these cassettes is easier as they lack of intron sequences. In eukaryotic sequences, these algorithms predict the exon – intron structure using either mRNA sequence data from RNA-Seq experiments or predicted intron/exon boundaries and propose gene models [82]. Compared to expression data, this methods overestimate the genes within a genome [83].

After determining putative gene models, their annotation can be assigned via a BLAST or DIAMOND search [84], [85]. These annotations only provide the potential name of a sequence compared to the sequence similarity to another characterized gene or protein. A detailed confirmation about functions of a protein can only be determined during experimental characterizations.

Functional annotation is a further step in the characterization of genes. Databases such as InterPro [86], Pfam [87], KEGG [88], GO [89] or COG [90] curate conserved regions and domains present in a set of genes. These patterns show similar activity and function in respective genes which allows a first glimpse in the genetic makeup and corresponding function of the genes in a new genome assembly. Putative gene models are annotated using above mentioned methods and a potential function is assigned. The resulting protein families can be a starting point for the search of promising candidates in drug discovery. For the identification of TS the InterProScan domain seed file IPR036965 can be used. Here, typical domains coding for TS activity can be located. Whereas the seed IPR01128 and IPR036396 is used as an indicator for CYPs.

1. 6. 4 **RNA-seq**

To provide an overview of transcribed data and improve the gene prediction, RNA-seq experiments can be performed. In addition to their sequenced nucleic acid (RNA for transcriptomic and DNA for genomic analysis), the applied sequencing method differs between transcriptomic and genomic data is. Whereas in genome sequencing experiments preferably long-read methods are used, short reads are still employed for transcriptome analysis out of economic reasons and for experiments were high coverage is needed. Thus, the quality and accuracy of each base called is higher compared to long-read datasets. Short-read data is usually used to further increase the quality of long read genomic data [91].

Common RNA-Seq experiment setups involve the use of single- or paired-end sequencing. The length of the sequenced RNA is mostly dependend on the method used. Illumina-based methods are generating short-reads and are overall more cost efficient for quantification experiments as well as *de novo* transcriptome assemblies. Long-read transcript sequencing, called IsoSeq (PacBio) has the potential to generate full length RNA sequences, thus rendering total RNA structures including splicing variants in a more comprehensive way [92], [93]. Therefore, differential gene expression analysis is a powerful tool used in molecular biology to compare the expression of genes across different samples or conditions. This analysis is important because it allows researchers to identify genes that are upregulated or downregulated in response to a particular stimulus, disease state, or environmental condition. By comparing gene expression patterns between two or more groups, differential gene expression analyses can reveal the molecular mechanisms underlying physiological and pathological processes.

Transcriptomic data is essential for differential gene expression analysis. Transcriptomics refers to the study of all RNA transcripts produced by a cell or tissue at a particular point of time. Transcriptomic data is obtained using high-throughput sequencing technologies, such as RNA-sequencing (RNA-seq), which allows the measurement of the abundance of each transcript in a sample. This information can then be used to identify genes that are differentially expressed between two or more groups. Transcriptomic data can also provide information about alternative splicing, RNA editing, and non-coding RNA expression, which are important regulators of gene expression.

Differential gene expression analysis has been used in many important scientific discoveries. For example, in the field of cancer research, differential gene expression analysis has been used to identify genes that are differentially expressed in cancer cells compared to normal cells. This has led to the identification of new targets for cancer therapy and the development of new diagnostic tools [94], [95]. In the field of developmental biology, differential gene expression analysis has been used to identify genes that are differentially expressed during embryonic development, leading to a better

understanding of the molecular mechanisms underlying development. Differential gene expression analysis has also been used in the study of infectious diseases, neurodegenerative diseases, and several other areas of research [96]–[99].

In conclusion, differential gene expression analysis is a powerful tool for understanding gene regulation in health and disease. Transcriptomic data is essential for this analysis, and advances in sequencing technologies have greatly facilitated its use. Differential gene expression analysis has played a critical role in many important scientific discoveries and will continue to be an essential tool in biomedical research.

2 Methods

Methods used in this thesis are summarized. A detailed overview and description can be found within the material and methods part of each submitted manuscript

2. 1 Chemicals and reagents

The chemicals utilized in the summarized projects were sourced from reputable suppliers and were of the highest available purity grade. Sequencing kits used for long-read sequencing were purchased from Pacific Biosciences (Menlo Park, CA, USA) and kits for short-read sequencing from Illumina (San Diego, CA, USA). TH Geyer (Renningen, Germany) or Roth Chemicals (Darmstadt, Germany) provided solvents, which were used in the extraction methods.

Table 1. Consumables for PacBio Sequencing

Consumables for sequencing	Vendor
SMRTbell Express Template Prep Kit 2.0	Pacific Bioscience, Menlo Park, CA, USA
SMRTbell Enzyme Cleanup Kit	Pacific Bioscience, Menlo Park, CA, USA
AMPure PB Beads	Pacific Bioscience, Menlo Park, CA, USA
Barcoded Overhang Adapter Kit 8B	Pacific Bioscience, Menlo Park, CA, USA
Sequel II Binding Kit 2.0 + Internal Control	Pacific Bioscience, Menlo Park, CA, USA
SMRT Cell 8M tray	Pacific Bioscience, Menlo Park, CA, USA
Sequel II sequencing kit 2.0	Pacific Bioscience, Menlo Park, CA, USA

2. 2 Extraction methods

2. 2. 1 Extraction of volatile compounds

Two ways were used to extract compounds out of plant samples. The first approach was to macerate the samples with liquid solvents such as hexane, methanol or ethyl acetate. These extracts can be used for regular GC-MS measurements. With the second approach, all volatile compounds in the plant sample are directly measured without the need of further extraction using GC-MS Headspace. For the projects in this thesis, only the latter approach was used.

2. 2. 2 Volatile compound analysis via GC-MS Headspace

“Fresh mature leaves were weighed in GC headspace vials and analyzed using a Trace GC-MS Ultra system with DSQII (Thermo Scientific, USA). Vials were incubated for 10 min at 100 °C and a TriPlus autosampler was used to inject 1 µl of the sample in split mode onto a SGE BPX5 column (30 m, I.D 0.25 mm, film 0.25 µm); an injector temperature of 280 °C was used. Initial oven temperature was kept

at 50 °C for 2.5 min. The temperature was increased with a ramp rate of 10 °C/min to 320 °C with a final hold for 3 min. Helium was used as carrier gas with a flow rate of 0.8 ml/min and a split ratio of 8. The MS chromatograms and spectra were recorded at 70 eV (EI). Masses were detected be-tween 50 m/z and 650 m/z in the positive mode [53]. Samples were measured in biological triplicates and the area average used to compare peaks. Compounds were identified by spectral comparison with a NIST/EPA/NIH MS library version 2.0. To provide insight in the differentiation between plant samples a PCA was conducted" [11].

2. 2. 3 **High molecular genomic DNA extraction**

To yield high-quality and pure high molecular weight genomic DNA for long-read sequencing, an extraction protocol was optimized according to the needs of plants [100]–[102]. The protocol can be used for other organisms too, however shows the best results when used with plants. Starting material is ground plant samples. This can be achieved either with mortar and pestle (under liquid nitrogen) or using technical applications such as the CryoMill (Retsch, Haan, Germany). In Table 2, the grinding parameters are listed.

Table 2. Summary of CryoMill parameters

Function	Frequency in Hz	Duration in min
Pre-cooling	5	6
Disruption	25	2:30
Cooling between cycles	5	0:30
Number of Cycles	2 to 3	

Extraction of gDNA for PacBio long-read sequencing

(CTAB / PCI method)

Extraction of the DNA

5 ml CTAB buffer is mixed with approximately 2% PVP and solved at 60 °C > cool down to 50 °C.

0,5 g plant sample is ground with liquid N_2 (Cryomill or using cooled mortar and pestle)

The ground sample is mixed in 5 ml CTAB

Clumps are homogenized with 5x short pulse vortexing

200 µl Proteinase K (Qiagen) is added

30 min incubation at 50 °C, invert by time to time > cool down to RT

During incubation, the centrifuge is cooled down to 10 °C

100 µl RNAse A is added (10 mg/ml > 1000 µg), 10 min incubation at RT

5 ml Phenol – Chloroform – Isoamylalkohol 125:24:1 (PCI) is added and Falcon inverted

Spin down at 10,000 xg for 5 min at 10 °C. Upper, aqueous phase is transferred in a new falcon

5 ml Chloroform is added and falcon inverted

Spin down at 10,000 xg for 5 min at 10 °C, upper, aqueous phase is transferred in a new falcon

Precipitation of the DNA

1 ml 30% PEG (6000) is added to 4 ml Probe (1:4)

30 min incubation at 4 °C

Spin down at 12,000 xg for 30 min, carefully discard the supernatant

Cleaning and Solving of the DNA

500 µl 70% Ethanol is added

Spin down at 5000 xg for 5 min at 10 °C, carefully discard the supernatant

Dry the pellet at 40 °C until ethanol is evaporated.

100 µl elution buffer is used to resolve the pellet

No flicking or vortexing. DNA resolves overnight

Quality Control:

Nanodrop: concentration and purity ratios

Qubit: concentration of intact high molecular weight DNA

Further cleaning can be performed with magnetic beads. Nanodrop and Qubit concentrations should not show deviations greater than 30 %.

50 ml Falcons

15 ml Falcons

2x

3x

3x

Additional information

PVP (polyvinylpyrrolidone)

RT (room temperature)

Do not vortex to mix different phases

When working with PCI, precaution is necessary

Precipitation with isopropanol/ethanol NaCl is also possible. However PEG yields cleaner DNA.

After Precipitation, the DNA pellet is clear, taking of the supernatant needs to be performed in a very careful manner. After first cleaning step, pellet will be whiteish.

Use freshly prepared 70% ethanol (made from 100% ethanol p.a.)

Adjust the volume of elution buffer according to the size of the pellet

Alternative grinding method: Cryomill (Precool 6 min, 5 Hz, Grind 2min 30sec, 25 Hz, Zwischenkühlen 30 sec, 5 Hz

CTAB buffer recipe:

2 % CTAB (cetyl trimethyl ammonium bromide)

100 mM Tris pH 8.0

20 mM EDTA

1.4 M NaCl

2. 2. 4 RNA extraction and quality check

To add more depth and accuracy to the proposed gene models, as well as to identify transcript abundancies, mRNA was sequenced. For RNA extraction, frozen, unthawed leaves were ground using a CryoMill and an RNeasy Plant Mini Kit (Qiagen, Venlo, The Netherlands). A Turbo DNA free Kit (Invitrogen, Waltham, USA) was used to further clean the RNA. Both kits were used according to manufacturer's recommendations. To ensure the integrity of RNA, for short-read sequencing the Bioanalyzer RNA 6000 assay kit (Agilent, Santa Clara, United States) was employed, whereas for long-read IsoSeq the Qubit RNA IQ kit was used to yield an average RNA Integrity Number of greater than 7. The high-quality RNA was used to perform an IsoSeq library prep using SMRTbell prep kit 3.0 and Sequel II Binding Kit 3.2. (Pacific Biosciences, Menlo Park, USA) for long-read transcript sequencing and for short-read sequencing, the library was prepared using the Illumina stranded mRNA prep kit with IDT for Illumina UD Indexes, Plate A. Corresponding adapter was the Illumina Nextera Adapter (CTGTCTCTTATACACATCT). Both library preparations were performed according to the manufacturer's protocol. For short-read sequencing a shortened fragmentation time from 8 min to 2 min was employed.

2. 3 Quality check and preparation of DNA for sequencing

After the extraction of the DNA the quality and size of the gDNA has to be assessed. For the quality, two parameters were assessed: Absorbance ratios (260 / 280 and 260 / 230) and the concentration. A Nanodrop photometer (Implen, Munich, Germany) was used to check the absorbance at respective wavelengths. DNA is acceptable with a 260 nm / 280 nm ratio of about 1.8. A lower ratio may be due to contaminations such as proteins, phenols or contaminations which absorb around 280 nm. The second ratio should be around 2.0. Here, a lower ratio indicates contamination with reagents absorbing at 230 nm such as typical salts such as EDTA, Guanidine HCl, or TRIzol, which are commonly used in DNA extractions kits [103]. The concentration can be also checked with the Nanodrop photometer however the Qubit dsDNA HS Kit (Thermo Scientific, Waltham, USA) specifically intercalates in double stranded DNA which can be used as control mechanism. Both concentrations should not differ greater than 50%. For further cleaning, an AMPure PB bead clean up or an electrophoretic clean up using a BluePippin system (Sage Science, Beverly, USA) can be performed. A last quality control step is employed with a Femto Pulse system (Agilent, Santa Clara, USA). The DNA is differentiated by its size and should contain a single peak at a high molecular weight (around 160 kb).

Ensuring a pure and well-sized DNA sample preparation can be performed. In a gTube (Covaris, Woburn, USA; 1,700× g), 5 µg HMW gDNA were sheared and used for whole genome library preparation using SMRTbell prep kit 3.0 (Pacific Biosciences, Menlo Park, USA) according to the

manufacturers' protocols. AMPure PB beads were used to determine a size cut-off and cleaner assemblies after sequencing. Libraries were stored at −20 °C. In a last step before sequencing, primer and polymerase were bound using the Sequel II Binding Kit 3.2 (Pacific Biosciences, Menlo Park, USA) as recommended by the manufacturer.

2. 4 Nucleic acid sequencing

Long-read sequencing was performed on a PacBio Sequel IIe (Pacific Biosciences, Menlo Park, CA, USA) with two hours of pre-extension, two hours of adaptive loading (target p1 + p2 = 0.95) yielding an on-plate concentration of 85 pM and 30-hour movie time. Short-read sequencing was performed at the Helmholtz Munich (HMGU) with the Genomics Core Facility on a NovaSeq6000 SP (2x150bp).

2. 5 Bioinformatic analyses

After internal refinement of the reads and de-multiplexing of barcoded samples, the sequences can be further analyzed. A vast pool of free and online-available tools can be employed. In the projects in this thesis tools from Galaxy project [75] were used in combination with self-developed python scripts and linux-based software. Statistical analysis was carried out using R.

2. 5. 1 Genome assembly

For genome assembly different methods can be taken in consideration. Two main methods are assisted / reference-based assembly or *de novo* assembly. As mentioned before, long-read data is suitable for *de novo* assembly due to the possibility to span over repetitive and GC-rich sequences [71]. The initial *de novo* genome assembly was performed with SMRT Link (v11.0.0+, Pacific Biosciences) which uses Improved Phased Assembly [104]. After polishing, the contigs were divided in primary and haplotype associated contigs using purge_dups [105]. This way polyploid genomes can be differentiated[106]. In a second refinement step, repetitive sequences were masked and contaminating sequences discarded using MetaBAT2 [107]. This assembler is usually used for metagenome analysis and divides the assembled reads into bins according to their taxonomy. To check the efficacy of the assembly and completeness of the genomes BUSCO can be used [108]–[111]. Here, gene sets which are specific for a taxonomy are searched in the assembly and according to their appearance, such as duplicated, single-copy, fragmented or missing, a completeness score is calculated. This score represents how comprehensive the assembly was. To estimate the correct genome size, jellyfish [112] and GenomeScope [113], [114] are employed. The depicted size is compared with the assembly size and BUSCO score to assess overall genome quality and completeness.

2. 5. 2 Gene model prediction and annotation

After genome assembly, the genes are predicted through AUGUSTUS [115]–[118]. Putative genes are detected regarding their sequence structure and can be further confirmed with hints from short or long-read transcriptomic data. After assessing the total count of models, their function can be analysed either by sequence homology with tools such as BLAST [84], [119] or DIAMOND [85] or using conserved functional domains using InterProScan [86], [120]. Here, for TPS activity the seed file IPR036965 and for cytochrome p450 enzymes IPR01128 are used. EggNOG [121], [122] is a combined functional annotation tool, which provides the annotation from COG , GO, KEGG, KEGG Pathways, InterPro, a BLAST description and further protein characteristics.

3 Research

3. 1 Summaries of included publications

Chapter I: Comparative Genome-Wide Analysis of Two Caryopteris x Clandonensis Cultivars: Insights on the Biosynthesis of Volatile Terpenoids

The article "Comparative Genome-Wide Analysis of Two Caryopteris x Clandonensis Cultivars: Insights on the Biosynthesis of Volatile Terpenoids" has been published in Plants in February 2023 (doi: https://doi.org/10.3390/plants12030632). The author of this thesis, Manfred J. Ritz and Nadim Ahmad contributed equally to the work and writing of this manuscript. Manfred J. Ritz identified the volatile compounds, developed the HMW gDNA extraction protocol for consecutive long-read sequencing and established the pipeline used for bioinformatics analysis.

Terpenoids represent the structurally most diverse natural product family encompassing over 80,000 characterized compounds with established antioxidant, anti-inflammatory, antiviral, antimalarial, antibiotic, or antitumor activities. These secondary metabolites are abundant throughout all domains of life. Especially flowering plants of the Angiospermae class display a vast diversity in terpenoids. Members of this class are *Caryopteris x Clandonensis*, which are so far only used as decorative plants. In these terpenoids are, among others, used as a defence mechanism against biotic (e.g. herbivores or pest) and abiotic influences (e.g. radiation or climate stress). Additionally, they function as attractors for pollinators or as a possibility for energy storage.

In this study, the genomes of two *Caryopteris x Clandonensis* cultivars are investigated by long read sequencing and comparative genomics to access their yet untapped potential in terpenoid biosynthesis. The focus is on genome assembly with subsequent identification of terpene synthases and cytochrome p450 enzymes. Therefore, various cultivars were investigated regarding their volatile compound composition to identify candidates with intriguing terpene spectra. Consequently, the two cultivars Dark Knight and Pink Perfection were selected for deep genome sequencing on a PacBio Sequell IIe, as these differentiate substantially in terpenoid composition based on a principle component analysis. For a comparative genomic approach, the prediction of gene models and a consecutive functional annotation using COG was conducted. Furthermore, a synteny analysis of the two genomes and a phylogenetic assessment of obtained terpene synthases was employed to highlight differences of the compared cultivars. In Dark Knight and Pink Perfection, 45 and 43 putative terpene synthases as well as 1316 and 1363 cytochrome p450 enzymes, were identified, respectively.

The presented results highlight *Caryopteris x Clandonensis* as a new and previously unexploited source of terpene synthases as well as cytochrome p450 enzymes for further studies on

terpenoid biosynthesis. Through genome mining, various other aspects of the endogenous biosynthesis pathways are made accessible for consecutive research in the field of natural products.

Chapter II: Differential RNA-Seq Analysis Predicts Genes Related to Terpenoid Tailoring in *Caryopteris x Clandonensis*

The article "Differential RNA-Seq Analysis Predicts Genes Related to Terpenoid Tailoring in Caryopteris x Clandonensis" has been published in Plants in May 2023 (doi: https://doi.org/10.3390/plants12122305). Manfred J. Ritz and Nadim Ahmad contributed equally to the work of this manuscript. The author of this thesis, Manfred J. Ritz conceived this manuscript and conducted respective short-read sequencing preparations, established the pipeline for analysis and identification of terpene modifying enzymes within the differentially expressed genes *of Caryopteris x clandonensis.*

In this study, the transcripts of six *Caryopteris x Clandonensis* cultivars are investigated by short read sequencing and mapped to our previously published genome. The used reference was subjected to a cleaning step and the assembly was refined. The 782 previously described scaffolds were decreased to 53 with keeping the completeness and contiguity at a high level using a binning method usually used for metagenomic data. Subsequent assignment of short reads revealed a high quality of sequencing and a mapping efficiency of about 88%. The identified transcripts indicated moderate variations in their functional patterns compared to the corresponding genome. Interestingly, differences in expression of terpenoid tailoring enzymes from the cytochrome p450 class were discovered which might address the cause for metabolic variations between these cultivars, especially those who modify limonene derived molecules.

The presented results highlight *Caryopteris x Clandonensis* as a new and previously unexploited source of cytochrome p450 enzymes for further studies on terpenoid tailoring. Through differential expression analysis, various aspects of the enzymatic variation between plant metabolites and their genes are made accessible for consecutive research in the field of natural products.

4 Full length publications

Comparative Genome-Wide Analysis of Two Caryopteris x Clandonensis Cultivars: Insights on the Biosynthesis of Volatile Terpenoids

4 Full length publications

Comparative Genome-Wide Analysis of Two Caryopteris x Clandonensis Cultivars: Insights on the Biosynthesis of Volatile Terpenoids

plants

IMPACT FACTOR 4.658

Indexed in: PubMed

Article

Comparative Genome-Wide Analysis of Two *Caryopteris x Clandonensis* Cultivars: Insights on the Biosynthesis of Volatile Terpenoids

Manfred Ritz, Nadim Ahmad, Thomas Brueck and Norbert Mehlmer

Special Issue

Applications of Bioinformatics in Plant Resources and Omics

Edited by

Dr. Noe Fernandez-Pozo and Prof. Dr. M. Gonzalo Claros

MDPI

https://doi.org/10.3390/plants12030632

plants

MDPI

Article

Comparative Genome-Wide Analysis of Two *Caryopteris x Clandonensis* Cultivars: Insights on the Biosynthesis of Volatile Terpenoids

Manfred Ritz †, Nadim Ahmad †, Thomas Brueck * and Norbert Mehlmer *

Werner Siemens Chair of Synthetic Biotechnology, Department of Chemistry, Technical University of Munich (TUM), 85748 Garching, Germany
* Correspondence: brueck@tum.de (T.B.); norbert.mehlmer@tum.de (N.M.)
† These authors contributed equally to this work.

Abstract: *Caryopteris x Clandonensis*, also known as bluebeard, is an ornamental plant containing a large variety of terpenes and terpene-like compounds. Four different cultivars were subjected to a principal component analysis to elucidate variations in terpenoid-biosynthesis and consequently, two representative cultivars were sequenced on a genomic level. Functional annotation of genes as well as comparative genome analysis on long read datasets enabled the identification of cultivar-specific terpene synthase and cytochrome p450 enzyme sequences. This enables new insights, especially since terpenoids in research and industry are gaining increasing interest due to their importance in areas such as food preservation, fragrances, or as active ingredients in pharmaceutical formulations. According to BUSCO assessments, the presented genomes have an average size of 355 Mb and about 96.8% completeness. An average of 52,090 genes could be annotated as putative proteins, whereas about 42 were associated with terpene synthases and about 1340 with cytochrome p450 enzymes.

Keywords: reference genome; terpene synthases; *Caryopteris x clandonensis*; plant volatiles; long read sequencing; TPS subfamilies

Citation: Ritz, M.; Ahmad, N.; Brueck, T.; Mehlmer, N. Comparative Genome-Wide Analysis of Two *Caryopteris x Clandonensis* Cultivars: Insights on the Biosynthesis of Volatile Terpenoids. *Plants* **2023**, *12*, 632. https://doi.org/10.3390/plants12030632

Academic Editors: Noe Fernandez-Pozo and M. Gonzalo Claros

Received: 21 December 2022
Revised: 25 January 2023
Accepted: 27 January 2023
Published: 1 February 2023

1. Introduction

Throughout the last decades, terpenes and terpenoids became more and more important in industrial applications. In the food industry terpenes are used, e.g., as flavoring compounds [1] or preservatives [2]. Due to its plant origin, the acceptance as a food additive is higher compared to chemical synthesis. In a pharmaceutical context the research and use of essential oils—with terpenes as their main components—range from anti-inflammatory [3], and immunomodulatory [4] to antiviral [5] and further indications [6–11]. The anti-cancer drug Taxol consists of a diterpenoid backbone [12] and is employed in different cancer treatments [13]. The success of this terpenoid surely is one of the reasons to further research terpenoids for pharmaceutical applications. Along with these applications, this class of molecules can be found throughout most organisms. Flowering plants show a vast diversity of terpenoids, which is a unique characteristic of the class *Angiospermae* [14]. In plants, they are used as a defense mechanism against biotic (e.g., herbivores or pests) and abiotic influences (e.g., radiation or climate stress) [15]. An example of a defense mechanism against biotic stress is the insect repellent activity of volatiles, such as p-menthane-3,8-diol from *Corymbia citriodora* [16,17]. This compound shows activity against the yellow fever mosquito *Aedes aegypti* [18]. *Caryopteris x clandonensis* essential oils also harbor a biological activity against these insects [19]. However, for these plants, the active agent is not yet identified. Additionally, terpenoids function as attractors for pollinators or as a possibility for energy storage [14].

The extensive diversity of natural terpenes derives from the conserved evolution of terpene synthases (TPS) and terpene-modifying enzymes, such as cytochrome p450

enzymes [20,21]. Terpenes are divided into different classes defined by their backbone. The basis is two building blocks, isopentenylpyrophosphate (IPP) and dimethylallyl diphosphate (DMAPP) which are synthesized in plants via the mevalonate pathway. IPP is the activated form of an isoprene unit consisting of five C-atoms (C5), also called hemiterpene. These are connected to larger units forming monoterpenes (C10), sesquiterpenes (C15), diterepenes (C20) and higher terpene structures [22]. Further steps of increasing terpenoid diversity involve the promiscuity of TPS as well as the subsequent modification by cytochrome p450 enzymes, which may encompass hydroxylation, carboxylation, acetylation or peroxide linkage. Examples include the biosynthesis of p-menthane-3,8-diol [17], gibberellin [23], taxol [24], and artemisinin [25], respectively. This results in a vast pool of natural compounds which account for a multitude of possible applications [14,26].

In general, plant TPS are divided into eight subfamilies which are grouped into classes I, II and III. This separation is based on functional assessment, sequence likelihood and architecture of genes. Class I is comprised of copalyl diphosphate synthases (TPS-c), *ent*-kaurene synthases (TPS-e), other diterpene synthases (TPS-f) and lycopod specific (TPS-h). TPS-d is only included in class II, which is specific for *Gymnosperms*. Lastly, class III consists of TPS-a, cyclic monoterpene synthases and hemi-TPS (TPS-b) and acyclic mono-TPS (TPS-g), which are *Angiosperm* specific [27].

With the advent of state-of-the-art bioinformatic technologies, deciphering the molecular mechanisms involved in the formation of terpenoids has become significantly easier [28]. Furthermore, the possibility to produce terpenes recombinantly by means of biotechnological production systems, rather than chemical synthesis, makes it an ecological and cost-effective technology for the increasing demand for terpenes in industrial applications, despite open challenges [29].

The combination of cutting-edge bioinformatics and next-generation sequencing technologies provided by Pacific Biosciences, Oxford Nanopore and Illumina allows for the rapid generation of draft genomes as well as the annotation of valid gene models. In this context, long-read sequencing technologies will be highlighted, as they exhibit no amplification biases. Consequently, these technologies provide a reliable basis for de novo whole-genome assemblies. Openly accessible bioinformatic tools enable cost-efficient assemblies, annotations and secondary downstream analyses for a broad range of scientists, and are publicly available via www.github.com (accessed on 11 December 2022) [30]. Two of these are the Quality Assessment Tool for *Genome* Assemblies (QUAST) [31] and Benchmarking Universal Single-Copy Orthologs (BUSCO). The latter is employed to assess the completeness of the obtained genome assemblies. Here, conserved and species-specific gene sequences are curated in databases and detected via a match-making algorithm to check for the gene set completeness of the evaluated taxonomic group [32]. An investigated genome is classified as complete if respective single-copy orthologs are present in the assembly.

In this work, we present the genomes of two *Caryopteris x clandonensis* cultivars (Dark Knight and Pink Perfection) from the *Lamiaceae* family in high quality employing long-read sequencing. These plants display a wide range of different metabolic pathways in regard to terpenoid biosynthesis, as also seen in other plants of the order *Lamiales*, e.g., in *Jasminum sambac* [33]. To elucidate variations between these multivariate datasets a principal component analysis (PCA) was conducted. Based on evident differences in volatile compound composition the two cultivars, Dark Knight and Pink Perfection were compared on a genomic level. This submission will be the 12th whole genome sequence within *Lamiaceae*, consisting of about 4788 further species, making it a source for gene sequences and further experimental basis in plant and natural product focused biosynthesis research.

2. Results and Discussion

2.1. PCA Analysis of Volatile Compounds

Differences between the volatile compounds of four cultivars were investigated using a GC-MS Headspace analysis. Ten main volatile components visible between the cultivars

were selected, predominantly monoterpenoids and sesquiterpenoids, which are listed in Table 1. It has already been shown that there is a variety of monoterpene synthases that are able to catalyze ionization and isomerization starting from geranyl diphosphate [34]. Furthermore, the analysis of the cultivars revealed that a switch between pinene and limonene-derived compounds took place, which was sparsely synthesized in the other plants. In Table 1, these compounds are marked with an asterisk, one (*) represents limonene-related terpenoids, and two (**) represents pinene-related terpenoids. This especially is visible in the C4-C6 shift compared to the limonene backbone as seen in pinene (C4 to C6, see Figure S1). Similar substances could be identified as investigated previously for this plant species [19].

Table 1. Ten main volatile compounds of four *Caryopteris x clandonensis* cultivars, visible between the cultivars were selected and are hierarchically listed (top: higher concentration, bottom: lower concentration). GC-MS Headspace was performed and an identification with a NIST/EPA/NIH MS library version 2.0 was conducted. * represents limonene-related terpenoids. ** represents pinene-related terpenoids.

Dark Knight	Good as Gold	Hint of Gold	Pink Perfection
α-pinene **	D-limonene *	D-limonene *	D-limonene *
trans-pinocarveol **	Cubebol	Cubebol	cis-p-mentha-1(7),8-dien-2-ol *
Pinocarvone **	Carvone *	trans-carveol *	trans-p-mentha-2,8-dien-1-ol *
Caryophyllene oxide	trans-carveol *	Carvone *	Caryophyllene oxide
β-pinene **	cis-p-mentha-1(7),8-dien-2-ol *	Caryophyllene oxide	trans-carveol *
(E,E)-α-farnesene	Caryophyllene oxide	trans-p-mentha-1(7),8-dien-2-ol *	cis-p-mentha-2,8-dien-1-ol *
α-campholenal	α-copaene	cis-p-mentha-1(7),8-dien-2-ol *	Carvone
α-copaene	β-pinene **	cis-p-mentha-2,8-dien-1-ol *	α-pinene **
Caryophyllene	cis-p-mentha-2,8-dien-1-ol *	α-copaene	β-pinene **
D-limonene *	trans-p-mentha-2,8-dien-1-ol *	trans-p-mentha-2,8-dien-1-ol *	Caryophyllene

As the plants are cultivars from *Caryopteris x clandonensis* a common base profile (e.g., caryophyllene, perillyl alcohol, sabinene, farnesene or campholenal) of volatiles was expected, see Table S1, as has been shown for other plants and their cultivars [35,36]. In this study, distinct differences between Dark Knight, Good as Gold, Hint of Gold, as well as Pink Perfection, can be shown.

To further investigate the variations in the compound profile found during the analysis, a principal component analysis (PCA) was performed (Figure 1). Good as Gold and Hint of Gold express high morphological and metabolic similarity (see Figure S2 and Table S1). This is also evident in Figure 1, as both cultivars are located close to each other. On the other hand, Dark Knight and Pink Perfection showed the highest deviation in volatile compound composition. Moreover, the switch between C1 and C6 as mentioned above results in an intriguing product spectrum. These data underline the variations between the cultivars and demonstrate a need for further investigations into the molecular makeup of underlying TPS and cytochrome p450 enzymes, which are key for generating the molecular diversity of plant-based terpenoid structures in plants [20]. Therefore, due to their distinct differences revealed in the PCA, the two cultivars, Dark Knight and Pink Perfection, were sequenced to elucidate genomic differences and identify unique and yet unknown genes.

2.2. Genome Sequencing and Quality Assessment

In Table 2, the sequencing metrics of the respective Sequel IIe runs are depicted. Details regarding sequencing quality reports can be found in Figure S3. Total bases were nearly twofold higher in Dark Knight than in Pink Perfection, the same as obtained HiFi reads and yield. However, the HiFi read length, read quality and number of passes are comparable in both sequencing runs. Deviations in sequencing parameters are closely related to utilized libraries and input DNA quality. As the read quality is well above Q20 both runs were subjected to further analyses.

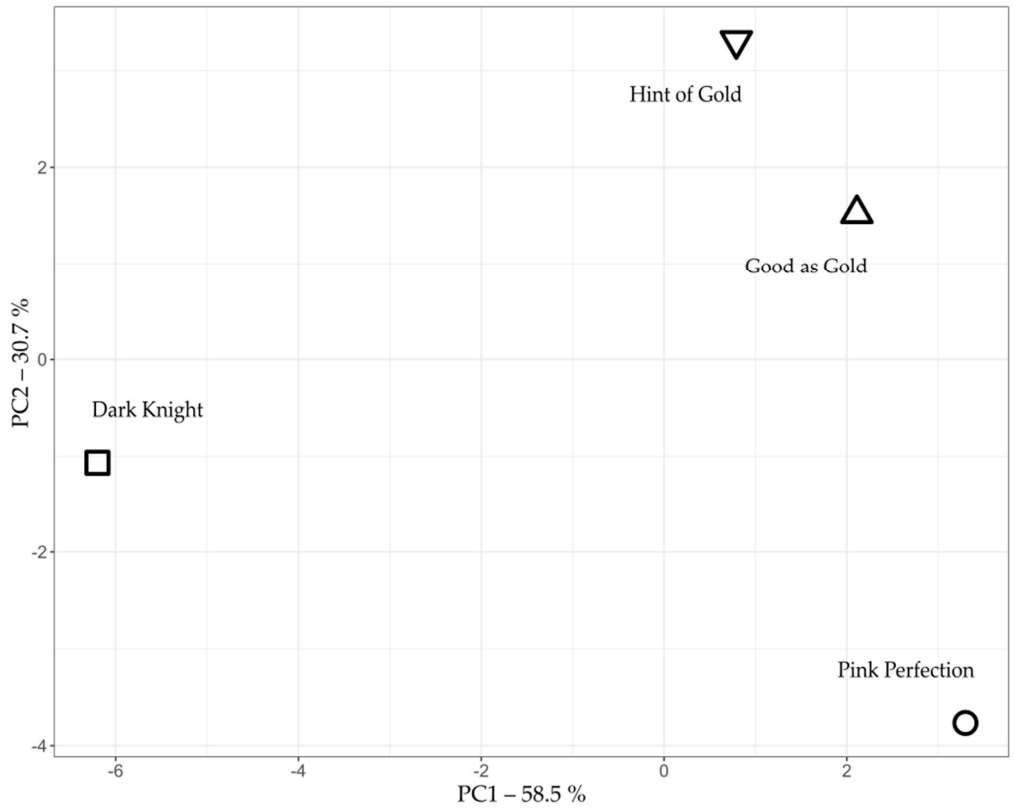


Figure 1. A principal component analysis of four different *Caryopteris x clandonensis* cultivars, Dark Knight, Good as Gold, Hint of Gold and Pink Perfection regarding the area of their volatile compounds analyzed by GC-MS Headspace.

Table 2. Sequencing parameters of the PacBio Sequel IIe runs of *Caryopteris x clandonensis* cultivars Dark Knight and Pink Perfection.

Analysis Metric	Dark Knight	Pink Perfection
Total Bases (Gb)	444.13	229.43
HiFi Reads	1,823,939	843,632
HiFi Yield (Gb)	27.28	12.92
HiFi Read Length (mean, bp)	14,954	15,312
HiFi Read Quality (median)	Q35	Q34
HiFi Number of Passes (mean)	12	13

In this study, both genomes of Dark Knight and Pink Perfection were assembled using the IPA assembler with a consecutive duplicate purging and phasing step. A QUAST analysis was conducted to assess assembly contiguity (see Table 3).

Table 3. Genome contiguity assessment based on statistics generated by using QUAST.

Assembly	Dark Knight	Pink Perfection
# contigs	1183	782
Largest contig	29,672,976	31,977,049
Total length	366,625,098	344,117,456
Estimated reference length	300,000,000	300,000,000
GC (%)	31.50	31.77
N50	8,177,750	7,086,741
L50	13	14
# N's per 100 kbp	0.41	0.44

The number of assembled contigs diverged in both candidates (see Table 3). However, respective L50 values were small (13 for Dark Knight and 14 for Pink Perfection) compared to obtained N50 (8.2 Mb and 7.1 Mb respectively), which assures gene integrity with only low or no fragmentation. The total contig length of complete genomes corresponds to their size, which is comparable (3.44 to 3.66 $\times$ 10^8 bp), and the same as seen for GC content (31.5% and 31.77%). Furthermore, genome size was calculated using a k-mer-based analysis, with a k-mer size of 20. Results support the haploid genome size of ~355 Mb and estimated a diploid genome, see Figures S4 and S5. Based on the calculated genome size the coverage of Dark Knight and Pink Perfection resembles 74 and 38, respectively.

To assess the genome completeness and reliability of both genome sequences, a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis was performed (see Figure 2). Both genomes were compared to the kingdom *Viridiplantae* and the clades *Embrophyta* and *Eudicotidae*, respectively. The selection of these lineages was based on the increasing grade of affiliation and the different accompanying BUSCO gene sets (in former order). For closer clades, more concise sequences are necessary in order to be identified as complete. In our case, even more affiliated clades show less deviation of completeness than expected in comparison to *Viridiplantae*. As the genomes were compared to different BUSCO datasets, the obtained results were depicted after normalization in Figure 2 to enable a concise comparison. Assessed genome completeness from the closest related clade (*Eudicotidae*) was 96.6% for Dark Knight and 96.8% for Pink Perfection, which were also compared to reference genomes of *Salvia splendens* (92.1%) [37] and *Sesamum indicum* (95.1%) [38]. The latter were only compared with the *Viridiplantae* database with BUSCO v2.0.1 and v3.0, whereas our data were analyzed by BUSCO v5.3.2. This may have caused the difference between 425 and 1440 BUSCO datasets, as frequent updates of the gene sets are necessary to improve BUSCO analysis [39]. The reference genomes were chosen due to the high prevalence in BLAST searches [40,41] using *Caryopteris x clandonensis* sequences. *S. splendens* appears to harbor mostly complete and duplicated BUSCOs, whereas *S. indicum* shows comparable results to the new genomes of Pink Perfection and Dark Knight with a majority of complete and single-copy BUSCOs. To interpret BUSCO results, it is necessary to understand duplicated BUSCOs and their nature, as these can be of biological or technical origin. In eukaryotic genomes, divergences in haplotypes often lead assemblers to form duplicates of high heterozygosity regions, resulting in contiguity issues and obstacles in further evaluation steps, such as gene annotation [42,43]. To circumvent these issues, tools such as "purge_dups" are utilized to remove duplicate regions (haplotigs) from the assembly to assure genome contiguity [42]. A consecutive polishing of obtained contigs and haplotigs using phasing results in increased genome quality. Of the newly assembled genomes only 0.24–0.69%/0.71–2.67% are fragmented or missing, respectively. The absence of some BUSCO genes may be due to a loss of true genes or these may be existing as true gene duplications [43].

2.3. Evaluation of Structural Differences between Genome Assemblies

To concisely compare genomes, the collinear gene order also known as synteny or syntheny blocks needs to be assessed [44]. It plays an important role in visualizing matches between organisms [45].

Investigating the synteny between cultivar genomes shows their close relation. Here, factors such as low contiguity and fragmentation have an effect on the analysis and lead to high error rates [46]. In our case, previously performed evaluations assured high contiguity and low fragmentation. Mauve was used to perform a multiple sequence alignment and applied to generate synteny blocks (Figure S6) [47]. Connections between these blocks reveal the high similarity within both genomes. This is typical for plant breeding, as specific traits are inherited from previous generations leading to inversions, duplications, or truncations in gene sets [48]. Furthermore, marker synteny can be used for phylogenetic analyses of cultivar evolution [49]. Thus, the plant samples seem to be closely related to the species *Caryopteris x clandonensis.*

2.4. Gene Models and Functional Annotation

Gene models were computed using the presented genome assembly and a long-read IsoSeq database as hints via AUGUSTUS [50–53]. As a training set *Solanum lycopersicum* was chosen due to its ancestral relation to *Lamiaceae*. For the cultivars, a total of 52,865 (Dark Knight), and 51,315 (Pink Perfection) genes were predicted and resemble putative proteins. The Cluster of Orthologous Groups (COG) and Gene Ontology (GO) terms were evaluated for all cultivars. It is to mention that only ~81% of the predicted genes were annotated using COG and GO databases. Out of these ~30% are poorly characterized (Figure 3E) and only a fraction (30%) of those can be assigned with GO terms. In regard to the complete genomes, nearly 20% of the proposed gene models remain without an assigned function. Figure 3 shows the COG counts for the following categories: (3B) cellular processes and signaling (3C) information storage and processing (3D) metabolism and (3E) poorly characterized. Figure 3A combines all the aforementioned categories. The obtained results emphasize a strong similarity in the compared cultivars. Further data in regard to the exact amount of COG per category can be found in Tables S2 and S3. This finding is a further indicator of the completeness of the presented genomes, as different cultivars have a similar set of genes, only varying in small nucleotide polymorphisms or other structural variants, which distinguish them [54,55].

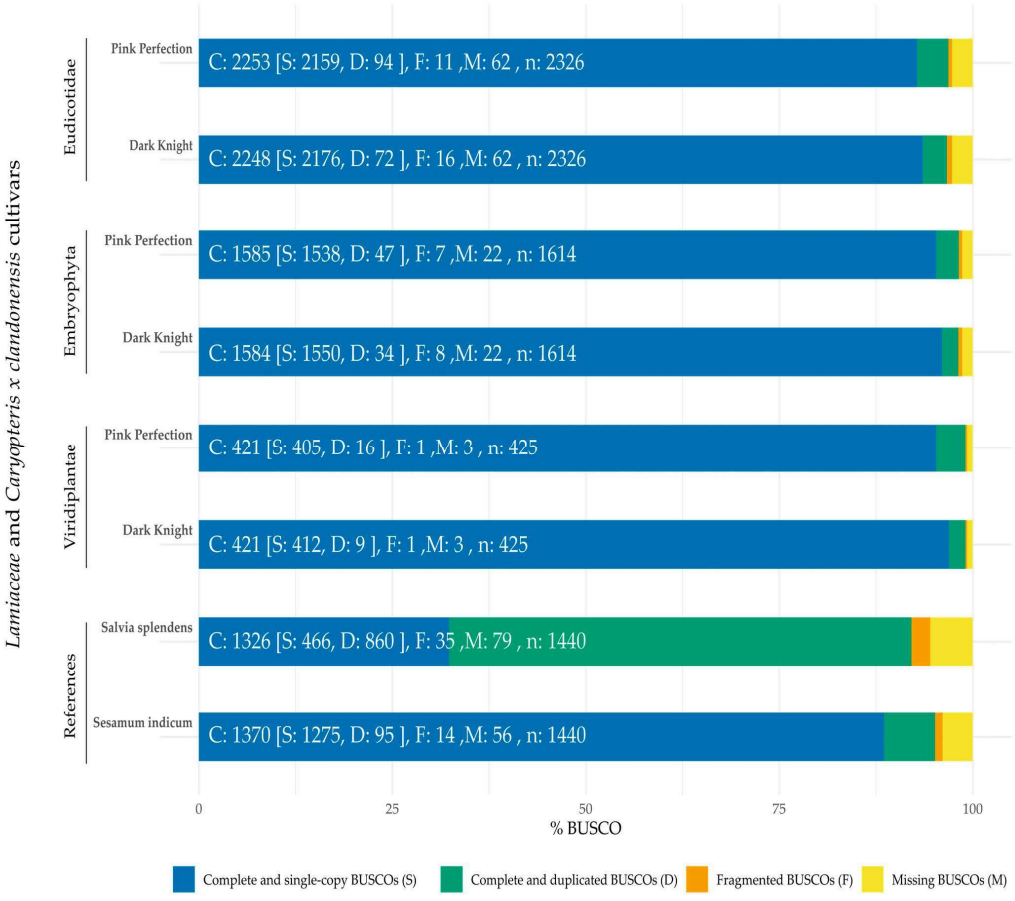


Figure 2. Comparison of BUSCO completeness of different cultivars of *Caryopteris x clandonensis* as well as *Salvia splendens* [37] and *Sesamum indicum* [38]. As the genomes were compared to other Benchmarking Universal Single-Copy Orthologs (BUSCO) datasets a normalization was performed to enable a comparison in genome completeness. Pink Perfection and Dark Knight were compared to the BUSCO datasets of *Viridiplantae*, *Embryophyta* and *Eudicotidae*, whereas *S. splendens* and *S. indicum* were compared to *Viridiplantae* only. Reference genomes were obtained from [37,38].

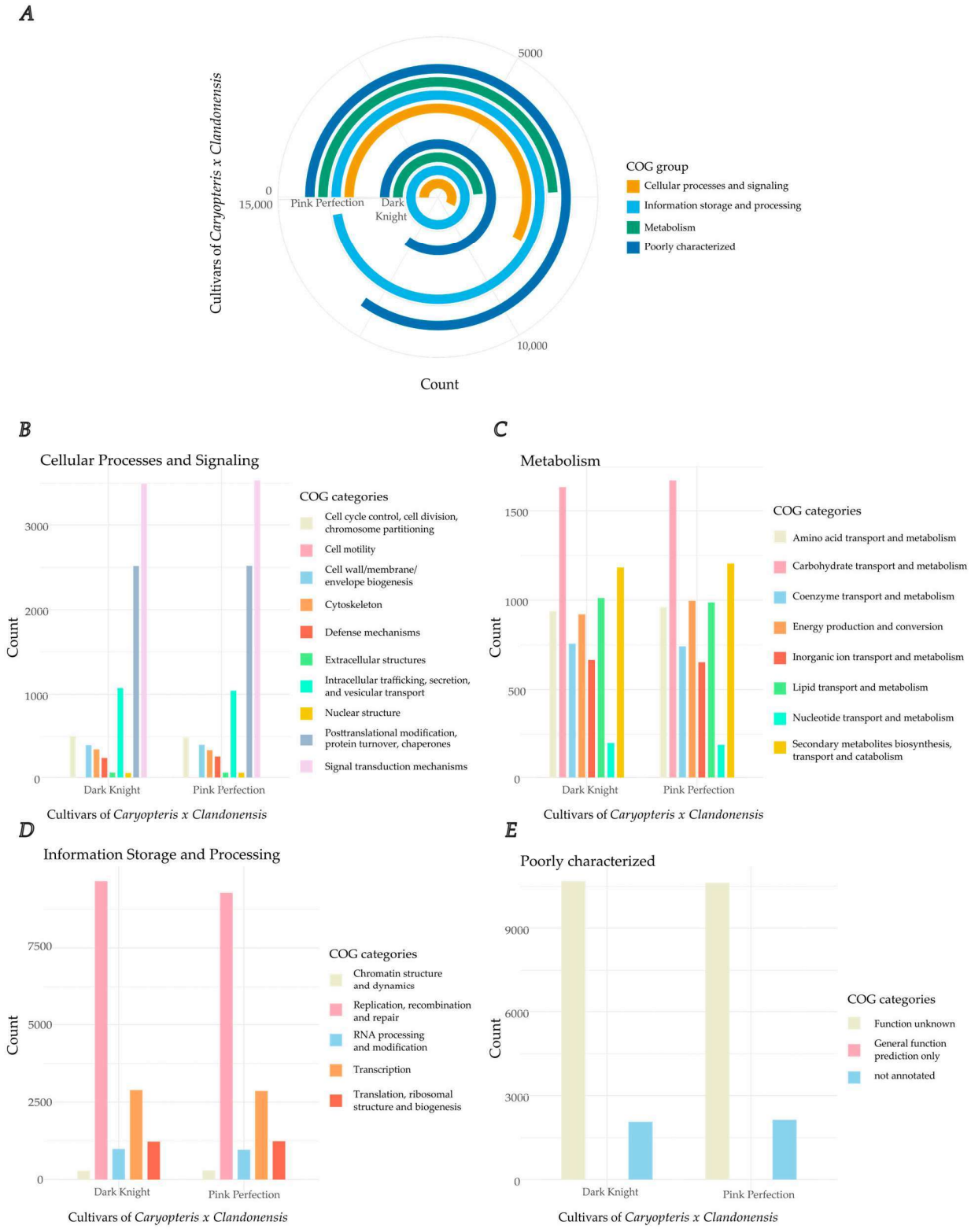


Figure 3. Annotation of gene sets for Cluster of Orthologous Groups (COG) for both cultivars, Dark Knight and Pink Perfection. (**A**) COG of two different cultivars of *Caryopteris x clandonensis*, Pink Perfection (outer ring) and Dark Knight (inner ring). Groups are divided in cellular processes and signaling, information storage and processing, metabolism, and a category for poorly characterized gene sets. (**B**) COG of cellular processes and signaling associated genes, total counts. (**C**) COG of metabolism-associated genes, total counts. (**D**) COG of information storage and processing associated genes, total counts. (**E**) COG of poorly characterized genes, total counts.

A closer look into the different groups reveals characteristic functions in the cultivars. Most genes identified and functionally annotated are associated with replication, recombination and repair, which make up about 20.5% of total annotated genes Figure 3D) followed by signal transduction mechanism (~8%) (Figure 3A). Plants are exposed to endogenous and exogenous stresses such as chemicals or UV-radiation which can significantly alter DNA, thus there is high importance for repair mechanisms [56]. High redundancy of those ensures the safe replication of DNA with almost no errors [57].

In Figure 3C, proteins related to the COG category secondary metabolites biosynthesis, transport and catabolism, rank in second place within metabolism (2.8%). This category harbors TPS and cytochrome p450 enzymes. However, proteins associated with carbohydrate transport and metabolism are most abundant in this group as they are important for general metabolism and backbone synthesis.

Compared to about 29,458 with COG functionally annotated genes, 11,118 unique GO terms were assigned to 14,280 different genes (27% of total gene models). COG terms are ancestrally conserved regions, GO terminology in contrast proposes functional annotation of each hypothetical gene. A gene-set enrichment analysis was conducted with GO terms as a source for gene sets [58]. The following figures show the GO term clustering regarding the three main categories in plants: biological process (Figure 4), molecular function (Figure 5) and cellular components (Figure 6). For all three an analysis was conducted based on GO terms identified in Pink Perfection. Detailed data for Dark Knight and Pink Perfection can be found in Tables S4 and S5. The GO analysis was visualized using REVIGO [59]. Respective cluster position within the semantic space is irrelevant, as similar semantic terms are located in vicinity of each other in the plot [58].

In Figure 4, GO terms related to biological processes are depicted with their respective prevalence (dot size). In addition, some clusters with similar functions were grouped by circles into the main function of these GO terms, as can be seen, e.g., with "translation" in the bottom right corner. Incorporated into this cluster are the terms: protein modification process, DNA metabolic process, nucleobase-containing compound metabolic process, and protein metabolic process. The cluster organelle organization includes cytoskeleton organization, cytoplasm organization, and mitochondrion organization. Clustered with transport: ion transport, protein transport. The last cluster response to stress contains the GO terms response to a biotic stimulus, response to an abiotic stimulus, response to an external stimulus, and response to an endogenous stimulus. GO terms without clustering but still strongly prevalent in the PCA are biological, metabolic and biosynthetic processes.

For the GO analysis of the category molecular function, only one larger cluster was formed, which is nucleic acid binding. It incorporates the functions of DNA binding, RNA binding, and nucleotide binding. The two main components in this category are molecular function and catalytic activity.

GO analysis in the category of cellular components yielded as the main results, intracellular anatomical structure and cellular components, as well as genes related to the cytoplasm. However, no semantic clustering was feasible based on the annotated GO terms.

2.5. Identification of Terpenoid Biosynthesis Enzymes

InterProScan predicts distinct protein domains and classifies them into families [60,61]. The seed files PF01397, PF03936 and IPR036965 are associated with TPS activity. In the annotated protein database, these seeds were used as homology motifs. For Dark Knight 43 and Pink Perfection 41 TPS were identified. The seed file IPR001128 is related to cytochrome p450 enzymes. Here, we were able to identify Dark Knight and Pink Perfection 1316 and 1363 sequences. Compared to other plants these findings are comparable, both for TPS and cytochrome p450 enzymes [62–65].

To investigate the similarity and the affiliation into TPS subfamilies regarding identified TPS, a phylogenetic tree was constructed. Analysis was based on multiple sequence alignment by Clustal Omega using default parameters (see Figure 7). To differentiate between TPS families, 55 selected sequences of representative plant species were utilized

as anchor sequences along with putative TPS from Dark Knight and Pink Perfection; the root was *Physcomitrella patens* (adapted from [66]). The multicolored clades belong to the different TPS subfamilies and are used as references, for a more detailed overview see the supplemental *lamiaceae* reference. Concise numbers of TPS subfamily distribution in both cultivars are shown in Table 4. The most prominent subfamilies are TPS-a (green), TPS-b (black) and TPS-c (purple), which is in line with the distribution in *Eudicots*, *Angiosperms* and land plants [67]. The subfamilies TPS-d and TPS-h are not present in the investigated cultivars. These findings are supported by the literature, as TPS-d clusters are derived from *Gymnosperm* species [63,68] and TPS-h are specific for *Selaginella moellendorffii* [67].

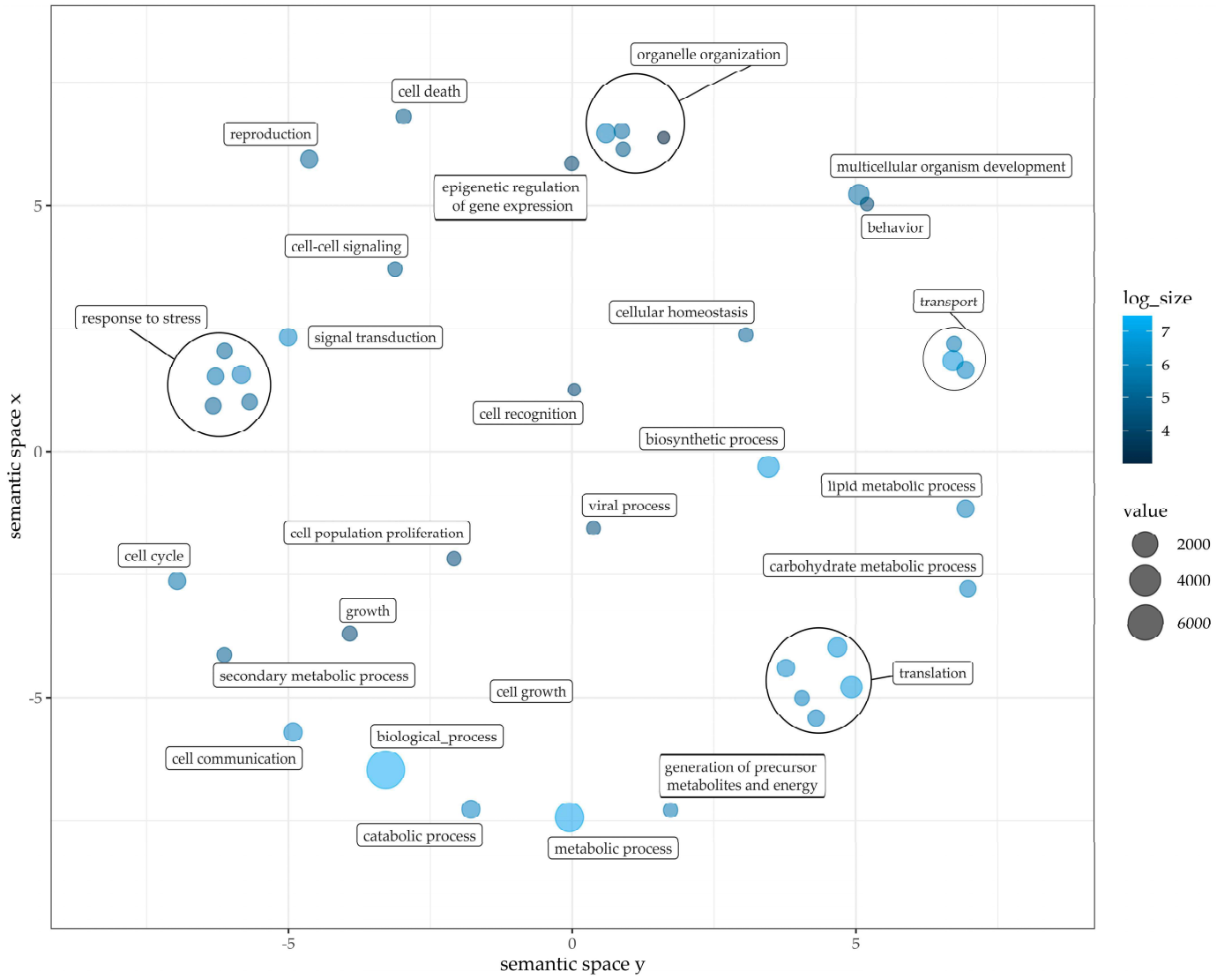


Figure 4. Gene Ontology term classification within biological processes of Pink Perfection. Clustered with response to stress: response to biotic stimulus, response to abiotic stimulus, response to external stimulus, response to endogenous stimulus. Clustered with translation: protein modification process, DNA metabolic process, nucleobase-containing compound metabolic process, protein metabolic process. Clustered with organelle organization: cytoskeleton organization, cytoplasm organization, mitochondrion organization. Clustered with transport: ion transport, protein transport. Figure was drafted employing REVIGO [59] and customized with R. Value and log size represents the counted GO terms across annotated gene models.

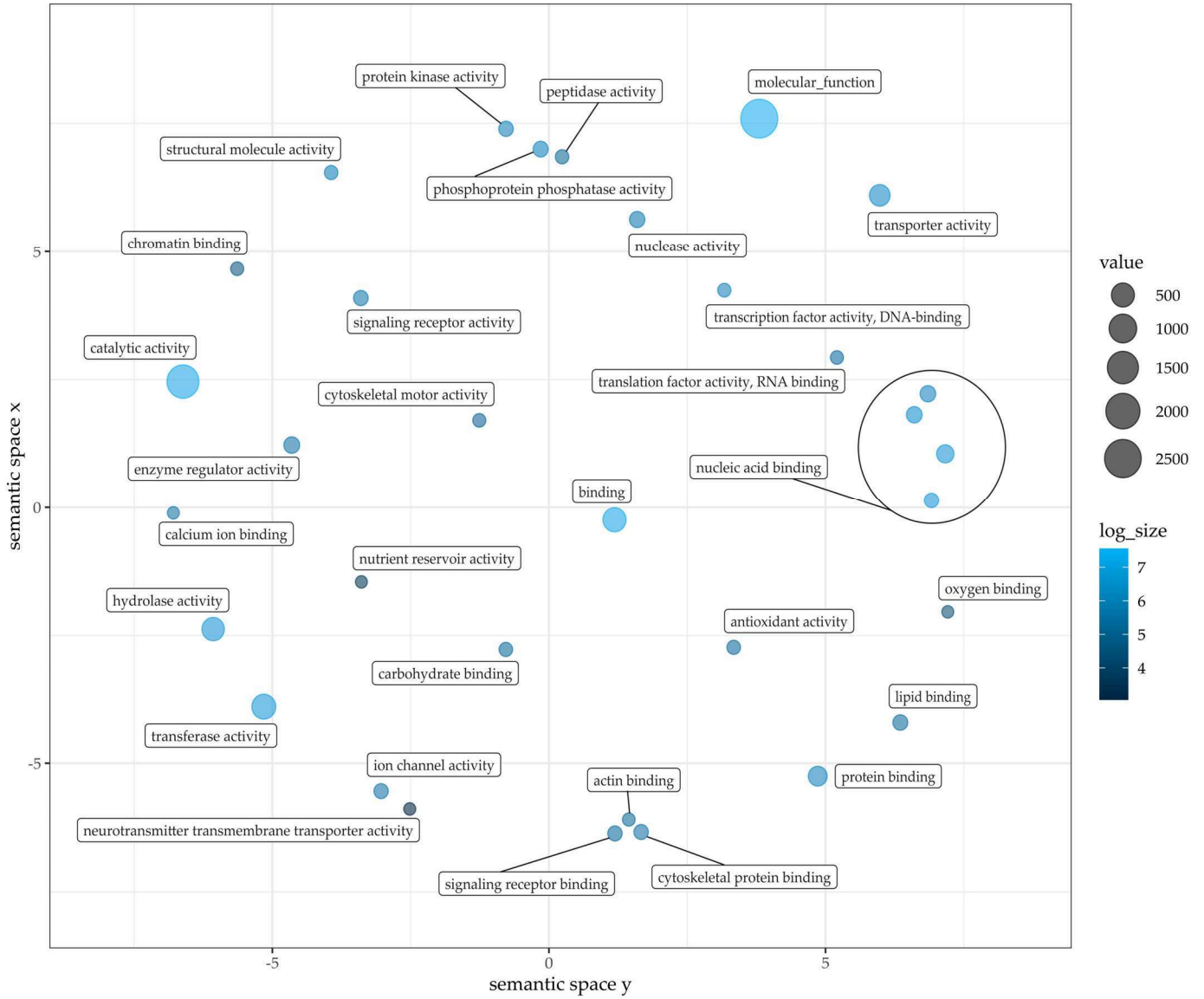


Figure 5. Gene Ontology term classification within molecular functions of Pink Perfection, clustered with nucleic acid binding: DNA binding, RNA binding, Nucleotide binding. Figure was drafted employing REVIGO [59] and customized with R. Value and log size represents the counted GO terms across annotated gene models.

Table 4. Terpene synthase (TPS) subfamilies and their distribution in the *Caryopteris x clandonensis* cultivars Dark Knight and Pink Perfection. TPS-a, -b and -c show the highest prevalence in both cultivars.

TPS Subfamily	Dark Knight	Pink Perfection
a (green)	16	14
b (black)	7	7
c (purple)	10	10
d (blue)	-	-
e (turquoise)	2	2
f (petrol)	5	5
g (red)	3	3
h (pink)	-	-

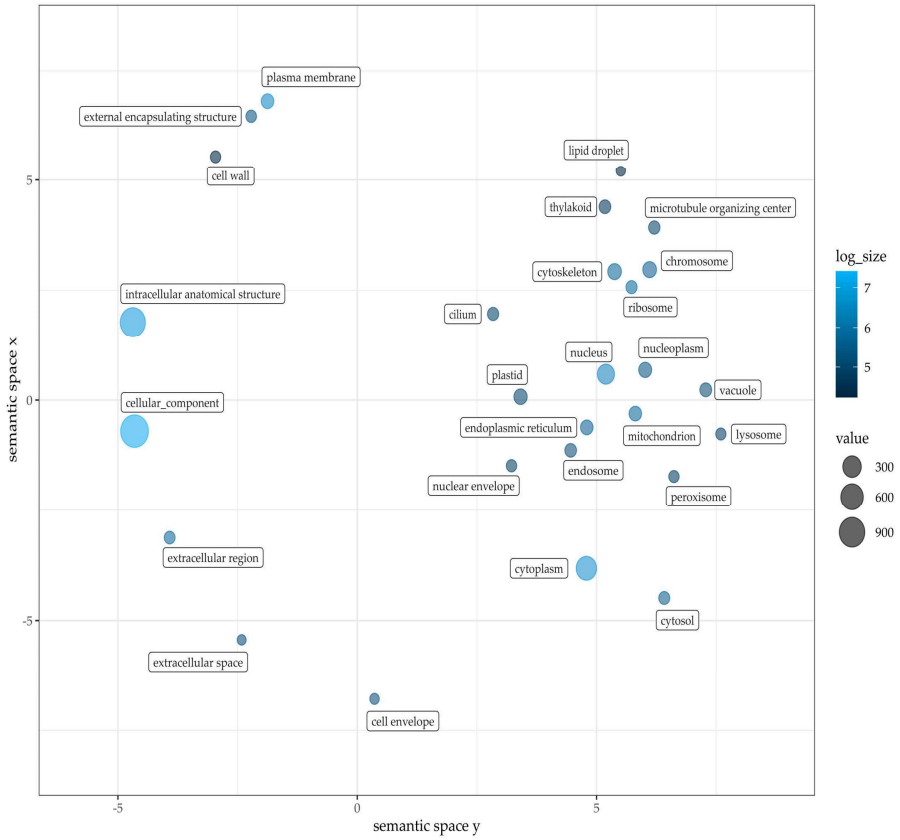


Figure 6. Gene Ontology term classification within cellular components of Pink Perfection. Figure was drafted employing REVIGO [59] and customized with R. Value and log size represent the counted GO terms across annotated gene models.

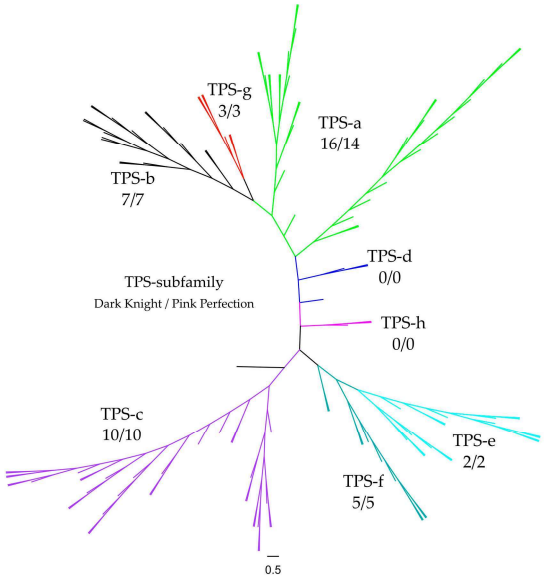


Figure 7. Phylogenetic tree of putative terpene synthases (TPS) within *Caryopteris x clandonensis* cultivars Dark Knight (DK) and Pink Perfection (PP). TPS-a (green), TPS-b (black), TPS-c (purple), TPS-d (blue), TPS-e (turquoise), TPS-f (petrol), TPS-g (red), TPS-h (pink). For phylogenetic tree construction, TPS a-h of selected plant species were included to assure correct classification of identified TPS. Numbers below the respective TPS subfamily indicate the count of predicted TPS in the genomes of the cultivars.

3. Materials and Methods

3.1. Plant Material

Four cultivars of *Caryopteris x clandonensis* were acquired from a local nursery (Foerstner Pflanzen GmbH, Bietigheim-Bissingen, Germany) and grown to maturity in the open in a warm, moderate climate zone. After maturity, healthy leaves and blossoms were sampled and snap frozen in liquid nitrogen and stored at −80 °C until preparation for transcriptome and genome sequencing. Fresh mature leaves were used for GC-MS headspace analysis of volatile compounds.

3.2. GC-MS Analysis of Volatile Compounds

Fresh mature leaves were weighed in GC headspace vials and analyzed using a Trace GC-MS Ultra system with DSQII (Thermo Scientific, Waltham, MA, USA). Vials were incubated for 30 min at 100 °C and a TriPlus autosampler was used to inject 500 µL of the sample in split mode onto a SGE BPX5 column (30 m, I.D 0.25 mm, film 0.25 µm); an injector temperature of 280 °C was used. The initial oven temperature was kept at 50 °C for 2.5 min. The temperature was increased with a ramp rate of 10 °C/min to 320 °C with a final hold for 5 min. Helium was used as a carrier gas with a flow rate of 1.2 mL/min and a split ratio of 8. The MS chromatograms and spectra were recorded at 70 eV (EI). Masses were detected between 50 m/z and 650 m/z in the positive mode [69]. Samples were measured in biological triplicates and the area average was used to compare peaks. Compounds were identified by spectral comparison with a NIST/EPA/NIH MS library version 2.0. To provide insight into the differentiation between plant samples a PCA was conducted.

3.3. High Molecular Weight DNA Extraction and Library Preparation

High molecular weight genomic DNA (HMW gDNA) suitable for long-read sequencing was achieved using a plant-optimized CTAB—PCI extraction method based on different protocols [70–72]; 1 g of frozen, unthawed plant leaves were ground using a CryoMill (Retsch, Haan, Germany; three cycles, 6 min precool at 5 Hz, disruption 2:30 min 25 Hz, cooling between cycles 0:30 min at 5 Hz). A CTAB extraction buffer (2% CTAB, 100 mM Tris pH 8.0, 20 mM EDTA, 1.4 M NaCl) was supplemented with 2% PVP prior to usage and solved at 60 °C. The unthawed fine powder was mixed with 5 ml buffer and incubated with 200 µL Proteinase K (Qiagen, Venlo, The Netherlands) for 30 min at 50 °C and occasionally inverted. At room temperature, 1 mg RNAse A (Thermo Scientific, Waltham, MA, USA) was added and incubated for 10 min. The mixture was washed twice, saving and reusing the aqueous upper phase, with one volume PCI (25:24:1) and three times with chloroform (10,000× g, 5 min, 10 °C). To pellet the HMW gDNA, 30% PEG was added to the aqueous phase (1:4), inverted, incubated for 30 min on ice and spun for 30 min at 12,000× g, 10 °C. The resulting shallow and colorless pellet was washed three times with 70% ethanol (5000× g, 5 min, 10 °C) and consequently, air dried at 40 °C and resuspended with 100 µL elution buffer (Qiagen, Venlo, The Netherlands). Quality and size of the gDNA were assessed using a Qubit dsDNA HS Kit (Thermo Scientific, Waltham, MA, USA), a Nanodrop photometer (Implen, Munich, Germany) and a Femto Pulse system (Agilent, Santa Clara, CA, USA), respectively. If variations in DNA concentration between Qubit and Nanodrop were > 50% an AMPure PB bead clean up or an electrophoretic clean up using a BluePippin system (Sage Science, Beverly, MA, USA) was performed; 5 µg HMW gDNA were sheared in a gTube (Covaris, Woburn, MA, USA; 1700× g) and used for whole genome library preparation using SMRTbell prep kit 3.0 (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's recommendations. Size selection of the resulting library was performed using AMPure PB beads. Libraries were stored at −20 °C. Prior sequencing, primer and polymerase were bound using a Sequel II Binding Kit 3.2 (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's recommendations.

3.4. Genome Sequencing and Assembly

Sequencing was performed on a Sequel IIe (Pacific Biosciences, Menlo Park, CA, USA) with two hours pre-extension, two hours adaptive loading (target p1 + p2 = 0.95) to an on-plate concentration of 85 pM, and 30 h movie time. The initial de novo genome assembly was performed using SMRT Link (v11.0.0+, Pacific Biosciences, Menlo Park, CA, USA) which uses Improved Phased Assembly (IPA) [73]. After polishing, the contigs were divided into primary and haplotype-associated contigs using purge_dups [74].

The assembled sequences can be found within the National Center for Biotechnology Information (NCBI). BioSample accession number: Dark Knight SAMN32308289, Pink Perfection SAMN32308290.

3.5. RNA Long Read IsoSeq

To increase the quality of the genome assembly, long-read transcripts were sequenced to add more depth and accuracy to the proposed gene models. For RNA extraction, frozen, unthawed leaves were ground using a CryoMill and an RNeasy Plant Mini Kit (Qiagen, Venlo, Niederlande). A Turbo DNA free Kit (Invitrogen) was used to further clean the RNA. The high-quality RNA was used to perform an IsoSeq library prep using SMRTbell prep kit 3.0 and Sequel II Binding Kit 3.2. (Pacific Biosciences, Menlo Park, CA, USA).

3.6. Bioinformatic and Statistical Analysis

Gene models were prepared through AUGUSTUS [50–53] using genomic data and long-read transcriptomic data as hints. Quality and completeness of the genome were estimated with QUAST (v5.2.0) [31] and BUSCO (v5.3.2) [39,43,75,76]. NCBI BLAST (v2.12.0+) [40,41] and InterProScan (v5.54-87) [60,61] were computed on a local computational unit. This analysis provided an annotation that was the basis for the determination of distinct protein families, in this case, terpene synthases and cytochrome p450 enzymes. EggNOG Mapper (v2.1.5) was used to determine COG and GO terms. Statistical analysis and figures were conducted using R (v4.2.1, revigo [59] and cateGOrizer [77]. Synteny analysis was performed using Mauve [47] (v2.4.0) and Geneious Prime (Geneious). For k-mer analysis jellyfish (v2.3.0) [78] was used (k-mer size: 20). GenomeScope [79,80] was used for the visualization of k-mer frequencies. The following analyses were conducted using galaxy project [81]: BUSCO, QUAST, EggNOG, Jellyfish, and GenomeScope. If not further specified default parameters were used for analysis.

3.7. Identification of TPS and Cytochrome p450 Enzymes

Genes associated with these protein classes were found using InterProScan and the domain seed files IPR036965 (TPS activity) and IPR01128 (cytochrome p450 enzymes). The phylogenetic tree was constructed using a global alignment with Blosum62. As a genetic distance model, Jukes–Cantor was chosen along with Neighbor-Joining as the Tree building method. The outlier was *Physcomitrella patens*, XP_024380398. Software used: Geneious Prime (Geneious).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/plants12030632/s1, Figure S1: Chemical structure of D-limonene backbone and difference to C6-C4 shift in α-pinene, Figure S2: Cultivars of *Caryopteris x clandonensis* used in this manuscript, Figure S3: PacBio sequencing quality reports of different *Caryopteris x clandonensis* cultivars, Figure S4: GenomeScope profile of k-mer analysis of Dark Knight, Figure S4: GenomeScope profile of k-mer analysis of Pink Perfection, Figure S6: Synteny evaluation between the *Caryopteris x clandonensis* cultivars, Table S1: GC-MS Headspace data of TOP30 identified compounds via NIST database Table S2: Data Pink Perfection COG, Table S3: Data Dark Knight COG, Table S4: Data Pink Perfection GO cluster, Table S5: Data Dark Knight GO cluster, Supplemental Lamiaceae Reference: Phylogenetic tree references in FASTA format.

Author Contributions: Conceptualization, M.R., N.A. and N.M.; methodology, M.R. and N.A.; software, M.R., N.A. and N.M.; validation, M.R. and N.A.; formal analysis, M.R. and N.A.; investigation, M.R., N.A.; resources, T.B.; data curation, M.R. and N.A.; writing—original draft preparation, M.R. and N.A.; writing—review and editing, M.R., N.A., N.M. and T.B.; visualization, M.R.; supervision, N.M. and T.B.; project administration, N.M. and T.B.; funding acquisition, T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Federal Ministry of Education and Research, grant number 031B0824A.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available in National Center for Biotechnology Information (NCBI). BioSample accession number: Dark Knight SAMN32308289, Pink Perfection SAMN32308290.

Acknowledgments: M.R., N.A., N.M. and T.B. gratefully acknowledge the support of colleagues at the Werner Siemens Chair for Synthetic Biotechnology during conducting experiments and writing this manuscript. Furthermore, all authors want to thank Foerstner Pflanzen GmbH for providing plant materials.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Caputi, L. Use of terpenoids as natural flavouring compounds in food industry. *Recent Pat. Food Nutr. Agric.* **2011**, *3*, 9–16. [CrossRef] [PubMed]
2. Masyita, A.; Sari, R.M.; Astuti, A.D.; Yasir, B.; Rumata, N.R.; Emran, T.B.; Nainu, F.; Simal-Gandara, J. Terpenes and terpenoids as main bioactive compounds of essential oils, their roles in human health and potential application as natural food preservatives. *Food Chem. X* **2022**, *13*, 100217. [CrossRef] [PubMed]
3. da Silva, G.L.; Luft, C.; Lunardelli, A.; Amaral, R.H.; Melo, D.A.D.S.; Donadio, M.V.; Nunes, F.B.; DE Azambuja, M.S.; Santana, J.C.; Moraes, C.M.; et al. Antioxidant, analgesic and anti-inflammatory effects of lavender essential oil. *An. Acad. Bras. Cienc.* **2015**, *87*, 1397–1408. [CrossRef] [PubMed]
4. Mediratta, P.; Sharma, K.; Singh, S. Evaluation of immunomodulatory potential of *Ocimum sanctum* seed oil and its possible mechanism of action. *J. Ethnopharmacol.* **2002**, *80*, 15–20. [CrossRef] [PubMed]
5. da Silva, J.K.R.; Figueiredo, P.L.B.; Byler, K.G.; Setzer, W.N. Essential oils as antiviral agents, potential of essential oils to treat SARS-CoV-2 infection: An in-silico investigation. *Int. J. Mol. Sci.* **2020**, *21*, 3426. [CrossRef] [PubMed]
6. Abdollahi, M.; Karimpour, H.; Monsef-Esfehani, H.R. Antinociceptive effects of *Teucrium polium* L. total extract and essential oil in mouse writhing test. *Pharmacol. Res.* **2003**, *48*, 31–35. [CrossRef] [PubMed]
7. Đorđević, S.; Petrović, S.; Dobrić, S.; Milenković, M.; Vučićević, D.; Žižić, S.; Kukić, J. Antimicrobial, anti-inflammatory, anti-ulcer and antioxidant activities of *Carlina acanthifolia* root essential oil. *J. Ethnopharmacol.* **2007**, *109*, 458–463. [CrossRef]
8. Cowen, D.; Wolf, A.; Paige, B.H. Toxoplasmic encephalomyelitis. *Arch. Neurol. Psychiatry* **1942**, *48*, 689–739. [CrossRef]
9. Jantan, I.; Ping, W.O.; Visuvalingam, S.D.; Ahmad, N.W. Larvicidal activity of the essential oils and methanol extracts of Malaysian plants on *Aedes aegypti*. *Pharm. Biol.* **2008**, *41*, 234–236. [CrossRef]
10. Cox-Georgian, D.; Ramadoss, N.; Dona, C.; Basu, C. Therapeutic and medicinal uses of terpenes. *Med. Plants Farm Pharm.* **2019**, *67*, 333–359. [CrossRef]
11. Sicora, O. The ethanolic stem extract of *Caryopteris* x *Clandonensis Posseses* antiproliferative potential by blocking breast cancer cells in mitosis. *Farmacia* **2019**, *67*, 1077–1082. [CrossRef]
12. Wani, M.C.; Taylor, H.L.; Wall, M.E.; Coggon, P.; Mcphail, A.T. Plant antitumor agents. VI. The isolation and structure of Taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J. Am. Chem. Soc.* **1971**, *93*, 2325–2327. [CrossRef]
13. Weaver, B.A. How Taxol/paclitaxel kills cancer cells. *Mol. Biol. Cell* **2014**, *25*, 2677–2681. [CrossRef]
14. Pichersky, E.; Raguso, R.A. Why do plants produce so many terpenoid compounds? *New Phytol.* **2018**, *220*, 692–702. [CrossRef]
15. Holopainen, J.K.; Himanen, S.J.; Yuan, J.S.; Chen, F.; Stewart, C.N. *Ecological Functions of Terpenoids in Changing Climates*; Ramawat, K., Mérillon, J.M., Eds.; Natural Products; Springer: Berlin/Heidelberg, Germany, 2013. [CrossRef]
16. Drapeau, J.; Rossano, M.; Touraud, D.; Obermayr, U.; Geier, M.; Rose, A.; Kunz, W. Green synthesis of para-Menthane-3,8-diol from *Eucalyptus citriodora*: Application for repellent products. *Comptes Rendus Chim.* **2011**, *14*, 629–635. [CrossRef]
17. Lee, S.Y.; Kim, S.H.; Hong, C.Y.; Park, S.Y.; Choi, I.G. Biotransformation of (-)-α-pinene and geraniol to α-terpineol and p-menthane-3,8-diol by the white rot fungus, *Polyporus brumalis*. *J. Microbiol.* **2017**, *53*, 462–467. [CrossRef]

18. Drapeau, J.; Verdier, M.; Touraud, D.; Kröckel, U.; Geier, M.; Rose, A.; Kunz, W. Effective insect repellent formulation in both surfactantless and classical microemulsions with a long-lasting protection for human beings. *Chem. Biodivers.* **2009**, *6*, 934–947. [CrossRef]
19. Blythe, E.; Tabanca, N.; Demirci, B.; Bernier, U.; Agramonte, N.; Ali, A.; Khan, I. Composition of the essential oil of Pink Chablis™ bluebeard (*Caryopteris* × *clandonensis* 'Durio') and its biological activity against the yellow fever mosquito *Aedes aegypti*. *Nat. Volatiles Essent. Oils* **2015**, *2*, 11–21.
20. Bathe, U.; Tissier, A. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. *Phytochemistry* **2019**, *161*, 149–162. [CrossRef]
21. Hernandez-Ortega, A.; Vinaixa, M.; Zebec, Z.; Takano, E.; Scrutton, N.S. A toolbox for diverse Oxyfunctionalisation of monoterpenes. *Sci. Rep.* **2018**, *8*, 1–8. [CrossRef]
22. Mabou, F.D.; Belinda, I.; Yossa, N. TERPENES: Structural classification and biological activities. *IOSR J. Pharm. Biol. Sci. e-ISSN* **2021**, *16*, 2319–7676.
23. Nett, R.S.; Montanares, M.; Marcassa, A.; Lu, X.; Nagel, R.; Charles, T.C.; Hedden, P.; Rojas, M.C.; Peters, R.J. Elucidation of gibberellin biosynthesis in bacteria reveals convergent evolution. *Nat. Chem. Biol.* **2016**, *13*, 69–74. [CrossRef] [PubMed]
24. Wang, T.; Li, L.; Zhuang, W.; Zhang, F.; Shu, X.; Wang, N.; Wang, Z. Recent research progress in Taxol biosynthetic pathway and acylation reactions mediated by *Taxus Acyltransferases*. *Molecules* **2021**, *26*, 2855. [CrossRef] [PubMed]
25. Wen, W.; Yu, R. Artemisinin biosynthesis and its regulatory enzymes: Progress and perspective. *Pharmacogn. Rev.* **2011**, *5*, 189–194. [CrossRef]
26. Gershenzon, J.; Dudareva, N. The function of terpene natural products in the natural world. *Nat. Chem. Biol.* **2007**, *3*, 408–414. [CrossRef]
27. Zhang, X.; Niu, M.; da Silva, J.A.T.; Zhang, Y.; Yuan, Y.; Jia, Y.; Xiao, Y.; Li, Y.; Fang, L.; Zeng, S.; 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.* **2019**, *19*, 115. [CrossRef]
28. Sharma, V.; Sarkar, I.N. Bioinformatics opportunities for identification and study of medicinal plants. *Brief. Bioinform.* **2013**, *14*, 238–250. [CrossRef]
29. Helfrich, E.J.N.; Lin, G.-M.; Voigt, C.A.; Clardy, J. Bacterial terpene biosynthesis: Challenges and opportunities for pathway engineering. *Beilstein J. Org. Chem.* **2019**, *15*, 2889–2906. [CrossRef]
30. Amarasinghe, S.L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020**, *21*, 1–16. [CrossRef]
31. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
32. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
33. Chen, G.; Mostafa, S.; Lu, Z.; Du, R.; Cui, J.; Wang, Y.; Liao, Q.; Lu, J.; Mao, X.; Chang, B.; et al. The Jasmine (*Jasminum sambac*) genome provides insight into the biosynthesis of flower fragrances and Jasmonates. *Genom. Proteom. Bioinform.* **2022**, *in press*. [CrossRef]
34. Degenhardt, J.; Köllner, T.G.; Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **2009**, *70*, 1621–1637. [CrossRef]
35. Zhu, X.; Li, Q.; Li, J.; Luo, J.; Chen, W.; Li, X. Comparative study of volatile compounds in the fruit of two banana cultivars at different ripening stages. *Molecules* **2018**, *23*, 2456. [CrossRef]
36. Cramer, A.-C.J.; Mattinson, D.S.; Fellman, J.K.; Baik, B.-K. Analysis of volatile compounds from various types of barley cultivars. *J. Agric. Food Chem.* **2005**, *53*, 7526–7531. [CrossRef]
37. Dong, A.-X.; Xin, H.-B.; Li, Z.-J.; Liu, H.; Sun, Y.-Q.; Nie, S.; Zhao, Z.-N.; Cui, R.-F.; Zhang, R.-G.; Yun, Q.-Z.; et al. High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *Gigascience* **2018**, *7*, giy068. [CrossRef]
38. Li, C.; Li, X.; Liu, H.; Wang, X.; Li, W.; Chen, M.-S.; Niu, L.-J. Chromatin architectures are associated with response to dark treatment in the oil crop *Sesamum indicum*, based on a high-quality genome assembly. *Plant Cell Physiol.* **2020**, *61*, 978–987. [CrossRef]
39. Manni, M.; Berkeley, M.R.; Seppey, M.; A Simão, F.; Zdobnov, E.M. BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **2021**, *38*, 4647–4654. [CrossRef]
40. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
41. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
42. Guan, D.; A McCarthy, S.; Wood, J.; Howe, K.; Wang, Y.; Durbin, R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **2020**, *36*, 2896–2898. [CrossRef] [PubMed]
43. Manni, M.; Berkeley, M.R.; Seppey, M.; Zdobnov, E.M. BUSCO: Assessing genomic data quality and beyond. *Curr. Protoc.* **2021**, *1*, e323. [CrossRef] [PubMed]

44. Tang, H.; Lyons, E.; Pedersen, B.; Schnable, J.C.; Paterson, A.H.; Freeling, M. Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinform.* **2011**, *12*, 102. [CrossRef] [PubMed]
45. Lee, J.; Hong, W.-Y.; Cho, M.; Sim, M.; Lee, D.; Ko, Y.; Kim, J. Synteny Portal: A web-based application portal for synteny block analysis. *Nucleic Acids Res.* **2016**, *44*, W35–W40. [CrossRef] [PubMed]
46. Liu, D.; Hunt, M.; Tsai, I.J. Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinform.* **2018**, *19*, 1–13. [CrossRef]
47. Darling, A.E.; Mau, B.; Perna, N.T. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **2010**, *5*, e11147. [CrossRef]
48. Arús, P.; Toshiya, Y.; Elisabeth, D.; Abbott, A.G. Synteny in the Rosaceae. In *Plant Breeding Reviews*; John and Wiley and Sons: Hoboken, NJ, USA, 2010; pp. 175–211.
49. Devos, K.M.; Moore, G.; Gale, M.D. Conservation of marker synteny during evolution. *Euphytica* **1995**, *85*, 67–372. [CrossRef]
50. Hoff, K.J.; Lomsadze, A.; Borodovsky, M.; Stanke, M. Whole-genome annotation with BRAKER. *Methods Mol. Biol.* **2019**, *1962*, 65–95. [CrossRef]
51. Hoff, K.J.; Lange, S.; Lomsadze, A.; Borodovsky, M.; Stanke, M. BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS: Table 1. *Bioinformatics* **2016**, *32*, 767–769. [CrossRef]
52. Brůna, T.; Hoff, K.J.; Lomsadze, A.; Stanke, M.; Borodovsky, M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **2021**, *3*, lqaa108. [CrossRef]
53. Stanke, M.; Schöffmann, O.; Morgenstern, B.; Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinform.* **2006**, *7*, 62. [CrossRef]
54. Yang, Z.; Ge, X.; Yang, Z.; Qin, W.; Sun, G.; Wang, Z.; Li, Z.; Liu, J.; Wu, J.; Wang, Y.; et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* **2019**, *10*, 2989. [CrossRef]
55. Liu, S.; An, Y.; Tong, W.; Qin, X.; Samarina, L.; Guo, R.; Xia, X.; Wei, C. Characterization of genome-wide genetic variations between two varieties of tea plant (*Camellia sinensis*) and development of InDel markers for genetic research. *BMC Genom.* **2019**, *20*, 1–16. [CrossRef]
56. Chatterjee, N.; Walker, G.C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.* **2017**, *58*, 235–263. [CrossRef] [PubMed]
57. Raina, A.; Sahu, P.K.; Laskar, R.A.; Rajora, N.; Sao, R.; Khan, S.; Ganai, R.A. Mechanisms of genome maintenance in plants: Playing it safe with breaks and bumps. *Front. Genet.* **2021**, *12*, 675686. [CrossRef]
58. Lim, C.; Pratama, M.Y.; Rivera, C.; Silvestro, M.; Tsao, P.S.; Maegdefessel, L.; Gallagher, K.A.; Maldonado, T.; Ramkhelawon, B. Linking single nucleotide polymorphisms to signaling blueprints in abdominal aortic aneurysms. *Sci. Rep.* **2022**, *12*, 20990. [CrossRef]
59. Supek, F.; Bošnjak, M.; Škunca, N.; Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **2011**, *6*, e21800. [CrossRef]
60. Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **2021**, *49*, D344–D354. [CrossRef]
61. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef]
62. Butler, J.B.; Freeman, J.; Potts, B.M.; Vaillancourt, R.; Grattapaglia, D.; Silva-Junior, O.B.; Simmons, B.; Healey, A.L.; Schmutz, J.; Barry, K.; et al. Annotation of the *Corymbia terpene* synthase gene family shows broad conservation but dynamic evolution of physical clusters relative to *Eucalyptus*. *Heredity* **2018**, *121*, 87–104. [CrossRef]
63. Warren, R.L.; Keeling, C.I.; Yuen, M.M.S.; Raymond, A.; Taylor, G.A.; Vandervalk, B.P.; Mohamadi, H.; Paulino, D.; Chiu, R.; Jackman, S.D.; et al. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J.* **2015**, *83*, 189–212. [CrossRef] [PubMed]
64. Jia, K.-H.; Liu, H.; Zhang, R.-G.; Xu, J.; Zhou, S.-S.; Jiao, S.-Q.; Yan, X.-M.; Tian, X.-C.; Shi, T.-L.; Luo, H.; et al. Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (*Lamiaceae*) genome. *Hortic. Res.* **2021**, *8*, 1–15. [CrossRef] [PubMed]
65. Chen, Z.; Vining, K.J.; Qi, X.; Yu, X.; Zheng, Y.; Liu, Z.; Fang, H.; Li, L.; Bai, Y.; Liang, C.; et al. Genome-wide analysis of terpene synthase gene family in *Mentha longifolia* and catalytic activity analysis of a single terpene synthase. *Genes* **2021**, *12*, 518. [CrossRef] [PubMed]
66. Hamilton, J.P.; Godden, G.T.; Lanier, E.; Bhat, W.W.; Kinser, T.J.; Vaillancourt, B.; Wang, H.; Wood, J.C.; Jiang, J.; Soltis, P.S.; et al. Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing *Lamiaceae* species, *Callicarpa americana*. *Gigascience* **2020**, *9*, giaa093. [CrossRef] [PubMed]
67. 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.* **2011**, *66*, 212–229. [CrossRef]
68. Shalev, T.J.; Yuen, M.M.S.; Gesell, A.; Yuen, A.; Russell, J.H.; Bohlmann, J. An annotated transcriptome of highly inbred Thuja plicata (*Cupressaceae*) and its utility for gene discovery of terpenoid biosynthesis and conifer defense. *Tree Genet. Genomes* **2018**, *14*, 35. [CrossRef]
69. Ringel, M.; Reinbold, M.; Hirte, M.; Haack, M.; Huber, C.; Eisenreich, W.; Masri, M.A.; Schenk, G.; Guddat, L.W.; Loll, B.; et al. Towards a sustainable generation of pseudopterosin-type bioactives. *Green Chem.* **2020**, *22*, 6033–6046. [CrossRef]

70. Inglis, P.W.; Pappas, M.; Resende, L.V.; Grattapaglia, D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLoS ONE* **2018**, *13*, e0206085. [CrossRef]
71. Healey, A.; Furtado, A.; Cooper, T.; Henry, R.J. Protocol: A simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **2014**, *10*, 21. [CrossRef]
72. Rogers, S.O.; Bendich, A.J. Extraction of total cellular DNA from plants, algae and fungi. *Plant Mol. Biol. Man.* **1994**, *2*, 183–190. [CrossRef]
73. GitHub—PacificBiosciences/pbipa: Improved Phased Assembler. Available online: https://github.com/PacificBiosciences/pbipa (accessed on 11 December 2022).
74. GitHub—dfguan/purge_dups: Haplotypic Duplication Identification Tool. Available online: https://github.com/dfguan/purge_dups (accessed on 11 December 2022).
75. Kriventseva, E.V.; Kuznetsov, D.; Tegenfeldt, F.; Manni, M.; Dias, R.; A Simão, F.; Zdobnov, E.M. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **2019**, *47*, D807–D811. [CrossRef] [PubMed]
76. Seppey, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **2019**, *1962*, 227–245. [CrossRef] [PubMed]
77. Hu, Z.-L.; Bao, J.; Reecy, J. CateGOrizer: A web-based program to batch analyze gene ontology classification categories. *Online J. Bioinform.* **2008**, *9*, 108–112. Available online: http://www.animalgenome.org/bioinfo/tools/catego/ (accessed on 11 December 2022).
78. Marçais, G.; Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **2011**, *27*, 764–770. [CrossRef]
79. Vurture, G.W.; Sedlazeck, F.J.; Nattestad, M.; Underwood, C.J.; Fang, H.; Gurtowski, J.; Schatz, M.C. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **2017**, *33*, 2202–2204. [CrossRef]
80. Ranallo-Benavidez, T.R.; Jaron, K.S.; Schatz, M.C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **2020**, *11*, 1432. [CrossRef]
81. Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W537–W544. [CrossRef]

Differential RNA-Seq Analysis Predicts Genes Related to Terpene Tailoring in *Caryopteris x Clandonensis*

plants

IMPACT FACTOR 4.658

Indexed in: PubMed

Article

Differential RNA-Seq Analysis Predicts Genes Related to Terpene Tailoring in *Caryopteris* × *clandonensis*

Manfred Ritz, Nadim Ahmad, Thomas Brueck and Norbert Mehlmer

Special Issue

Recent Advances in Plant Genomics and Transcriptome Analysis

Edited by

Dr. Nam-Soo Kim

MDPI

https://doi.org/10.3390/plants12122305

plants

MDPI

Article

Differential RNA-Seq Analysis Predicts Genes Related to Terpene Tailoring in *Caryopteris* × *clandonensis*

Manfred Ritz †, Nadim Ahmad †, Thomas Brueck * and Norbert Mehlmer *

Werner Siemens Chair of Synthetic Biotechnology, Department of Chemistry, Technical University of Munich (TUM), 85748 Garching, Germany; manfred.ritz@tum.de (M.R.); nadim.ahmad@tum.de (N.A.)
* Correspondence: brueck@tum.de (T.B.); norbert.mehlmer@tum.de (N.M.)
† These authors contributed equally to this work.

Abstract: Enzymatic terpene functionalization is an essential part of plant secondary metabolite diversity. Within this, multiple terpene-modifying enzymes are required to enable the chemical diversity of volatile compounds essential in plant communication and defense. This work sheds light on the differentially transcribed genes within *Caryopteris* × *clandonensis* that are capable of functionalizing cyclic terpene scaffolds, which are the product of terpene cyclase action. The available genomic reference was subjected to further improvements to provide a comprehensive basis, where the number of contigs was minimized. RNA-Seq data of six cultivars, Dark Knight, Grand Bleu, Good as Gold, Hint of Gold, Pink Perfection, and Sunny Blue, were mapped on the reference, and their distinct transcription profile investigated. Within this data resource, we detected interesting variations and additionally genes with high and low transcript abundancies in leaves of *Caryopteris* × *clandonensis* related to terpene functionalization. As previously described, different cultivars vary in their modification of monoterpenes, especially limonene, resulting in different limonene-derived molecules. This study focuses on predicting the cytochrome p450 enzymes underlying this varied transcription pattern between investigated samples. Thus, making them a reasonable explanation for terpenoid differences between these plants. Furthermore, these data provide the basis for functional assays and the verification of putative enzyme activities.

Keywords: terpene biosynthesis; cytochrome p450; *Caryopteris* × *clandonensis*; long read sequencing; transcriptomics; chemical diversity; volatile compound

Citation: Ritz, M.; Ahmad, N.; Brueck, T.; Mehlmer, N. Differential RNA-Seq Analysis Predicts Genes Related to Terpene Tailoring in *Caryopteris* × *clandonensis*. *Plants* **2023**, *12*, 2305. https://doi.org/10.3390/plants12122305

Academic Editors: Andreas W. Ebert and Nam-Soo Kim

Received: 25 April 2023
Revised: 17 May 2023
Accepted: 7 June 2023
Published: 13 June 2023

1. Introduction

Caryopteris × *clandonensis* is an ornamental plant, also known as "Bluebeard", which is phylogenetically classified in the *Lamiaceae* family. It is easily cultivated and rich in volatile compounds. These, and other molecules detected and described, are terpenes, e.g., α-copaene, limonene, or δ-cadinene [1], terpene derivates, e.g., keto-glycosides, clandonosides, and harpagides [2], as well as the pyrano-juglon derivate α-caryopteron [3]. The species' essential oil was found to display mosquito-repellent activity; however, the active agent for this mode of action was not yet detected [4]. The *Lamiaceae* family is known to harbor an interesting and valuable profile in secondary metabolites, including terpenoids, flavonoids, and phenylpropanoids [5–7]. These compounds play important roles in the plant's interaction with its environment [8,9] as for the defense against abiotic and biotic stresses [10]. They also harbor potential in pharmaceutical or industrial applications, as seen for taxol [11], menthol [12], malvidin [13], isoliquiritigenin [14] or umbelliferone [15]. In general, terpenes and terpenoids are a molecule class, which is produced in vast varieties by flowering plants [16] and is involved in a wide range of biological activities. Essential oils and their monoterpenes, such as α-pinene and limonene, were investigated in terms of their anti-inflammatory and virucidal activity in recent studies [17–19]. Moreover, other terpenoids employ antibacterial properties [20] while others act as insecticides [4], are used

as allelochemicals [21], or as attractants for pollinators [22]. The backbone of plant-derived terpenes is produced via the mevalonate pathway. For this, the precursors dimethylallyl diphosphate (DMAPP) and the functional isomer isopentenyl pyrophosphate (IPP) can be connected via isoprenyl diphosphate synthases (IDS) to form larger units of terpenes. IPP consists of five C-atoms (hemiterpene) whereas, through condensation of IPP and DMAPP via IDS monoterpenes (C10), sesquiterpenes (C15), diterpenes (C20), and higher terpene structures are built [23]. Further tailoring of these basic terpenes is conducted by terpene synthases (TPS) and cytochrome p450 enzymes (CYPs). Plant TPS mediate complex carbocation reactions, resulting in various cyclic structures of higher terpenes [24,25]. These can be divided into eight subfamilies (TPSa-h) which can be clade- or even species-specific [26]. The first step in tailoring monoterpenes is hydroxylation. Subsequently, CYPs are mediating a plethora of further reactions to enhance the functionalization (carboxylation, acetylation or forming peroxides) [27,28]. Due to their promiscuity towards substrates, only a few enzymes are necessary to yield various terpenoid structures and, therefore, differences in their functions and modality [29]. Multiple sequences of different source organisms are available in curated databases [30,31]. These allow easy access to the genetic information on these enzymes. With CYPs occurring in all living organisms [30], the enzymes, similarly to TPS, are divided for better identification, whereas specific CYP families are reserved for each type of organism. Plant CYP families can be found in CYP71-99 and CYP701-999, and in a four-digit scheme from CYP7001-9999 [32]. The categorization into these classes is dependent on sequence similarity. The same family (Arabic number) needs matching amino acids $\geq$40% and the subfamily (Arabic letter) $\geq$55% [33]. Therefore, the CYP76S40 [34] is the 40th individual enzyme from the CYP76S subfamily and the CYP76 family. This way, after annotation, contaminating sequences can be discarded solely due to their classification in a non-plant CYP family.

One approach to elucidate variations in the enzymatic makeup and investigate the sequences underlying terpene diversity is to compare differentially expressed gene (DEG) products at a quantitative level using modern bioinformatics tools. Differences in the metabolite profile exist during different stages of plant growth [35]. Different genes are regulated from seedlings to mature plants to translate their genomic information into proteins and interact in plant differentiation, protection or communication, depending on their developmental state [36]. During plant breeding, deletion, duplications, mutations or fragmentations can occur. Therefore, a distinct set of genes varies in its nucleotide code and their transcription or translation rate, resulting in different phenotypes in the mature plant [37]. The data can be levied and evaluated regarding efficacy to investigate these differences. The number of transcripts does not solely result in higher protein outcome, but also in, respectively, higher concentrations of secondary metabolites. Therefore, differential expression analysis can identify genes or gene products responsible for either the stress response mechanisms observed for abiotic stressors, such as drought or radiation, or as has been shown for biotic stressors, such as pests and plant reactions to herbivores [38]. Typical DEG experiments harness the up- and down-regulation of genes after induction or shock, e.g., during exposure to chemicals [39] or different environments [40]. Another possibility is the investigation of specific traits of plant cultivars due to their variations between hybrid plants [41]. Previously, the variations in *Caryopteris* × *clandonensis*' volatile compound setup was investigated, and a difference in the synthesis of limonene-derived molecules (LDM) was observed [1]. The cultivar Dark Knight was detected to harbor a low amount, whereas Pink Perfection shows high amounts of LDM. These variations were discovered without a distinct change in their TPS or CYP makeup.

To that end, we show that the identification of terpene variety between different plant cultivars can be pursued on a molecular level using a quantitative bioinformatics method such as RNA-Seq analysis. Furthermore, we focus on terpene functionalization enzymes, especially cytochrome p450 enzymes, to elucidate the mechanisms behind the variations in monoterpene modifications as seen for limonene [1].

2. Results and Discussion

2.1. RNA Sequencing and Mapping Quality

Samples subjected to short-read sequencing were taken from leaves of six *Caryopteris* × *clandonensis* cultivars known to show differences in their LDM profile, Dark Knight (DK), Grand Bleu (GB), Good as Gold (GG), Hint of Gold (HG), Sunny Blue (SB), and Pink Perfection (PP). Sequencing was performed using an Illumina NovaSeq platform, which generated about 20 million raw reads in bases for each sample. The reads were processed to remove low-quality reads, bases, and adapter sequences, resulting in the clean reads used for downstream analysis. After this purification step, a loss of 5.0 to 14.9 million bases was seen between the samples. In Table 1, the run as well as cleaning and mapping statistics are summarized. The Q20 and Q30 scores indicate the sequencing quality, with Q30 indicating a lower error rate than Q20. This experiment's high Q20 and Q30 scores suggest that the sequencing quality was highly sufficient, with only a few sequencing errors. Moreover, the clean reads exhibit a slight increase in quality scores, persistent throughout all samples.

Table 1. Statistics of short Illumina reads used for mapping on the reference genome (NCBI SAMN32308290 (Pink Perfection, PP)). A paired-end run was employed on a NovaSeq6000 SP (2 × 150 bp) for sequencing.

Caryopteris × *clandonensis* Cultivar		Raw Reads in Bases		Q20 in %	Q30 in %	Clean Reads in Bases		Q20 in %	Q30 in %	Totally Mapped in %	Uniquely Mapped in %
		Unique	Duplicate			Unique	Duplicate				
Dark Knight	R1	24,501,785	19,238,555	99.95	94.76	13,072,273	29,945,355	99.99	95.08	87.8	79.3
	R2	26,380,719	17,359,621	99.25	87.90	16,204,470	26,813,158	99.46	88.25		
Grand Bleu	R1	17,917,215	51,971,129	99.85	93.75	11,552,426	57,659,626	99.98	94.21	85.8	76.8
	R2	18,808,258	51,080,086	99.51	92.12	13,260,446	55,951,606	99.68	92.42		
Good as Gold	R1	22,797,327	27,074,692	99.60	94.52	13,160,322	31,359,084	99.95	95.08	86.7	75.4
	R2	25,142,438	24,729,581	99.35	89.41	16,112,061	28,407,345	99.54	89.76		
Hint of Gold	R1	20,547,645	20,953,229	99.89	94.67	15,165,071	33,935,373	99.98	95.06	86.4	77.2
	R2	23,044,700	18,456,174	99.38	88.40	18,084,814	31,015,630	99.56	88.81		
Sunny Blue	R1	20,535,582	25,022,034	99.96	94.96	13,908,152	27,181,140	99.99	95.35	87.0	80.5
	R2	22,771,085	22,786,531	99.39	88.30	16,573,745	24,515,547	99.56	88.60		
Pink Perfection	R1	25,751,312	28,610,858	99.96	94.23	12,295,625	26,046,846	99.99	94.60	87.7	82.0
	R2	29,512,685	24,849,485	99.40	90.42	14,649,539	23,692,932	99.58	90.70		

The available genome sequences from *Caryopteris* × *clandonensis* PP [1] were subjected to further cleaning and improvement steps to curb the influence of contamination. A binning algorithm, MetaBAT2 [42], usually used for metagenomic data, was used on the long-read assembly of the genome and differentiated into 40 bins. The completeness and contiguity were checked and, in summary, the 782 scaffolds/848 contigs, which add up to 344 Mb with a genome completeness score of 96.8%, were reduced to 53 scaffolds/88 contigs, which add up to 298 Mb and a BUSCO score of 96.5%. The utilized BUSCO gene sets belonged to the closest affiliate *Eudicotidae*. Detailed information can be found in Table S1. This refined genome was used as a reference for mapping the short-read sequences. A preliminary mapping of DK transcripts on the respective long-read genomic data, compared to mapping the transcripts on the PP genomic data, revealed an increased assignment of unique reads. Thus, the genome of *Caryopteris* × *clandonensis* PP was chosen as a mapping reference for both cultivars, DK and PP, resulting in a more comprehensive downstream analysis. The exact mapping counts for the different methods can be found in Table S2.

The percentages of reads mapped to the reference genome, as seen in Table 1, indicate the data accuracy and low presence of contaminating DNA. The amount of uniquely mapped reads is also an important metric, as it indicates the proportion of reads that map to a unique location in the reference genome. A high percentage of uniquely mapped reads (greater than 70%) is desirable, reducing the possibility of mapping errors or ambiguous mapping locations [43]. In our setting, we were able to accurately map between 85.8% and 87.8% of the sequences, indicating that a large proportion of the reads were successfully located on the provided genome. Furthermore, the percentage of uniquely mapped reads ranged from 75.4% to 82.0%, which is reasonably high and suggests that the quality of the sequencing reads was sufficient to allow for exact mapping and is suited for downstream

analysis. The observed duplication rates varied between 5.7% and 11.3%, and are well-known in plant transcript mapping due to transcript isoforms [44].

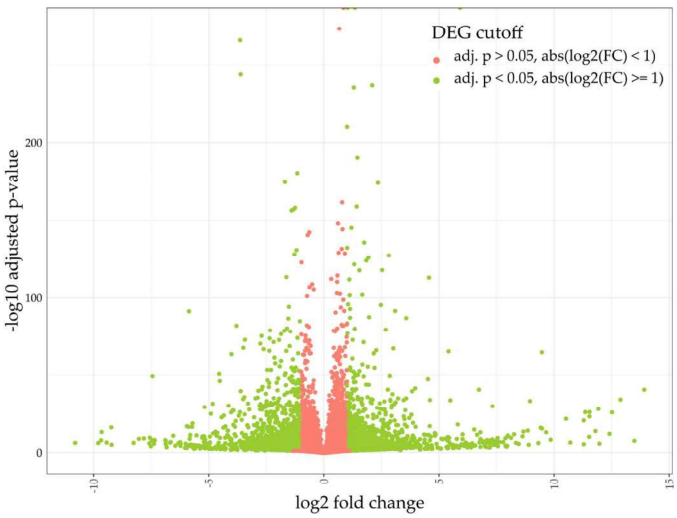
2.2. Identification of DEG

To identify the mechanism behind the modification of LDM, we wanted to focus on the DEGs between the cultivars of *Caryopteris* × *clandonensis*. Therefore, the mapping data were subset and pooled into highly LDM-positive (SB, PP) and highly LDM-negative (DK, GB) cultivars. The cultivars GG and HG were neither highly LDM-positive nor highly LDM-negative, therefore both were disregarded during the initial DEG analysis. From the 29,210 predicted genes in the mapping reference, 23,477 were observed to map in all investigated sets. The DEGs were filtered using a log2 fold-change cutoff of absolute values greater than 1, and an adjusted *p*-value of a minimum of 0.05, thereby the values for each cultivar were transcribed at least two-fold. The values fitting these parameters are highlighted in green; those which were disregarded during further analysis, because of not fitting the parameters, are shown in red. Compared to the genes close to the middle, there are a few genes with high fold-changes in LDM-positive plants, compared to LDM-negative and those with significantly higher or lower transcript abundance. After filtering the DEGs between LDM-positive and LDM-negative cultivars, 3305 genes were identified, as seen in Figure 1A. For 100 genes, no Pfam class [45] and, for a further 168, no EggNOG [46] description, could be assigned. Regarding the DEGs, a closer look reveals the 20 most diverged genes, which can be seen in Figure 1B,C. Half of the annotated genes are still uncharacterized, or their distinct function is unknown, according to the cluster of orthologous groups. Interestingly, the genes associated with metal transport and metal binding are differentially transcribed, as seen for g4372, g9694, and g8497. These functions are known to be responsible for catalyzing redox reactions in plants [47,48]. Examining DEGs further, g14432 is associated with the protein argonaute family and g1887 is a zinc finger-like protein, whereas g3464 is a thioredoxin/disulfide isomerase. These proteins regulate biological processes [49], as well as responses to abiotic stresses such as drought stress [50,51]. In general, these DEGs describe the effects on the primary metabolism and stress response of plants; however, they do not show any direct participation in tailoring secondary metabolites within the plants. CYPs, in particular, are iron-binding; however, a connection between the upregulation of metal-transporting proteins and CYPs cannot be drawn from this data. The biosynthesis of LDM is not artificially induced in one cultivar or silenced in the other. Thus, a specific and significant transcription of related terpene-tailoring genes cannot be observed. To elucidate these mechanisms, it is necessary to take a closer look into the DEGs of CYPs [28,52].

2.3. Terpene Tailoring through CYPs between Plant Cultivars

The identified 3305 DEGs can be further filtered into genes related to CYPs due to conserved domains and the corresponding CYP Pfam class. Here, the domain PF00067 was integrated into IPR001128. Both domains are indicators for sequences associated with the cytochrome p450 superfamily (IPR036396) [53]. This homology-based search allowed the identification of 70 putative sequences with different total lengths. Assuming a minimum size of 29 kDa for a CYP, 61 genes remain. From a statistical point of view, the average size of this pool amounts to a median of 1485 nucleotides, corresponding to the average size of a translated protein of 54.5 kDa. This is also reported in the literature, with an average plant CYP molecular mass between 45 and 62 kDa [54,55]. In regards to the identification of LDM-modifying enzymes, this subset is necessary to obtain a detailed overview into CYPs. These enzymes are known to play a huge part in terpene diversity in plants [56]. They are able to catalyze the hydroxylation of different backbones due to their substrate promiscuity [29,57]. Therefore, the transcript abundance of specific CYPs may reveal the mechanism behind LDM variances in this plant.

A



B

Significant gene	BLAST hit	Per. Ident	Accession
g648	Dihydropyrimidine dehydrogenase	92.5	XP_011073145.1
g3464	Thioredoxin/protein disulfide isomerase	91.5	PIN09134.1
g3715	Uncharacterized protein	-	-
g4372	Probable metal-nicotianamine transporter YSL7	85.9	XP_011072072.1
g4404	Uncharacterized protein	-	-
g7016	Histone acetyltransferase	70.6	PIN11203.1
g8497	Calcineurin-like metallo-phosphoesterase	86.0	KAH6806096.1
g10387	Methylcrotonoyl-CoA carboxylase beta chain	62.3	GEW45479.1
g10571	Class I glutamine amidotransferase-like	89.3	KAH6773205.1
g10574	Vacuolar sorting protein VPS1, dynamin	78.0	PIN18656.1
g9694	Probable metal-nicotianamine transporter YSL7	86.1	CAA2977926.1
g11688	Uncharacterized protein	-	-
g11801	Glutathione transferase gst 23	77.0	GFP83366.1
g14432	Protein argonaute 2	66.7	GFQ05617.1
g14500	Berberine bridge enzyme-like 18	72.0	XP_011078509.1
g18873	Cys-trp domain-cont. zinc finger protein 7-like	68.0	XP_047972904.1
g20543	Auxin response factor 18 isoform X2	84.0	XP_011093958.1
g21268	Uncharacterized protein	-	-
g28042	Xanthine/uracil permease family protein	91.3	KAH6834545.1
g29181	Uncharacterized protein	-	-

C

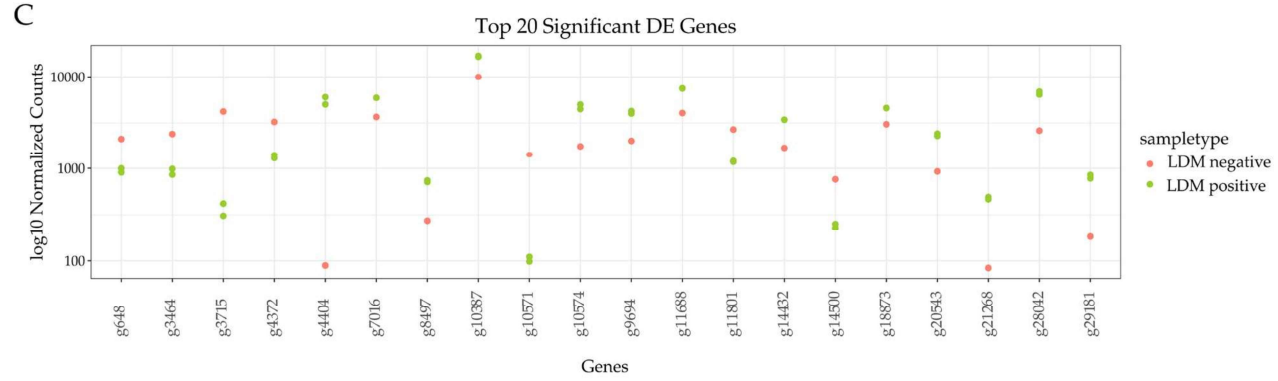


Figure 1. Differential expressed genes (DEGs) of *Caryopteris* × *clandonensis* cultivars highly producing limonene-derived molecules (LDM-positive) and cultivars which produce lower amounts of LDM (LDM-negative). Cultivars used for LDM-positive subset: Sunny Blue and Pink Perfection, and for LDM-negative subset Dark Knight and Grand Bleu. (**A**) The volcano plot of DEG was identified between the LDM-positive vs. LDM-negative plant cultivar subsets. Absolute log2 fold-change cutoff was set to 1 and an adjusted *p*-value of 0.05 was used to assign the DEGs; values fitting these parameters are highlighted in green and those which were disregarded during further analysis are shown in red. (**B**) Top 20 most significantly transcribed genes and their respective description, including BLAST search percentage identity and determined accession for the putative assignment. (**C**) log10 normalized counts of the top 20 significant DEG in this setup. Genes from LDM-positive samples are displayed in green, those corresponding to LDM-negative samples are highlighted in red.

Out of all the 23,477 mapped genes, 221 CYPs were detected, whereas 61 showed differences in transcript abundance. In Figure 2, all identified CYPs are visualized in an unrooted phylogenetic tree. CYPs with high transcript abundance in LDM-positive cultivars are highlighted in green, whereas CYPs with low transcript abundance are represented in red.

To allocate the putative CYPs to their distinct family or subfamily, the Pfam-classified CYP sequences were subjected to a BLAST search using a custom CYP database [54]. The sequences were assigned to the same subfamily if the percent identity was above 55%, and to the same family if greater than 40%. Eight CYP clans were highlighted within the found enzymes, CLAN51, CLAN71, CLAN72, CLAN74, CLAN710, CLAN85, CLAN86, and CLAN 97. This highlights that the major classes 71 and 72 are found to be involved, primarily, in the terpene tailoring of different terpene classes [28]. For CYP71, a variety of monoterpene modifications are described [34,58–61]. In our setting, most DEGs were observed in this clan. The enzymes related to CYP72 are described as tailoring triterpenoids

as saponins, characterized within plant defensive mechanisms against biotic stressors such as herbivores or microbes [38,62].

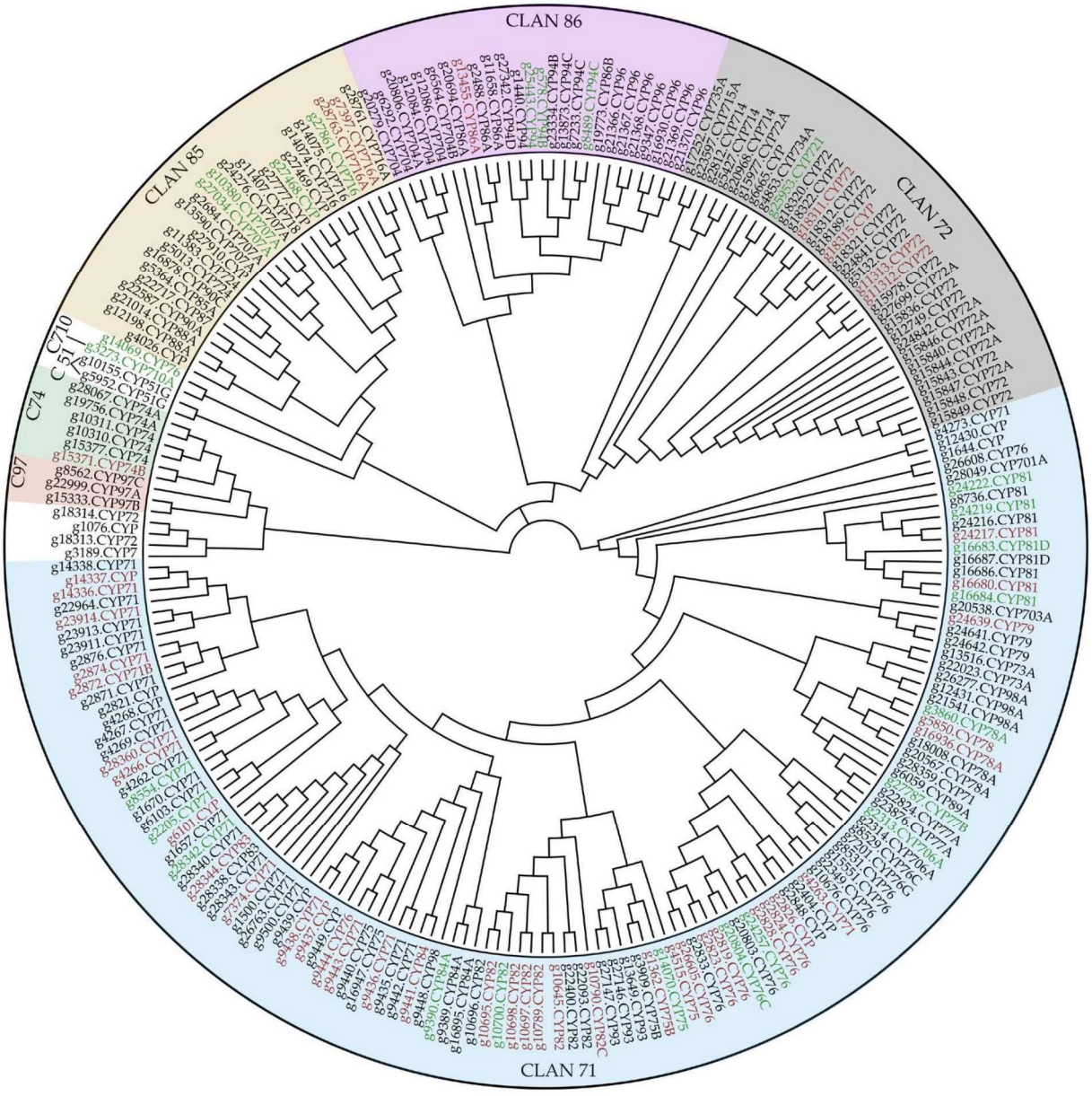


Figure 2. Phylogenetic tree of all transcribed cytochrome p450 (CYP) enzymes within the six investigated cultivars. Clan localization is highlighted on the outer ring. Differentially expressed genes (DEGs) were marked in green for a high transcript abundance in limonene-derived-molecules-positive cultivars and red for low transcript abundance, as seen in their fold-change differences. The tree was constructed using the following parameters: Global alignment with a Blosum62 cost matrix, Genetic distance model Jukes-Cantor, Neighbor-Joining and no outgroup was used, gap open penalty was set to 12, and gap extension penalty to 3 during pairwise alignments.

DEGs with high transcript abundance in LDM-positive samples were used to compare the genes between all sequenced cultivars. PP and SB were considered highly LDM-positive, whereas DK and GB were LDM-negative. GG and HG were in between and, therefore, were excluded in the initial DEG analysis. For the comparison of CYPs between the four previously mentioned samples and the two latter samples, the CYPs found in LDM-positive and LDM-negative samples were searched in GG and HG, and the normalized counts of all samples were compared. PP was chosen, due to its LDM profile, as a setpoint to compare

the transcript abundance between all samples. In, the results of a comparative approach are visualized. The phylogenetic distance between the identified CYPs is shown in 3A. Three clusters can be differentiated, with the first seen in the upper part consisting of 4 genes (g25953, g25443, g578, g8489), the second in the middle (g3273, g10380, g27034, g27468, g27861), and the third cluster with 14 genes (g24222, g2313, g27787, g24257, g20804, g3860, g10700, g16684, g24219, g9390, g14070, g28342, g8554, g2205) at the bottom. In Section 3B, the fold-change between the cultivars is visualized; boxes marked with X were transcripts with no mapping results in the respective cultivar. The clusters do not share a similar transcript abundance pattern, nor do the genes that are closely related. However, investigating the recurring, fixed-length patterns inside the sequences led to the discovery of five motifs shared among all sequences. Figure 3C visualizes the motifs and their distribution in the sequence. The exact motif sequences are presented in Table S3. A closer look also reveals distinct recurring, CYP-specific domains [63]. The conserved regions were reviewed extensively [38] and can be confirmed in this dataset. Starting with the proline-rich membrane hinge (motif 8), which is part of the membrane anchor, another conserved motif, which is important for the correct function of CYPs, is the site for oxygen binding and activation, A/G-G-X-E/D-T-T/S (motif 3). This is followed by the E-R-R triad and P(E)R(F) domain. Furthermore, the heme-binding site, with cysteine as the main ligand to the heme, C-X-G (motif 2), which is necessary for the typical redox reaction of CYPs [64], as well as the ERR triad (motif 6) and the (P(E)R(F)) sequence (motif 6), can be differentiated among the discovered 10 motifs.

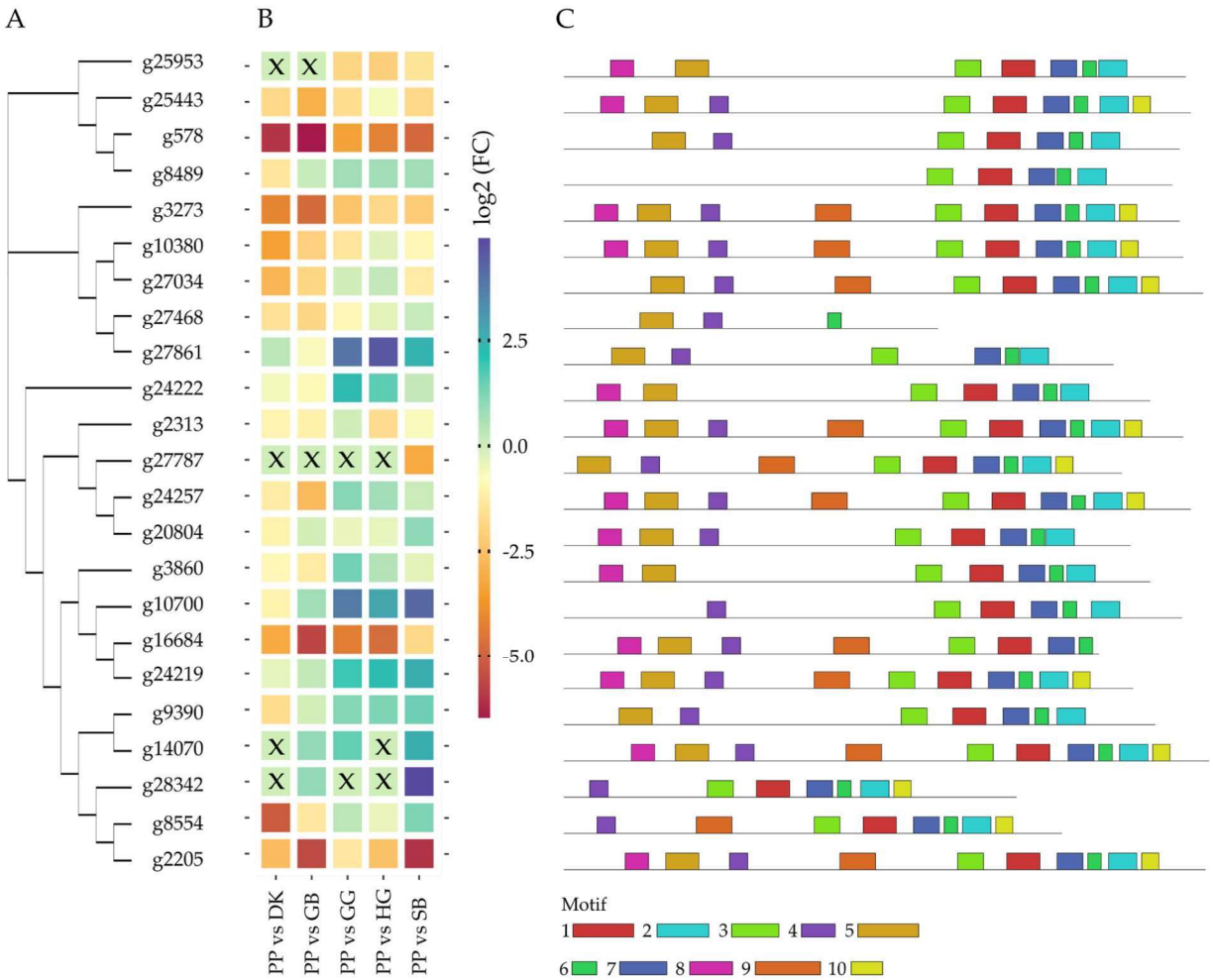


Figure 3. Analysis of differentially expressed cytochrome p450 enzymes (CYP) in different plant cultivars of *Caryopteris* × *clandonensis*, Dark Knight (DK), Grand Bleu (GB), Good as Gold (GG), Hint

of Gold (HG), Sunny Blue (SB), and Pink Perfection (PP). (**A**) Phylogenetic analysis of CYP sequences with highly abundant transcripts regarding limonene-derived molecules (LDM) within the cultivars, using Neighbor-joining method. (**B**) Heatmap of normalized transcript counts between distinct cultivars. X represents enzymes with no transcripts in respective cultivars. The color palette displays genes with high transcripts abundance in red to light-yellow colors, high transcript abundance is depicted in light-green to blue (**C**) Identification of recurring, fixed-length patterns (motifs) identified in LDM-positive transcripts. Motifs 1 to 10 are illustrated as colored boxes, to distinguish the motifs between the different genes. Sequences can be found in Table S3.

Regarding the production of LDM, the genes g8554, g27861, g10700, and g24222 show an interesting pattern compared to the highly LDM-positive cultivar PP, which makes them candidates for further functional characterization to prove their LDM-producing potential.

The candidate genes were further investigated in terms of their putative function. The initial estimates, using sequence and structural homology, consider g2422 and g8554 to be involved in the hydroxylation of cinnamic acid, whereas g27861 and g10700 display unknown activity towards flavonoids, sterols, and ferruginol. This substrate promiscuity is known for CYPs, as they are able to catalyze different ligands [57,65], thus making functional characterization using prokaryotic, yeast, or plant expression systems indispensable to support claims on putative functions.

3. Materials and Methods

3.1. Plant Material

Cultivars of *Caryopteris* × *clandonensis*, DK, GB, GG, HG, SB, and PP, were acquired from a local nursery (Foerstner Pflanzen GmbH, Bietigheim-Bissingen). DK and GB were investigated to show a highly LDM-negative profile, whereas SB and PP show a highly LDM-positive profile. GG and HG showed a non-conclusive profile in between. After growing to maturity in the open in a warm, moderate climate zone, healthy leaves were sampled and snap-frozen in liquid nitrogen and stored at −80 °C until RNA preparation for RNA-Seq.

3.2. Genomic Resource

The reference genome of *Caryopteris* × *clandonensis* used in this study was obtained from NCBI SAMN32308290 (PP). The raw data were assembled as previously described [1] and subjected to further refinements. For further processing, the reference was cleared from possible contaminations, and scaled down from 783 contigs to 53 contigs using Metabat2 (v2.15) [42], keeping the genome completeness with 96.5% at a high level according to BUSCO (v5.3.2) [66] analysis (2326 BUSCO groups, lineage dataset: *Eudicotidae*). Gene model prediction was conducted using AUGUSTUS [67–70]. To detect repetitive sequences, such as tandem repeats or transposable elements, soft masking was employed using Red (v2018.09.10) [71].

3.3. RNA Preparation and Short Read Sequencing

High-quality RNA was extracted using the RNeasy Plant Mini Kit (Qiagen, Venlo, The Netherlands) according to the manufacturer's protocol. To ensure RNA integrity, the Bioanalyzer RNA 6000 assay kit (Agilent, Santa Clara, CA, USA) was employed to yield an average RNA Integrity Number of 7.7. The library preparation was performed using the Illumina stranded mRNA prep kit with IDT for Illumina UD Indexes, Plate A. Corresponding adapter was the Illumina Nextera Adapter (CTGTCTCTTATACACATCT). Library preparation was performed according to the manufacturer's protocol with a shortened fragmentation time from 8 min (protocol) to 2 min (this study). Sequencing was performed at the Helmholtz Munich (HMGU) by the Genomics Core Facility on a NovaSeq6000 SP (2 × 150 bp). For each sample, two lanes were loaded and an average of 22 Mio fragments were yielded. The corresponding lanes of each sample were concatenated tail-to-head (v8.25) [72]. The combined short reads were subjected to comprehensive quality control steps. Every step was analyzed with FastQC (v0.11.9) [73] and the necessity of another trimming step was evaluated. Sequences shorter than 20 bp minimum length and with a

quality phred score beneath 20 were extracted from the paired-end read data. The Illumina Nextera Adapter was used to trim each read pair using Cutadapt (v.4.0) [74]. The first 10 bases were cut from the sequences, due to their sequence GC content, using Trimmomatic (v0.38) and headcrop parameter [75].

3.4. Mapping and Annotation of Aligned Reads

Refined short reads were mapped on the clean reference genome using STAR (v2.7.10b) [76], 140 bases were chosen as the length of the genomic sequence around annotated junctions. EggNOG (v2.1.5) [46,77] was employed to evaluate the function of the differentially expressed genes using Pfam, GO, and COG databases. MEME suite (v5.5.1) [78] was used for identification of motifs within sequences of interest. Visualizations were built in R. Except for STAR; all sequencing analyses were conducted using galaxy project [79]. Analysis was based on reference-based RNA-Seq data analysis [80,81]. The detection of CYPs was performed using a homology-based search, using the conserved domain PF00067, which was integrated to IPR001128. Both domains are indicators for a sequence association with the cytochrome p450 superfamily (IPR036396) [53]. CYP-family classification was performed using a BLAST search [82] and a custom database [83].

3.5. Evaluation of Differential Gene Expression between Aerial Plant Parts

Aligned transcripts were counted using FeatureCounts (v3.16) [84], normalized, and differentially investigated with DESeq2 (v1.34.0) [85–87]. An adjusted p-value below 0.05, and a fold-change greater than 2 and below 0.5, was used to determine the most differentially expressed genes in this dataset.

4. Conclusions

This study provides a basis for further CYP research in *Caryopteris* × *clandonensis*, especially regarding LDM. Furthermore, the reference genome was subjected to a cleaning step, resulting in a decrease from 782 scaffolds to 53 scaffolds. Six cultivars were subjected to an RNA analysis, which gradually neared the prediction of 4 possible LDM tailoring CYPs out of 24, which were differentially expressed, and showed high transcript abundance, compared to the other cultivars. Furthermore, the classification and phylogenetic analysis of all mapped CYPs were conducted and they showed a distinct clustering in CYP CLAN71 and 72. All essential and conserved motifs could be recognized within these sequences. However, experimentally focused research for functional characterization needs to be conducted in order to identify the exact predicted function of these enzymes. A further in silico step can include the prediction of docking and catalysis sites within a three-dimensional structural model, as well as through molecular dynamic techniques and free energy calculations [88,89].

In general, this approach can be used to detect further mechanisms and pathways in plants, which show valuable medicinal effects. The biotechnological production of artemisinin [90] and taxol [11] is a popular example of the possibilities in medicinal plant research. There are already several approaches used, which combine omics approaches to identify substances of interest [91–93].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/plants12122305/s1, Table S1: BUSCO assessment and assembly statistics, Table S2: Mapping statistics on the genomic reference of Pink Perfection, Table S3: Motif sequences of identified reoccurring patterns.

Author Contributions: Conceptualization, M.R., N.A. and N.M.; methodology, M.R. and N.A.; software, M.R., N.A. and N.M.; validation, M.R. and N.A.; formal analysis, M.R. and N.A.; investigation, M.R. and N.A.; resources, T.B.; data curation, M.R. and N.A.; writing—original draft preparation, M.R.; writing—review and editing, M.R., N.A., N.M. and T.B.; visualization, M.R.; supervision, N.M. and T.B.; project administration, N.M. and T.B.; funding acquisition, T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Federal Ministry of Education and Research, grant number 031B0824A.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in a publicly accessible repository. The refined genome data presented in this study are openly available at the National Center for Biotechnology Information (NCBI). BioSample accession number: Pink Perfection SAMN32308290.

Acknowledgments: The authors want to gratefully acknowledge the support of Christine Wurmser, (Chair of Animal Physiology and Immunology, TUM School of Life Sciences, Technical University of Munich) for her support in the library preparation and the handling of Illumina sequencing, and Foerstner Pflanzen GmbH, for providing plant materials. Furthermore, the authors want to acknowledge the support of the following colleagues at the Werner Siemens-Chair for Synthetic Biotechnology: Nathanael Arnold, Kevin Heieck, Zora Rerop, Selina Engelhart-Straub, and further colleagues for their support during conducting experiments and writing this manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ritz, M.; Ahmad, N.; Brueck, T.; Mehlmer, N. Comparative Genome-Wide Analysis of Two Caryopteris x Clandonensis Cultivars: Insights on the Biosynthesis of Volatile Terpenoids. *Plants* **2023**, *12*, 632. [CrossRef] [PubMed]
2. Hannedouche, S.; Jacquemond-Collet, I.; Fabre, N.; Stanislas, E.; Moulis, C. Iridoid keto-glycosides from Caryopteris × Clandonensis. *Phytochemistry* **1999**, *51*, 767–769. [CrossRef]
3. Matsumoto, T.; Mayer, C.; Eugster, C.H. α-Caryopteron, ein neues Pyrano-juglon aus Caryopteris clandonensis. *Helv. Chim. Acta* **1969**, *52*, 808–812. [CrossRef]
4. Blythe, E.K.; Tabanca, N.; Demirci, B.; Bernier, U.R.; Agramonte, N.M.; Ali, A.; Baser, H.C.; Khan, I.A. Composition of the essential oil of Pink Chablis™ bluebeard (Caryopteris ×clandonensis 'Durio') and its biological activity against the yellow fever mosquito Aedes aegypti. *Nat. Volatiles Essent. Oils* **2015**, *2*, 11–21.
5. Abdelaty, N.A.; Attia, E.Z.; Hamed, A.N.E.; Desoukey, S.Y. A review on various classes of secondary metabolites and biological activities of Lamiaceae (*Labiatae*) (2002–2018). *J. Adv. Biomed. Pharm. Sci.* **2021**, *4*, 16–31. [CrossRef]
6. Siciliano, T.; Bader, A.; Vassallo, A.; Braca, A.; Morelli, I.; Pizza, C.; De Tommasi, N. Secondary metabolites from Ballota undulata (*Lamiaceae*). *Biochem. Syst. Ecol.* **2005**, *33*, 341–351. [CrossRef]
7. Mimica-Dukic, N.; Bozin, B.; Mentha, L. Species (*Lamiaceae*) as Promising Sources of Bioactive Secondary Metabolites. *Curr. Pharm. Des.* **2008**, *14*, 3141–3150. [CrossRef]
8. Kliebenstein, D.J. Secondary metabolites and plant/environment interactions: A view through Arabidopsis thaliana tinged glasses. *Plant. Cell Environ.* **2004**, *27*, 675–684. [CrossRef]
9. Boncan, D.A.T.; Tsang, S.S.K.; Li, C.; Lee, I.H.T.; Lam, H.M.; Chan, T.F.; Hui, J.H.L. Terpenes and Terpenoids in Plants: Interactions with Environment and Insects. *Int. J. Mol. Sci.* **2020**, *21*, 7382. [CrossRef]
10. Holopainen, J.K.; Himanen, S.J.; Yuan, J.S.; Chen, F.; Stewart, C.N. Ecological functions of terpenoids in changing climates. *Nat. Prod.* **2013**, *1*, 2913–2940.
11. Wang, T.; Li, L.; Zhuang, W.; Zhang, F.; Shu, X.; Wang, N.; Wang, Z. Recent Research Progress in Taxol Biosynthetic Pathway and Acylation Reactions Mediated by Taxus Acyltransferases. *Molecules* **2021**, *26*, 2855. [CrossRef] [PubMed]
12. Kamatou, G.P.P.; Vermaak, I.; Viljoen, A.M.; Lawrence, B.M. Menthol: A simple monoterpene with remarkable biological properties. *Phytochemistry* **2013**, *96*, 15–25. [CrossRef]
13. Khoo, H.E.; Azlan, A.; Tang, S.T.; Lim, S.M. Anthocyanidins and anthocyanins: Colored pigments as food, pharmaceutical ingredients, and the potential health benefits. *Food Nutr. Res.* **2017**, *61*, 1361779. [CrossRef]
14. Selvaraj, B.; Kim, D.W.; Huh, G.; Lee, H.; Kang, K.; Lee, J.W. Synthesis and biological evaluation of isoliquiritigenin derivatives as a neuroprotective agent against glutamate mediated neurotoxicity in HT22 cells. *Bioorg. Med. Chem. Lett.* **2020**, *30*, 127058. [CrossRef]
15. Mazimba, O. Umbelliferone: Sources, chemistry and bioactivities review. *Bull. Fac. Pharm. Cairo Univ.* **2017**, *55*, 223–232. [CrossRef]
16. Pichersky, E.; Raguso, R.A. Why do plants produce so many terpenoid compounds? *New Phytol.* **2018**, *220*, 692–702. [CrossRef] [PubMed]
17. Lešnik, S.; Furlan, V.; Bren, U. Rosemary (*Rosmarinus officinalis* L.): Extraction techniques, analytical methods and health-promoting biological effects. *Phytochem. Rev.* **2021**, *20*, 1273–1328. [CrossRef]

18. Furlan, V.; Bren, U. Helichrysum italicum: From Extraction, Distillation, and Encapsulation Techniques to Beneficial Health Effects. *Foods* **2023**, *12*, 802. [CrossRef]
19. Fadilah, N.Q.; Jittmittraphap, A.; Leaungwutiwong, P.; Pripdeevech, P.; Dhanushka, D.; Mahidol, C.; Ruchirawat, S.; Kittakoop, P. Virucidal Activity of Essential Oils From Citrus x aurantium L. Against Influenza A Virus H1N1: Limonene as a Potential Household Disinfectant Against Virus. *Nat. Prod. Commun.* **2022**, *17*, 1934578X211072713. [CrossRef]
20. Chassagne, F.; Samarakoon, T.; Porras, G.; Lyles, J.T.; Dettweiler, M.; Marquez, L.; Salam, A.M.; Shabih, S.; Farrokhi, D.R.; Quave, C.L. A Systematic Review of Plants With Antibacterial Activities: A Taxonomic and Phylogenetic Perspective. *Front. Pharmacol.* **2021**, *11*, 2069. [CrossRef]
21. Islam, A.K.M.M.; Suttiyut, T.; Anwar, M.P.; Juraimi, A.S.; Kato-Noguchi, H. Allelopathic Properties of Lamiaceae Species: Prospects and Challenges to Use in Agriculture. *Plants* **2022**, *11*, 1478. [CrossRef]
22. Byers, K.J.R.P.; Bradshaw, H.D.; Riffell, J.A. Three floral volatiles contribute to differential pollinator attraction in monkeyflowers (*Mimulus*). *J. Exp. Biol.* **2014**, *217*, 614–623. [CrossRef] [PubMed]
23. Nagel, R.; Schmidt, A.; Peters, R.J. Isoprenyl diphosphate synthases: The chain length determining step in terpene biosynthesis. *Planta* **2018**, *249*, 9–20. [CrossRef]
24. Dickschat, J.S. Bacterial Diterpene Biosynthesis. *Angew. Chem. Int. Ed.* **2019**, *58*, 15964–15976. [CrossRef]
25. Bohlmann, J.; Meyer-Gauen, G.; Croteau, R. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 4126–4133. [CrossRef] [PubMed]
26. 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.* **2011**, *66*, 212–229. [CrossRef]
27. Karunanithi, P.S.; Zerbe, P. Terpene Synthases as Metabolic Gatekeepers in the Evolution of Plant Terpenoid Chemical Diversity. *Front. Plant Sci.* **2019**, *10*, 1166. [CrossRef]
28. Liu, X.; Zhu, X.; Wang, H.; Liu, T.; Cheng, J.; Jiang, H. Discovery and modification of cytochrome P450 for plant natural products biosynthesis. *Synth. Syst. Biotechnol.* **2020**, *5*, 187. [CrossRef]
29. Foti, R.S.; Honaker, M.; Nath, A.; Pearson, J.T.; Buttrick, B.; Isoherranen, N.; Atkins, W.M. Catalytic vs. Inhibitory Promiscuity in Cytochrome P450s: Implications for Evolution of New Function. *Biochemistry* **2011**, *50*, 2387. [CrossRef]
30. Fischer, M.; Knoll, M.; Sirim, D.; Wagner, F.; Funke, S.; Pleiss, J.; Bateman, A. The Cytochrome P450 Engineering Database: A navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics* **2007**, *23*, 2015–2017. [CrossRef] [PubMed]
31. Nelson, D.R. The Cytochrome P450 Homepage. *Hum. Genom.* **2009**, *4*, 59. [CrossRef] [PubMed]
32. Nelson, D.R. Cytochrome P450 nomenclature, 2004. *Methods Mol. Biol.* **2006**, *320*, 1–10. [CrossRef] [PubMed]
33. Rasool, S.; Mohamed, R. Plant cytochrome P450s: Nomenclature and involvement in natural product biosynthesis. *Protoplasma* **2015**, *253*, 1197–1209. [CrossRef] [PubMed]
34. Krause, S.T.; Liao, P.; Crocoll, C.; Boachon, B.; Förster, C.; Leidecker, F.; Wiese, N.; Zhao, D.; Wood, J.C.; Buell, C.R.; et al. The biosynthesis of thymol, carvacrol, and thymohydroquinone in Lamiaceae proceeds via cytochrome P450s and a short-chain dehydrogenase. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2110092118. [CrossRef] [PubMed]
35. Gupta, P.; Geniza, M.; Naithani, S.; Phillips, J.L.; Haq, E.; Jaiswal, P. Chia (*Salvia hispanica*) Gene Expression Atlas Elucidates Dynamic Spatio-Temporal Changes Associated With Plant Growth and Development. *Front. Plant Sci.* **2021**, *12*, 667678. [CrossRef]
36. Li, H.; Li, J.; Dong, Y.; Hao, H.; Ling, Z.; Bai, H.; Wang, H.; Cui, H.; Shi, L. Time-series transcriptome provides insights into the gene regulation network involved in the volatile terpenoid metabolism during the flower development of lavender. *BMC Plant Biol.* **2019**, *19*, 313. [CrossRef]
37. Lichman, B.R.; Godden, G.T.; Buell, C.R. Gene and genome duplications in the evolution of chemodiversity: Perspectives from studies of Lamiaceae. *Curr. Opin. Plant Biol.* **2020**, *55*, 74–83. [CrossRef]
38. Bak, S.; Beisson, F.; Bishop, G.; Hamberger, B.; Höfer, R.; Paquette, S.; Werck-Reichhart, D. Cytochromes P450. *Arab. Book* **2011**, *9*, e0144. [CrossRef]
39. Xie, Y.; Ye, S.; Wang, Y.; Xu, L.; Zhu, X.; Yang, J.; Feng, H.; Yu, R.; Karanja, B.; Gong, Y.; et al. Transcriptome-based gene profiling provides novel insights into the characteristics of radish root response to Cr stress with next-generation sequencing. *Front. Plant Sci.* **2015**, *6*, 202. [CrossRef]
40. Manzano, A.; Carnero-Diaz, E.; Herranz, R.; Medina, F.J. Recent transcriptomic studies to elucidate the plant adaptive response to spaceflight and to simulated space environments. *iScience* **2022**, *25*, 104687. [CrossRef]
41. Howlader, J.; Robin, A.H.K.; Natarajan, S.; Biswas, M.K.; Sumi, K.R.; Song, C.Y.; Park, J.-I.; Nou, I.-S. Transcriptome Analysis by RNA–Seq Reveals Genes Related to Plant Height in Two Sets of Parent-hybrid Combinations in Easter lily (*Lilium longiflorum*). *Sci. Rep.* **2020**, *10*, 9082. [CrossRef] [PubMed]
42. Kang, D.D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **2019**, *7*, e7359. [CrossRef]
43. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szcześniak, M.W.; Gaffney, D.J.; Elo, L.L.; Zhang, X.; et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **2016**, *17*, 13. [CrossRef] [PubMed]
44. Chaudhary, S.; Khokhar, W.; Jabre, I.; Reddy, A.S.N.; Byrne, L.J.; Wilson, C.M.; Syed, N.H. Alternative splicing and protein diversity: Plants versus animals. *Front. Plant Sci.* **2019**, *10*, 708. [CrossRef]
45. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2015**, *44*, D279–D285. [CrossRef] [PubMed]

46. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2019**, *47*, D309. [CrossRef]
47. Curie, C.; Cassin, G.; Couch, D.; Divol, F.; Higuchi, K.; Le Jean, M.; Misson, J.; Schikora, A.; Czernic, P.; Mari, S. Metal movement within the plant: Contribution of nicotianamine and yellow stripe 1-like transporters. *Ann. Bot.* **2009**, *103*, 1–11. [CrossRef]
48. Ishimaru, Y.; Masuda, H.; Bashir, K.; Inoue, H.; Tsukamoto, T.; Takahashi, M.; Nakanishi, H.; Aoki, N.; Hirose, T.; Ohsugi, R.; et al. Rice metal-nicotianamine transporter, OsYSL2, is required for the long-distance transport of iron and manganese. *Plant J.* **2010**, *62*, 379–390. [CrossRef]
49. Li, Z.; Li, W.; Guo, M.; Liu, S.; Liu, L.; Yu, Y.; Mo, B.; Chen, X.; Gao, L. Origin, evolution and diversification of plant ARGONAUTE proteins. *Plant J.* **2022**, *109*, 1086–1097. [CrossRef]
50. Zhang, Z.; Liu, X.; Li, R.; Yuan, L.; Dai, Y.; Wang, X. Identification and functional analysis of a protein disulfide isomerase (AtPDI1) in Arabidopsis thaliana. *Front. Plant Sci.* **2018**, *9*, 913. [CrossRef]
51. Finkelstein, R. Abscisic Acid Synthesis and Response. *Arab. Book* **2013**, *11*, e0166. [CrossRef] [PubMed]
52. Liao, W.; Zhao, S.; Zhang, M.; Dong, K.; Chen, Y.; Fu, C.; Yu, L. Transcriptome assembly and systematic identification of novel cytochrome P450s in taxus chinensis. *Front. Plant Sci.* **2017**, *8*, 1468. [CrossRef]
53. Degtyarenko, K.N. Structural domains of P450-containing monooxygenase systems. *Protein Eng. Des. Sel.* **1995**, *8*, 737–747. [CrossRef] [PubMed]
54. Vasav, A.P.; Barvkar, V.T. Phylogenomic analysis of cytochrome P450 multigene family and their differential expression analysis in *Solanum lycopersicum* L. suggested tissue specific promoters. *BMC Genom.* **2019**, *20*, 116. [CrossRef] [PubMed]
55. Wegrzyn, G.; Schachner, M.; Gabbiani, G.; Minerdi, D.; Savoi, S.; Sabbatini, P. Role of Cytochrome P450 Enzyme in Plant Microorganisms’ Communication: A Focus on Grapevine. *Int. J. Mol. Sci.* **2023**, *24*, 4695. [CrossRef]
56. Bathe, U.; Tissier, A. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. *Phytochemistry* **2019**, *161*, 149–162. [CrossRef]
57. Hansen, C.C.; Nelson, D.R.; Møller, B.L.; Werck-Reichhart, D. Plant cytochrome P450 plasticity and evolution. *Mol. Plant* **2021**, *14*, 1244–1265. [CrossRef]
58. Haudenschild, C.; Schalk, M.; Karp, F.; Croteau, R. Functional Expression of Regiospecific Cytochrome P450 Limonene Hydroxylases from Mint (*Mentha* spp.) in Escherichia coli and Saccharomyces cerevisiae. *Arch. Biochem. Biophys.* **2000**, *379*, 127–136. [CrossRef]
59. Lupien, S.; Karp, F.; Wildung, M.; Croteau, R. Regiospecific cytochrome P450 limonene hydroxylases from mint (*Mentha*) species: cDNA isolation, characterization, and functional expression of (-)-4S-limonene-3-hydroxylase and (-)-4S-limonene-6-hydroxylase. *Arch. Biochem. Biophys.* **1999**, *368*, 181–192. [CrossRef]
60. Chen, X.; Zhang, C.; Too, H.P. Multienzyme Biosynthesis of Dihydroartemisinic Acid. *Molecules* **2017**, *22*, 1422. [CrossRef]
61. Wu, Y.; Hillwig, M.L.; Wang, Q.; Peters, R.J. Parsing a multifunctional biosynthetic gene cluster from rice: Biochemical characterization of CYP71Z6 & 7. *FEBS Lett.* **2011**, *585*, 3446. [CrossRef]
62. Sawai, S.; Saito, K. Triterpenoid Biosynthesis and Engineering in Plants. *Front. Plant Sci.* **2011**, *585*, 3446–3451. [CrossRef] [PubMed]
63. Chen, Z.; Qi, X.; Yu, X.; Zheng, Y.; Liu, Z.; Fang, H.; Li, L.; Bai, Y.; Liang, C.; Li, W. Genome-Wide Analysis of Terpene Synthase Gene Family in Mentha longifolia and Catalytic Activity Analysis of a Single Terpene Synthase. *Genes* **2021**, *12*, 518. [CrossRef] [PubMed]
64. Zhang, W.; Liu, Y.; Yan, J.; Cao, S.; Bai, F.; Yang, Y.; Huang, S.; Yao, L.; Anzai, Y.; Kato, F.; et al. New reactions and products resulting from alternative interactions between the P450 enzyme and redox partners. *J. Am. Chem. Soc.* **2014**, *136*, 3640–3646. [CrossRef] [PubMed]
65. Hernandez-Ortega, A.; Vinaixa, M.; Zebec, Z.; Takano, E.; Scrutton, N.S. A Toolbox for Diverse Oxyfunctionalisation of Monoterpenes OPEN. *Sci. Rep.* **2018**, *8*, 14396. [CrossRef]
66. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
67. Hoff, K.J.; Lomsadze, A.; Borodovsky, M.; Stanke, M. Whole-Genome Annotation with BRAKER. *Methods Mol. Biol.* **2019**, *1962*, 65. [CrossRef]
68. Hoff, K.J.; Lange, S.; Lomsadze, A.; Borodovsky, M.; Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **2016**, *32*, 767–769. [CrossRef]
69. Brůna, T.; Hoff, K.J.; Lomsadze, A.; Stanke, M.; Borodovsky, M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **2021**, *3*, lqaa108. [CrossRef]
70. Stanke, M.; Schöffmann, O.; Morgenstern, B.; Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinform.* **2006**, *7*, 62. [CrossRef] [PubMed]
71. Girgis, H.Z. Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinform.* **2015**, *16*, 227. [CrossRef]
72. Grüning, B.; Yusuf, D.; Houwaart, T.; Anika; Miladi, M.; Gu, Q.; Batut, B.; Soranzo, N.; Gamaleldin, H.; Von Kuster, G.; et al. *Bgruening/Galaxytools: September Release 2019*; Zenodo: Geneva, Switzerland, 2018. [CrossRef]

73. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 24 March 2023).
74. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10. [CrossRef]
75. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
76. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef] [PubMed]
77. Huerta-Cepas, J.; Forslund, K.; Coelho, L.P.; Szklarczyk, D.; Jensen, L.J.; Von Mering, C.; Bork, P. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **2017**, *34*, 2115–2122. [CrossRef] [PubMed]
78. Bailey, T.L.; Johnson, J.; Grant, C.E.; Noble, W.S. The MEME Suite. *Nucleic Acids Res.* **2015**, *43*, W39–W49. [CrossRef] [PubMed]
79. Afgan, E.; Baker, D.; Batut, B.; Van Den Beek, M.; Bouvier, D.; Ech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W537–W544. [CrossRef]
80. Batut, B.; Freeberg, M.; Heydarian, M.; Erxleben, A.; Videm, P.; Blank, C.; Doyle, M.; Soranzo, N.; van Heusden, P.; Delisle, L. Reference-Based RNA-Seq Data Analysis (Galaxy Training Materials). Available online: https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html#citing-this-tutorial (accessed on 25 March 2023).
81. Batut, B.; Hiltemann, S.; Bagnacani, A.; Baker, D.; Bhardwaj, V.; Blank, C.; Bretaudeau, A.; Brillet-Guéguen, L.; Čech, M.; Chilton, J.; et al. Community-Driven Data Analysis Training for Biology. *Cell Syst.* **2018**, *6*, 752–758.e1. [CrossRef] [PubMed]
82. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
83. Kweon, O.; Kim, S.J.; Kim, J.H.; Nho, S.W.; Bae, D.; Chon, J.; Hart, M.; Baek, D.H.; Kim, Y.C.; Wang, W.; et al. CYPminer: An automated cytochrome P450 identification, classification, and data analysis tool for genome data sets across kingdoms. *BMC Bioinform.* **2020**, *21*, 160. [CrossRef]
84. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [CrossRef] [PubMed]
85. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
86. Zhu, A.; Ibrahim, J.G.; Love, M.I. Heavy-tailed prior distributions for sequence count data: Removing the noise and preserving large differences. *Bioinformatics* **2019**, *35*, 2084–2092. [CrossRef] [PubMed]
87. Bioinformatics Training at the Harvard Chan Bioinformatics Core. Available online: https://hbctraining.github.io/main/ (accessed on 19 April 2023).
88. Pantiora, P.; Furlan, V.; Matiadis, D.; Mavroidi, B.; Perperopoulou, F.; Papageorgiou, A.C.; Sagnou, M.; Bren, U.; Pelecanou, M.; Labrou, N.E. Monocarbonyl Curcumin Analogues as Potent Inhibitors against Human Glutathione Transferase P1-1. *Antioxidants* **2023**, *12*, 63. [CrossRef] [PubMed]
89. Kores, K.; Kolenc, Z.; Furlan, V.; Bren, U. Inverse Molecular Docking Elucidating the Anticarcinogenic Potential of the Hop Natural Product Xanthohumol and Its Metabolites. *Foods* **2022**, *11*, 1253. [CrossRef]
90. Wen, W.; Yu, R. Artemisinin biosynthesis and its regulatory enzymes: Progress and perspective. *Pharmacogn. Rev.* **2011**, *5*, 189. [CrossRef]
91. Sun, W.; Xu, Z.; Song, C.; Chen, S. Herbgenomics: Decipher molecular genetics of medicinal plants. *Innovation* **2022**, *3*, 100322. [CrossRef]
92. Alami, M.M.; Ouyang, Z.; Zhang, Y.; Shu, S.; Yang, G.; Mei, Z.; Wang, X. The Current Developments in Medicinal Plant Genomics Enabled the Diversification of Secondary Metabolites' Biosynthesis. *Int. J. Mol. Sci.* **2022**, *23*, 15932. [CrossRef]
93. Cheng, Q.Q.; Ouyang, Y.; Tang, Z.Y.; Lao, C.C.; Zhang, Y.Y.; Cheng, C.S.; Zhou, H. Review on the Development and Applications of Medicinal Plant Genomes. *Front. Plant Sci.* **2021**, *12*, 2981. [CrossRef]

5 Discussion

5. 1 Use-cases of genome data and proposed gene models

The yielded gene models can further be investigated using wet-lab techniques or *in silico* modelling. To characterize the distinct function of each possible enzyme, each gene needs to be extracted from a mRNA library of the plant. This way one can ensure the correct sequence and also measure detailed transcript abundancies using qPCR. After successful extraction and reverse transcription of the cDNA, suitable vector systems for both cloning and expression need to be identified. In Figure 4, a typical workflow is depicted schematically. A first start begins with literature research, to determine the different values and metabolites of a plant. Already in ancient texts, plant extracts are known to have medicinal and psychoactive properties to treat different ailments [123]–[125]. Residues of this knowledge are still used, as seen in traditional Chinese medicine [126]. However, as seen in the big genome databases, only little is known about the exact genes behind the mode of action or even the active agent of plant extracts [127]–[129]. Therefore, the next step is the extraction of high-molecular genomic information out of the plant of interest and consecutive long-read sequencing. The yielded reads are preferably *de novo* assembled to a reference genome of the plant. The genome can now be used for the elucidation of gene structures and assign putative functions. As the first interest was the active agent and its mode of action seen in a medicinal plant, the next step is, as depicted, either the expression of a desired protein or whole-cell bio catalysis of a desired product. Using this workflow, a high-throughput platform can be established to provide a comprehensive and fast screening method for genes to elucidate the best candidate for further mutagenesis and improvement of product formation, as already seen in the literature [130], [131]. Here, the potential of still unknown active agents of medicinal plants bares a huge possibility in further plant genome research to exploit the natural compounds for human health.

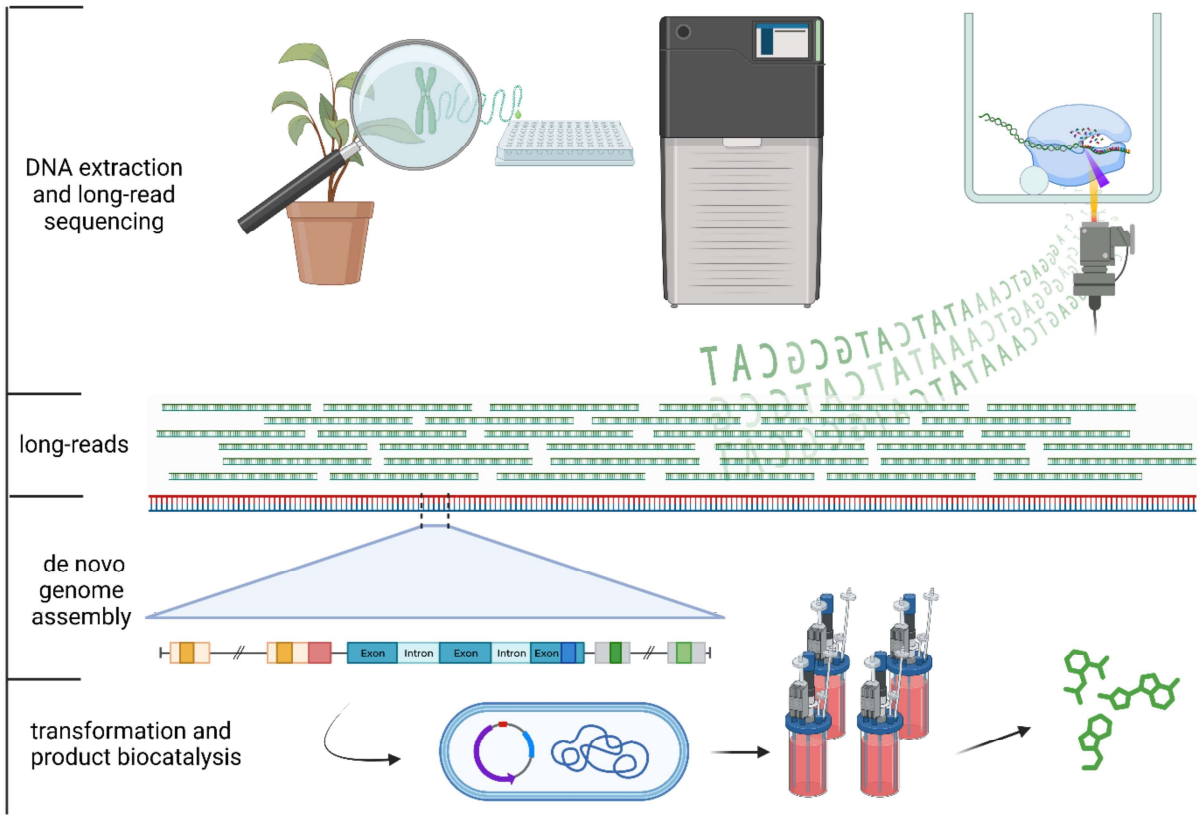


Figure 4. Schematic overview on the elucidation of a plants genetic make-up. The plants DNA is extracted and subjected a long-read sequencing analysis. A Zero-mode waveguide depicts the sequencing mechanism behind a PacBio flow cell. Assembly of the reads yields a reference genome which harbors the gene structures necessary for cloning and further expression and product biosynthesis.

5. 2 Challenges in biotechnological synthesis of plant metabolites

In general, the recombinant production of proteins and metabolites can result in a few challenges during biotechnological production. Especially for molecules originating from plants, as seen for known compounds such as taxol, derived from *Taxus brevifolia*, and Artemisinin, derived from *Artemisia annua*. Among others, the precursor supply, the biochemistry of desired products and the expression system itself needs to be optimized for high yields. In the next chapters, these challenges are investigated in a closer detail.

5. 2. 1 Precursor supply

To produce in high yields, the host flux needs to be optimized to provide the necessary precursors for a balanced biosynthesis. Especially for plant metabolites, a complex metabolic pathway needs to be engineered in production organisms, which require a variety of precursors, co-factors, and enzymes [132]–[134]. These are often found in low concentrations in plants, as high concentration can alter the productivity or lead to cell toxicity [135], [136]. Furthermore, a possible external feed is often not feasible as these molecules are not readily available in pure form.

To overcome these challenges, researchers have developed various strategies for precursor supply in whole-cell biocatalysis. One approach is to use alternative precursors, that are structurally similar to the natural precursors, but are more readily available or easier to synthesize. Another approach is to optimize the expression of genes involved in precursor biosynthesis in the host cell, or to engineer the metabolic pathway to increase precursor flux [137], [138]. The balance in expression is important as mentioned above.

Overall, the availability of precursors for plant metabolite synthesis remains a significant challenge in whole-cell biosynthesis. Addressing this challenge will require a combination of innovative strategies, metabolic engineering approaches, and advances in precursor synthesis and purification.

5. 2. 2 Product biosynthesis

Within this paragraph, one challenge to produce the compound is highlighted in more detail. The product formation often occurs in specialized compartments inside the plant, such as the periplasm, vacuoles or different organelles. Missing these compartments in a heterologous host can result in a bioprocess with low yield or even toxicity for the host [136]. To overcome this issue, one promising approach is the encapsulation of a specific process. A pathway step with specific needs to its environment or a potential toxic side product can be encapsulated using naturally membrane or protein based systems [139]. A further way to tackle this issue is to reduce the necessity of this pathway itself by means of synthetic biology pathway engineering. This way, a substitute can be put in place to overcome the challenge of specialized compartments [140], [141]. Both approaches need to be evaluated depending on the desired molecule and its natural biosynthesis pathway.

Plant *in vitro* cultures are another possibility to overcome this challenge [136]. The plant can be transformed with the desired pathway genes using *Agrobacterium tumefaciens* or direct DNA transformation methods [142], [143]. However, both tools have limitation towards efficacy and limit the scalability and sustainability of product biosynthesis.

5. 2. 3 Production hosts

One example of a microorganism used for whole-cell biocatalysis of plant metabolites is *Escherichia coli*. It has already been used to produce a variety of plant metabolites, including flavonoids, alkaloids, and terpenoids [144]–[146]. The wealth of molecular biology techniques to modify and transform *E. coli* makes it an easy host for first proof-of-concept experiments. However, the expression of plant enzymes in *E. coli* can be challenging due to differences in post-translational modifications, redox balances and compartment specializations, as described above. This is also true for other host organisms such as *Saccharomyces cerevisiae*, commonly known as baker's yeast. It has also been used for the synthesis of plant metabolites. For example, it has been engineered to produce

artemisinic acid, a precursor to the antimalarial drug artemisinin [53], [147] or the precursor taxadiene for the anti-cancer drug taxol [148], [149].

No matter of the system of choice, *in vitro* plant cell culture or a microorganism as expression host, there is always a need for pathway engineering, host engineering or enzyme engineering to achieve a high yielding expression system for plant natural products.

5.3 Artificial intelligence based investigation of genome databases

A last outlook shall be depicted towards the possibilities of artificial intelligence in combination with genome sequencing. In the past decades, more and more sequencing projects have been launched to allow a comprehensive overview on different genomes of the worlds' flora and fauna. The pan genome project of the human genome wants to bridge the gap between population and ethnicity gaps and yield a more diverse and accurate genome which spans over multiple individuals. This can be used for further bioinformatics investigations especially for gene – disease association studies[150]. A further project is the 100K Pathogen Genome Project with its goal to strengthen the knowledge and broaden the research in host pathogen interaction to secure food supplies and global health. Already a plethora of important pathogen genomes were made available through this project such as *Shigella, Salmonella, Helicobacter* and many more [151]. Further sequencing projects are the 1k and 10k plant genome project [152] and the vertebrate genomes project which aims to completely sequence and provide more than 70,000 fully annotated reference genomes [153].

Advancements in the identification of functional gene structures and the possibility of simulating splicing variants as well as *in silico* protein structure modelling enable researcher to identify biomarkers without the necessity of wet-lab experiments [154]. In context with artificial intelligence systems, this means the possibility to use the provided datasets to investigate interactions between different organisms. As depicted in Figure 5, there are plenty of interactions between different organisms which influence each other. Metabolites produced by bacteria can influence the metabolism of plants [155]. For a few toxins the mode of action is already elucidated however a multifold of interactions between virus compounds and human cells is not known. This is also true for protein and metabolite interaction between humans and animal compounds as well as for plants and animals and so on. With the availability of computational centers and bioinformatics tools, possible interactions can be investigated and functional biomarkers identified through extensive structural modelling and ligand docking [156]–[158]. This way, new technologies, such as the emerging use of artificial intelligence can support research and leap the understanding of protein-protein and protein-ligand interaction to the next level. However, without a comprehensive genome database this will not be possible.

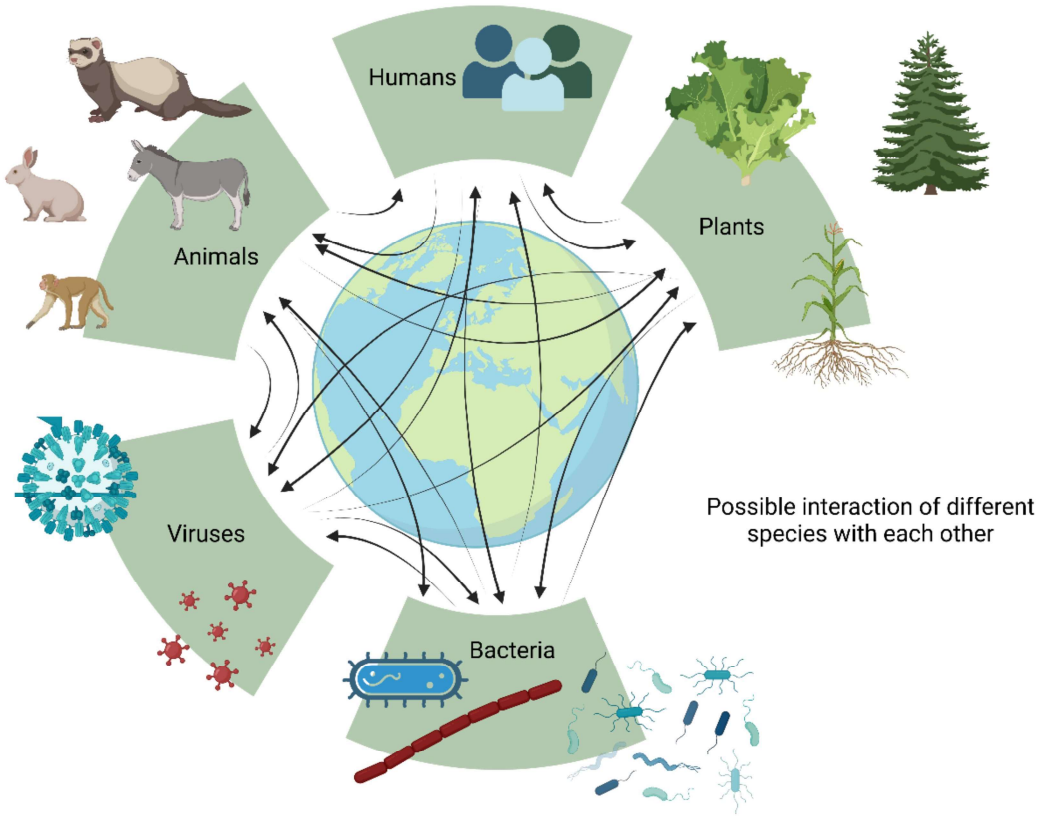


Figure 5. Schematic overview on the interaction between different organisms. Each synthesized protein and metabolite is able to influence their surrounding and also close organisms. Structural modelling and ligand docking allows the prediction of possible interaction and therefore possible biomarker identification without wet-lab experiments

6 Conclusion

This thesis, including their two genome and transcriptome projects is building a basis for further research on Caryopteris x Clandonensis and their valuable profile in limonene-derived molecules. Economical value can be drawn from the various possibilities in further exploitation of putative enzymes synthesizing secondary metabolites. In addition to the project goals, we were able to establish a long-read sequencing platform employing the PacBio Sequell IIe to foster a new pillar of expertise for further projects. From the extraction protocol, yielding suitable amounts of high-molecular genomic DNA, to internal library preparation optimization over to the bioinformatics analysis pipeline. In addition, the investigated genomes are made publicly available, thus making it accessibly for a manifold of interested researcher in the field of Lamiaceae genomes and a focus on terpenoid research.

7 List of Publications

Ritz, Manfred, Nadim Ahmad, Thomas Brueck, and Norbert Mehlmer. 2023. “Comparative Genome-Wide Analysis of Two Caryopteris x Clandonensis Cultivars: Insights on the Biosynthesis of Volatile Terpenoids.” Plants 2023, Vol. 12, Page 632 12 (3). MDPI: 632. doi:10.3390/PLANTS12030632.

Ritz, Manfred, Nadim Ahmad, Thomas Brueck, and Norbert Mehlmer. 2023. “Differential RNA-Seq Analysis Predicts Genes Related to Terpene Tailoring in Caryopteris × Clandonensis.” Plants (Basel, Switzerland) 12 (12). Plants (Basel): 2305. doi:10.3390/PLANTS12122305

Ahmad, Nadim, **Manfred Ritz**, Anjuli Calchera, Jürgen Otte, Imke Schmitt, Thomas Brueck, and Norbert Mehlmer. 2023. “Biosynthetic Potential of Hypogymnia Holobionts: Insights into Secondary Metabolite Pathways.” Journal of Fungi 9 (5). MDPI: 546. doi:10.3390/JOF9050546/S1.

Szabo, Edina K., Christina Bowhay, Namratha Badawadagi, Beatrice Fung, Camila Gaio, Kayla Bailey, **Manfred Ritz**, Anupama Ariyaratne, Joel Bowron, and Constance A. M. Finney. 2021. “Early Th2 Cytokine Production in Heligmosomoides Polygyrus and Toxoplasma Gondii Co-Infected Mice Is Associated with Reduced IFNγ Production, Increased Parasite Loads and Increased Mortality.” BioRxiv, May. Cold Spring Harbor Laboratory, 2021.05.27.445631. doi:10.1101/2021.05.27.445631.

In process of publication:

Ahmad, Nadim, **Manfred Ritz**, Anjuli Calchera, Jürgen Otte, Imke Schmitt, Thomas Brueck, and Norbert Mehlmer. 2023. “Biosynthetic Gene Cluster Synteny - Orthologous Polyketide Synthases in the genus Parmeliaceae” Journal of Fungi.

Vasic, Vedran, Can Buldun, **Manfred Ritz**, Steffen Dickopf, Guy J. Georges, Christian Spick, Alessa Peuker, Thomas Meier, Klaus Mayer, Ulrich Brinkmann. 2023 “Targeted chain-exchange-mediated reconstitution of a split type-I cytokine for conditional immunotherapy” MAbs.

8 Reprint Permission

No special permission is required to reuse all or part of article published by MDPI, including figures and tables. For articles published under an open access Creative Common CC BY license, any part of the article may be reused without permission provided that the original article is clearly cited. Reuse of an article does not imply endorsement by the authors or MDPI.

Copyright and Licensing

For all articles published in MDPI journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.
(https://www.mdpi.com/openaccess; accessed:03-02-2023)

9 Figures & Tables

9.1 List of Figures

9.2 List of Tables

10 References

[1] C. L. Schoch *et al.*, "NCBI Taxonomy: A comprehensive update on curation, resources and tools," *Database*, vol. 2020, 2020.

[2] M. W. Chase *et al.*, "An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV," *Bot. J. Linn. Soc.*, vol. 181, no. 1, pp. 1–20, May 2016.

[3] F. Zhao *et al.*, "An updated tribal classification of Lamiaceae based on plastome phylogenomics," *BMC Biol. 2021 191*, vol. 19, no. 1, pp. 1–27, Jan. 2021.

[4] L. Maleci Bini and C. Giuliani, "The glandular trichomes of the Labiatae. A review," *Acta Hortic.*, vol. 723, pp. 85–90, 2006.

[5] C. M. Uritu *et al.*, "Medicinal plants of the family Lamiaceae in pain therapy: A review," *Pain Res. Manag.*, vol. 2018, 2018.

[6] K. Carović-Stanko *et al.*, "Medicinal plants of the family Lamiaceae as functional foods - a review," *Czech J. Food Sci.*, vol. 34, no. 5, pp. 377–390, Oct. 2016.

[7] W. Dhifi, S. Bellili, S. Jazi, N. Bahloul, and W. Mnif, "Essential Oils' Chemical Characterization and Investigation of Some Biological Activities: A Critical Review," *Medicines*, vol. 3, no. 4, p. 25, Sep. 2016.

[8] D. Aebisher, J. Cichonski, E. Szpyrka, S. Masjonis, and G. Chrzanowski, "Essential Oils of Seven Lamiaceae Plants and Their Antioxidant Capacity," *Molecules*, vol. 26, no. 13, Jun. 2021.

[9] L. R. Ramos Da Silva *et al.*, "Lamiaceae Essential Oils, Phytochemical Profile, Antioxidant, and Biological Activities," *Evid. Based. Complement. Alternat. Med.*, vol. 2021, 2021.

[10] E. K. Blythe *et al.*, "Composition of the essential oil of Pink Chablis™ bluebeard (Caryopteris ×clandonensis 'Durio') and its biological activity against the yellow fever mosquito Aedes aegypti," *Nat. Volatiles Essent. Oils*, vol. 2, no. 1, pp. 11–21, 2015.

[11] M. Ritz, N. Ahmad, T. Brueck, and N. Mehlmer, "Comparative Genome-Wide Analysis of Two Caryopteris x Clandonensis Cultivars: Insights on the Biosynthesis of Volatile Terpenoids," *Plants 2023, Vol. 12, Page 632*, vol. 12, no. 3, p. 632, Feb. 2023.

[12] S. Park, M. J. Son, C. S. Yook, C. Jin, Y. S. Lee, and H. J. Kim, "Chemical constituents from aerial parts of Caryopteris incana and cytoprotective effects in human HepG2 cells," *Phytochemistry*, vol. 101, pp. 83–90, 2014.

[13] M. Dumaa *et al.*, "Two new alkaloids from the aerial parts of Caryopteris mongolica Bunge," *Mong. J. Chem.*, vol. 13, pp. 41–45, Sep. 2014.

[14] T. Matsumoto, C. Mayer, and C. H. Eugster, "α-Caryopteron, ein neues Pyrano-juglon aus Caryopteris clandonensis," *Helv. Chim. Acta*, vol. 52, no. 3, pp. 808–812, 1969.

[15] S. Hannedouche, I. Jacquemond-Collet, N. Fabre, E. Stanislas, and C. Moulis, "Iridoid keto-glycosides from Caryopteris × Clandonensis," *Phytochemistry*, vol. 51, no. 6, pp. 767–769, Jul. 1999.

[16] S. C. Sha, Z. L. Qi, Z. Ligang, S. Du Shu, and L. L. Zhi, "Chemical composition and toxic activity of essential oil of Caryopteris incana against Sitophilus zeamais," *African J. Biotechnol.*, vol. 10, no. 42, pp. 8416–8480, Aug. 2011.

[17] O. Sicora and M. A. Naghi, "THE ETHANOLIC STEM EXTRACT OF CARYOPTERIS X

CLANDONENSIS POSSESES ANTIPROLIFERATIVE POTENTIAL BY BLOCKING BREAST CANCER CELLS IN MITOSIS," *Farmacia*, vol. 67, p. 6, 2019.

[18] J. K. R. da Silva, P. L. B. Figueiredo, K. G. Byler, and W. N. Setzer, "Essential Oils as Antiviral Agents, Potential of Essential Oils to Treat SARS-CoV-2 Infection: An In-Silico Investigation," *Int. J. Mol. Sci. 2020, Vol. 21, Page 3426*, vol. 21, no. 10, p. 3426, May 2020.

[19] G. L. da Silva *et al.*, "Antioxidant, analgesic and anti-inflammatory effects of lavender essential oil," *An. Acad. Bras. Cienc.*, vol. 87, no. 2, pp. 1397–1408, Aug. 2015.

[20] P. K. Mediratta, K. K. Sharma, and S. Singh, "Evaluation of immunomodulatory potential of Ocimum sanctum seed oil and its possible mechanism of action," *J. Ethnopharmacol.*, vol. 80, no. 1, pp. 15–20, Apr. 2002.

[21] M. C. Wani, H. L. Taylor, M. E. Wall, P. Coggon, and A. T. Mcphail, "Plant Antitumor Agents.VI.The Isolation and Structure of Taxol, a Novel Antileukemic and Antitumor Agent from Taxus brevifolia2," *J. Am. Chem. Soc.*, vol. 93, no. 9, pp. 2325–2327, May 1971.

[22] M. Zaynab, M. Fatima, Y. Sharif, M. H. Zafar, H. Ali, and K. A. Khan, "Role of primary metabolites in plant defense against pathogens," *Microb. Pathog.*, vol. 137, p. 103728, Dec. 2019.

[23] C. Fang, A. R. Fernie, and J. Luo, "Exploring the Diversity of Plant Metabolism," *Trends Plant Sci.*, vol. 24, no. 1, pp. 83–98, Jan. 2019.

[24] J. K. Holopainen, S. J. Himanen, J. S. Yuan, F. Chen, and C. N. Stewart, "Ecological functions of terpenoids in changing climates," *Nat. Prod. Phytochem. Bot. Metab. Alkaloids, Phenolics Terpenes*, pp. 2913–2940, Jan. 2013.

[25] J. Hong, L. Yang, D. Zhang, and J. Shi, "Plant Metabolomics: An Indispensable System Biology Tool for Plant Science," *Int. J. Mol. Sci.*, vol. 17, no. 6, Jun. 2016.

[26] J. Kroymann, "Natural diversity and adaptation in plant secondary metabolism," *Curr. Opin. Plant Biol.*, vol. 14, no. 3, pp. 246–251, Jun. 2011.

[27] H.-W. Heldt and B. Piechulla, *Plant Biochemistry - Polysaccharides Are Storage and Transport Forms of Carbohydrates Produced by Photosynthesis*, 5th ed. Academic Press, 2021.

[28] H.-W. Heldt and B. Piechulla, *Plant Biochemistry - Lipids Are Membrane Constituents and Function as Carbon Stores*, 5th ed. Academic Press, 2021.

[29] C. Stasolla, R. Katahira, T. A. Thorpe, and H. Ashihara, "Purine and pyrimidine nucleotide metabolism in higher plants," *J. Plant Physiol.*, vol. 160, no. 11, pp. 1271–1295, Jan. 2003.

[30] R. Zrenner, M. Stitt, U. Sonnewald, and R. Boldt, "PYRIMIDINE AND PURINE BIOSYNTHESIS AND DEGRADATION IN PLANTS," *https://doi.org/10.1146/annurev.arplant.57.032905.105421*, vol. 57, pp. 805–836, May 2006.

[31] B. A. Moffatt and H. Ashihara, "Purine and Pyrimidine Nucleotide Synthesis and Metabolism," *Arabidopsis Book*, vol. 1, p. e0018, Jan. 2002.

[32] I. Andersson and A. Backlund, "Structure and function of Rubisco," *Plant Physiol. Biochem.*, vol. 46, no. 3, pp. 275–291, Mar. 2008.

[33] C. Maurel, Y. Boursiac, D. T. Luu, V. Santoni, Z. Shahzad, and L. Verdoucq, "Aquaporins in plants," *Physiol. Rev.*, vol. 95, no. 4, pp. 1321–1358, Sep. 2015.

[34] S. Khare *et al.*, "Plant secondary metabolites synthesis and their regulations under biotic and

abiotic constraints," *J. Plant Biol. 2020 633*, vol. 63, no. 3, pp. 203–216, Apr. 2020.

[35] H. N. Matsuura and A. G. Fett-Neto, "Plant Alkaloids: Main Features, Toxicity, and Mechanisms of Action," *Plant Toxins*, pp. 1–15, 2015.

[36] J. Ziegler and P. J. Facchini, "Alkaloid Biosynthesis: Metabolism and Trafficking," *Annu. Rev. Plant Biol.*, vol. 59, pp. 735–769, Apr. 2008.

[37] N. Francenia Santos-Sánchez, R. Salas-Coronado, B. Hernández-Carlos, and C. Villanueva-Cañongo, "Shikimic Acid Pathway in Biosynthesis of Phenolic Compounds," *Plant Physiol. Asp. Phenolic Compd.*, Sep. 2019.

[38] V. Chowdhary, S. Alooparampil, R. V. Pandya, and J. G. Tank, "Physiological Function of Phenolic Compounds in Plant Defense System," in *Phenolic Compounds - Chemistry, Synthesis, Diversity, Non-Conventional Industrial, Pharmaceutical and Therapeutic Applications*, IntechOpen, 2021.

[39] M. C. Dias, D. C. G. A. Pinto, and A. M. S. Silva, "Plant Flavonoids: Chemical Characteristics and Biological Activity," *Molecules*, vol. 26, no. 17, Sep. 2021.

[40] D. Treutter, "Significance of flavonoids in plant resistance: A review," *Environ. Chem. Lett.*, vol. 4, no. 3, pp. 147–157, Aug. 2006.

[41] E. Pichersky and R. A. Raguso, "Why do plants produce so many terpenoid compounds?," *New Phytol.*, vol. 220, no. 3, pp. 692–702, Nov. 2018.

[42] E. J. N. Helfrich, G.-M. Lin, C. A. Voigt, and J. Clardy, "Bacterial terpene biosynthesis: challenges and opportunities for pathway engineering," *Beilstein J. Org. Chem*, vol. 15, pp. 2889–2906, 2019.

[43] M. Rohmer, "The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants†," *Nat. Prod. Rep.*, vol. 16, no. 5, pp. 565–574, Jan. 1999.

[44] K. W. George, J. Alonso-Gutierrez, J. D. Keasling, and T. S. Lee, "Erratum to: Isoprenoid drugs, biofuels, and chemicals—artemisinin, farnesene, and beyond [Adv Biochem Eng Biotechnol, DOI: 10.1007/10_2014_288]," *Adv. Biochem. Eng. Biotechnol.*, vol. 148, p. 469, 2015.

[45] H. Karlic and F. Varga, "Mevalonate Pathway," in *Encyclopedia of Cancer*, Academic Press, 2019, pp. 445–457.

[46] D. Tholl, "Biosynthesis and biological functions of terpenoids in plants," *Adv. Biochem. Eng. Biotechnol.*, vol. 148, pp. 63–106, 2015.

[47] D. Tritsch, A. Hemmerlin, T. J. Bach, and M. Rohmer, "Plant isoprenoid biosynthesis via the MEP pathway: In vivo IPP/DMAPP ratio produced by (E)-4-hydroxy-3-methylbut-2-enyl diphosphate reductase in tobacco BY-2 cell cultures," *FEBS Lett.*, vol. 584, no. 1, pp. 129–134, Jan. 2010.

[48] J. Bohlmann, G. Meyer-Gauen, and R. Croteau, "Plant terpenoid synthases: Molecular biology and phylogenetic analysis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 8, pp. 4126–4133, Apr. 1998.

[49] D. W. Christianson, "Structural and Chemical Biology of Terpenoid Cyclases," *Chem. Rev.*, vol. 117, no. 17, pp. 11570–11648, Sep. 2017.

[50] F. Chen, D. Tholl, J. Bohlmann, and E. Pichersky, "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.*, vol. 66, no. 1, pp. 212–229, Apr. 2011.

[51] X. Liu, X. Zhu, H. Wang, T. Liu, J. Cheng, and H. Jiang, "Discovery and modification of cytochrome P450 for plant natural products biosynthesis," *Synth. Syst. Biotechnol.*, vol. 5, no. 3, p. 187, Sep. 2020.

[52] T. Wang *et al.*, "Recent Research Progress in Taxol Biosynthetic Pathway and Acylation Reactions Mediated by Taxus Acyltransferases," *Mol. 2021, Vol. 26, Page 2855*, vol. 26, no. 10, p. 2855, May 2021.

[53] W. Wen and R. Yu, "Artemisinin biosynthesis and its regulatory enzymes: Progress and perspective," *Pharmacogn. Rev.*, vol. 5, no. 10, p. 189, Jul. 2011.

[54] M. Fischer *et al.*, "The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family," vol. 23, no. 15, pp. 2015–2017, 2007.

[55] D. R. Nelson, "Cytochrome P450 nomenclature, 2004.," *Methods Mol. Biol.*, vol. 320, pp. 1–10, 2006.

[56] S. Rasool and R. Mohamed, "Plant cytochrome P450s: nomenclature and involvement in natural product biosynthesis," *Protoplasma 2015 2535*, vol. 253, no. 5, pp. 1197–1209, Sep. 2015.

[57] S. T. Krause *et al.*, "The biosynthesis of thymol, carvacrol, and thymohydroquinone in Lamiaceae proceeds via cytochrome P450s and a short-chain dehydrogenase," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 52, p. e2110092118, Dec. 2021.

[58] S. Bak *et al.*, "Cytochromes P450," *Arabidopsis Book*, vol. 9, p. e0144, Jan. 2011.

[59] W. Zhang *et al.*, "New reactions and products resulting from alternative interactions between the P450 enzyme and redox partners," *J. Am. Chem. Soc.*, vol. 136, no. 9, pp. 3640–3646, Mar. 2014.

[60] K. H. C. Baser and G. Buchbauer, Eds., *Handbook of Essential Oils - History and Sources of Essential Oil Research*, 1st ed. Boca Raton: CRC Press, 2009.

[61] H. S. Elshafie and I. Camele, "An Overview of the Biological Effects of Some Mediterranean Essential Oils on Human Health," *Biomed Res. Int.*, vol. 2017, 2017.

[62] N. Ben Derbassi, M. C. Pedrosa, S. Heleno, M. Carocho, I. C. F. R. Ferreira, and L. Barros, "Plant volatiles: Using Scented molecules as food additives," *Trends Food Sci. Technol.*, vol. 122, pp. 97–103, Apr. 2022.

[63] L. Caputi and E. Aprea, "Use of Terpenoids as Natural Flavouring Compounds in Food Industry," *Recent Patents Food, Nutr. Agric.*, vol. 3, no. 1, pp. 9–16, Oct. 2012.

[64] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, p. 5463, 1977.

[65] D. B. Neale *et al.*, "Assembled and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary adaptive potential and investigating hexaploid origin," *G3 Genes|Genomes|Genetics*, vol. 12, no. 1, Jan. 2022.

[66] A. D. Scott *et al.*, "A Reference Genome Sequence for Giant Sequoia," *G3 Genes|Genomes|Genetics*, vol. 10, no. 11, pp. 3907–3919, Nov. 2020.

[67] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, "Overview of Next Generation Sequencing Technologies," *Curr. Protoc. Mol. Biol.*, vol. 122, no. 1, p. e59, Apr. 2018.

[68] S. Goodwin, J. D. McPherson, and W. R. McCombie, “Coming of age: ten years of next-generation sequencing technologies,” *Nat. Rev. Genet. 2016 176*, vol. 17, no. 6, pp. 333–351, May 2016.

[69] T. Hu, N. Chitnis, D. Monos, and A. Dinh, “Next-generation sequencing technologies: An overview,” *Hum. Immunol.*, vol. 82, no. 11, pp. 801–811, Nov. 2021.

[70] M. T. Pervez, M. J. U. Hasnain, S. H. Abbas, M. F. Moustafa, N. Aslam, and S. S. M. Shah, “A Comprehensive Review of Performance of Next-Generation Sequencing Platforms,” *Biomed Res. Int.*, vol. 2022, 2022.

[71] T. J. Treangen and S. L. Salzberg, “Repetitive DNA and next-generation sequencing: computational challenges and solutions,” *Nat. Rev. Genet. 2011 131*, vol. 13, no. 1, pp. 36–46, Nov. 2011.

[72] H. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb, “Zero-mode waveguides for single-molecule analysis at high concentrations,” *Science (80-.).*, vol. 299, no. 5607, pp. 682–686, Jan. 2003.

[73] J. Eid *et al.*, “Real-time DNA sequencing from single polymerase molecules,” *Science (80-.).*, vol. 323, no. 5910, pp. 133–138, Jan. 2009.

[74] “Bioinformatics Training at the Harvard Chan Bioinformatics Core.” [Online]. Available: https://hbctraining.github.io/main/. [Accessed: 19-Apr-2023].

[75] E. Afgan *et al.*, “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update,” *Nucleic Acids Res.*, vol. 46, no. W1, pp. W537–W544, Jul. 2018.

[76] B. Batut *et al.*, “Community-Driven Data Analysis Training for Biology,” *Cell Syst.*, vol. 6, no. 6, p. 752–758.e1, Jun. 2018.

[77] B. Grüning *et al.*, “bgruening/galaxytools: September release 2019,” Sep. 2018.

[78] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, “Assembly of long, error-prone reads using repeat graphs,” *Nat. Biotechnol. 2019 375*, vol. 37, no. 5, pp. 540–546, Apr. 2019.

[79] C. S. Chin *et al.*, “Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data,” *Nat. Methods 2013 106*, vol. 10, no. 6, pp. 563–569, May 2013.

[80] S. Nurk *et al.*, “HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads,” *Genome Res.*, vol. 30, no. 9, pp. 1291–1305, Oct. 2020.

[81] G. Angenon, M. Van Montagu, and A. Depicker, “Analysis of the stop codon context in plant nuclear genes,” vol. 271, no. 1, pp. 144–146, 1990.

[82] A. S. N. Reddy, M. F. Rogers, D. N. Richardson, M. Hamilton, and A. Ben-Hur, “Deciphering the plant splicing code: Experimental and computational approaches for predicting alternative splicing and splicing regulatory elements,” *Front. Plant Sci.*, vol. 3, no. FEB, p. 19249, Feb. 2012.

[83] C. Wei and M. R. Brent, “Using ESTs to improve the accuracy of de novo gene prediction,” *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–10, Jul. 2006.

[84] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990.

[85] B. Buchfink, K. Reuter, and H. G. Drost, “Sensitive protein alignments at tree-of-life scale using DIAMOND,” *Nat. Methods 2021 184*, vol. 18, no. 4, pp. 366–368, Apr. 2021.

[86] P. Jones *et al.*, "InterProScan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, May 2014.

[87] R. D. Finn *et al.*, "The Pfam protein families database: towards a more sustainable future.," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279-85, Dec. 2015.

[88] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[89] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, p. 25, May 2000.

[90] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, no. 5338, pp. 631–637, Oct. 1997.

[91] A. X. Dong *et al.*, "High-quality assembly of the reference genome for scarlet sage, Salvia splendens, an economically important ornamental plant," *Gigascience*, vol. 7, no. 7, pp. 1–10, Jul. 2018.

[92] M. L. Gonzalez-Garay, "Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq)," pp. 141–160, 2016.

[93] R. Stark, M. Grzelak, and J. Hadfield, "RNA sequencing: the teenage years," *Nat. Rev. Genet. 2019 2011*, vol. 20, no. 11, pp. 631–656, Jul. 2019.

[94] W. Yasui, N. Oue, R. Ito, K. Kuraoka, and H. Nakayama, "Search for new biomarkers of gastric cancer through serial analysis of gene expression and its clinical implications," *Cancer Sci.*, vol. 95, no. 5, pp. 385–392, May 2004.

[95] K. Hibbs *et al.*, "Differential Gene Expression in Ovarian Carcinoma: Identification of Potential Biomarkers," *Am. J. Pathol.*, vol. 165, no. 2, pp. 397–414, Aug. 2004.

[96] R. Rodriguez-Esteban and X. Jiang, "Differential gene expression in disease: A comparison between high-throughput studies and the literature," *BMC Med. Genomics*, vol. 10, no. 1, pp. 1–10, Oct. 2017.

[97] J. Bouquet *et al.*, "Longitudinal transcriptome analysis reveals a sustained differential gene expression signature in patients treated for acute Lyme disease," *MBio*, vol. 7, no. 1, Feb. 2016.

[98] J. Cooper-Knock, J. Kirby, L. Ferraiuolo, P. R. Heath, M. Rattray, and P. J. Shaw, "Gene expression profiling in human neurodegenerative disease," *Nat. Rev. Neurol. 2012 89*, vol. 8, no. 9, pp. 518–530, Aug. 2012.

[99] Y. Xie *et al.*, "Transcriptome-based gene profiling provides novel insights into the characteristics of radish root response to Cr stress with next-generation sequencing," *Front. Plant Sci.*, vol. 6, p. 202, Mar. 2015.

[100] P. W. Inglis, R. P. Marilia de Castro, L. V. Resende, and D. Grattapaglia, "Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications," *PLoS One*, vol. 13, no. 10, p. e0206085, Oct. 2018.

[101] A. Healey, A. Furtado, T. Cooper, and R. J. Henry, "Protocol: A simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species," *Plant Methods*, vol. 10, no. 1, pp. 1–8, Jun. 2014.

[102] S. O. Rogers and A. J. Bendich, "Extraction of total cellular DNA from plants, algae and fungi,"

Plant Mol. Biol. Man., pp. 183–190, 1994.

[103] "T042-TECHNICAL BULLETIN NanoDrop Spectrophotometers."

[104] "GitHub - PacificBiosciences/pbipa: Improved Phased Assembler." [Online]. Available: https://github.com/PacificBiosciences/pbipa. [Accessed: 11-Dec-2022].

[105] "GitHub - dfguan/purge_dups: haplotypic duplication identification tool." [Online]. Available: https://github.com/dfguan/purge_dups. [Accessed: 11-Dec-2022].

[106] K. H. Jia *et al.*, "Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (Lamiaceae) genome," *Hortic. Res. 2021 81*, vol. 8, no. 1, pp. 1–15, Sep. 2021.

[107] D. D. Kang *et al.*, "MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 2019, no. 7, p. e7359, Jul. 2019.

[108] M. Seppey, M. Manni, and E. M. Zdobnov, "BUSCO: Assessing Genome Assembly and Annotation Completeness," *Methods Mol. Biol.*, vol. 1962, pp. 227–245, 2019.

[109] E. V. Kriventseva *et al.*, "OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D807–D811, Jan. 2019.

[110] M. Manni, M. R. Berkeley, M. Seppey, and E. M. Zdobnov, "BUSCO: Assessing Genomic Data Quality and Beyond," *Curr. Protoc.*, vol. 1, no. 12, Dec. 2021.

[111] M. Manni, M. R. Berkeley, M. Seppey, F. A. Sim~ Ao, and E. M. Zdobnov, "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes," *Mol. Biol. Evol*, 2021.

[112] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, Mar. 2011.

[113] G. W. Vurture *et al.*, "GenomeScope: fast reference-free genome profiling from short reads," *Bioinformatics*, vol. 33, no. 14, pp. 2202–2204, Jul. 2017.

[114] T. R. Ranallo-Benavidez, K. S. Jaron, and M. C. Schatz, "GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes," *Nat. Commun. 2020 111*, vol. 11, no. 1, pp. 1–10, Mar. 2020.

[115] T. Brůna, K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, "BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database," *NAR Genomics Bioinforma.*, vol. 3, no. 1, pp. 1–11, Jan. 2021.

[116] M. Stanke, O. Schöffmann, B. Morgenstern, and S. Waack, "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–11, Feb. 2006.

[117] K. J. Hoff and M. Stanke, "Predicting Genes in Single Genomes with AUGUSTUS," *Curr. Protoc. Bioinforma.*, vol. 65, no. 1, Mar. 2019.

[118] K. J. Hoff, A. Lomsadze, M. Borodovsky, and M. Stanke, "Whole-Genome Annotation with BRAKER," *Methods Mol. Biol.*, vol. 1962, p. 65, 2019.

[119] C. Camacho *et al.*, "BLAST+: Architecture and applications," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–9, Dec. 2009.

[120] M. Blum *et al.*, "The InterPro protein families and domains database: 20 years on.," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D344–D354, Jan. 2021.

[121] J. Huerta-Cepas *et al.*, “Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper,” *Mol. Biol. Evol.*, vol. 34, no. 8, pp. 2115–2122, Aug. 2017.

[122] J. Huerta-Cepas *et al.*, “eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses,” *Nucleic Acids Res.*, vol. 47, no. Database issue, p. D309, Jan. 2019.

[123] J. P. Allen, *The Art of Medicine in Ancient Egypt - James P. Allen, Metropolitan Museum of Art (New York, N.Y.) - Google Books*, 2nd ed., vol. 1. New York: The Metropolitan Museum of Art, 2005.

[124] A. M. Metwaly *et al.*, “Traditional ancient Egyptian medicine: A review,” *Saudi J. Biol. Sci.*, vol. 28, no. 10, pp. 5823–5832, Oct. 2021.

[125] F. J. Carod-Artal, “Psychoactive plants in ancient Greece,” *Neurosci. Hist.*, vol. 1, no. 1, pp. 28–38, 2013.

[126] M. He *et al.*, “The NCI Library of Traditional Chinese Medicinal Plant Extracts – Preliminary Assessment of the NCI-60 Activity and Chemical Profiling of Selected Species,” *Fitoterapia*, vol. 137, p. 104285, Sep. 2019.

[127] “Call for papers - Medicinal plant genomics.” [Online]. Available: https://www.biomedcentral.com/collections/MPG. [Accessed: 23-Jun-2023].

[128] Q. Q. Cheng *et al.*, “Review on the Development and Applications of Medicinal Plant Genomes,” *Front. Plant Sci.*, vol. 12, p. 2981, Dec. 2021.

[129] M. M. Alami *et al.*, “The Current Developments in Medicinal Plant Genomics Enabled the Diversification of Secondary Metabolites’ Biosynthesis,” *Int. J. Mol. Sci.*, vol. 23, no. 24, p. 15932, Dec. 2022.

[130] R. Lauchli *et al.*, “High-Throughput Screening for Terpene-Synthase-Cyclization Activity and Directed Evolution of a Terpene Synthase,” *Angew. Chemie Int. Ed.*, vol. 52, no. 21, pp. 5571–5574, May 2013.

[131] N. G. H. Leferink *et al.*, “An automated pipeline for the screening of diverse monoterpene synthase libraries,” *Sci. Rep.*, vol. 9, no. 1, Dec. 2019.

[132] G. Naseri, “A roadmap to establish a comprehensive platform for sustainable manufacturing of natural products in yeast,” *Nat. Commun. 2023 141*, vol. 14, no. 1, pp. 1–13, Apr. 2023.

[133] Y. Wei, B. Ji, R. Ledesma-Amaro, T. Chen, and X. J. Ji, “Editorial: Engineering Yeast to Produce Plant Natural Products,” *Front. Bioeng. Biotechnol.*, vol. 9, p. 798097, Dec. 2021.

[134] D. Romero-Suarez, J. D. Keasling, and M. K. Jensen, “Supplying plant natural products by yeast cell factories,” *Curr. Opin. Green Sustain. Chem.*, vol. 33, p. 100567, Feb. 2022.

[135] T. L. Sivy, R. Fall, and T. N. Rosenstiel, “Evidence of Isoprenoid Precursor Toxicity in Bacillus subtilis,” *Biosci. Biotechnol. Biochem.*, vol. 75, no. 12, pp. 2376–2383, Dec. 2011.

[136] C. A. Espinosa-Leal, C. A. Puente-Garza, and S. García-Lara, “In vitro plant tissue culture: means for production of biological active compounds,” *Planta*, vol. 248, no. 1, p. 1, Jul. 2018.

[137] B. Lin and Y. Tao, “Whole-cell biocatalysts by design,” *Microb. Cell Fact.*, vol. 16, no. 1, pp. 1–12, Jun. 2017.

[138] Y. Cao *et al.*, “Manipulation of the precursor supply for high-level production of longifolene by metabolically engineered Escherichia coli,” *Sci. Reports 2019 91*, vol. 9, no. 1, pp. 1–10, Jan.

2019.

[139] T. W. Giessen and P. A. Silver, “Encapsulation as a Strategy for the Design of Biological Compartmentalization,” *J. Mol. Biol.*, vol. 428, no. 5, pp. 916–927, Feb. 2016.

[140] X. Zhu *et al.*, “Synthetic biology of plant natural products: From pathway elucidation to engineered biosynthesis in plant cells,” *Plant Commun.*, vol. 2, no. 5, Sep. 2021.

[141] S. M. Pearsall, C. N. Rowley, and A. Berry, “Advances in Pathway Engineering for Natural Product Biosynthesis,” *ChemCatChem*, vol. 7, no. 19, pp. 3078–3093, Oct. 2015.

[142] S. Anami, E. Njuguna, G. Coussens, S. Aesaert, and M. Van Lijsebettens, “Higher plant transformation: principles and molecular tools,” *Int. J. Dev. Biol.*, vol. 57, no. 6-7–8, pp. 483–494, Sep. 2013.

[143] S. B. Gelvin, “Agrobacterium-Mediated Plant Transformation: the Biology behind the ‘Gene-Jockeying’ Tool,” *Microbiol. Mol. Biol. Rev.*, vol. 67, no. 1, p. 16, Mar. 2003.

[144] A. Cravens, J. Payne, and C. D. Smolke, “Synthetic biology strategies for microbial biosynthesis of plant natural products,” *Nat. Commun. 2019 101*, vol. 10, no. 1, pp. 1–12, May 2019.

[145] H. Schäfer and M. Wink, “Medicinally important secondary metabolites in recombinant microorganisms or plants: Progress in alkaloid biosynthesis,” *Biotechnol. J.*, vol. 4, no. 12, pp. 1684–1703, Dec. 2009.

[146] K. T. Watts, B. N. Mijts, and C. Schmidt-Dannert, “Current and Emerging Approaches for Natural Product Biosynthesis in Microbial Cells,” *Adv. Synth. Catal.*, vol. 347, no. 7–8, pp. 927–940, Jun. 2005.

[147] L. Zhao *et al.*, “From Plant to Yeast—Advances in Biosynthesis of Artemisinin,” *Molecules*, vol. 27, no. 20, Oct. 2022.

[148] B. Engels, P. Dahm, and S. Jennewein, “Metabolic engineering of taxadiene biosynthesis in yeast as a first step towards Taxol (Paclitaxel) production,” *Metab. Eng.*, vol. 10, no. 3–4, pp. 201–206, May 2008.

[149] B. Nowrouzi *et al.*, “Enhanced production of taxadiene in Saccharomyces cerevisiae,” *Microb. Cell Fact.*, vol. 19, no. 1, pp. 1–12, Dec. 2020.

[150] T. Wang *et al.*, “The Human Pangenome Project: a global resource to map genomic diversity,” *Nat. 2022 6047906*, vol. 604, no. 7906, pp. 437–446, Apr. 2022.

[151] B. C. Weimer, “100K Pathogen Genome Project,” *Genome Announc.*, vol. 5, no. 28, 2017.

[152] S. Cheng *et al.*, “10KP: A phylodiverse genome sequencing plan,” *Gigascience*, vol. 7, no. 3, pp. 1–9, Mar. 2018.

[153] A. Rhie *et al.*, “Towards complete and error-free genome assemblies of all vertebrate species,” *Nat. 2021 5927856*, vol. 592, no. 7856, pp. 737–746, Apr. 2021.

[154] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021.

[155] G. A. Strobel, “BACTERIAL PHYTOTOXINS,” *Annu. Rev. Microbiol.*, vol. 31, pp. 205–224, Nov. 2003.

[156] A. Jongejan, C. de Graaf, N. P. E. Vermeulen, R. Leurs, and I. J. P. de Esch, “The role and application of in silico docking in chemical genomics research.,” *Methods Mol. Biol.*, vol. 310, pp. 63–91, 2005.

[157] O. Trott and A. J. Olson, “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading,” *J. Comput. Chem.*, vol. 31, no. 2, p. 455, Jan. 2010.

[158] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, “AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings,” *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 3891–3898, Aug. 2021.

Good luck is a residue of preparation.

| Jack Youngblood