CHAPTER THREE

Genomics-Based Discovery of Plant Genes for Synthetic Biology of Terpenoid Fragrances: A Case Study in Sandalwood oil Biosynthesis

J.M. Celedon, J. Bohlmann[1]

Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada
[1]Corresponding author: e-mail address: bohlmann@msl.ubc.ca

Contents

Abstract

Terpenoid fragrances are powerful mediators of ecological interactions in nature and have a long history of traditional and modern industrial applications. Plants produce a great diversity of fragrant terpenoid metabolites, which make them a superb source of biosynthetic genes and enzymes. Advances in fragrance gene discovery have enabled new approaches in synthetic biology of high-value speciality molecules toward

Methods in Enzymology, Volume 576
ISSN 0076-6879
http://dx.doi.org/10.1016/bs.mie.2016.03.008

applications in the fragrance and flavor, food and beverage, cosmetics, and other industries. Rapid developments in transcriptome and genome sequencing of nonmodel plant species have accelerated the discovery of fragrance biosynthetic pathways. In parallel, advances in metabolic engineering of microbial and plant systems have established platforms for synthetic biology applications of some of the thousands of plant genes that underlie fragrance diversity. While many fragrance molecules (eg, simple monoterpenes) are abundant in readily renewable plant materials, some highly valuable fragrant terpenoids (eg, santalols, ambroxides) are rare in nature and interesting targets for synthetic biology. As a representative example for genomics/transcriptomics enabled gene and enzyme discovery, we describe a strategy used successfully for elucidation of a complete fragrance biosynthetic pathway in sandalwood (*Santalum album*) and its reconstruction in yeast (*Saccharomyces cerevisiae*). We address questions related to the discovery of specific genes within large gene families and recovery of rare gene transcripts that are selectively expressed in recalcitrant tissues. To substantiate the validity of the approaches, we describe the combination of methods used in the gene and enzyme discovery of a cytochrome P450 in the fragrant heartwood of tropical sandalwood, responsible for the fragrance defining, final step in the biosynthesis of (*Z*)-santalols.

1. INTRODUCTION

Terpenoids constitute one of the largest and most diverse classes of plant specialized metabolites, with a wide range of chemoecological functions (Gershenzon & Dudareva, 2007) and industrial applications (Arendt, Pollier, Callewaert, & Goossens, 2016; Bohlmann & Keeling, 2008). Industrial applications of specialized plant terpenoids include medicinal compounds such as the anticancer drug taxol (Croteau, Ketchum, Long, Kaspera, & Wildung, 2006) and the antimalarial agent artemisin (Paddon, Westfall, Pitera, et al., 2013), as well as fragrance compounds such as ambroxides (Zerbe & Bohlmann, 2015a) or santalols (Jones et al., 2011). Terpene fragrances are produced in many different plant tissues and in all of the major plant organs. For example, terpenes represent the largest class of floral volatiles (Knudsen, Eriksson, Gershenzon, & Ståhl, 2006; Muhlemann, Klempien, & Dudareva, 2014). Flowers of species, such as snapdragon, petunia, or roses, have been explored for the discovery of fragrance genes (Dudareva et al., 2003; Magnard, Roccia, Caissard, et al., 2015; Qualley, Widhalm, Adebesin, Kish, & Dudareva, 2012), while organs such as roots, rhizomes, or mature heartwood represent another less explored and valuable source of such genes. Biosynthesis of terpenoid fragrances may occur in highly specialized cell types, such as floral glandular trichomes in lavender (Demissie et al., 2012; Demissie, Erland, Rheault, & Mahmoud,

2013) or clary sage (Caniard et al., 2012), or may be present in epidermal or parenchymatic tissues. The sequestration and release of terpenoid fragrances are thought to involve both active and passive processes (Widhalm, Jaini, Morgan, & Dudareva, 2015).

The chemical diversity of terpenoid fragrance compounds is the results of a modular pathway system that is present with variations in all plant species (Chen, Tholl, Bohlmann, & Pichersky, 2011). Terpenoid precursors of 10, 15, or 20 carbon atoms, namely, geranyl diphosphate (GPP, C_{10}), neryl diphosphate (NPP, C_{10}), farnesyl diphosphate (FPP, C_{15}), geranylgeranyl diphosphate (GGPP, C_{20}), and nerylneryl diphosphate (NNPP, C_{20}), are assembled by prenyltransferases from the two isomeric 5-carbon building blocks isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). IPP and DMAPP are produced in the mevalonic acid (MEV) pathway, which is located across the cytosol, peroxisome, and endoplasmic reticulum, or in the plastidial methylerythritol phosphate (MEP) pathway. DMAPP, GPP, NPP, FPP, GGPP, and NNPP are the substrates of terpene synthases (TPSs) that produce acyclic and cyclic terpene olefins and alcohols, which may undergo additional regio- and stereo-specific oxidations, peroxidations, acylations, methylations, or glycosylations. The many different variations of producing terpenoid scaffolds by TPSs, combined with subsequent modification by other enzymes, most commonly involving cytochrome P450-dependent reactions, result in the many different terpenoid fragrance profiles of different plant species. In most plant species, TPS and P450 enzymes are encoded by large gene families, and the many possible combinations of these enzymes represent a nearly unlimited resource for the natural biosynthesis, and synthetic biology, of known and novel terpenoid molecules as demonstrated, for example, for the diterpenoids (Andersen-Ranberg, Kongstad, Nielsen, et al., 2016; Zerbe & Bohlmann, 2015b; Zerbe, Hamberger, Yuen, et al., 2013).

Here, we describe general approaches for the discovery and characterization of terpenoid fragrance genes and enzymes in nonmodel plants, which are defined here as plant species that have traditionally not been easily accessible with molecular and genetic tools, and therefore have been difficult to explore for gene discovery. Most of the world's plant species that are of interest for fragrances are nonmodel species, and the same is true for most medicinal plants. Some of these species, such as sandalwood (*Santalum* ssp.), may have been overexploited for extraction of terpenoids. High-throughput DNA sequencing, most commonly applied for transcriptome sequencing, has made nonmodel plant species accessible to gene discovery,

and by extension has enabled new opportunities for synthetic biology of specialized plant metabolites, including fragrances (Arendt et al., 2016; Facchini, Bohlmann, Covello, et al., 2012; Zerbe & Bohlmann, 2015a). In general, for nonmodel plant systems, gene discovery by transcriptome sequencing has substantial advantages over genome sequencing. Specifically, transcriptomes are generally less costly to produce and easier to assemble than genome sequences, sequence coverage of transcriptomes is focused on expressed genes and not distributed across excessive nonexpressed or nonprotein-coding genomic regions, and protein-coding open reading frame sequences can be predicted from transcriptome assemblies without consideration of intron/exon structures. To illustrate general concepts and methods with specific examples, we describe the discovery of a novel P450 gene and the encoded enzyme that is critical in the biosynthesis of the fragrance defining sesquiterpenoid alcohol components of sandalwood oil, namely, (*Z*)-α-santalol, (*Z*)-β-santalol, (*Z*)-α-*exo*-bergamotol, and (*Z*)-*epi*-β-santalol (Celedon et al., 2016). The principles described here are broadly applicable to the discovery of specific genes in large gene families, such as TPSs and P450s.

2. PRIOR KNOWLEDGE OF SANDALWOOD TPSs AND P450S AND DEVELOPMENT OF A HYPOTHESIS

The first report of sandalwood (*Santalum album*) TPS gene discovery (Jones et al., 2008) used a PCR cloning strategy with degenerate primers targeting conserved TPS motifs followed by RACE to recover full-length cDNAs. This approach resulted in the cDNA cloning and functional characterization of the *SaMonoTPS1* and *SaSesquiTPS1* genes that produced, respectively, (+)-α-terpineol and (−)-limonene, and germacrene D-4-ol and helminthogermacrene as the main products. In a subsequent and different approach, cDNAs of TPS genes encoding santalene synthases (SSy) were discovered in three different *Santalum* species, *S. album* (*SaSSy*), *S. spicatum* (*SpiSSy*), and *S. austrocaledonicum* (*SauSSy*) (Jones et al., 2011). The breakthrough discovery of SaSSy was based on candidate TPS identification in Sanger-sequenced EST libraries from *S. album* wood cores followed by functional characterization, which identified SaSSy as a multiproduct sesquiterpene synthase that produces α-santalene, β-santalene, *epi*-β-santalene, and α-*exo*-bergamotene. These sesquiterpene olefins are the specific precursors for the major sesquiterpene alcohols of sandalwood oil, (*Z*)-α-santalol,

(*Z*)-β-santalol, (*Z*)-*epi*-β-santalol, and (*Z*)-α-*exo*-bergamotol (Jones et al., 2011).

The presence of a hydroxyl group at the C12 position of α-santalol, β-santalol, *epi*-β-santalol, and α-*exo*-bergamotol was indicative of P450-dependent oxidation of SaSSy products. Mining the same Sanger EST library and additional 454 sequences for P450 candidates of the CYP71 clan, Diaz-Chavez, Moniodis, Madilao, et al. (2013) identified nine different members of the sandalwood CYP76F subfamily, which were expressed in baker's yeast (*Saccharomyces cerevisiae*) and shown to produce predominantly the (*E*) stereoisomers of α-santalol, β-santalol, *epi*-β-santalol, and α-*exo*-bergamotol (Diaz-Chavez et al., 2013). Since sandalwood oil is mostly composed of the opposite (*Z*) stereoisomers of these sesquiterpene alcohols, it was hypothesized that the ability of P450s to hydroxylate α-santalene, β-santalene, *epi*-β-santalene, and α-*exo*-bergamotene evolved in different subfamilies of the P450 gene family in *S. album* and that different P450s preferentially produce either the (*Z*)- or the (*E*)-stereoisomers of α-santalol, β-santalol, *epi*-β-santalol, and α-*exo*-bergamotol. It was further hypothesized that transcripts of P450s that produce the (*Z*) isomers would be coexpressed (ie, similar spatial and temporal patterns of gene expression) with the early santalol pathway genes, including SaSSy, FPP synthase, and the MEV pathway.

With the goal to discover the elusive P450 enzyme involved in (*Z*) selective sesquiterpene oxidation, these hypotheses were tested by tissue-specific RNA-seq analysis across the gradient of developing *S. album* wood tissues, including the outer sapwood (SW), the intermediate transition zone (TZ), and the inner oil-accumulating heartwood (HW). To rule out possible temporal or developmental separation of the TPS and P450 steps in the biosynthesis of sandalwood oil, targeted sesquiterpene metabolite analysis was performed to track pathway intermediates and final products in each tissue (Celedon et al., 2016).

3. REPLICATION, SAMPLING, AND STATISTICAL DESIGN

Replication and experimental conditions are critical for statistical analysis of variation for RNA-seq-based gene discovery if comparison of multiple conditions and treatments is key to the analysis. The selection of an appropriate number of biological replicates and the control of sources of variability (eg, intraspecific genetic variation, tissue type and developmental variation, environmental and life-history variation) are important aspects

when analyzing transcriptomes of plants, as many different factors may influence gene expression as well as metabolite biosynthesis and accumulation. This is of particular importance when working with nonmodel plant species collected from natural populations or grown under field conditions. A common observation, although not always the case, is that a high abundance of a metabolite of interest (eg, santalols) in a given tissue is correlated with, or preceded in time by, a high abundance of the corresponding biosynthetic transcripts. Therefore, comparison of transcriptomes of tissues that are contrasted for low and high levels of metabolite accumulation may lead to successful gene discovery via identification of differentially expressed transcripts. The general objective of the statistical design of tissue sampling is to minimize and control sources of variability (noise) and maximize the chances of capturing the natural and developmental variation of gene expression in fragrance biosynthesis.

The sandalwood trees in our case study were all of the same age of 15 years grown in a field plantation in the Kununurra area in Northern Australia. Six trees grown under the same conditions and exhibiting a similar visual phenotype (ie, foliage, trunk diameter, no signs of disease, or major herbivore damage) were selected for tissue sampling, and all samples were kept separate throughout the analyses to maintain proper biological replication and allow for statistical assessment of variations between samples.

4. DEFINING TEMPORAL AND SPATIAL VARIABLES FOR TISSUE SAMPLING

Given their roles in plant ecological interactions with other organisms, the biosynthesis (and transcript expression) of fragrance molecules may be restricted in space to specialized cell types, specific tissues, or organs. Biosynthesis may also be restricted in time to specific developmental stages, time of day, or response to external stimuli. Knowledge of the biological system of interest therefore can be critical in defining the relevant samples for transcriptome-based discovery of fragrance biosynthetic genes. At the minimum, metabolite profiling of different tissues and developmental stages is recommended to identify temporal and spatial patterns, as well as conditions, under which biosynthetic pathway genes are mostly likely to be expressed.

In sandalwood trees, the fragrant oil accumulates in the inner HW tissue of stems and roots, typically at a tissue age of 15 years and older. Contributions of the outer and younger SW, as well as the TZ (the tissue between SW and HW) to oil formation were previously not well characterized. For these

reasons, we sampled all three tissues to assess their sesquiterpene content and composition and to gain insight into the potential differences with regard to gene expression of sesquiterpene biosynthesis.

5. TISSUE SAMPLING

The logistics of tissue sampling for RNA isolation range from straight forward, if plants can be grown in growth chambers or greenhouses in a laboratory or research facility; or they may require major planning and large efforts, if plants have to be accessed in remote field sites. Commercial reagents are available that may allow for less stringent conditions of sample handling. However, to preserve RNA quality for transcriptome sequencing we consider it essential that all instruments and reagents that come in contact with the plant tissue are treated to be RNase free and that tissues are immediately flash-frozen in liquid nitrogen, followed by cryogenic transportation to the laboratory. Storage of samples at −80°C is recommended especially if samples are not immediately extracted. Dry ice or nitrogen dry shippers are frequently used for temporary storage and transportation of field-collected samples. Special care must be taken when sampling involves procedures that damage cell integrity, as this increases exposure to RNA degrading enzymes.

Sampling of sandalwood tissues was performed in a remote tropical plantation, requiring prior shipment of liquid nitrogen cooled dry shippers over 3000 km by airplane and return shipment of samples in dry shippers back to the laboratory. Wood tissues were sampled from the stem of living sandalwood trees. The high value of these trees required nondestructive and minimally invasive sampling, which was performed with a manual drill early in the morning to avoid the high midday temperatures of Northern Australia. Trees were sampled at 50 cm height from the base of the trunk. Samples from the outer SW, intermediate TZ, and inner HW were visually separated by color (SW: white/yellow; TZ: pink/red; HW: red/dark-brown) as they were collected and immediately frozen in liquid nitrogen.

6. METABOLITE PROFILING

Metabolite profiles are often used with one or both of two main goals in a gene discovery study, both upstream and downstream of transcriptome sequencing: First, to select the appropriate tissue and conditions for sampling and second, to identify candidate genes based on correlations between gene expression patterns and patterns of metabolite distributions across tissues and

conditions. Volatile fragrance compounds are generally profiled by gas chromatography–mass spectrometry (GC–MS) using liquid extractions from isolated tissue samples or headspace analysis from intact living plants or dissected plant samples. Solvent extraction of tissues allows quantitative analysis of small amounts of isolated tissues, providing a desired resolution of tissue- or even cell-type-specific metabolite profiles. Solvent extraction typically results in complex metabolite profiles, requiring additional data analysis efforts for identification of target compounds in the recorded GC–MS data. Cell-type-specific analysis can be achieved by isolation of surface cells such as those of glandular or nonglandular trichomes and epidermis (Lange, 2016), or by laser microdissection of inner tissues that accumulate terpenes (Abbott, Hall, Hamberger, & Bohlmann, 2010; Hamberger, Ohnishi, Hamberger, Séguin, & Bohlmann, 2011). Headspace analysis of the volatile emissions allows monitoring of fragrances emitted from whole organs (ie, flowers, leaves) or intact plants, making possible, for example, to measure temporal patterns of volatile emission profiles or changes in response to external stimuli. Metabolites represented in GC–MS profiles are identified by comparison of their mass spectra with reference mass spectral libraries, comparison of retention index with reference data, and whenever possible by comparison of mass spectra and retention indices with those of authentic standards.

In our case study, sandalwood metabolites were extracted from four biological replicates for each tissue type (SW, TZ, HW) in technical duplicates using ~50 mg of ground tissue with 1 mL of pentane spiked with isobutyl benzene (0.1 μg mL^{-1}) as internal standard. Extractions were performed in glass vials by end-over-end mixing for 24 h at room temperature (RT). Samples were centrifuged at $1000 \times g$ for 15 min, and the pentane phase was transferred to a new GC vial. GC–MS analysis of sesquiterpenoids was performed on an Agilent 7890A/5975C GC–MS system operating in electron ionization selected ion monitoring (SIM) scan mode and equipped with a DB-Wax fused silica column (30 m, 250 μm ID, 0.25-μm film thickness). The conditions for a typical analysis of sandalwood sesquiterpenoids were: Injector was operated in pulsed splitless mode with the injector temperature kept at 250°C. Helium was used as the carrier gas with a flow rate of 0.8 mL min^{-1} and pulsed pressure set at 25 psi for 0.5 min. Scan range: m/z 40–500; SIM: m/z 93, 94, 105, 107, 119, 122, and 202 (dwell time 50 ms). The oven program was: 40°C for 3 min; ramp of 10°C min^{-1} to 130°C, ramp of 2°C min^{-1} to 200°C, ramp of 50°C min^{-1} to 250°C; hold at 250°C for 15 min. ChemStation software was used for data acquisition

and processing. Metabolite identification was done by comparison of mass spectra with the NIST/EPA/NIH mass spectral library v2.0. Relative quantities of metabolites were calculated by manual integration of peak areas followed by normalization to the internal standard and dry weight of the tissue used in each extraction.

7. ISOLATION OF HIGH-QUALITY RNA FROM RECALCITRANT TISSUES

Challenges associated with isolation of high-quality RNA are highly dependent on the nature of the plant material. Comparison of different RNA isolation protocols, including available commercial kits, is highly recommended to achieve optimal results. Plant tissues containing high amounts of phenolic metabolites are particularly challenging for isolation of RNA with standard methods. A CTAB-based method (Kolosova et al., 2004) was specifically developed to isolate RNA from plant tissue with high content of polysaccharides, phenolics, and terpenoids. This protocol has been tested in recalcitrant tissues yielding high-quality RNA. PureLink® Plant RNA Reagent (Invitrogen, USA) is a fast, simple, and robust method to isolate RNA from most plant tissues yielding large amounts of high-quality RNA from small amounts of tissue (100 mg). Depending on the tissue of interest, this protocol can be modified to find an optimal ratio of reagent to tissue. The quality of isolated RNA should be assessed with quantitative methods such as the RIN number (Schroeder, Mueller, Stocker, et al., 2006) obtained with an Agilent 2100 Bioanalyzer or the RQI index obtained with an Experion™ Bio-Rad. Regardless of the method used, RNA samples should be analyzed at the temperature at which cDNA synthesis will be performed to represent the quality of the input RNA used for library construction.

The SW, TZ, and HW samples of sandalwood represent extremely difficult tissues for RNA isolation, since much of the material consists of highly lignified dead cells making up the wood core of the trunk of the tree, plus phenolic parenchyma cells. Frozen sandalwood tissue was ground to a fine powder for RNA extraction using PureLink® Plant RNA Reagent. Glycogen was added in the final precipitation step at a final concentration of 3.3 ng μL^{-1} to increase stability of the RNA pellet. Given that only very small amounts of RNA could be extracted from HW tissue, the addition of glycogen was also important as it made the final RNA pellet apparent by eye, which reduced the risk of accidentally discarding the RNA.

RNA quality and concentration were assessed on a 2100 Bioanalyzer after incubation of the isolated RNA at 70°C for 2 min as required in the Illumina cDNA library preparation protocol.

8. TRANSCRIPTOME SEQUENCING AND DE NOVO ASSEMBLY

The reduced costs for DNA sequencing and reduced amounts of RNA required for library construction have made transcriptome sequencing a common approach for gene discovery of secondary metabolism in non-model plant species (Facchini et al., 2012). The use of microfluidics in the preparation of cDNA libraries has reduced the amounts of nucleic acid required to as low as 25 ng of total RNA (NeoPrep™ System, Illumina), which makes it possible to obtain transcriptomes from small samples, such as those obtained by laser microdissection of individual cell types, or isolated trichomes and secretory glands. Removal of ribosomal RNA prior to library preparation improves the quality of the downstream transcriptome sequence assembly and increases sequence coverage of low-abundance transcripts. Presently, one of the most commonly used sequencing platforms used for transcriptome sequencing is the Illumina platform due to competitive per-base cost, sequence quality, and coverage.

To sequence the transcriptomes of different wood samples of sandalwood, we aimed for deep coverage that would allow us not only to identify expressed biosynthetic genes but also transcription factors and other low-abundance transcripts with potential regulatory functions in the developmental progression from SW to HW. cDNA libraries were constructed after removal of ribosomal RNA and ligation of strand-specific adapters. Paired-end (PE) reads, 100 bp long with an average insert size of 300 bases, were chosen for sequencing. Libraries were multiplexed, three libraries per sequencing lane, with the goal to achieve a sequencing depth of ~66 million reads per library. For reference, an RNA-seq study in *Arabidopsis* showed that 50 million reads represented a near-saturation coverage of expressed genes (Van Verk, Hickman, Pieterse, & Wees, 2013). Read quality of sandalwood RNA-seq data was inspected with FASTQC. Bases with a quality lower than 3 and adapter sequences were removed with Trimmomatic (Lohse et al., 2012). Overlapping PE-reads were merged with BBmerge (BBmap software; http://sourceforge.net/projects/bbmap/) improving the quality of the assembly and reducing time and computational resources required. Merged and unmerged reads from 12 libraries (four

independent biological replicates each for SW, TZ, and HW) were combined and assembled with Trinity (Haas, Papanicolaou, Yassour, et al., 2013). Other de novo assemblers including Trans-ABySS (Robertson, Schein, Chiu, et al., 2010) and Velvet-Oases (Schulz, Zerbino, Vingron, & Birney, 2012) should also be considered. Detailed comparisons of assemblers can be found in several excellent review articles (Martin & Wang, 2011; Zhao et al., 2011). Prediction of protein-coding transcripts was done with TransDecoder (http://transdecoder.github.io/). To assess the quality of a de novo transcriptome assembly, consideration should be given to a number of parameters including: transcript median length, number of predicted peptides and peptide average length, and the representation of the Core Eukaryotic Genes Mapping Approach (CEGMA) with *Arabidopsis* sequences (Parra, Bradnam, & Korf, 2007). Specifically for the quality assessment of sandalwood transcriptome assemblies as a resource for discovery of terpenoid pathway genes, we also assessed the representation of the core terpenoid biosynthetic genes of the MEP and MEV pathways.

9. TRANSCRIPTOME MINING AND ANNOTATION

Using the catalogue of predicted protein-coding sequences in the transcriptome assemblies, transcripts are initially annotated based on their similarity to other annotated genes in databases such as those of the NCBI and EMBL. In addition, for discovery of genes of biosynthetic pathways, expert annotation of extracted gene sets has proven extremely valuable. This latter approach requires biochemical knowledge of metabolic pathways, enzymes, and gene families. For example, Zerbe et al. (2013) developed expert reference databases used for the annotation of TPS and P450 genes of terpenoid biosynthesis in transcriptomes of nonmodel plant species allowing for the focused and rapid discovery of members of these two large gene families in newly established transcriptomes and genomes. This approach can be extended to other gene families commonly involved in terpenoid and other secondary metabolism, such as glycosyl transferases, acyltransferases, reductases, or transporters. Additional whole-transcriptome annotations can be performed using tools such as Blast2go (Conesa et al., 2005) or the Trinotate software suite (http://trinotate.sourceforge.net/) of the Trinity assembler. When appropriate, annotated genes of gene families of interest can be further classified into subfamilies or classes by phylogenetic analysis or presence of subfamily-specific motifs (Zerbe et al., 2013). A priori prediction of specific gene functions based on sequence homology alone, ie,

prediction of specific substrates or products of enzymes, is not currently possible for TPS and P450 gene family members. This is mostly due to the fact that as a feature of divergent evolution in terpenoid secondary metabolism, minor variations in sequence of duplicated genes can have major effects on enzyme functions (Chen et al., 2011; Nelson & Werck-Reichhart, 2011).

The combined sandalwood SW, TZ, and HW transcriptomes were annotated for candidate P450s and TPS using expert curated databases (Zerbe et al., 2013). Sandalwood P450s longer than 400 aa were tentatively assigned to CYP families and subfamilies by phylogenetic analysis and bidirectional best BLAST hit. All genes in the MEP and MEV pathway were annotated using *Arabidopsis* reference sequences.

10. EXPRESSION ANALYSIS AND CANDIDATE GENE SELECTION

Differential expression (ie, transcript abundance) across tissues, developmental stages, or under different conditions can provide critical information to identify candidate genes associated with the biosynthesis of fragrance compounds. Methods to quantify gene expression based on kmer counting algorithms provide a rapid and resource-efficient approach for RNA-seq data (Patro, Mount, & Kingsford, 2014). This alignment-free method relies on creating an index of kmers in the assembled transcripts followed by counting of kmers in the raw reads. To detect statistically significant differences in gene expression, several R packages, including EdgeR and DESeq, have been developed and specifically adapted to analyze RNA-seq data (Anders & Huber, 2010; Robinson, McCarthy, & Smyth, 2010). Annotated transcripts and their expression can be used to create pathway expression maps that provide a broader picture of transcriptome dynamics across general and secondary metabolism. Similarly, all transcripts annotated in a gene family of interest can be used to generate family-specific expression maps. Unbiased analysis of gene expression data, such us hierarchical clustering, can be used to identify groups of genes and tissues that have similar overall expression patterns and include new gene families for further analysis and characterization. Candidate gene selection is accomplished by overlaying expression data with annotations and metabolite profiles. Top candidate genes should ideally be supported by relevant tissue-specific expression, metabolite profiles, and ontology.

Expression analysis of sandalwood transcripts identified a large number of them being differentially expressed between SW, TZ, and HW. These included transcripts annotated with core terpenoid pathways (eg, MEP

and MEV pathways genes, FPP synthase), TPSs and P450s. Examination of HW-specific P450 transcripts identified members of the CYP736, CYP74, and CYP98 subfamilies as the most highly expressed P450 genes in sandalwood HW. Differential expression of MEP and MEV pathway genes across SW, TZ, and HW was captured in MEP and MEV pathway expression diagrams, revealing an informative developmental transcriptome profile with the MEP pathway (involved in mono- and diterpene biosynthesis) being preferentially expressed in SW, and the MEV pathway (involved in sesquiterpenoid biosynthesis) being preferentially expressed in HW (Celedon et al., 2016). In addition, expression analysis of TPS transcripts showed SaSSy as the overall most highly HW expressed TPS gene with a strong HW specificity of the expression. The combined and integrated analysis of metabolite profiles and gene expression identified the HW transcriptome (but not SW and TZ) as the target sequence source for candidate gene selection of P450s catalyzing the (*Z*)-hydroxylation of santalols.

11. FUNCTIONAL CHARACTERIZATION OF CANDIDATE GENES

Functional characterization of candidate fragrance genes of nonmodel systems typically requires cloning into suitable expression vectors and selection of a heterologous expression host system. Candidate genes can be cloned from cDNA or obtained by gene synthesis. Reduced costs of gene synthesis make this a time-efficient option with additional advantages of simultaneous codon and sequence optimization (ie, modification of targeting sequences) for expression in the host of choice. Different yeast strains have been engineered for the expression and characterization of genes of terpenoid fragrances and have been successfully employed in synthetic biology studies (Diaz-Chavez et al., 2013; Hansen, Møller, Kock, et al., 2009; Ignea, Pontini, Maffei, Makris, & Kampranis, 2014; Ignea et al., 2012). *Nicotiana benthamiana* and *Physcomitrella patens* are alternative plant expression platforms also successfully used for characterization of fragrance biosynthetic genes (Cankar et al., 2015; Zhan, Zhang, Chen, & Simonsen, 2014). Key advantages of using a plant host include utilization of plant tRNAs for translation, suitable ER for proper P450 and CPR insertion, and appropriate protein folding and trafficking machineries. Independent of the expression host, two complementary approaches exist to functionally characterize candidate genes: In vivo assays where whole pathways are reconstructed in a given host organism and in vitro assays where expressed

proteins, such as TPSs or P450s, are first isolated and then tested with target substrates in enzyme assays under controlled conditions.

11.1 Yeast In Vivo Assays

Initial screening of sandalwood candidate P450s was done using in vivo assays in the yeast strain AM94 (Ignea et al., 2012) coexpressing sandalwood FPP Synthase (SaFPPS), SaSSy, and cytochrome P450 reductase (SaCPR2) (Celedon et al., 2016; Diaz-Chavez et al., 2013). The host yeast strain was engineered for improved production of sesquiterpenes and resulted in intracellular accumulation of SaSSy products, facilitating the functional screening of P450 candidates. Synthetic, codon-optimized P450 sequences were subcloned into the expression vector pYEDP60, which has been successfully used for yeast expression of many plant P450s (Duan & Schuler, 2006; Hamann & Møller, 2007). SaFPPS, SaSSy, and SaCPR2 were expressed in standard pESC vectors. All vectors were transformed into yeast following the procedure listed below:

Yeast transformation

1. Grow 5 mL overnight (ON) cultures from 3–5 yeast colonies in YPD media at 28°C.
2. Dilute cultures to $OD_{600} \sim 0.2$ in a total volume of 50 mL YPD and adenine. Grow yeast at 28°C to an OD_{600} 0.6–0.8.
3. Pellet the cells by centrifugation at $2000 \times g$ for 2 min and discard supernatant.
4. Resuspend pellet in 25 mL sterile water.
5. Pellet the cells by centrifugation as before and discard supernatant.
6. Resuspend in 1 mL 100 m*M* LiAc and transfer to an Eppendorf tube.
7. Pellet the cells by centrifugation at $1000 \times g$ for 1 min and discard supernatant.
8. Resuspend cells in 400 μL 100 m*M* LiAc.
9. Aliquot 50 μL of resuspended cells into Eppendorf tubes.
10. Pellet cells as before and remove supernatant.
11. In the following order add:
    240 μL PEG (MW 3350, 50% w/v)
    35 μL 1.0 *M* LiAc
    25 μL denatured Salmon sperm DNA (2.0 mg mL^{-1}, average size 5–7 kb)
    50 μL water with 0.5 μg plasmid DNA
12. Vortex cells vigorously until in suspension or for up to 1 min.

13. Shake tubes slowly at 28°C for 30 min.
14. Heat shock cells at 42°C, 15–25 min (optimal time may vary with strain).
15. Pellet cells by centrifugation at 1000 × *g* for 1 min and discard supernatant carefully.
16. Gently resuspend in 500 μL sterile water.
17. Apply 50 μL or more to selective medium (SD—selective aa) and incubate plates at 30°C for 3–4 days in order to see transformed colonies.

Depending on the number and the length of the genes to be transformed, options are to cotransform all of them at the same time or to perform sequential transformations. In the case of long genes and large plasmids, sequential transformation may be advised. Once all pathway genes are transformed into AM94 cells, expression cultures are prepared, and extraction of metabolites, as read out of the candidate P450 activities, is performed according to the following two consecutive protocols:

Expression cultures for in vivo assays

1. Start 5 mL ON culture from a single colony in selection media at 28°C, 220 RPM
2. Inoculate 50 mL culture (in 250-mL baffled flask) to a starting OD_{600} of 0.2
3. Incubate at 28°C, 220 RPM to an OD_{600} of 0.6–0.8
4. Centrifuge cultures in 50-mL Falcon tubes at 1000 × *g* for 5 min at RT
5. Resuspend cell pellet in 50 mL SG selective media to induce expression and incubate ON (12–16 h) at 28°C, 220 RPM
6. Centrifuge the cultures as before and transfer the supernatant to a new falcon tube for analysis of products potentially released to the media
7. Resuspend the pellet in 5 mL sterile water and transfer to ice
8. Transfer resuspended cells to preweighed a glass test tube and keep cells on ice
9. Centrifuge resuspended cells at 1000 × *g* for 5 min at 4°C
10. Remove and discard the supernatant and weigh pellets for normalization of metabolite amounts

Extraction of terpenoid metabolites

11. Start extraction of metabolites from weighted cell pellets by adding 250 μL of small acid-washed glass beads followed by 2 mL of organic solvent
12. Vortex for 1 min and decant the solvent to a new glass tube
13. Repeat the extraction a second time and pool the solvent fractions in the same tube

14. Add 250 μL sodium sulfate (anhydrous) to the tube containing the pooled solvents to remove residual water in the sample
15. Concentrate to ~1 mL under gentle flow of nitrogen gas and transfer to a GC vial
16. Store concentrated extracts at −80°C or proceed to GC–MS analysis

Extracts from in vivo assays are analyzed by GC–MS as described in Section 6.

11.2 Microsomes In Vitro Assays

Following the identification of P450 candidates which yield the target terpenoid metabolite (eg, santalols) in the yeast in vivo screening, in vitro assays with isolated microsomes containing the candidate P450 and CPR proteins allow for a more detailed biochemical characterization of P450 proteins including determination of kinetic constants, substrate specificity, optimum pH and temperature, activity with different CPR and cytB5 partners, etc. Microsomal membranes represent a mixture of native yeast protein and heterologously expressed proteins and, therefore, a negative control with WT microsomes should be included. Microsomes are isolated using the following procedures:

Microsome expression cultures

1. Start 5 mL cultures from a 3–5 colonies in SD selection media and incubate at 28°C, 220 RPM for 24 h.
2. Transfer cultures to 50 mL SD media (in 250-mL baffled flask) and grow ON at 28°C, 220 RPM.
3. Inoculate 200 mL YPD-E media with 50 mL culture and incubate for another 24 h at 28°C, 220 RPM.
4. Centrifuge cultures at 1000 × *g* for 10 min at RT.
5. Resuspend cell pellets in 200 mL YPG to induce expression and incubate ON (12–16 h) at 28°C, 220 RPM.

Microsome isolation

1. Centrifuge cells for 10 min at 4°C, 3000 × *g*, discard the supernatant, and keep cells on ice.
2. Resuspend the pellet with 5 mL of cold TEK buffer (50 m*M* Tris–HCl, pH 7.5, 1 m*M* EDTA, 100 m*M* KCl) and transfer to a Falcon tube.
3. Centrifuge cells again for another 10 min at 4°C, 3000 × *g* and discard the supernatant.
4. Resuspend pellet in 5 mL cold TES2 buffer (50 m*M* Tris–HCl, pH 7.5, 1 m*M* EDTA, 600 m*M* sorbitol, 5 m*M* DTT, 0.25 m*M* PMSF).
5. Add cold glass beads to the resuspended cells to ~5 mm from surface.

6. In a cold room, break cells by hand shaking vigorously in four cycles of 30 s shaking and 1 min rest on ice.
7. Transfer broken cells to a 50-mL Oakridge tube trying to avoid glass beads.
8. Wash beads twice with TES2 and pool with the supernatant from step 7. Aim for a final pooled volume of 25 mL.
9. Centrifuge for 15 min at 4°C, 10,000 × *g* to remove unbroken cells and glass beads.
10. Gently collect the supernatant through a nylon cloth and into an ultracentrifuge tube just until the surface of the pellet starts to loosen. Weigh tubes and balance with TES2 buffer.
11. Ultracentrifuge for 1 h at 4°C, 100,000 × *g* to pellet microsome membranes.
12. Discard the supernatant and wash the microsome pellet twice with TES buffer (50 m*M* Tris–HCl, pH 7.5, 1 m*M* EDTA, 600 m*M* sorbitol), and once in TEG (50 m*M* Tris–HCl, pH 7.5, 1 m*M* EDTA, 30% (v/v) glycerol) taking care to not loosen the pellet.
13. Resuspend pellet in 1 to 2 mL TEG buffer with a Potter-Elmer.
14. Microsomes can be directly used in assays or flash-frozen in liquid nitrogen and stored at −80°C.

The concentration of active P450 proteins in the microsome membranes can be estimated spectroscopically by the difference CO spectra method according to Guengerich, Martin, Sohl, and Cheng (2009).

Microsome assays

1. Microsome assays contain 50 m*M* potassium phosphate (pH 7.5), 1 m*M* NADPH, and 100 μ*M* sesquiterpene olefins (eg, SaSSy products) as substrates. Combine assay components in a GC vial and start reactions by adding 20 μL of purified microsomes (10 μg protein μL^{-1}) in a final volume of 400 μL.
2. Incubate reactions at 30°C for 20 min with gentle shacking at 30 RPM.
3. Terminate reactions by adding 0.5 mL of hexane spiked with isobutyl benzene as internal standard and vortexing immediately.
4. Transfer the hexane layer to a new GC vial for GC–MS analysis as described earlier.

12. PRODUCT IDENTIFICATION

Establishing the chemical identity of the products formed in an enzymatic reaction either in vitro or in vivo is critical for correct assignment of

gene function and for our understanding of secondary metabolic pathways and networks. GC–MS profiles of all candidate gene assays should be compared with negative controls to rule out endogenous yeast activities that may lead to false-positive peaks or that may alter the profile generated by the target enzyme. These controls are critical both for in vivo assays and for microsome in vitro assays, both of which contain proteins from the host organism. The two most commonly used methods for product identification of fragrance compounds are described later.

12.1 MS–MS

In the case of previously characterized fragrance compounds, identification may rely on comparison of MS–MS spectra with reference mass spectral libraries, comparison of retention index with reference data, and whenever possible comparison of mass spectra and retention indices with those of authentic standards. We generally confirm metabolite identification by GC–MS analysis by performing GC injections at different temperatures to exclude artifacts of metabolites that are unstable at high temperature.

12.2 Nuclear Magnetic Resonance

Definitive identification of a reaction product is performed by nuclear magnetic resonance (NMR) analysis. This may be critical when the reaction product does not have a perfect match in reference libraries, or when the compound is a novel structure. In these cases, NMR analysis of purified reaction products can provide accurate determination of the chemical structure of the fragrance molecule. Depending on the yield and purity of the reactions, it may be necessary to perform large-scale cultures or enzyme assays to produce sufficient material for NMR analysis. However, this may not be a trivial task, since enzyme assays of TPS or terpenoid modifying P450s do not always yield sufficient amounts or purity of product for NMR analysis.

In our case study, the products of a newly discovered sandalwood P450, SaCYP736A167, were identified by GC–MS as (*Z*)-α-santalol, (*Z*)-β-santalol, (*Z*)-α-*exo*-bergamotol, and (*Z*)-*epi*-β-santalol (Celedon et al., 2016), successfully concluding the genomics enabled discovery of the final step in sandalwood fragrance biosynthesis. Yeast cells coexpressing SaFPPS, SaSSy, SaCPR, and SaCYP736A167 produced (*Z*)-α-santalol, (*Z*)-β-santalol, (*Z*)-α-*exo*-bergamotol, and (*Z*)-*epi*-β-santalol in proportions that resemble their relative abundance in authentic sandalwood oil.

ACKNOWLEDGMENTS

Research on sandalwood in the laboratory of J.B. was supported with funds from Evolva Inc., Allylix Inc., and the Natural Sciences and Engineering Research Council of Canada and was in collaboration with researchers from the University of Western Australia, including Elizabeth L. Barbour, Patrick M. Finnegan, Christopher Jones, and Julie Plummer. Their contributions are recognized with coauthorship on the original research papers on sandalwood that formed some of the foundation of this chapter. We acknowledge the expert technical assistance of Angela Chiang, Lina Madilao, and Macaire Yuen with the development or optimization of methods in J.B.'s laboratory described in this chapter, and Karen Reid for general project management support.

REFERENCES

Abbott, E., Hall, D., Hamberger, B., & Bohlmann, J. (2010). Laser microdissection of conifer stem tissues: Isolation and analysis of high quality RNA, terpene synthase enzyme activity and terpenoid metabolites from resin ducts and cambial zone tissue of white spruce (*Picea glauca*). *BMC Plant Biology*, *10*, 1–16.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106.

Andersen-Ranberg, J., Kongstad, K. T., Nielsen, M. T., et al. (2016). Expanding the landscape of diterpene structural diversity through stereochemically controlled combinatorial biosynthesis. *Angewandte Chemie International Edition in English*, *55*, 1–6.

Arendt, P., Pollier, J., Callewaert, N., & Goossens, A. (2016). Synthetic biology for production of natural and new-to-nature terpenoids in photosynthetic organisms. *The Plant Journal*, 1–22.

Bohlmann, J., & Keeling, C. I. (2008). Terpenoid biomaterials. *The Plant Journal*, *54*, 656–669.

Caniard, A., Zerbe, P., Legrand, S., Cohade, A., Valot, N., Magnard, J.-L., et al. (2012). Discovery and functional characterization of two diterpene synthases for sclareol biosynthesis in *Salvia sclarea* (L.) and their relevance for perfume manufacture. *BMC Plant Biology*, *12*, 119.

Cankar, K., Jongedijk, E., Klompmaker, M., Majdic, T., Mumm, R., Bouwmeester, H., et al. (2015). (+)-Valencene production in *Nicotiana benthamiana* is increased by down-regulation of competing pathways. *Biotechnology Journal*, *10*, 180–189.

Celedon, J. M., Chiang, A., Yuen, M. M. S., Diaz-Chavez, M. L., Madilao, L. L., Finnegan, P. M., et al. (2016). Heartwood specific transcriptome and metabolite signatures of tropical sandalwood (*Santalum album*) reveal the final step of (*Z*)-santalol fragrance biosynthesis. *The Plant Journal*. http://dx.doi.org/10.1111/tpj.13162. Published online.

Chen, F., Tholl, D., Bohlmann, J., & Pichersky, E. (2011). The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *The Plant Journal*, *66*, 212–229.

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

Croteau, R., Ketchum, R. E. B., Long, R. M., Kaspera, R., & Wildung, M. R. (2006). Taxol biosynthesis and molecular genetics. *Phytochemistry Reviews*, *5*, 75–97.

Demissie, Z. A., Cella, M. A., Sarker, L. S., Thompson, T. J., Rheault, M. R., & Mahmoud, S. S. (2012). Cloning, functional characterization and genomic organization of 1,8-cineole synthases from *Lavandula*. *Plant Molecular Biology*, *79*, 393–411.

Demissie, Z. A., Erland, L. A. E., Rheault, M. R., & Mahmoud, S. S. (2013). The biosynthetic origin of irregular monoterpenes in *Lavandula*. *The Journal of Biological Chemistry*, *288*, 6333–6341.

Diaz-Chavez, M. L., Moniodis, J., Madilao, L. L., et al. (2013). Biosynthesis of sandalwood oil: *Santalum album* CYP76F cytochromes P450 produce santalols and bergamotol. *PloS One*, *8*, e75053.

Duan, H., & Schuler, M. A. (2006). Heterologous expression and strategies for encapsulation of membrane-localized plant P450s. *Phytochemistry Reviews*, *5*, 507–523.

Dudareva, N., Martin, D., Kish, C. M., Kolosova, N., Gorenstein, N., Fäldt, J., et al. (2003). (E)-β-ocimene and myrcene synthase genes of floral scent biosynthesis in snapdragon: Function and expression of three terpene synthase genes of a new terpene synthase subfamily. *Plant Cell*, *15*, 1227–1241.

Facchini, P. J., Bohlmann, J., Covello, P. S., et al. (2012). Synthetic biosystems for the production of high-value plant metabolites. *Trends in Biotechnology*, *30*, 127–131.

Gershenzon, J., & Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nature Chemical Biology*, *3*, 408–414.

Guengerich, F. P., Martin, M. V., Sohl, C. D., & Cheng, Q. (2009). Measurement of cytochrome P450 and NADPH-cytochrome P450 reductase. *Nature Protocols*, *4*, 1245–1251.

Haas, B. J., Papanicolaou, A., Yassour, M., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*, 1494–1512.

Hamann, T., & Møller, B. L. (2007). Improved cloning and expression of cytochrome P450s and cytochrome P450 reductase in yeast. *Protein Expression and Purification*, *56*, 121–127.

Hamberger, B., Ohnishi, T., Hamberger, B., Séguin, A., & Bohlmann, J. (2011). Evolution of diterpene metabolism: Sitka spruce CYP720B4 catalyses multiple oxidations in resin acid biosynthesis of conifer defense against insects. *Plant Physiology*, *157*, 1677–1695.

Hansen, E. H., Møller, B. L., Kock, G. R., et al. (2009). De novo biosynthesis of vanillin in fission yeast (*Schizosaccharomyces pombe*) and baker's yeast (*Saccharomyces cerevisiae*). *Applied and Environmental Microbiology*, *75*, 2765–2774.

Ignea, C., Pontini, M., Maffei, M. E., Makris, A. M., & Kampranis, S. C. (2014). Engineering monoterpene production in yeast using a synthetic dominant negative geranyl diphosphate synthase. *ACS Synthetic Biology*, *3*, 298–306.

Ignea, C., Trikka, F. A., Kourtzelis, I., Argiriou, A., Kanellis, A. K., Kampranis, S. C., et al. (2012). Positive genetic interactors of HMG2 identify a new set of genetic perturbations for improving sesquiterpene production in Saccharomyces cerevisiae. *Microbial Cell Factories*, *11*, 162.

Jones, C. G., Keeling, C. I., Ghisalberti, E. L., Barbour, E. L., Plummer, J. A., & Bohlmann, J. (2008). Isolation of cDNAs and functional characterisation of two multi-product terpene synthase enzymes from sandalwood, *Santalum album* L. *Archives of Biochemistry and Biophysics*, *477*, 121–130.

Jones, C. G., Moniodis, J., Zulak, K. G., Scaffidi, A., Plummer, J. A., Ghisalberti, E. L., et al. (2011). Sandalwood fragrance biosynthesis involves sesquiterpene synthases of both the terpene synthase (TPS)-a and TPS-b subfamilies, including santalene synthases. *The Journal of Biological Chemistry*, *286*, 17445–17454.

Knudsen, J. T., Eriksson, R., Gershenzon, J., & Ståhl, B. (2006). Diversity and distribution of floral scent. *The Botanical Review*, *72*, 1–120.

Kolosova, N., Miller, B., Ralph, S., Ellis, B. E., Douglas, C., Ritland, K., et al. (2004). Isolation of high-quality RNA from gymnosperm and angiosperm trees. *Biotechniques*, *36*, 821–824.

Lange, B. M. (2016). Online resources for gene discovery and biochemical research with aromatic and medicinal plants. *Phytochemistry Reviews*, 1–22.

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., et al. (2012). RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, *40*, W622–W627.

Magnard, J.-L., Roccia, A., Caissard, J.-C., et al. (2015). Biosynthesis of monoterpene scent compounds in roses. *Science*, *349*, 81–83.

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews. Genetics*, *12*, 671–682.

Muhlemann, J. K., Klempien, A., & Dudareva, N. (2014). Floral volatiles: From biosynthesis to function. *Plant, Cell & Environment*, *37*, 1936–1949.

Nelson, D., & Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. *The Plant Journal*, *66*, 194–211.

Paddon, C. J., Westfall, P. J., Pitera, D. J., et al. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*, 528–532.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*, 1061–1067.

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, *32*, 462–464.

Qualley, A. V., Widhalm, J. R., Adebesin, F., Kish, C. M., & Dudareva, N. (2012). Completion of the core β-oxidative pathway of benzoic acid biosynthesis in plants. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 16383–16388.

Robertson, G., Schein, J., Chiu, R., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*, 909–912.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–140.

Schroeder, A., Mueller, O., Stocker, S., et al. (2006). The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, *7*, 3.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*, 1086–1092.

Van Verk, M. C., Hickman, R., Pieterse, C. M. J., & Wees, S. C. M. Van. (2013). RNA-Seq: Revelation of the messengers. *Trends in Plant Science*, *18*, 175–179.

Widhalm, J. R., Jaini, R., Morgan, J. A., & Dudareva, N. (2015). Rethinking how volatiles are released from plant cells. *Trends in Plant Science*, *20*, 545–550.

Zerbe, P., & Bohlmann, J. (2015a). Enzymes for synthetic biology of ambroxide-related diterpenoid fragrance compounds. In J. Schrader & J. Bohlmann (Eds.), *Advances in biochemical engineering/biotechnology: Vol. 148* (pp. 427–447).

Zerbe, P., & Bohlmann, J. (2015b). Plant diterpene synthases: Exploring modularity and metabolic diversity for bioengineering. *Trends in Biotechnology*, *33*, 419–428.

Zerbe, P., Hamberger, B., Yuen, M. M. S., et al. (2013). Gene discovery of modular diterpene metabolism in nonmodel systems. *Plant Physiology*, *162*, 1073–1091.

Zhan, X., Zhang, Y.-H., Chen, D.-F., & Simonsen, H. T. (2014). Metabolic engineering of the moss *Physcomitrella patens* to produce the sesquiterpenoids patchoulol and α/β-santalene. *Frontiers in Plant Science*, *5*, 636.

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., & Hao, P. (2011). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: A comparative study. *BMC Bioinformatics*, *12*, S2.