The evolution and engineering of Asteraceae oxidosqualene cyclases

Honghao Su

A thesis submitted to the University of East Anglia in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Abstract

Plants produce a diverse range of secondary metabolites that, while not essential for normal growth and development, play vital roles in communication, defence, and reproduction. Triterpenoids are one of the largest and most structurally diverse classes of secondary metabolites of which many are of interest to medicine and industry. The first step of triterpenoid biosynthesis is the cyclisation of 2,3-oxidosqualene into triterpene scaffolds by oxidosqualene cyclases (OSCs). This thesis investigates the evolution, specificity, regulation and engineering of taraxasterol synthases (TXSSs), a class of OSCs involved in the biosynthesis of triterpenoids reported to have a range of bioactivities, including being responsible for the anti-inflammatory activity of *Calendula officinalis* (pot marigold). In Chapter 3, I provide evidence that TXSSs likely evolved via gene duplication and neofunctionalisation of a mixed amyrin synthase soon after the emergence of the Asteraceae (the daisy family). This work also revealed that TXSSs have been maintained across the Asteraceae, however, TXSSs found in different lineages produce varying ratios of two scaffolds: ψ-taraxasterol and taraxasterol. Evolutionary studies described in Chapter 4 revealed that the likely predominant product of ancestral TXSSs was ψ-taraxasterol and the change to preferential production of taraxasterol in one Asteraceae tribe likely conferred a selected advantage. I identify two residues in the active site likely to be under positive selection and, using structure guided studies and computational simulations, characterise residues important for product specificity. In Chapter 5, I describe the identification of an R2R3-MYB-family transcription factor that contributes to the floral-specific transcriptional regulation of *TXSS* in *C. officinalis.* Finally, with the goal of providing new routes for heterologous biosynthesis, I explore rational and deep learning guided protein design methods to engineer a membrane detached TXSS (Chapter 6). Together, this work provides insights into the evolution and function of OSCs as well as new opportunities for biosynthesis.

Acknowledgement
This work is sponsored by Biotechnology and Biological Sciences Research Council Norwich Research Park Doctoral Training Partnership scholarship (BB/T008717/1 Project No. 2578291).

First and foremost, I would like to thank my primary supervisor, Dr Nicola Patron. Through her supervision, I have learnt how to conduct rigorous scientific research and communicate with others. I am grateful that she encouraged me to explore interdisciplinary work in collaboration with others. Her mentorship was invaluable in helping me navigate through challenges during the past four years of research.
I would also like to thank my secondary supervisor, Dr Wilfried Haerty, for bringing in his unique perspectives to my research, advising me to work smartly and taking care of administration duties.

I would like to thank the support I received from past and present members of Patron group both in Norwich and Cambridge. In Norwich, thanks to Dr Melissa Salmon for starting the pot marigold project and taking care of it throughout, and Dr Daria Cuthbert (Golubova) and Dr Connor Tansley for working collaboratively on it. Thanks to Dr Sam Witham and Dr Tufan Oz for teaching me molecular biology and cloning when I started, and Dr Yaomin Cai for showing me some interesting experimental techniques and chatting with me over interesting synthetic biology and social topics during teatime. Thanks to Ms Rachel Shen, Dr Mahima Mahima and Dr Sarah Guiziou for engaging in helpful scientific discussions and being great lab companions. In Cambridge, thanks to Dr Zhengao Di for helping transfer materials and move flat. Thanks to Dr Juanjuan Wang for helping with lettuce protoplast assays. Thanks to Dr Konstantina Beritza for her valuable advice in confocal imaging and helping me with thesis format conversion on the submission day. Thanks to Ms Catarina Almeida, Dr Caroline Faessler, Ms Sophie Beck and Ms Assia Ibrahim for their company and delightful conversations. I would also like to Prof Alison Smith and members of Smith group, who I shared lab space with and received lab support from, including Mr Daniel Zhang, Dr Lorraine Archer, Dr Katrin Geisler, Dr Pawel Mordaka, Dr Tatiana Zuniga-Burgos, Dr Payam Mehrshahi and Dr Kostas Papadopoulos. Also, thanks to Dr Facundo Romani for his support in confocal microscopy.

I would also like to thank the scientific support from collaborators. Thanks to Dr Dave Wright and Dr Will Nash at the Earlham Institute for their advice on evolutionary and phylogenetic analysis. Thanks to Prof Ilia Leitch and Ms Sahr Mian at Royal Botanic Gardens, Kew for help with genome size and ploidy estimation, Prof Andrew Hemmings at University of East Anglia for help with molecular modelling and docking, Dr Marc van der Kamp at University of Bristol for helpful discussions on MD and QM/MM simulations and Prof Dek Woolfson and Mr Rokas Petrenas at University of Bristol for advice in protein design.

Thanks to horticultural service both in Norwich and in Cambridge for taking care of the plants which I sacrificed for this thesis. Thanks to John Innes Centre metabolomic platform for their support in GC-MS.

On personal level, I would like to thank my friends and companions both in Norwich and Cambridge. In Norwich, thanks to Dr Josh Colmer for our good time playing League of Legends and chatting over gaming and machine learning. Thanks to Ms Ruoxi Lin for our scientific exchange on plant metabolism and bearing with my rants on failed experiments. Thanks to Dr Amr El-Demerdash and Dr Abdul Kader Alabdullah for regularly refreshing my Arabic and their scientific and life advice. In Cambridge, thank Mr Eddie Xiao, Mr Anfu Wang, Mr Khaleefa Al Dhaheri and Ms Asma Ibrahim for their long-standing friendship since undergrad. Thanks to Dr Farahnoz Khojayori for providing valuable feedback on my thesis. Thanks to Mr Asad Ibn Saifuallah for our regular weekend lunch catch-up and sharing funny travel stories. Thanks to everyone who I had delightful conversations with at Plant Sciences tearoom. I would also like to thank Mr Thomas Qu at University of Chicago for accompanying with me on a trip to Egypt and discussing MD simulations along the way. Thank Mr Gabriel Ong at Scripps Research for constantly chatting about protein design and teaming up in protein design competitions. I have stayed in quite a few different houses throughout the four years, and I would like to thank my amazing housemates for our good time, especially Mr Arif Ahmed and Mr Omar Shabana. Apologise for having a bad memory and not being able to name all, but thanks to all my friends who have supported me on this journey.

Finally, I would like to thank my family, especially my parents Mr Hu Su and Ms Qingchun Zhang for their unwavering support and visiting me regularly both in Norwich and Cambridge. Thanks to my grandparents, aunts and uncles, and cousins Mr Yunrui Li, Mr Jingyuan Wan and Ms Qingqing Sheng, and other members of the Su and the Zhang family who have supported me.

List of abbreviations

ACT	Acyltransferase
AMBER	Assisted Model Building with Energy Refinement
ANOVA	Analysis of variance
BEAST	Bayesian Evolutionary Analysis Sampling Trees
bHLH	Basic Helix-Loop-Helix
CDS	Coding Sequence
CRE	Cis-Regulatory Element
cryo-EM	Cryogenic Electron Microscopy
CYP	Cytochrome P450
DMAPP	Dimethylallyl diphosphate
DMSO	Dimethyl sulfoxide
DOF	DNA binding with One Finger
ER	Endoplasmic reticulum
FAE	Fatty acid ester
FIMO	Find Individual Motif Occurrences
FPP	Farnesyl diphosphate
GAFF	General AMBER Force Field
GC-MS	Gas chromatography-mass spectrometry
GPP	Geranyl diphosphate
HMGR	3-hydroxy,3-methylglutaryl-CoA reductase
IPP	Isopentenyl diphosphate
LC-MS	Liquid chromatography–mass spectrometry
LucF	Firefly Luciferase
LucN	NanoLuc Luciferase
MAS	Mixed Amyrin Synthase
McLS	*Methylococcus capsulatus* lanosterol synthase
MD	Molecular Dynamics
MEP	2-C-methyl-D-erythritol 4-phosphate
MPNN	Message-Passing Neural Network
MRCA	Most Recent Common Ancestor
MVA	Mevalonic acid
Mya	Million years ago
NLS	Nuclear Localisation Signal
NPT	Constant-temperature, constant-pressure ensemble
NVT	Constant-temperature, constant-volume ensemble
OSC	Oxidosqualene cyclase
PAML	Phylogenetic Analysis by Maximum Likelihood
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PFM	Position frequency matrix
pLDDT	predicted Local Distance Difference Test
QM/MM	Combined Quantum Mechanics and Molecular Mechanics
RMSD	Root Mean Square Deviation

SASA	Solvent Accessible Surface Area
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TXSS	Taraxasterol Synthase
UDP	Uridine Diphosphate
UGT	UDP-Glycosyltransferase
WGD	Whole Genome Duplication
WHAM	The Weighted Histogram Analysis Method
YFP	Yellow Fluorescence Protein

Table of Contents

Table of Figures

Table of Tables

1. Chapter 1 General Introduction

1.1 The origin, evolution and biogeography of Asteraceae

1.1.1 Asteraceae phylogenetics

Asteraceae, also known as the aster family or sunflower family, is one of the largest families of flowering plants containing an estimated 23,000 to 35,000 species (Mandel et al., 2019). Asteraceae has a cosmopolitan distribution with members found on every continent of the world including Antarctica (Lewis Smith and Richardson, 2011) and in all types of habitats, with most occurring in deserts and semi-deserts. A head-like inflorescence (capitulum), packed with tiny individual florets protected by involucres, is the diagnostic feature of Asteraceae (Figure 1.1). Variations in capitulum features, such as petal colour, size, and shape contribute to the diversification and evolutionary success of Asteraceae.

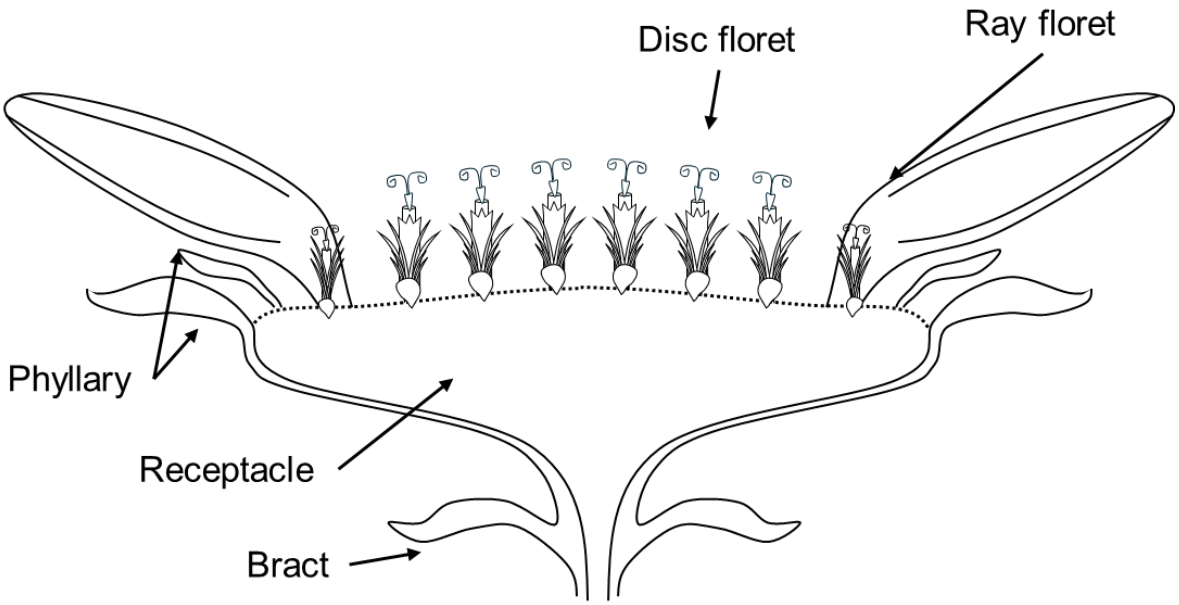


Figure 1.1 Graphical representation of Asteraceae capitulum. The numerous small flowers within the capitulum (flower head) are classified as either ray florets or disc florets.

In early phylogenetic studies, based on differences in morphology, anatomy and chromosomal numbers, the Asteraceae was classified into two subfamilies: Cichorioideae and Asteroideae, with Cichorioideae diverging earlier and more closely related to non-Asteraceae (Carlquist, 1976; Wagenitz, 1976). Subsequent studies on chloroplast genomes observed a 22kb inversion in all sampled Asteraceae lineages except for those in Barnadesiinae, a subtribe with morphological features similar to families closely related to Asteraceae, such as bilabiate flowers and woody habitats (Jansen and Palmer, 1987). Hence, Barnadesiinae likely represented the earliest evolutionary split and was established as a new subfamily, Barnadesioideae, sister to the rest of the Asteraceae (Bremer and Jansen, 1992).

The resolution of the phylogenetic tree was improved by analysis of the chloroplast gene *ndhF*, sampled from 89 Asteraceae and 5 outgroup species (Kim and Jansen, 1995). This analysis supported the monophyly of Barnadesioideae and Asteroideae. However, subfamily Cichorioideae and two tribes within Cichorioideae, Mutisieae and Cardueae, were found to be paraphyletic and two new subfamilies, Carduoideae and Mutisioideae, were established (Bremer, 1996; Katinas et al., 2008).

With the emergence of next-generation sequencing technologies, access to genetic information became easier and cheaper. This enabled phylogenetic analysis using multiple genetic loci, which improved resolution and accuracy. Mandel et al. built an Asteraceae species phylogenetic tree based on a concatenated alignment of multiple nuclear genetic loci from 256 species (Mandel et al., 2019). This phylogenetic tree resolved the evolutionary relationships of the basal subfamilies, supported the paraphyly of Cichorioideae, and corroborated on the monophyly of Asteroideae (Figure 1.2).

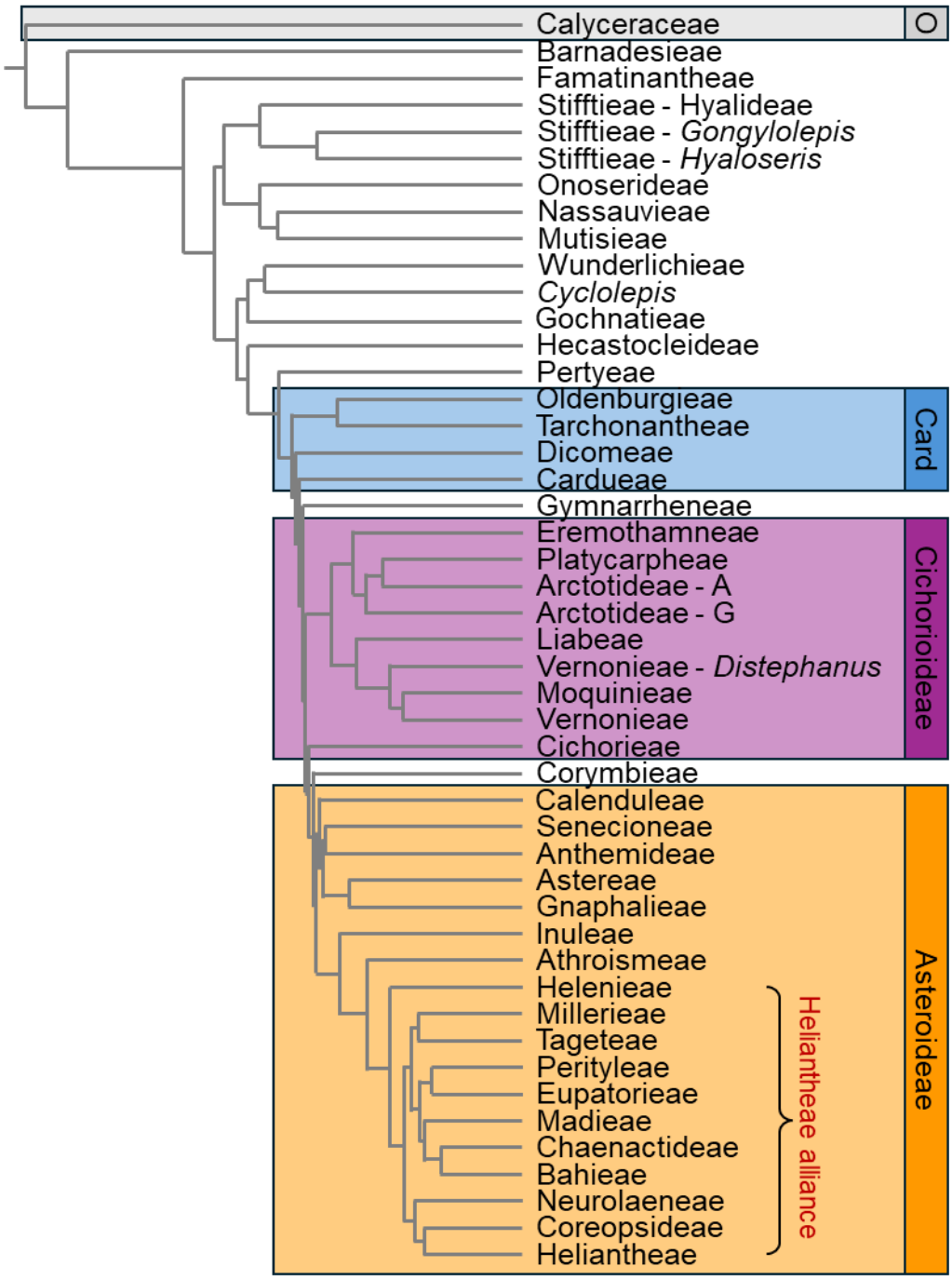


Figure 1.2 Phylogenetic relationships of tribes in the Asteraceae. Tribes belonging to the three main superfamilies, Carduoideae, Cichorioideae and Asteroideae are labelled in blue, purple and orange respectively. Tribes belonging to supertribe Heliantheae alliance are labelled in a bracket. O, outgroup. Card, Carduoideae. Figure adapted from Mandel et al., 2019.

1.1.2 The emergence and spread of Asteraceae

Barreda et al. first reported fossilised pollen grains from *Tubulifloridites lilliei* type A (likely a relative of the Barnadesioideae) preserved in Antarctica sedimentary rocks which contain fossilised dinosaur remains and dated them to around 76-66 million years ago (Mya) in the late Cretaceous (Barreda et al., 2015). Based on the age of

those fossil grains, the most recent common ancestor (MRCA) of Asteraceae was estimated to originate at 80 Mya. In subsequent studies, Panero and Crozier placed the date for MRCA of Asteraceae slightly later at around 69.5 Mya (still late Cretaceous) by imposing nine absolute time constraints (Panero and Crozier, 2016). They also estimated that the MRCA of extant Asteraceae originated around 64.75 Mya in the early Paleocene. Nevertheless, Mandel et al. pushed the date of Asteraceae origin back to around 83 Mya (95% confidence interval 64 to 91 Mya) by imposing fossil constraints on species phylogenetic trees built on concatenated nuclear genetic loci (Mandel et al., 2019).

Barnadesioideae, the basal Asteraceae subfamily, was documented to possess at least eight genera centred in the northern Andes, leading to the hypothesis that Asteraceae originated in montane South America (Jansen and Palmer, 1987). This was corroborated by an ancestral range estimate by Mandel et al., which concluded that Asteraceae likely originated in southern South America, before extending into the north and central Andes and spreading into the Guiana Shield region and Brazil (Mandel et al., 2019).

Given Asteraceae originated after the breakup of Gondwana when South America was already an isolated landmass, the spread of Asteraceae from its origin to other parts of the world involved inter-continental dispersal and, potentially, transoceanic events. Two routes were proposed to explain how Asteraceae spread across the world: the South America-North America-Asia route and the South America-Africa route (**Figure 1.3**) (Panero and Funk, 2008).

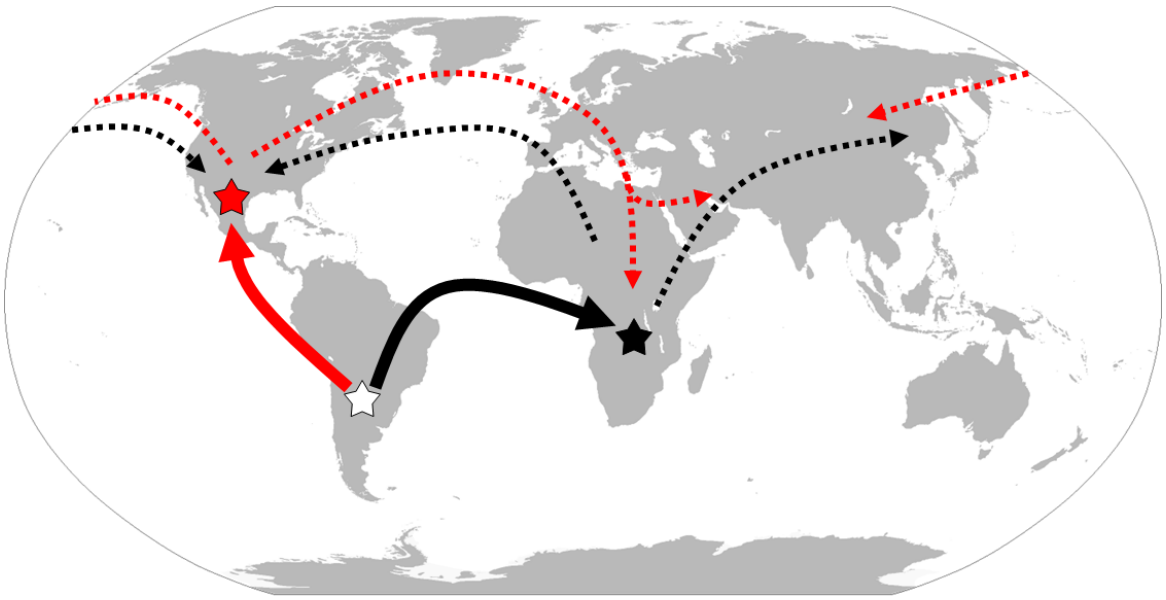


Figure 1.3 Dispersal routes of Asteraceae out of South America. The red arrow represents the South America-North America-Asia route. The black arrow represents the South America-Africa route. The South America-North America-Asia route hypothesised that Asteraceae dispersed to North America first before spreading to Asia, the Middle East and North Africa via two different routes. The South America-Africa route hypothesised that Asteraceae first reached Africa, where adaptive radiation occurred, before reaching North America and Eurasia via two different routes.

1.1.3 Whole genome duplication in Asteraceae

Whole-genome duplication (WGD) is a process in which an individual acquires another copy of its genome via chromosome non-disjunction during meiosis (autopolyploidisation) or hybridisation (allopolyploidisation). Post-WGD, genomes are usually unstable, and extensive genome rearrangements, gene loss and activation of transposable elements are observed (Otto, 2007). In the long term, homeologous chromosomes (chromosome pairs that arose from WGD) can diverge and form two pairs of bivalents, ultimately leading to diploidisation (reversion to diploid but with another set of chromosomes). However, an extra copy of the genome enables a lineage to evolve phenotypic novelty using the additional gene copies for sub- and neo-functionalisation. Consequently, adaptive radiations and increases in the rate of diversification are often observed after WGD events (Fawcett et al., 2009).

WGD events occurred throughout the evolution of Asteraceae, contributing to its adaptive radiation. A WGD shared by the Calyceraceae and Asteraceae and a WGD shared by all core Asteraceae occurred around the Cretaceous-Paleogene boundary (K-Pg boundary), a period when around 67% of species went extinct (Huang et al., 2016; Jablonski et al., 1994). With the environmental and ecological shift, there was a rapid radiation of Asteraceae subfamilies (Barreda et al., 2015). Similarly, a WGD shared by members of the Heliantheae alliance occurred around the time of the Grande Coupure, an extinction event at the Eocene-Oligocene boundary in Europe (Hooker et al., 2004; Huang et al., 2016). After this event, there was an eight-fold increase in net diversification, leading to the diversification of the Heliantheae alliance (Huang et al., 2016).

1.2 Plant secondary metabolism

1.2.1 Plant secondary metabolites

Plants produce a wide range of metabolites that play different roles throughout their life cycles. Plant metabolites are generally classified into primary metabolites and secondary metabolites. Primary metabolites, including sugars, lipids and amino acids, are essential for plant growth and development. Secondary metabolites, including phenolic compounds, non-protein amino acids, alkaloids, sulphur-containing compounds and terpenoids, are not essential for plant growth and development but play a role in interactions with the environment. Secondary metabolites are derived from primary metabolites and many of them are only found in specific lineages. Within these lineages they are often only produced in specific tissues or even cell types.

1.2.2 Biological functions of secondary metabolites

Plant secondary metabolites have been documented to have anti-pathogen, anti-herbivory, allelopathic and pollinator attracting properties.

Some secondary metabolites demonstrate activity against the growth of pathogenic bacteria, fungi and oomycetes. For example, isoflavanoid pterocarpans produced by the soybean family inhibit fungal germ tube and hyphae elongation (Weston and

Mathesius, 2013). Secondary metabolites could be produced either upon pathogen attack (phytoalexins) or constitutively to form a barrier against pathogens (phytoanticipins) (VanEtten et al., 1994).

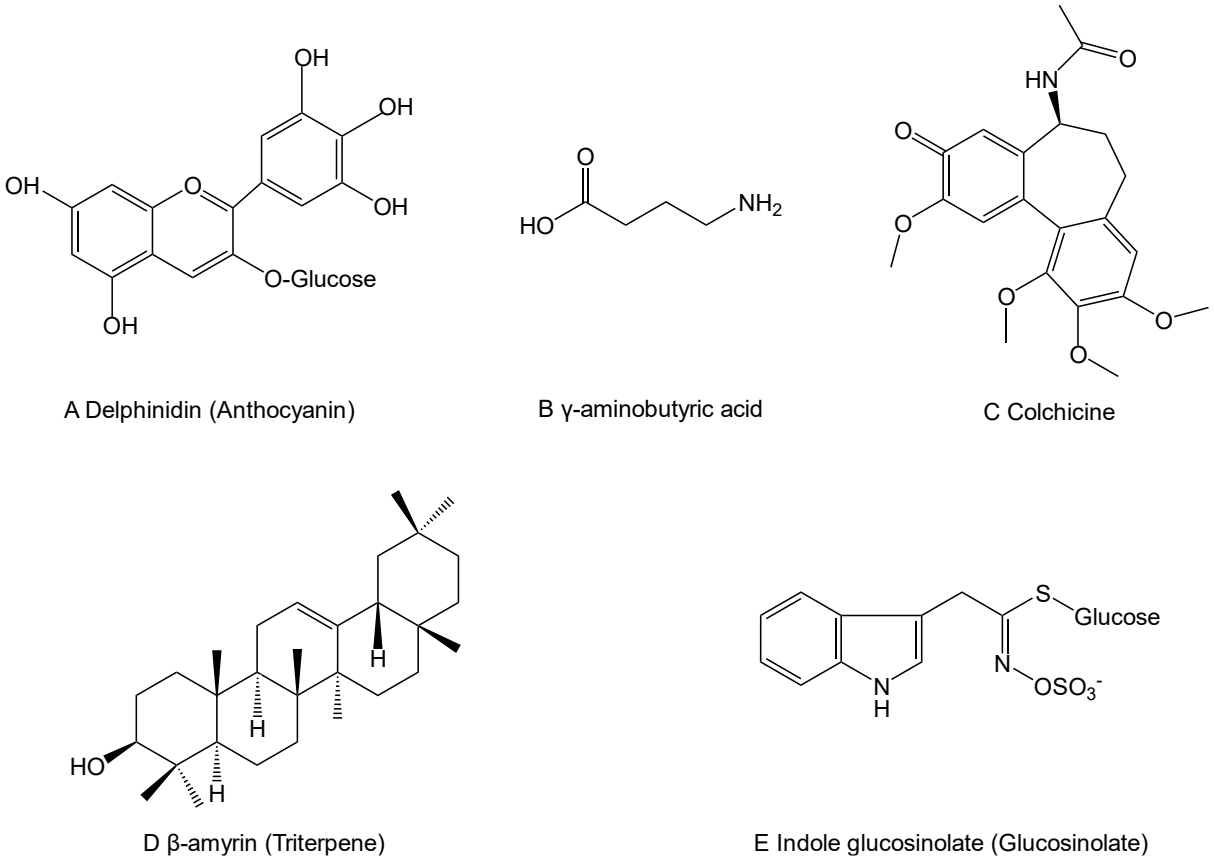
Unpalatable and toxic secondary metabolites are used by plants to deter and defend against herbivory. For example, cassava plants store cyanogenic glycosides in their vacuoles. Upon herbivory, the vacuole membrane breaks and the cyanogenic glycosides would react with cytoplasmic β-glucosidase and α-hydroxynitrilase to generate hydrogen cyanide, which is poisonous to herbivores (Vetter, 2000).

Allelopathy refers to the release of secondary metabolites (usually phytotoxic compounds) to influence the germination, growth and survival of neighbouring plants (Schandry and Becker, 2020). For example, Gramineous species, such as wheat, maize and rye, release benzoxazinoid compounds into the soil to inhibit the germination of competing species (Schandry and Becker, 2020).

Flowers in most angiosperm lineages accumulate anthocyanin pigments in their vacuoles, which confer floral colouration (Glover and Martin, 2012). Anthocyanins are composed of an anthocyanidin ring and conjugated sugar(s). Variation in the conjugated sugar and pH within the vacuole influences anthocyanin colour (Khoo et al., 2017). Different floral colours attract different types of pollinators (Glover and Martin, 2012). For example, bees have a strong preference for yellow flowers while hummingbirds more frequently visit red flowers (Handelman and Kohn, 2014; Wenzell et al., 2025)

1.2.3 Classification of plant secondary metabolites

Plant secondary metabolites are generally classified into five categories: phenolic compounds, non-protein amino acids, alkaloids, sulphur-containing compounds and terpenoids (**Figure 1.4**).



A Delphinidin (Anthocyanin)

B γ-aminobutyric acid

C Colchicine

D β-amyrin (Triterpene)

E Indole glucosinolate (Glucosinolate)

Figure 1.4 Examples of the five main types of plant secondary metabolite. (A) phenolic compounds, (B) non-protein amino acids, (C) alkaloids, (D) sulphur-containing compounds and (E) terpenoids.

Phenolic compounds contain at least one aromatic ring with at least one hydroxyl group. They perform a wide range of biological functions, including general defence (cutins, suberins, tannins), pollinator attraction (anthocyanins) and UV protection (flavonoids). Most phenolic compounds are derived from shikimic acids, with phenylalanine being the main precursor.

Non-protein amino acids, such as L-DOPA, GABA and canavanine, are produced by several legume and grass lineages. They play a role in defence against herbivores by disrupting metabolism and nervous system function (Huang et al., 2011).

Alkaloids are large and highly diverse group of compounds that contain nitrogen and aromatic groups. Examples include nicotine from tobacco, caffeine from coffee and tea plants and morphine from opium poppy. They generally act as agonists or antagonists of the animal central nervous systems and are thus produced as defence against herbivores. Alkaloid biosynthesis is varied, with compounds produced from amino acid, purine and terpene precursors.

Glucosinolates, derived from glucose and an amino acid (mostly methionine), are the most common sulphur-containing compounds. These compounds are widespread in the Brassicaceae family. Following cell damage, glucosinolates come into contact with myrosinases, which convert them into isothiocyanates that have fungicidal, bactericidal and anti-proliferative properties (Barba et al., 2016).

Terpenoids include terpenes, hydrocarbon molecules composed of 5C isoprene units, and terpene derivatives, including oxygenated terpenes. Terpenoids are involved in both primary metabolism (such as sterols) and secondary metabolism (such as artemisinin). Their biosynthesis is discussed in further details in Chapter 1.3.

1.2.4 The evolution and diversification of secondary metabolites

More than 200,000 plant secondary metabolites have been documented, including more than 25,000 terpenoids, 12,000 alkaloids and 8000 phenolic compounds (Yonekura-Sakakibara and Saito, 2009). However, the precursors of these diverse secondary metabolites come from a restricted set of compounds produced by primary metabolic pathways including glycolysis, the tricarboxylic acid cycle (TCA) and the shikimate pathway (Wink, 2016). To enable the production of a wide range of compounds from a restricted set of precursors, enzymes with different substrate- and product-specificity evolved, mostly through gene duplication and subsequent functional divergence.

Gene duplication can occur via several mechanisms, including tandem duplication, retrotransposition and WGD. Mutations that accumulate in the coding sequences of duplicates can lead to functional changes, including changes in substrate and/or product specificity. For example, DIMBOA (2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one) biosynthesis in Gramineae requires four cytochrome P450s (CYPs), which perform four consecutive hydroxylation reactions and one ring expansion reaction. The genes encoding them are tandemly arranged and were formed by tandem duplication (Frey et al., 1997). A whole-genome duplication in the Brassicales, enabled duplicated glucosinolate synthases with specificity for phenylalanine and branched-chain aliphatic amino acids to acquire specificity for indole and methionine (Barco and Clay, 2019). After gene duplication, mutations in the regulatory sequences of the duplicated genes could enable expression in different tissues and environmental conditions (Pichersky and Gang, 2000).

1.3 Terpene biosynthesis

Terpenes and their derivatives are the most diverse class of plant secondary metabolites of which there are more than 25,000 different compounds (Yonekura-Sakakibara and Saito, 2009). Terpenes are broadly classified into monoterpenes (10C), sesquiterpenes (15C), diterpenes (20C), triterpenes (30C) and tetraterpenes (40C). All terpenes share the same precursor, isopentenyl pyrophosphate (IPP), produced from either the cytoplasmic mevalonate (MVA) pathway or from the plastidial 2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate (MEP) pathway (**Figure 1.5**). IPP synthesised from the MVA pathway is used to produce phytosterols, triterpenes and some sesquiterpenes such as phytoalexins, whereas IPP synthesised in the MEP pathway is used to produce monoterpenes, diterpenes, carotenoids and other sesquiterpenes. IPP crosstalk between cytoplasm and plastid is thought to be limited, with at most 1% of IPP transferred between compartments (Rodríguez-Concepción and Boronat, 2002).

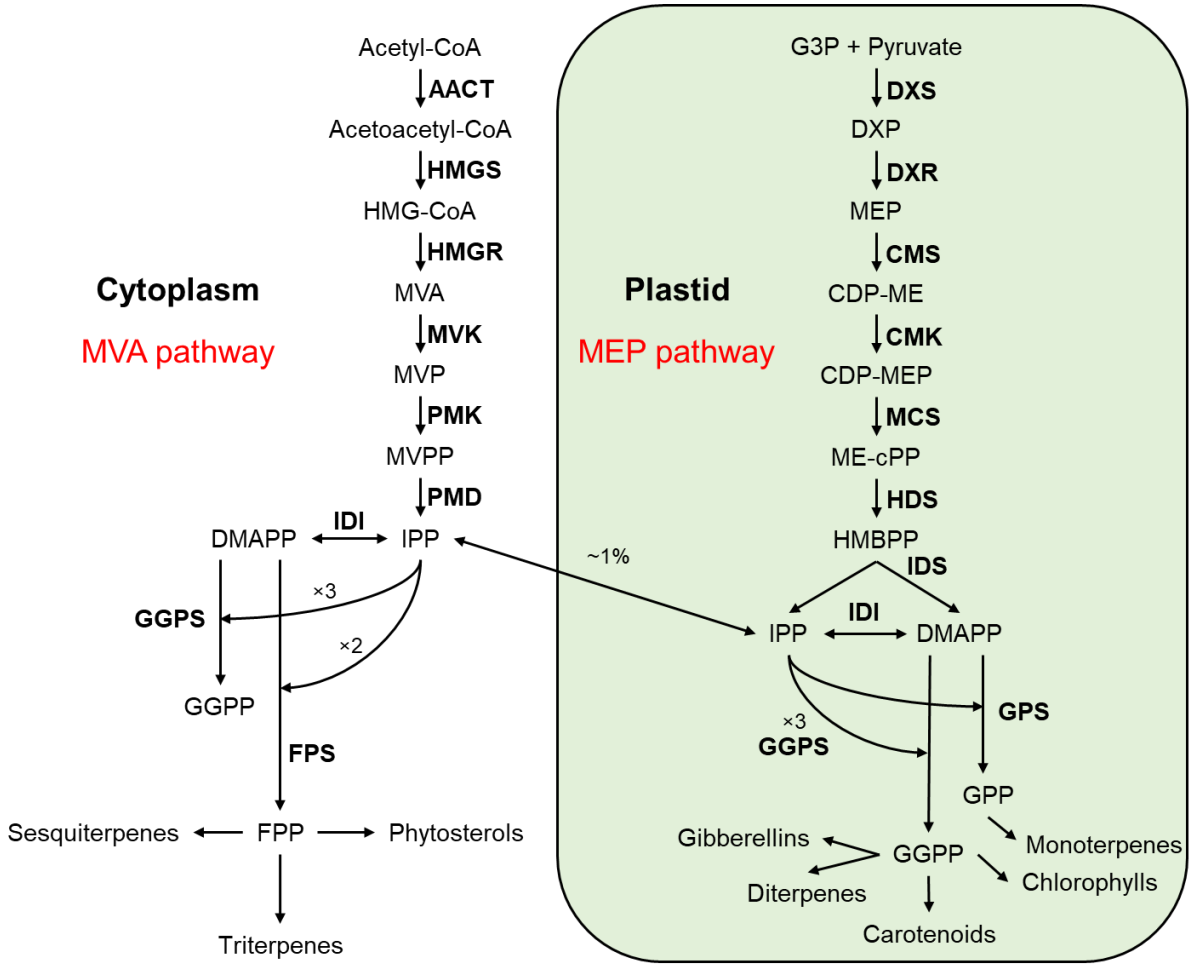


Figure 1.5 The biosynthesis pathway of IPP in plants. ***Enzymes in the cytosolic MVA pathway:*** AACT, acetoacetyl-CoA thiolase; HMGS, HMG-CoA synthase; HMGR, HMG-CoA reductase; MVK, MVA kinase; PMK, MVP kinase; PMD, MVPP decarboxylase; IDI, IPP isomerase; FPS, FPP synthase; GGPS, GGPP synthase. ***Compounds in the cytosolic MVA pathway*****:** HMG-CoA, Hydroxymethylglutaryl CoA; MVA, mevolonate; MVP, 5-phosphomevalonate; MVPP, 5-diphosphomevalonate; IPP, isopentenyl pyrophosphate; DMAPP, dimethylallyl pyrophosphate; FPP, farnesyl diphosphate; GGPP, geranylgeranyl pyrophosphate. ***Enzymes in the plastidial MEP pathway*****:** DXS, DXP synthase; DXR, DXP reductoisomerase; CMS, CDP-ME synthase; CMK, CDP-ME kinase; MCS, ME-cPP synthase; HDS, HMBPP synthase; IDS, IPP/DMAPP synthase; GPS, GPP synthase. ***Compounds in the plastidial MEP pathway*****:** G3P, glyceraldehyde 3-phosphate; DXP, deoxyxylulose 5-phosphate; MEP, methylerythritol 4-phosphate; CDP-ME, 4-diphosphocytidyl-2-C-methylerythritol; CDP-MEP, 4-diphosphocytidyl-2-C-methyl-D-erythritol 2-phosphate; ME-cPP, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate; HBMPP, hydroxymethylbutenyl 4-diphosphate; GPP, geranylpyrophosphate. (Figure adapted from Rodríguez-Concepción & Boronat 2002)

IPP is first converted to dimethylallyl pyrophosphate (DMAPP), which then loses its pyrophosphate group and forms a carbocation (**Figure 1.6**). Catalysed by GPP synthase, another IPP performs a nucleophilic attack on the cation to form geranyl pyrophosphate (GPP) (Burke et al., 1999). Further addition of IPP units by corresponding synthases gives rise to farnesyl pyrophosphate (FPP, 15C) and geranylgeranyl pyrophosphate (GGPP, 20C), the precursors of sesquiterpenes and diterpenes respectively. Pyrophosphate-to-pyrophosphate coupling of two FPPs and

two GGPPs generates squalene and phytoene, which are precursors of triterpenes and tetraterpenes respectively.

Catalysed by different types of terpene synthases, these precursors are cyclised into terpene scaffolds before further modifications are made. Dependent on their structure and catalytic mechanism, terpene synthases are classified into type I and type II terpene synthases. Type I terpene synthases contain a DDxx(D,E) motif to coordinate a metal cofactor that abstracts the diphosphate group to form the intermediary carbocation. Type II terpene synthases contain a DxDD motif that directly protonates the substrate and initiates carbocation cyclisation (D. Rudolf and Chang, 2020).

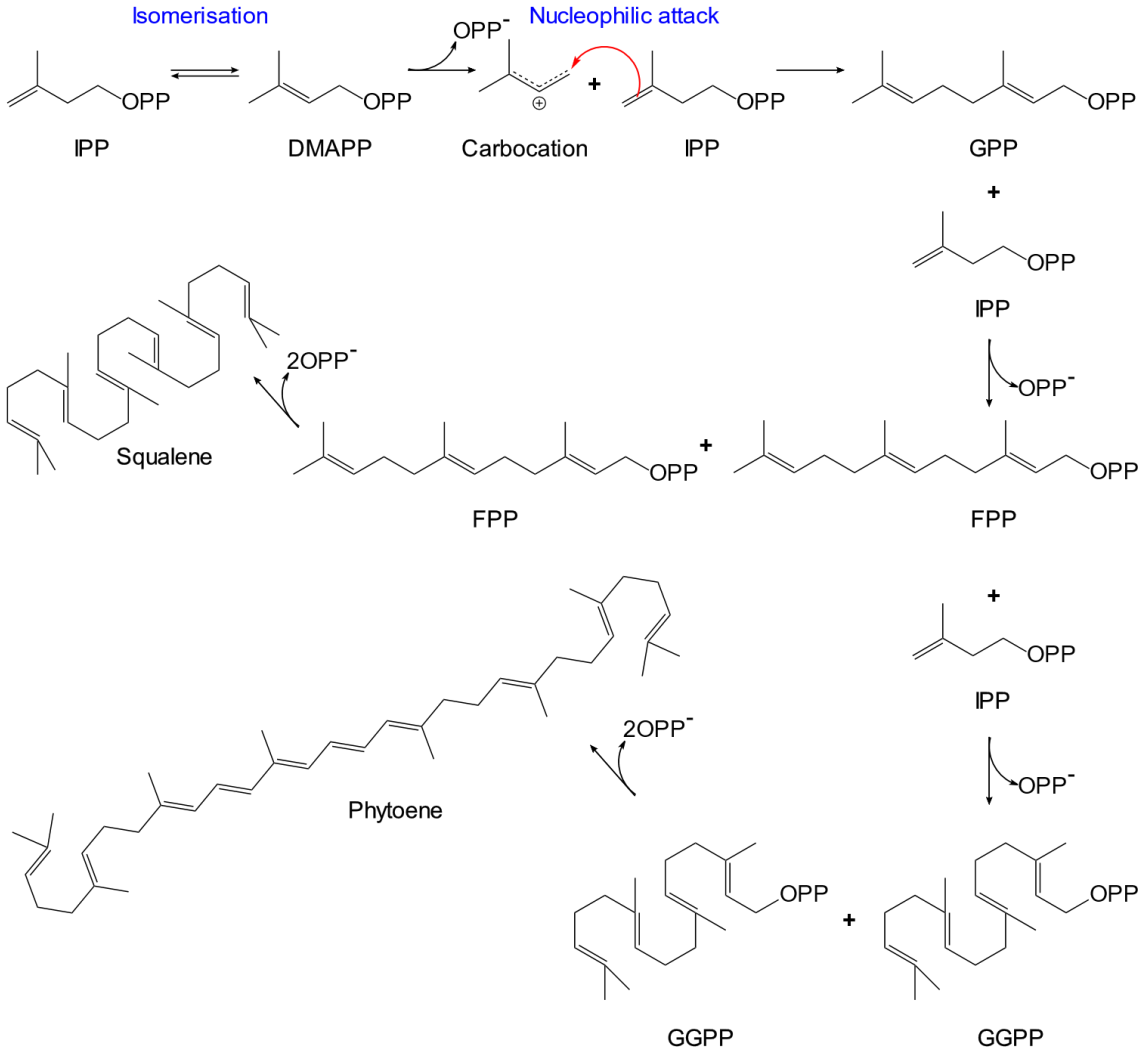


Figure 1.6 Polymerisation of IPP gives rise to the precursors of different terpenoids.

1.4 Triterpene biosynthesis

Triterpenes consist of six isoprene units with the molecular formula $C_{30}H_{48}$. In plants, triterpenes are derived from the cytoplasmic MVA pathway. Squalene synthase condenses two FPPs through pyrophosphate-to-pyrophosphate coupling into the 30-carbon squalene. An endoplasmic reticulum (ER)-localised squalene epoxidase

further converts squalene into 2,3-oxidosqualene, which is the precursor of both sterols and triterpenes (Chappell, 2002). Subsequently, oxidosqualene cyclases (OSCs), type II terpene synthases located at the ER outer membrane, cyclise 2,3-oxidosqualene into different sterol or triterpene scaffolds through a complex cascade of reactions.

When 2,3-oxidosqualene substrate enters the OSC active site, it becomes folded into a ‘chair-boat-chair’ conformation for sterol synthesis or a ‘chair-chair-chair’ conformation for triterpene synthesis (**Figure 1.7**) (Thimmappa et al., 2014). The epoxide group of 2,3-oxidosqualene is then protonated by the aspartate residue of the conserved DCTAE motif to initiate ring cyclisation, which leads to the formation of a dammarenyl cation (Thimmappa et al., 2014). Subsequent shifts in positive charge drive ring expansion through intermediate cations. After the formation of a pentacyclic carbocation, further methyl and hydride shifts move the positive charge across different carbons on the carbocation, generating a series of carbocation intermediates. Finally, specific carbocation intermediate(s) are deprotonated by the OSC, forming the final triterpene scaffold. Triterpene scaffolds are modified by tailoring enzymes that add hydroxyl, methyl, epoxy, glycosyl, and acyl groups (Thimmappa et al., 2014).

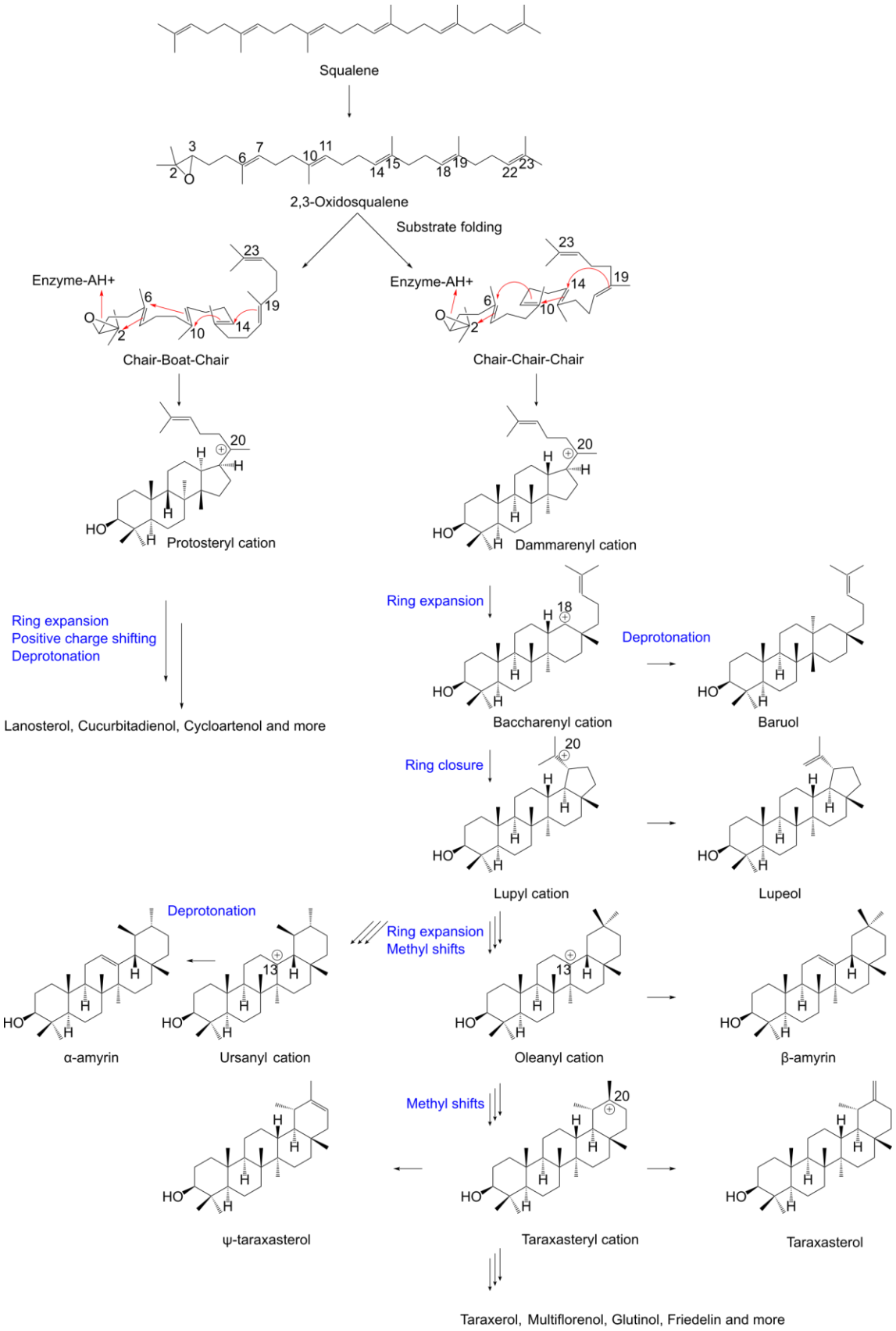


Figure 1.7 The biosynthesis of sterols and triterpenes from 2,3-oxidosqualene. 2,3-oxidosqualene is folded into chair-boat-chair conformation by sterol synthase and chair-chair-chair conformation by triterpene synthase. OSC performs substrate folding, reaction initiation by epoxide protonation, ring expansion, carbocation migration and deprotonation.

1.5 The structure and mechanism of oxidosqualene cyclases (OSCs)

1.5.1 Crystal structure of human lanosterol synthase

Due to the technical difficulties of purifying membrane-associated proteins, no crystal structures of plant OSCs have been acquired. A bacterial squalene-hopene cyclase from *Alicyclobacillus acidocaldarius* (AaSHC), which is structurally similar to OSCs, was structurally characterised using X-ray crystallography (1SQC) (Wendt et al., 1997). The only crystal structure of an OSC is of human lanosterol synthase (1W6K) (Thoma et al., 2004). In previous studies, this was used as a template for homology modelling of plant OSCs (Almeida et al., 2018; Dokarry, 2010; Salmon et al., 2016; Xue et al., 2018a).

The crystal structure of human lanosterol synthase was resolved at 2.1 Angstroms resolution (Thoma et al., 2004). Structural features including two α-α barrel domains connected by loops and β-structures, an active site located at the molecular centre, five QW-motifs and a nonpolar channel in domain 2 connected to a non-polar plateau at the surface (Figure 1.8). Three membrane-embedded helices were also observed.

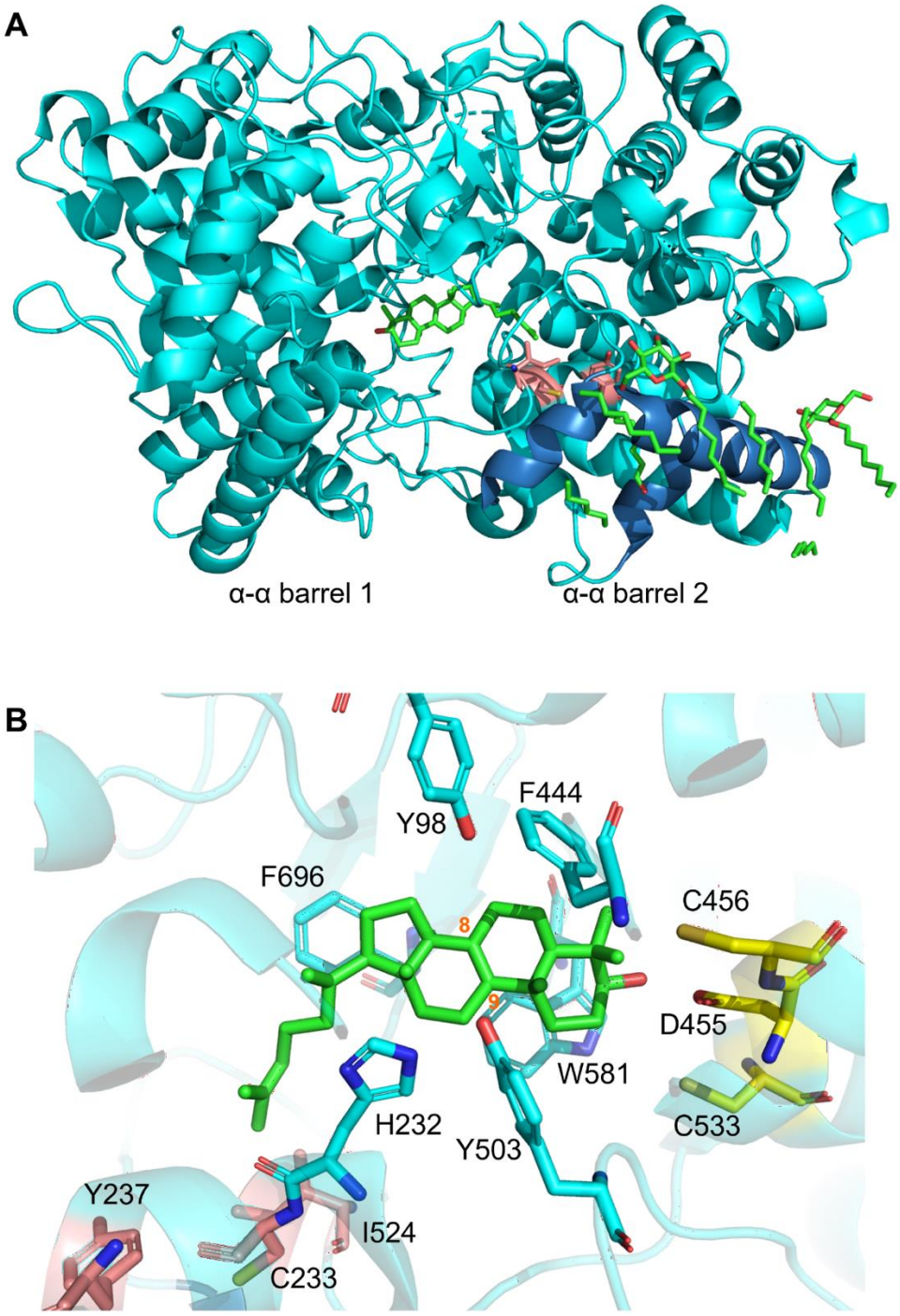


Figure 1.8 Crystal structure of human lanosterol synthase (1W6K). (A) Architecture of human lanosterol synthase. Three membrane-embedded helices are in dark blue. Lanosterol in the active site and detergents surrounding membrane embedded helices are in green. Residues gating substrate entrance are in pink. (B) A closer look at the active site of human lanosterol synthase. Y237, C233 and I524 (pink) gate the entrance to the active site. Y98 forces a chair-boat-chair conformation for 2,3-oxidosqualene. D455, stabilised by C456 and C533 (yellow), protonates the epoxide group on C2/C3. F444, Y503, W581, H232 and F696 are involved in the stabilisation of reaction intermediates. H232 acts as the final proton acceptor in the deprotonation reaction.

The hydrophobic substrate entrance channel for 2,3-oxidosqualene is gated by Y237, C233 and I524 and two rearrangeable strained loops (Figure 1.8). Within the active site, Y98 puts a constraint on the substrate forcing a 'chair-boat-chair' conformation. D455 in the DCTAE motif, stabilised by hydrogen bonding with C456 and C533, protonates the epoxide group to initiate the reaction. F444, Y503 and W581 stabilise intermediate cations through π-cation interaction during A-ring and B-ring formation. Subsequently, H232 and F696 stabilise the cation during C-ring formation. Once the five-carbon D-ring has formed, the positive charge shifts to C8/C9 (B ring and C ring boundary) during skeletal rearrangement. Finally, H232 deprotonates the cation at C8/C9 to terminate the reaction.

1.5.2 Cryo-EM structure of *Tripterygium wilfordii* (thunder god vine) friedelin synthase

T. wilfordii friedelin synthase (TwOSC) is the only plant OSC for which a structure has been experimentally determined using cryogenic electron microscopy (cryo-EM) (8J5Z) (Luo et al., 2023). The cryo-EM structure was resolved at 4.75 Angstroms with an overall tetrameric structure. In similarity with human lanosterol synthase, each monomer consists of two α-α barrel domains connected by loops and β-structures in addition to an N-terminal structure (Figure 1.9). A flat plane on one barrel domain, which consists of a loop, an α-helix and a 3_{10} helix, is embedded into the membrane. Within the active site, D488 both initiates the cyclisation reaction by protonating 2,3-oxidosqualene and terminates it by deprotonating the friedelyl cation to form friedelin. QM/MM simulation suggested that Y262, F477, F731, W615 and W420 are involved in stabilising the intermediates while saturation mutagenesis suggested the importance of L486, G536 and the K732-M733 dyad in product specificity.

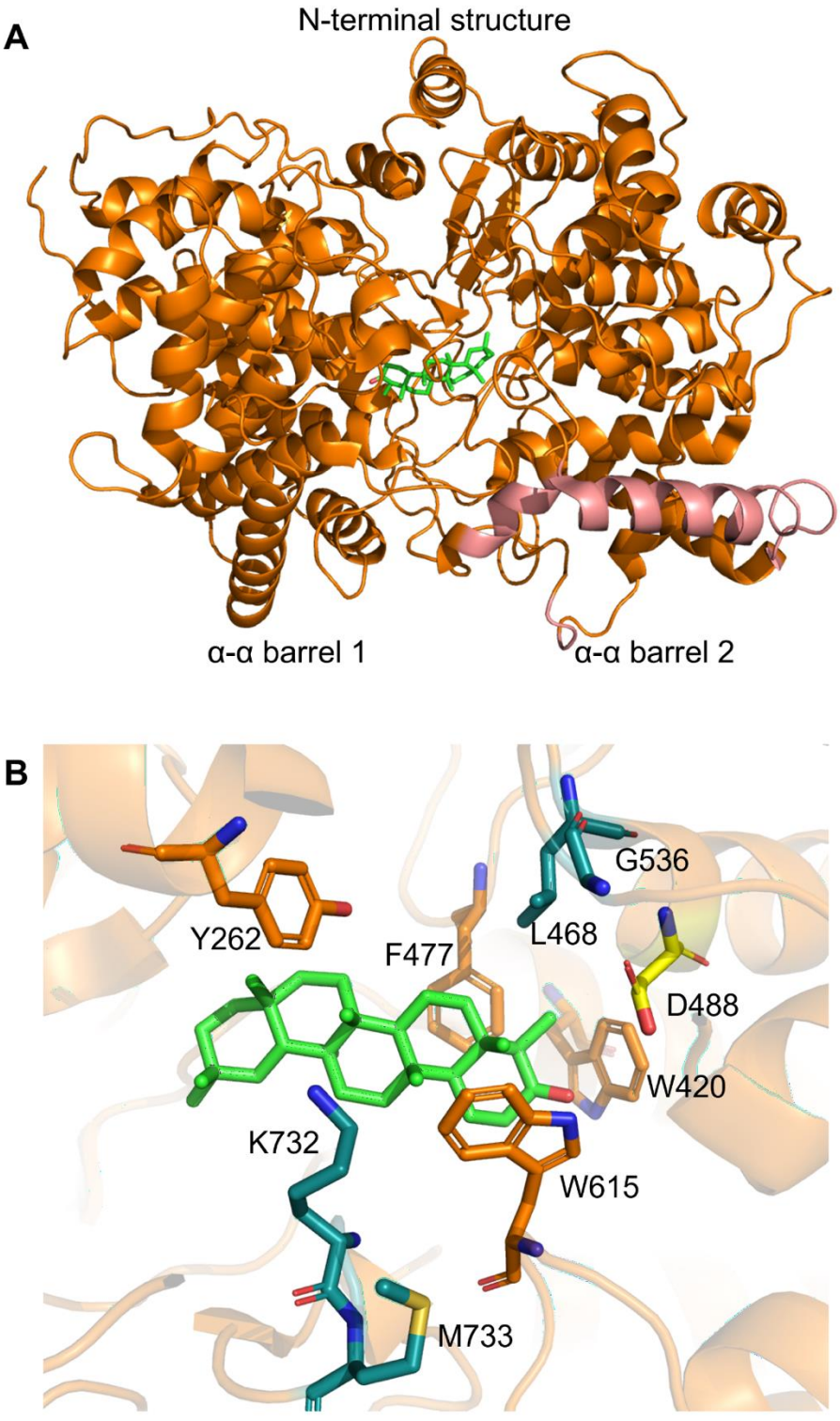


Figure 1.9 Cryo-EM structure of TwOSC monomer (8J5Z). (A) Architecture of a TwOSC monomer. Membrane embedded structures are in pink. Friedelin docked in the active site via Autodock Vina (Eberhardt et al., 2021) is in green. (B) A closer look at the TwOSC active site. D488 (in yellow) protonates the epoxide group of 2,3-oxidosqualene and deprotonates the friedelyl cation. Y262, F477, F731, W615 and W420 are involved in the stabilisation of reaction intermediates. L486, G536 and dyad K732-M733 are involved in determining product specificity (dark green).

1.6 Bioactivity of plant triterpenoids

Anti-microbial activity has been demonstrated for several types of triterpenoids. For example, avenacins, found in the roots of *Avena strigosa,* inhibit the growth of root-infecting fungi (Papadopoulou et al., 1999). Compared to wild-type plants, *A. strigosa* avenacin deficient mutants develop black lesions on their roots during *Gaeumannomyces graminis* var *tritici* infection. They also showed increased susceptibility to *Fusarium culmorum* and *F. avenaceum*. Subsequent work demonstrated that the anti-fungal activity of avenacin is likely attributed to the C12/13 epoxide (Geisler et al., 2013).

Some plant triterpene and derivatives have been documented to affect the microbiome. For example, the *Arabidopsis thaliana* triterpenoids thalianin, thalianyl fatty acid esters and arabidin play a role in shaping and maintaining the rhizosphere microbiota, with deficient mutants accumulating significantly more Bacteroidetes and significantly fewer Deltaproteobacteria in rhizospheres (Huang et al., 2019). Root triterpenoids were also found to promote the abundance of Proteobacteria and inhibit Actinobacteria *in vitro* (Huang et al., 2019). Similar studies have found that soyasaponin Bb secreted by *Glycine max* root promotes potentially beneficial bacteria *Novosphingobium* (Fujimatsu et al., 2020) and cucurbitacin B secreted by *Cucumis melo* root promotes *Enterobacter* and *Bacillus* and leads to resistance against pathogenic *Fusarium oxysporum* (Zhong et al., 2022).

Many plant triterpenoids have been shown to have anti-microbial activity against human pathogens, as well as anti-proliferative and anti-inflammatory activity on mammalian cells. For example, betulin, a lupeol derivative, extracted from the bark of *Croton macrostachys* inhibited the growth of six human pathogens with *Staphylococcus aureus* (bacteria) and two *Candida* species (fungi) being the most sensitive (Tene et al., 2009). Several triterpene fatty acid esters (TFAEs) extracted from *Celastrus rosthornianus* have also demonstrated cytotoxicity against HeLa cells (Wang, 2007).

Finally, triterpenes have been shown to function as growth regulators. In *Oryza sativa*, poaceatapetol and its fatty acid derivatives were shown to participate in pollen wall formation, and friedelin was shown to regulate grain development (Ma et al., 2024; Xue et al., 2018b). Further, in *Sorghum bicolor*, triterpenoids found in leaf wax were shown to contribute to cuticle formation, which confers drought tolerance (Busta et al., 2021).

1.7 Production of faradiol fatty acid esters (FAEs) in *C. officinalis*

This thesis investigates an OSC identified in *C. officinalis*, a member of the Asteroideae subfamily. This plant was used in traditional medicine for its wound healing and anti-inflammatory activities for many centuries (Colombo et al., 2015). Its floral extracts had previously been shown to be enriched with triterpene fatty acid esters (TFAEs) (Niżyński et al., 2015). During the preparation of this thesis, my colleague Dr Daria Golubova demonstrated that the anti-inflammatory activity of these extracts is conferred by C:16 hydroxylated TFAEs (faradiol FAEs) by

supressing the phosphorylation of STAT3. This leads to a reduction in the production of the anti-inflammatory cytokine, interleukin 6 in lipopolysaccharide (LPS)-activated human monocytic (THP-1) cells.

Before and during the collection of data for this thesis, complementary work to investigate the genetic basis of TFAE biosynthesis was conducted by colleagues in the Patron Laboratory (Dr Melissa Salmon, Dr Daria Golubova, and Dr Connor Tansley), by Professor Maria O'Connell (University of East Anglia), and by the core bioinformatics team at the Earlham institute, which is led by Dr David Swarbreck (Earlham Institute). First, GC-MS and LC-MS analysis (available at http://doi.org/10.5281/zenodo.13869958) by Dr Melissa Salmon, confirmed that faradiol myristate and faradiol palmitate are particularly abundant in this species. Next, to enable the elucidation of the biosynthetic pathway of these compounds, the *C. officinalis* genome was sequenced using Illumina, PacBio, Chromium linked-read, and OmniC technologies and assembled (GCA_964273985.1; https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_964273985.1/). In addition, the transcriptomes of rays, discs and leaves were sequenced using Illumina and PacBio Isoseq and assembled (PRJEB80524; https://www.ebi.ac.uk/ena/browser/view/PRJEB80524), and differential gene expression analysis was performed (available at http://doi.org/10.5281/zenodo.13869958).

Finally, the biosynthetic pathway of faradiol FAEs was elucidated (Figure 1.10) by Dr Melissa Salmon, Dr Daria Golubova and Dr Connor Tansley. This work identified that 2,3-oxidosqualene is cyclised into ψ-taraxasterol by *C. officinalis* taraxasterol synthase (CoTXSS). Two cytochrome P450s (CoCYP716A392 and CoCYP716A393) were identified and characterised that add a hydroxyl group to ψ-taraxasterol C16 converting ψ-taraxasterol into faradiol. In this thesis, they are named as CoCYP1 and CoCYP2 respectively. Finally, two pairs of acyl transferases were characterised that add myristate or palmitate either to faradiol to form faradiol FAEs (CoACT1 and CoACT2) or to ψ-taraxasterol for form ψ-taraxasterol FAEs (CoACT4 and CoACT5).

Notably, CoTXSS was observed to be multi-functional, predominantly producing ψ-taraxasterol, but also taraxasterol, β-amyrin and lupeol (Figure 1.10 and 1.11). This was particularly interesting as orthologues of CoTXSS that had been previously identified from *Taraxacum kok-saghyz* and *T. coreanum* predominantly produced taraxasterol (Pütter et al., 2019; Han et al., 2019). Further, Dr Daria Golubova's inflammatory bioassays showed that while ψ-taraxasterol, faradiol and their FAEs were anti-inflammatory, taraxasterol is pro-inflammatory, increasing the abundance of the inflammatory cytokine IL-6 in human monocytic (THP-1) cells.

The above work, together with much of the work described in Chapters 3 and 4 of this thesis has been published (Golubova et al., 2025).

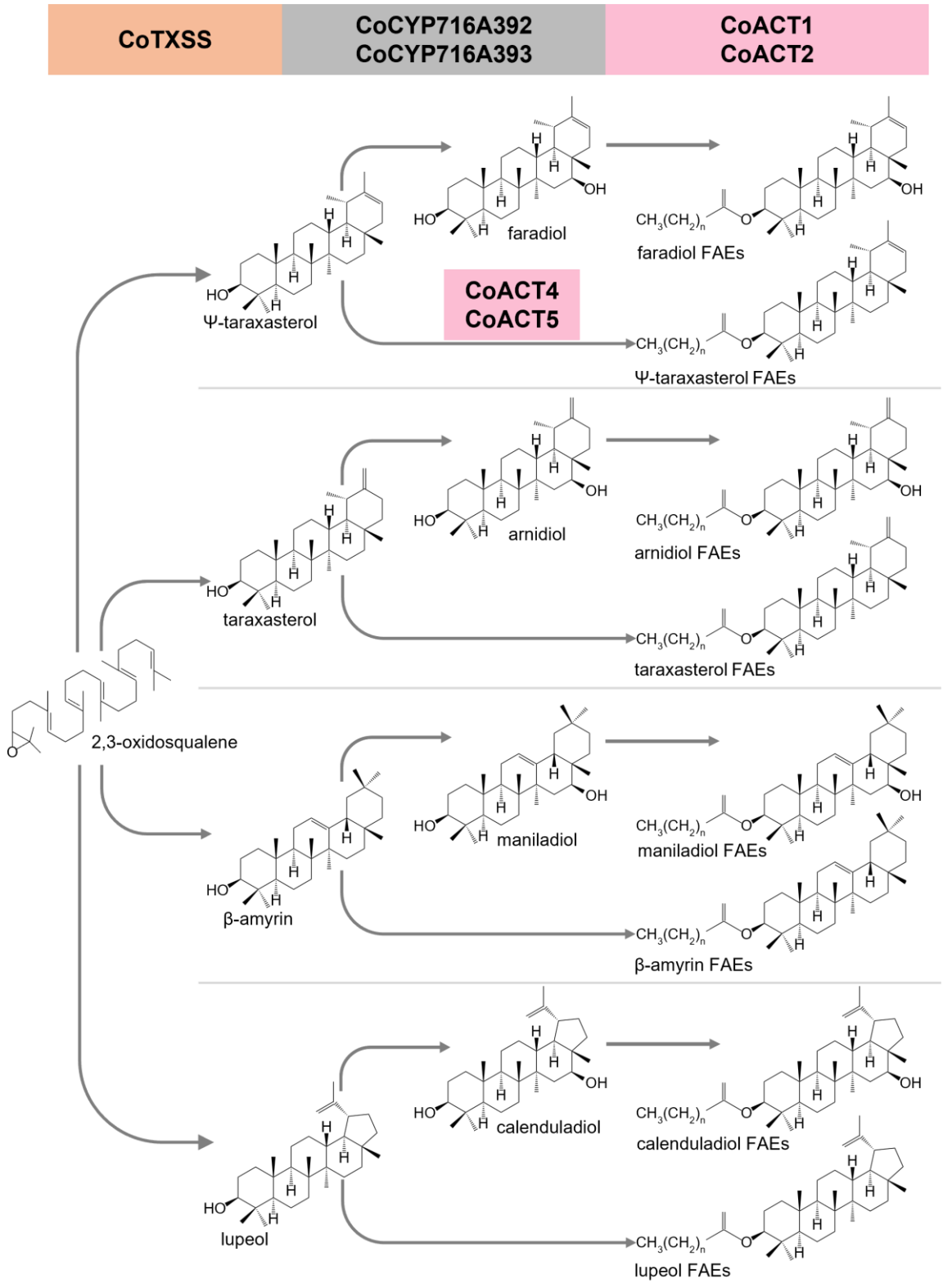


Figure 1.10 The biosynthesis of triterpene FAEs in *C. officinalis*. An oxidosqualene cyclase (CoTXSS) converts 2,3-oxidosqualene into ψ-taraxasterol (major product) and taraxasterol, β-amyrin and lupeol (minor products). These are hydroxylated by cytochrome p450s (CoCYP716A392 and CoCYP716A393). C16 hydroxylated compounds are acylated by acyl transferases (CoACT1 and CoACT2). Triterpene scaffolds are acylated by acyl transferases (CoACT4 and CoACT5). Figure adapted from Golubova et al., 2025.

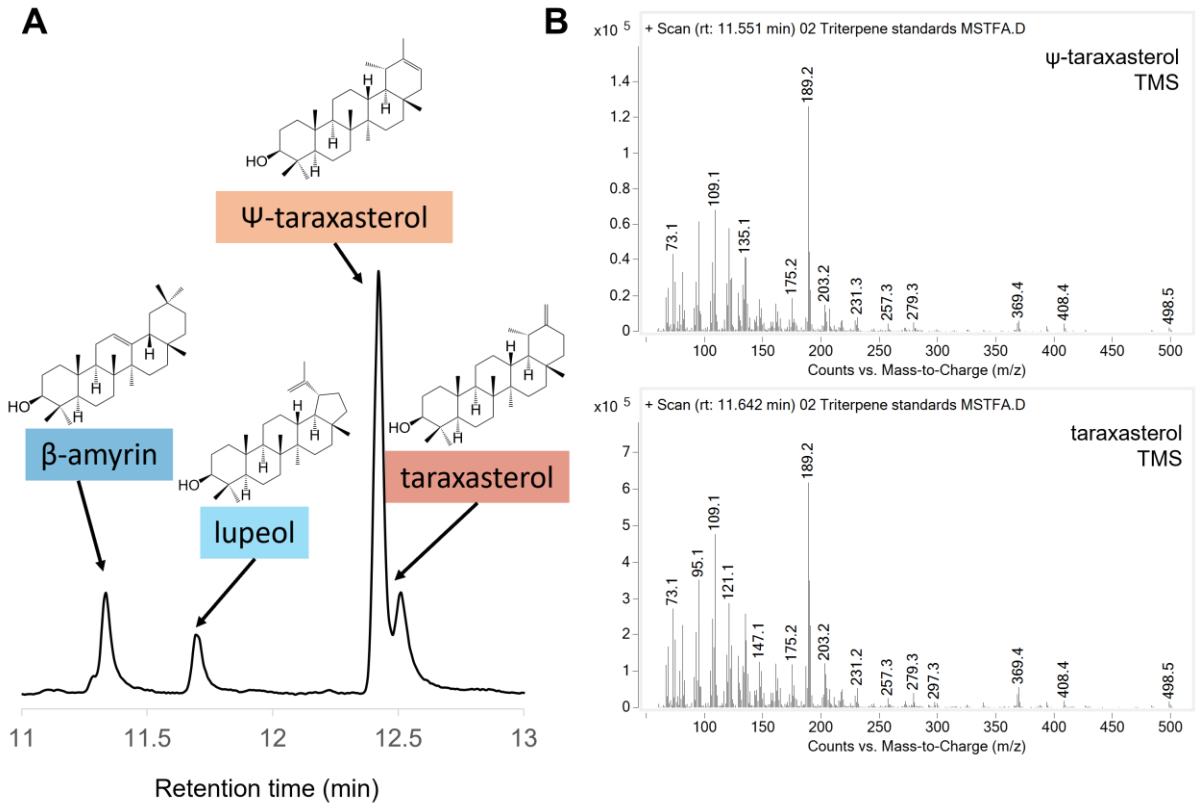


Figure 1.11 GC-MS analysis of *N. benthamiana* leaves transiently expressing CoTXSS. (A) Representative total ion chromatogram of trimethylsilyl (TMS) derivatised extracts of *N. benthamiana* leaves transiently expressing CoTXSS. (B) Mass spectra for TMS derivatised ψ-taraxasterol and taraxasterol standards. ψ-taraxasterol has lower counts of ions with mass-to-charge ratio of 109.1 than taraxasterol.

1.8 Aims of this thesis

Gene mining techniques enable the identification of enzymes involved in metabolite production. However, the structural basis of the substrate/product specificity of enzymes and the subcellular locations of enzymes involved in such pathways often remain unknown. Such knowledge is critical as it can serve as the basis for enzyme engineering to change or improve enzyme specificity or activity. For example, mutation of a single residue (F726T) was able to convert *Alisma orientale* cycloartenol synthase into a protostadienol synthase (Zhang et al., 2023). In addition, although pathway products often show tissue or cell-type specific accumulation, knowledge of how this specificity is regulated often remains elusive. Identification of transcriptional regulatory factors can be used to manipulate pathway expression. For example, overexpression of transcription factor SlMYB75 in *Solanum lycopersicum* (tomato) was used to enhance anthocyanin accumulation and aroma production in fruits (Jian et al., 2019).

In this thesis, I investigate CoTXSS, a key enzyme in the production of anti-inflammatory triterpenoids found in *C. officinalis*. I explore the evolution, specificity, sub-cellular localisation and regulation of CoTXSS. The aims of this thesis are to:

1. Uncover the evolutionary history of TXSSs determining when and how these enzymes evolved (Chapter 3).
2. Understand the molecular basis of TXSS product specificity (including whether the production of different products has been selected or is a product of genetic drift) and explore the mechanism of catalysis (Chapter 4).
3. Identify regulators of CoTXSS (and faradiol FAE) expression (Chapter 5).
4. Investigate the molecular basis of CoTXSS localisation to the outer membrane of the ER and explore if this can be altered by the application of protein design methods (Chapter 6).

2. Chapter 2 General Methods

2.1 Genomics, transcriptomics and evolution bioinformatics

2.1.1 Identification of candidate taraxasterol synthases (TXSSs)

The genomes or transcriptomes of 60 Asteraceae species, 10 non-Asteraceae Asterales species and 11 non-Asterales species that have previously been reported to accumulate taraxasterol were downloaded from NCBI WGS, the Chinese National Genomics Data Center, NCBI TSA, orcAE and the 1kp project databases (Supplementary Table 1). Prior to the start of this project, the *Calendula officinalis* genome was sequenced with Illumina, PacBio, Chromium linked-read, and OmniC technologies and assembled and annotated at the Earlham Institute and deposited under accession number GCA_964273985.1. Transcriptomes of leaf and flower tissues from five Asteraceae species (*Achillea millefolium*, *Calendula arvensis*, *Eupatorium cannabinum*, *Inula britannica* and *I. ensifolia*) were also sequenced and assembled at the Earlham Institute.

To identify candidate TXSSs, the peptide sequence of *C. officinalis* taraxasterol synthase (CoTXSS) was used as a query in BLASTp to search the translated coding sequences of the genomes and transcriptomes described above. Sequences with more than 50% query coverage and an E-value of less than 0.1 were used to construct a neighbour-joining phylogenetic tree together with the protein sequences of 165 previously characterised OSCs. Sequences within the same clade as previously characterised TXSSs were identified as candidate TXSSs.

2.1.2 Construction and visualisation of an Asteraceae species tree

The phylogenetic relationships of all Asteraceae species described in Chapter 2.1.1 were acquired from Lifemap (Vienne, 2016) and exported as a newick file. The resulting species tree was visualised using Interactive Tree Of Life (iTOL) (Letunic and Bork, 2024).

2.1.3 Construction of a maximum likelihood phylogenetic tree

To construct a phylogenetic tree of TXSSs (Figure 3.4), peptide sequences of greater than 500 residues were selected. All TXSSs together with 12 outgroup non-TXSS OSCs were aligned using MAFFT with 1000 rounds of iterative refinement from global pairwise alignments (Flags: --globalpair --maxiterate 1000) (Katoh and Standley, 2013). Gap were trimmed using trimAl with an automatic method (flag: -gappyout) (Capella-Gutiérrez et al., 2009). A Maximum-likelihood tree of all sequences was built using iqtree2 (Minh et al., 2020) with a JTTDC-Mut+G+I model and 1000 rounds of Ultrafast Bootstrap Approximation (Minh et al., 2013). The tree was visualised in iTOL (Letunic and Bork, 2024).

Equivalent methods were used to construct a phylogenetic tree of plant OSCs (Figure 3.7), mixed amyrin synthases (MASs) and TXSSs (Figure 3.9) and MYB DNA binding domains (Figure 5.3, 5.5, 5.7, 5.8).

2.1.4 Detection of positive selection

A maximum likelihood phylogenetic tree of selected TXSSs was constructed as described in Chapter 2.1.3. The corresponding coding sequences (CDSs) were mapped to the sequence alignment using pal2nal (Suyama et al., 2006), and all gaps in the CDS alignment were removed.

CodeML from the PAML package was used to fit different codon substitution models (M0, M1, M2, M3, M7 and M8) to the data and to estimate parameters using the maximum likelihood method (Yang, 2007, p. 200). Likelihood ratio tests (LRTs) were conducted to compare log likelihood values of each pair of models.

CodeML was also used for branch-site model testing. Non-basal Cichorieae TXSSs were selected as the foreground with all other TXSSs constituting the background. A branch-site model A null model, where omega is allowed to vary between foreground and background (in job control file: model = 2) while only neutral and purifying selections are allowed in foreground (fix_omega = 1, omega = 1), and alternative model, where everything stays the same except that the foreground branch is allowed to be under positive selection (fix_omega = 0, omega = 1), were fitted to the data. LRT was performed to compare log likelihood values of foreground and background. Sites with more than 50% probability under positive selection in foreground were retrieved by Bayes Empirical Bayes method (Yang et al., 2005).

2.1.5 Bayesian phylogenetic analysis

The CDSs of selected sequences were codon-aligned, and all gaps in the alignment were removed. BEAUTi2 was used to split the alignment into three partitions and generate an input file. The clock models and tree were linked among the three partitions. Analysis was performed using BEAST2 with a GTR model and an *Artemisia* fossil and a *Cichorium intybus* type fossil as calibrators (Bouckaert et al., 2014; Tremetsberger et al., 2013; WANG, 2004). A strict molecular clock with rate=1 was used. A Yule Model was used as a prior model and 10,000,000 rounds of MCMC were run with 10% of burn-in. The ESS of each parameter was verified using Tracer (Rambaut et al., 2018) and the maximum clade credibility tree was generated with TreeAnnotator. 95% length HPD was used to represent the branch length range.

2.1.6 Genome synteny analysis

Genome synteny was analysed using the progressive Mauve algorithm and visualised using Mauve Plugin for Geneious Prime 2024.0.5 (Darling et al., 2010). *C. officinalis* contigs 1, 2, 11 and 13 were compared to *Cynara cardunculus* chromosome LG14, *Lactuca sativa* chromosome 7 and *Helianthus annuus* chromosome 4 sequences, acquired from NCBI under accessions NC_037541.1, NC_056629.2 and NC_035436.2 (Badouin et al., 2017; Reyes-Chin-Wo et al., 2017; Scaglione et al., 2016).

2.1.7 Identification of candidate transcription factor binding sites

The promoters of all *C. officinalis* gene were arbitrarily defined as the region 1000 bp upstream of the transcription start site. These sequences were analysed together with the 5' UTRs. To calculate the background frequencies of all bases, the promoters of all *C. officinalis* genes were acquired using a custom shell script and the fasta-get-markov function of FIMO (Grant et al., 2011).

The Position Frequency Matrices (PFMs) of plant transcription factors were acquired from JASPAR 2024 (Rauluseviciute et al., 2024) and used to search the promoter regions of target genes (*pCoTXSS, pCoCYP1, pCoCYP2, pCoACT1 and pCoACT2*) using FIMO. Hits with p-value<10^{-4} and q-value<0.05 were retained.

To identify if there were additional candidate R2R3-MYB TFBSs undetected by FIMO, the target promoters were analysed using the degenerate search function in Benchling with the search strings "NNRTTHGGY" for MYB24, "YAMCWAMY" for MYB4 and MYB111 and "NNNWWUTTAGGTNNN" for MYB27.

2.1.8 Co-expression analysis

Raw read counts from the *C. officinalis* transcriptomes of ray, disc and leaf tissues (described in Chapter 2.1.1) were normalised by log2 transformation using log2(counts + 1) function in R and stored in a separate data frame normalised_counts. Then, Pearson correlations between expressed genes were calculated through cor(normalised_counts, method="pearson") function in R and Pearson correlations between genes involved in faradiol fatty acid ester biosynthesis were acquired.

2.1.9 Identification of R2R3-MYB candidates

Peptide sequences of all *Arabidopsis thaliana* R2R3-MYBs (AtMYBs) were downloaded from the TAIR database (Reiser et al., 2024; Stracke et al., 2001). Based on previous annotations, the first 130 residues of all MYB sequences were retained as putative DNA binding domains (Wang et al., 2020). These DNA binding domains were used as queries in BLASTp to search the translated protein sequences of *C. officinalis* genes more highly expressed in rays and discs than in leaves as determined by differential gene expression analysis conducted prior to the start of this project (Golubova et al., 2025). Hits with an E-value of less than 0.01 were retained as candidate CoMYBs and ranked by rays-to-leaves log2FC values. The first 130 residues of candidate CoMYBs were also destinated as putative DNA binding domains. A maximum likelihood phylogenetic tree of candidate CoMYBs and AtMYBs putative DNA binding domains was constructed using the methods described in Chapter 2.1.3.

2.2 Structural modelling, molecular dynamic simulation and protein design

2.2.1 Structural modelling with AlphaFold2

Structural models of CoMAS, CoTXSS and TkTXSS were generated using AlphaFold2 (Jumper et al., 2021) on the ADA HPC cluster at University of East Anglia. The top ranked relaxed model for each enzyme (pLDDT>0.95) was selected

for further investigation. Structural models were visualised and annotated in PyMOL 3.1 (Schrödinger).

2.2.2 Substrate docking

A taraxasteryl cation structural model was created using ChemDraw and geometrically optimised with forcefield HF/6-31G* in Gaussian 16 (Frisch et al., 2016). This cation model was manually docked in the active site of the CoTXSS and TkTXSS models with reference to the position of lanosterol in the human lanosterol synthase crystal structure (1W6K) (Thoma et al., 2004). The orientations of the hydroxyl group and A-, B- and C-rings were consistent between the taraxasteryl cation and lanosterol. Energy minimisation was performed on the enzyme-intermediate complex using AMBER 22 (Case et al., 2022).

2.2.3 Classical molecular dynamic simulation

The taraxasteryl cation forcefield was defined using General Amber Force Field (GAFF) and the TXSS forcefield was defined using ff19SB. The TXSS-taraxasteryl cation complex was dissolved in a 12 Angstrom x 12 Angstrom x 12 Angstrom TIP3P water box and the charge was balanced by adding sodium ions. Subsequently, 0.15 M sodium and chloride ions were added to form a buffer solution. The initial topology and parameter files for the dissolved complex were generated by tleap function of AMBER 22.

AMBER 22 was used to perform energy minimisation and molecular dynamic simulation. Three rounds of energy minimisation were performed on the solvated complex in which all solvents were constrained in the first round, backbone atoms were constrained in the second round and no atoms wert constrained in the third round. Then, 100 ps of NVT was performed on the system to heat it up from 0 K to 300 K. Next, 100 ps of NPT was performed on the system at 300 K with a target pressure of 1 atm. Finally, 100 ns of NPT production simulation was performed on the system at 300 K to produce a trajectory. During simulations, the SHAKE algorithm was applied, and a 12 Angstrom cutoff was imposed on both van der Waal and electrostatic interactions. A periodic boundary condition was applied to the system. Six replicates were run for each enzyme-ligand complex. The last 70 ns of production simulation of each replicate was used for analysis with cpptraj.

2.2.4 Combined quantum mechanics and molecular mechanics (QM/MM)

Starting structures for CoTXSS–taraxasteryl cation QM/MM simulations were chosen from MD simulation frames based on the distance between D385 and proton (< 3 Angstroms) and the orientation of taraxasteryl cation in the active site. Starting structures were minimised at QM/MM level. Semi-empirical method PM6 was applied to taraxasteryl cation, Y273 and D385 (Stewart, 2007). ff19SB forcefield was applied to the rest of the complex (Tian et al., 2020). Subsequently, QM/MM umbrella sampling (US) was performed along a reaction coordinate of the difference between the length of breaking C-H bond (on taraxasteryl cation) and the forming O-H bond (between D385 and taraxasteryl cation). PM6 was applied to taraxasteryl cation, Y273 and D385. ff19SB forcefield was applied to the rest of the complex. The

sampling time in each window was 10 ps. All simulations were performed with AMBER 22. The weighted histogram analysis method (WHAM) was used to construct the free energy profile for each deprotonation reaction (Kumar et al., 1992).

2.2.5 Computational protein design

In the CoTXSS structural model, three helices that are predicted to embedded into the plant ER membrane were identified based on the membrane-embedded helices of human lanosterol synthase (HsLS) crystal structure (1W6K) (Thoma et al., 2004). Exposed hydrophobic residues on helices 1 (254-258) and 3 (555-560) were mutated to E. For helix 2 (331-350), SolubleMPNN was run on either all exposed residues or on the entire helix with a sampling temperature of 0.1 and backbone Gaussian noise of 0.1 Angstrom (Goverde et al., 2024). K, R, H, and C were excluded in SolubleMPNN. In addition, as a standalone rational design method, exposed hydrophobic residues on helix 2 were substituted to either E or Q (Woolfson, 2023).

Structural models of redesigned CoTXSS variants were built using AlphaFold2 and the pLDDT scores for top-ranked models were checked (pLDDT>0.95) (Jumper et al., 2021). Rosetta filter ExposedHydrophobics was used to calculate the hydrophobic solvent accessible surface areas of CoTXSS variants and Rosetta filter RMSD was used to calculate RMSD between CoTXSS and its variants (Fleishman et al., 2011).

2.3 Molecular biology methods

2.3.1 Assembly of standardised DNA parts

The coding sequence (CDS) of *Tragopogon dubius* TXSS was codon optimised for *Nicotiana benthamiana* using the Twist Bioscience codon optimisation tool (https://www.twistbioscience.com/resources/digital-tools/codon-optimization-tool). The resulting sequence (without the stop codon) was flanked by inverted BsaI sites and chemically synthesised and cloned into pTwist by Twist Bioscience (Twist Bioscience, South San Francisco, CA, USA) resulting in Level 0 parts compatible with the phytobrick standard (Patron et al., 2015).

The CDSs of CoMYB24 and CoMYB111 and a sequence encoding a signal peptide of OsRAmy3A (Engler et al., 2014) were chemically synthesised flanked by sequences to enable SapI mediated cloning into pUAP4 (Sauret-Güeto et al., 2020) (Addgene #136079) (Integrated DNA Technologies, Leuven, Belgium). The CDS of *Methylococcus capsulatus* lanosterol synthase (McLS) (Lamb et al., 2007) was codon optimised for *Nicotiana benthamiana* using the Twist Bioscience codon optimisation tool. The resulting sequence was chemically synthesised by Twist Bioscience (Twist Bioscience, South San Francisco, CA, USA). To make Level 0 parts with and without stop codons, the McLS CDS was amplified using primers with 5' extensions to enable SapI mediated cloning into pUAP4. The C-terminal region of McLS CDS was amplified from using primers with 5' extensions to enable SapI mediated cloning into pUAP4. CoTXSS variants were amplified from their

corresponding pENTRY constructs (pEPHSdGM0055 to pEPHSdGM0060; Chapter 2.3.4) using primers with 5' extensions to enable SapI mediated cloning into pUAP4. The mCherry CDS was amplified from pICSL80007 (Engler et al., 2014) (Addgene #50323) using primers with 5' extensions to introduce a HDEL ER retention signal and stop codon, and to enable SapI mediated cloning into pUAP4. A Level 0 part containing the CoTXSS CDS was obtained from Dr Melissa Salmon (pEPMS1CB0001; Addgene #227533). A version without a stop codon was created by amplification using primers with 5' extensions to enable SapI mediated cloning into pUAP4. A YFP CDS was amplified from pICSL30004 (Engler et al., 2014) (Addgene #50302) using primers with 5' extensions to introduce a 5 x GS linker and sequences to enable SapI mediated cloning into pUAP4. The promoter regions *pCoTXSS*, *pCoCYP2*, *pCoACT1* and *pCoACT2* were amplified from *C. officinalis* genomic DNA using gene specific primers with 5' extensions to enable cloning into pUAP1 (Patron et al., 2015) (Addgene#63674) or pUAP4. *pCoACT1* was amplified in two fragments using primer to mutate -320 G→C to remove an internal BpiI recognition site. The sequences of all primers used in cloning are provided in Supplementary Table 2. Linear synthetic fragments and amplicons were cloned into pUAP4 using the reaction and conditions described in Tables 2.1 and 2.2, resulting in standardised Level 0 parts compatible with the phytobrick standard. The completed resections were used to transform *E. coli* as described in chapter 2.3.5. A table of all Level 0 parts used in this thesis is provided in Supplementary Table 3.

Component	Concentration	Volume
pUAP4/pUAP1	66 ng/µL	0.25 µL
Insert	30 - 100 ng/µL	variable
NEB rCutSmart Buffer	10X	1 µL
ATP	100 mM	0.2 µL
SapI/BpiI	10 U/µL	0.5 µL
NEB T4 ligase	200 U/µL	0.2 µL
Water		Up to 10 µL

Table 2.1 Reaction mixes of Golden Gate assembly reactions used to clone linear DNA fragments into Level 0 acceptors.

Steps	Temperature	Time	Cycles
Digestion	37 °C	5 min	35x
Ligation	16 °C	5 min	
Final digestion	55 °C	10 min	1x
Inactivation	80 °C	5 min	1x
Hold	4 °C	Infinite	1x

Table 2.2 Reaction conditions for Golden Gate assembly of Level 0 DNA parts.

2.3.2 Assembly of Level 1 plant expression constructs

Transcriptional units were assembled from Level 0 standard parts by Golden Gate assembly into the Level 1 acceptor pICH47732 (Engler et al., 2014) (Addgene#48000), pCk1 (Sauret-Güeto et al., 2020) (Addgene#136695), or pCk2

(Sauret-Güeto et al., 2020) (Addgene#136696). Constructs were assembled in the reaction provided in Table 2.2, which were then incubated in the conditions described in Table 2.3. A table of all Level 1 constructs used in this thesis is provided in Supplementary Table 4.

Component	Concentration	Volume
Acceptor	66 ng/µL	variable
Insert(s)	66 ng/µL	variable
NEB T4 ligase Buffer	10X	2 µL
Bovine Serum Albumin	2 mg/mL	1 µL
BsaI	10 U/µL	1 µL
NEB T4 ligase	200 U/µL	1 µL
Water		Up to 20 µL

Table 2.3 Reaction conditions for Golden Gate assembly of Level 1 constructs.

2.3.3 Site-directed mutagenesis

To introduce mutations into the CDSs of OSC sequences, site-directed mutagenesis was performed on Level 1 constructs: pEPMS1CB0003 (CoMAS), pEPMS1CB0001 (CoTXSS) and pEPMS1CB0008 (TkTXSS) and on the pENTRY clone pEPHSdGM0050 (CoTXSS). Mutations were introduced using a modified QuikChange mutagenesis PCR protocol in which mutations are introduced in overlapping PCR primers (Liu and Naismith, 2008). A table of primers used for mutagenesis is provided in Supplementary Table 2. The reaction mix used for QuikChange is provided in Table 2.4 and the reaction conditions are provided in Table 2.5.

Component	Concentration	Volume
NEB Q5 buffer	5X	5 µL
dNTP mix	10 mM	0.5 ng
Forward primer	10 µM	0.125 µL
Reverse primer	10 µM	0.125 µL
Q5 polymerase	2 U/µL	0.25 µL
Template	10 ng/µL	1 µL
Water		Up to 25 µL

Table 2.4 Reaction mixes for site-directed mutagenesis.

Steps	Temperature	Time	Cycles
Initial denaturation	98 °C	5 min	1x
Denaturation	98 °C	15 seconds	30 x
Annealing	55–65 °C	15 seconds	
Extension	72 °C	30 seconds per kb + 1 min	
Final extension	72 °C	10 min	1x
Hold	4 °C	Infinite	1x

Table 2.5 Reaction conditions for site-directed mutagenesis.

The resulting PCR amplicons were visualised on 0.8% to 1.5% agarose gel containing SYBR-Safe (1:20,000). The remaining PCR reaction was incubated with 10 U DpnI (NEB) at 37 °C for 2 hours to digest the methylated plasmid template.

2.3.4 Cloning into *E. coli* expression constructs

Gateway cloning was to assemble *E. coli* expression constructs. The CDSs of McLS and CoTXSS were amplified using primers with 5' overhands to introduce attB1 and attB2 sites. PCR mix and reaction conditions are the same as Table 2.4 and Table 2.5. The resulting amplicons were cloned into pDONR207 using BP clonase (ThermoFisher, Waltham, MA). In each BP reaction, 150 ng attB-PCR product, 150 ng pDONR207, 2 µL Invitrogen BP Clonase II enzyme mix, and TE buffer (pH=8.0) were mixed in a 10 µL reaction mix and incubated at 25 degrees for 1 hour. Subsequently, 2 µg proteinase K was added and incubated at 37 °C for 10 mins to terminate the reaction. Details of the resulting pENTRY clones are provided in Supplementary Table 5. Site-directed mutagenesis was performed on CoTXSS pENTRY clones as described in Chapter 2.3.3

The resulting pENTRY clones were used in LR clonase reactions (ThermoFisher). with the pDESTINATION vector pH9GW (Reece-Hoyes and Walhout, 2018). In each LR reaction, 150 ng pENTRY clone, 150 ng pH9GW, 2 µL Invitrogen LR Clonase II enzyme mix, and TE buffer (pH=8.0) were mixed in a 10 µL reaction mix and incubated at 25 degrees for 1 hour. Subsequently, 2 µg proteinase K was added and incubated at 37 °C for 10 mins to terminate the reaction. Details of the resulting pEXPRESSION clones are provided in Supplementary Table 5.

2.3.5 Transformation of *E. coli*

An aliquot (1.5 µL) of Golden-Gate digestion ligation, site-directed mutagenesis, BP clonase and LR-clonase reactions was added to 5 µL competent *E. coli* NEB 5-alpha (for plasmid amplification) or NEB SHuffle T7 express (for protein expression). Cells were incubated on ice for 20-30 mins, transferred to 42 °C for 30 seconds, then returned to ice 5 mins. Pre-heated SOC media (80 µL) was added to the cells before incubation at 37 °C with 200 rpm shaking for 1 hour. An aliquot (40 µL) of the transformation mix was spread on LB agar plates containing appropriate antibiotics

for plasmid selection, 1 mM IPTG and 20 mg/mL X-gal (for blue-white selection) and incubated at 37 °C overnight.

2.3.6 Validation of constructs

One to three colonies were selected and used to inoculate 4 mL LB with appropriate antibiotics. Cultures were incubated at 37 °C; 220 rpm overnight. Plasmid DNA was extracted using the QIAprep Spin Miniprep Kit (Qiagen). Plasmid concentrations were measured using a NanoDrop One Microvolume UV-Vis Spectrophotometer (Thermo Scientific) before dilution to 66 ng/μL for Sanger sequencing with appropriate sequencing primers (GENEWIZ from Azenta).

2.3.7 Protein expression and extraction from *E. coli*

NEB SHuffle T7 express colonies containing protein expression constructs were cultured in 10 mL LB with 50 μg/mL kanamycin at 37 °C, 220 rpm shaking overnight. 2 mL saturated culture was added to 50 mL LB with 50 μg/mL kanamycin in 250 mL Erlenmeyer flask and grown at 37 °C; 220 rpm until OD600 reached 0.4 to 0.8. protein expression was induced by adding 20 μL 1ML IPTG and cultures were incubated at 18/24/37 °C; 220 rpm for 20 hours.

E. coli cells were harvested by centrifugation at 5000x *g* at 4 °C for 10 min. Whole cell samples were resuspended in suspension buffer (water, 50 mM Tris-HCl pH=8, 500 mM NaCl, 20 mM imidazole, 10% glycerol, and 1% Tween-20). Samples for cell lysis were resuspended in lysis buffer (water, 50 mM Tris-HCl pH=8, 500 mM NaCl, 20 mM imidazole, 10% glycerol, 1% Tween-20, 1 mg/mL lysozyme, 1 mM EDTA, and 10 mM β-mercaptoethanol) and lysed at 30 degrees with 220 rpm shaking for 40 min. Lysed samples were centrifuged at 13,000x g for 10 mins and supernatants were retained as soluble fractions.

2.3.8 SDS-PAGE and Western Blot

10 μL protein extract (whole cell or supernatant) was mixed with 5 μL Invitrogen NuPAGE LDS Sample Buffer and 5 μL water and boiled at 98 degrees for 15 min. Protein samples were separated on 4–12% NuPAGE Bis-Tris gels (Invitrogen) in 1x MOPS buffer (Invitrogen) at 120 V. Proteins were transferred to a preconditioned nitrocellulose membrane using an iBlot 3 Western Blot Transfer Device (Broad range protocol, 25 V, 6 min). The membrane was blocked with 5% dried skimmed milk powder in 1x TBST buffer for one hour with 50 rpm shaking before washing three times in 1x TBST buffer for 5 mins/wash. HRP (horseradish peroxidase) conjugated to anti 6xHis antibody (ThermoFisher) in 1x TBST buffer (1:5,000) was then applied to the membrane overnight. The membrane was washed three times in 1x TBST for 10 min/wash and ECL HRP chemiluminescent substrate (Invitrogen) was applied to the membrane for 1 min.

2.3.9 qPCR

cDNA synthesised from RNA extracted from rays and discs during young bud stage (S1 in Chapter 5.4.2) and corresponding leaves were obtained from Dr Daria Golubova. Each 10 μL qPCR reaction contained 1× SYBR Green JumpStart Taq

ReadyMix, 0.2 µM forward and reverse primer and 6 ng cDNA template. To analyse primer efficiency, two technical replicates were performed. For gene expression analysis, two technical replicates and three biological replicates from three different plants were performed. No reverse transcriptase controls and no template controls were included. qPCR was performed in 96-well plates using a Bio-Rad CFX 96 machine using the programme provided in Table 2.6. The Delta-delta Ct method was used to quantify the expression levels of genes of interest against housekeeping gene *CoSAND*. The sequences of all primers are provided in Supplementary Table 2.

Steps	Temperature	Time	Cycles
Initial denaturation	94 °C	5 min	1x
Denaturation	94 °C	15 seconds	32 x
Annealing and Extension	58 °C	1 min	
Melting curve	58 °C – 94 °C		

Table 2.6 Reaction conditions for qPCR of *C. officinalis* genes

2.4 Plant growth

2.4.1 Cultivation of Asteraceae

The sources of seeds used in this study are described in Table 2.7.

Species	Source	Accession number
Carthamus lanatus	GRIN-Global	PI20272876i
Silybum marianum	Chiltern Seeds	1175E
Tragopogon dubius	Millennium Seed Bank	201151
Scolymus hispanicus	Cambridge University Botanic Garden	20100269
Taraxacum kok-saghyz	GRIN-Global	W6351562010i
Cichorium endivia	GRIN-Global	NSL31369
Crepis capillaris	Cambridge University Botanic Garden	20080266
Calendula officinalis	Chiltern Seeds	1507
Achillea millefolium	Millennium Seed Bank	639853
Matricaria chamomilla	Millennium Seed Bank	59341
Eupatorium cannabinum	Millennium Seed Bank	70672

Table 2.7 Seed sources of Asteraceae species used in this study.

Seeds were sown in 9 cm pots with Levington F2 starter compost. After germination, seedlings were transferred to 11 cm pots with John Innes Multipurpose soil containing grit. For plants used in root metabolite extraction, seedlings were transferred to 11 cm pots with perlite and grown in a flood and drain hydroponic facility with a 1.5 m x 0.5 m water tank. The water tank was connected to a water reservoir beneath with constant water circulation. Plants were grown in an unheated glasshouse with natural temperature and day length.

For *T. dubius* and *T. kok-saghyz*, two-week old plants were vernalised at 4 °C for eight weeks before transferral back to the glasshouse.

2.4.2 Cultivation of *Nicotiana benthamiana*

N. benthamiana seeds were sown in Levington F2 starter compost for two weeks. Seedlings were transferred to individual 9 cm pots with Levington F2 starter compost and grown for another three to four weeks prior to infiltration. All *N. benthamiana* plants were grown in a controlled environment room with 25°C/22°C 16h/8h day/night cycles.

2.4.3 Cultivation of *Lactuca sativa*

L. sativa seeds were surface sterilised with 70% ethanol and 30% bleach and washed with sterile water. Seeds were sown onto ½ MS media and grown in a controlled growth cabinet with 50% humidity, 22°C/20°C, 16h/8h day/night cycle. The leaves of four-to-eight-week-old seedlings were harvested for protoplast extraction.

2.5 Plant transformation, extraction, assay and analysis

2.5.1 Transformation of *Agrobacterium tumefaciens*

A 40 µL aliquot of competent *Agrobacterium tumefaciens* GV3101 was incubated on ice with 200 ng plasmid DNA for 2 mins before transferral to pre-chilled electroporation cuvettes (Geneflow). Agrobacteria were electroporated using a BioRad MicroPulser Electroporator at 2.2 kV with Agrobacterium mode selected. 400 µL SOC media was added and transferred to a 2 mL microcentrifuge tube. Cultures were incubated at 28 °C and 200 rpm for 2 hours before spreading on LB-agar plates containing 50 µg/mL carbenicillin/kanamycin, 20 µg/mL gentamicin and 100 µg/mL rifampicin. Plates were incubated at 28 °C for two days.

2.5.2 Agroinfiltration of *Nicotiana benthamiana*

Single colonies (or a glycerol stock) of *A. tumefaciens* were cultured in 10 mL LB with 50 µg/mL carbenicillin/kanamycin, 20 µg/mL gentamicin and 100 µg/mL rifampicin at 28°C with 200 rpm shaking for 2 days. On the day of *N. benthamiana* infiltration, cultures were centrifuged at 3400 x *g* for 20 mins and cells were resuspended in 10 mL infiltration media (10 mM MgCl2, 10 mM MES pH 5.6, 0.2 mM acetosyringone). OD600 were measured using an Eppendorf Biospectrophotometer and cells were incubated at 28 °C, 200 rpm before dilution to OD600=0.8. *A. tumefaciens* strains used for co-infiltration were mixed in equal volumes.

28 to 32-day old *N. benthamiana* plants were used for agroinfiltration. The abaxial side of each leaf was first pierced with a needle before 0.5-1 mL Agrobacteria cultures were infiltrated using 1 mL needless syringe. Infiltrated plants were kept in a growth chamber at 16h/8h; 25°C/22°C day/night for three to five days.

2.5.3 Total protein extraction from *N. benthamiana* leaves

After five days, five leaf discs (approximate weight = 120 mg) were harvested and snap frozen in liquid nitrogen. Leaf discs were grounded using tungsten beads using a Qiagen TissueLyser II (25 Hz for 1 min). Protein was extracted in 700 µL 1x Promega Passive Lysis Buffer with cOmplete EDTA-free protease inhibitor cocktail (Merck). After centrifugation at 13,000 x *g* for 10 min at 4 °C, 500 µL supernatants were retained. 10 µg 2,3-oxidosqualene was added to each sample and incubated at

37 °C for one hour. An equal volume of ethyl acetate was added to each sample to extract triterpenes.

2.5.4 Metabolic profiling using gas chromatography-mass spectrometry (GC-MS)

To analyse Asteraceae samples, 30 mg fresh tissue was freeze-dried for two days and then ground using tungsten beads with a Qiagen TissueLyser II (25 Hz for 1 min). 1 mL ethyl acetate was added to each sample, incubated at 40 °C with 700 rpm shaking for two hours, and left at room temperature for two days. Samples were centrifuged at 13,000 x *g* for 5 mins to pellet tissue debris and 700 µL of supernatant was transferred to GC-MS vials. GC was performed using an Agilent 7890B fitted with a Zebron ZB5-HT Inferno column -007 (Phenomenex). 2 µL of each sample was injected to each inlet pre-heated to 325 °C in pulsed splitless mode (10 psi pulse pressure). A 27-minute GC programme was run on each sample: the oven temperature was first maintained at 150 °C for 30 seconds before increasing to 360 °C at a rate of 20 °C per mins and held at 360 °C for 12.5 mins. Mass spectrometry was performed by Agilent 5977B Mass Selective Detector coupled with GC oven set to scan mode over a range of 60-800 mass units. The solvent delay time was set to 3 mins. Peaks representing different compounds eluted were identified and quantified using the MassHunter workstation software (v B.08.00; Agilent). Compounds were identified by comparing mass spectra of the peaks with those of known TFAE compounds.

To analyse infiltrated *N. benthamiana* leaves five leaf discs (approximate mass =120 mg) were harvested and freeze-dried for two days. Freeze-dried tissue was ground using tungsten beads with a Qiagen TissueLyser II (25 Hz for 1 min). 500 µL ethyl acetate with 20 ng/µL friedelin was added to each sample, incubated at 40 °C with 700 rpm shaking for two hours and then left at room temperature for two days. Samples were centrifuged at 13,000 x *g* for 5 mins to pellet tissue debris and 450 µL of supernatant transferred to GC-MS vials. To derivatise metabolites, a 30 µL aliquot was taken and ethyl acetate solvent was evaporated by incubation at room temperature in a fume hood. Samples were dissolved in 30 µL MFTSA (Sigma-Aldrich) and incubated at 40 degrees for 30 mins. GC was performed using an Agilent 7890B fitted with a Zebron ZB5-HT Inferno column -007 (Phenomenex). 2 µL of each sample was injected to each inlet pre-heated to 325 °C in pulsed splitless mode (10 psi pulse pressure). A 15.5-minute GC programme was run on each sample: the oven temperature was first maintained at 150 °C for 30 seconds before increasing to 270 °C at a rate of 20 °C per mins and held at 270 °C for 9 mins. Mass spectrometry was performed by Agilent 5977B Mass Selective Detector coupled with GC oven set to scan mode over a range of 60-800 mass units. The solvent delay time was set to 3 mins. Peaks representing different compounds eluted were identified and quantified using the MassHunter workstation software (v B.08.00; Agilent). Compounds were identified by comparing mass spectra of the peaks with those of known TFAE compounds.

Infiltrated *N. benthamiana* samples in Chapter 6 were prepared as described above except without the addition of friedelin or derivatisation. Solvents were removed from these samples and from *in vitro* enzyme assays by evaporation and the samples were resuspended in hexane before GC analysis (Thermo Scientific Trace 1300) with a Zebron CD-5MS column (30 m × 0.25 mm × 0.25 µm). 1 µL sample was injected to the inlet preheated to 300 °C in pulsed splitless mode (10 psi pulse pressure). Helium was used as a carrier gas at a constant flow of 1.2 mL·min-1. A 31-minute GC programme was run on each sample: the oven temperature was first maintained at 150 °C for 30 seconds before increasing to 250 °C at a rate of 20 °C per mins, then increasing to 300 °C at a rate of 4°C per min, followed by increasing to 320 °C at a rate of 40 °C per min and holding for 12.5 min. Mass spectrometry was performed by ISQ 7000 Single Quadrupole Mass Spectrometer coupled with GC oven set to scan mode over a range of 60-800 mass units. The solvent delay time was set to 3 mins. The detector temperature was 320 °C. Peaks representing different compounds eluted were identified with Chromeleon software.

2.5.5 *L. sativa* protoplast extraction, transfection and dual luciferase assay

A protoplast extraction and transfection protocol was adapted from Yoo et al. with modifications, where W5 solution (154mM NaCl, 125mM $CaCl_2$, 5mM KCl, 2mM MES (pH=5.7)) was used for overnight incubation and plasmid ratio was 1:1:1 (Yoo et al., 2007). To summarise the protocol, *L. sativa* leaves from 4 to 8-week-old plants were cut into strips and digested in enzyme digestion solutions for 4-6 hours at room temperature with 50 rpm shaking. After digestion, protoplasts were diluted in W5 solution, passed through a 100 µm filter, washed in W5 twice and resuspended in MMG solution (0.6 mM mannitol, 15mM $MgCl_2$, 4mM MES (pH=5.7)) to a final concentration of 1 x 10^6 cells/mL. Plasmids pEPSW1KN0034, pEPHS1KN0048 or pEPHS1KN0049 or pEPOR1CB0068, and pEPHS1KN0042 to pEPHS1KN0045 were mixed in 1:1:1 ratio (5 µg: 5 µg: 5 µg). Plasmids were mixed with 1 x 10^5 *L. sativa* protoplasts and PEG solution (40% m/v PEG 4000, 0.2M mannitol, 0.2M $CaCl_2$) and incubated at room temperature for 15 min. After transfection, protoplasts were diluted in W5 solution and washed twice with W5 before transferral to 5% BSA coated plates for 24 h culture in a growth cabinet (50% humidity, 22°C/20°C, 16h/8h day/night cycle). Fluorescence from protoplasts transfected with pEPOR1CB0068 (CaMV35s:YFP) plasmids was observed using epifluorescence microscopy (Leica DFC425 C, 10x magnification, YFP filter set) to confirm transfection efficiency. Efficiency higher than 70% was acceptable.

Protoplasts were harvested by centrifugation (4 °C, 100x g, 3 mins) and resuspended in 1x Promega passive lysis buffer with 1 x protease inhibitor cocktail. After 15 min on ice, samples were centrifuged and 40 µL supernatant was collected. Supernatant was mixed with 40 µL Promega One-Glo EX luciferase reagents on a 4titude 96-well black polystyrene flat bottom microplate and incubated at room temperature for 10 min. The Firefly luciferase (LucF) luminescence signal was read in a CLARIOstar Plus plate reader with 10 seconds read time and 1 seconds settling time. Next, 40 µL Promega NanoDLR Stop & Glo Reagent was added to each

sample and incubated at room temperature for 10 min before nano luciferase (LucN) luminescence signal was read. The LucN/LucF ratio of each sample was calculated by dividing nano luciferase signal by firefly luciferase signal. Measurement of all samples was done in the same batch.

2.5.6 Laser scanning confocal microscopy

Infiltrated *N. benthamiana* leaves were harvested three days post infiltration. Leaf discs were immersed in a water droplet on a glass slide and covered with a cover slip. An upright Leica SP8X confocal microscope equipped with a 460–670 nm supercontinuum white light laser, two continuous wavelength laser lines of 405 nm and 442 nm and a five-channel spectral scanhead (four hybrid detectors and one photomultiplier) was used for imaging. Imaging was performed using a 40× water immersion objective (HC PL APO CS2 40×/1.10 WATER) to visualise the subcellular localisation of OSC:YFP and mCherryER. YFP was exited 515 nm and fluorescence emissions were detected at 525–550 nm. mCherry was excited at 585 nm and fluorescence emissions were detected at 600-635 nm. Images were exported as tiff files and processed with Fiji (2.14.0) and fluorescent intensity was acquired using the plot profile function.

2.5.7 Flow cytometry

The relative genome sizes of *A. millefolium* and *I. britannica* were measured by propidium iodide flow cytometry following a previously described one-step protocol (Doležel et al., 2007). *Pisum sativum* (2C = 9.09 pg) and *Petroselinum crispum* (2C = 4.5 pg) were used as the references for *A. millefolium* and *I. britannica* respectively. Before measurement with a flow cytometer, fresh leaves of each sample were sampled, and nuclei were isolated using the nuclei extraction buffer from the CyStain PI Absolute P kit (Sysmex UK).

Relative genome size estimation was done using a Sysmex CyFlow Space (Sysmex Europe GmbH, Norderstedt, Germany) flow cytometer fitted with a 100 mW green solid-state laser at The Royal Botanic Gardens, Kew. For each sample, three replicates were run on the flow cytometer and the measurements were made after more than 1000 nuclei from the reference and sample were recorded. Measurements with CV less than 5% were retained. The relative genome sizes were calculated by the following formula:

$$Sample\ genome\ size = \frac{Sample\ mean\ fluorescence}{Reference\ mean\ fluorescence} \times Reference\ genome\ size$$

The ploidy level of *A. millefolium* was assigned by comparing against published genome sizes and chromosomal counts of the same species deposited in Kew Plant DNA C-values database (http://data.kew.org/cvalues). However, since neither genome size estimation nor chromosomal count of *I. britannica* has been made before, the ploidy level of *I. britannica* needed to be measured separately by chromosome count.

2.5.8 Chromosome counts

Thick, white and fleshy root tips of *I. britannica* were picked and cleaned in tubes with distilled water, before bromonaphthalene was added to arrest mitosis. After chilling at 4 °C for 24 hours, root tips were transferred to a tube filled with Carnoy's Fixative (ethanol: acetic acid = 3: 1) for 24 hours. Root tips were washed in distilled water and hydrolysed in 1M HCl solution. During hydrolysis, the sample was incubated at 65 ˚C for ~5 mins. Then, the root tips were transferred to aceto-orcein, where they absorbed the stain for at least 20 mins. When the stain had been absorbed, the root tips were transferred to a drop of 4.5% acetic acid on a slide, covered by a coverslip and the number of chromosomes was counted under a microscope.

2.6 Statistical tests

Non-parametric Kruskal–Wallis tests followed by pairwise Wilcoxon rank sum test with Benjamini-Hochberg correction were used to compare the amounts of metabolites produced by wild-type CoMAS, CoTXSS, TkTXSS and their corresponding mutants. Six biological replicates were tested for each sample.

One-way ANOVA followed with Dunnett's test was used to compare the LucN/LucF ratio of *L. sativa* protoplasts co-transfected with CaMV35s:LucF, CaMV35s:CoMYB, and promoter:LucN and those co-transfected with CaMV35s:LucF, CaMV35s:YFP, and promoter:LucN. Four biological replicates (independent transfections) were tested for each sample.

3. Chapter 3 The emergence and evolution of TXSS

3.1 Introduction

Taraxasterol (18α,19α-Urs-20(30)-en-3β-ol) is a pentacyclic triterpene (Figure 3.1) with a molecular weight of 426.72 g/mol and a boiling point of 221-222 °C (Jiao et al., 2022). Taraxasterol often co-occurs with its structural isomer ψ-taraxasterol (Figure 3.1), which differs in the position of one double bond. In taraxasterol, there is a double bond connecting C20 and C29 and, in ψ-taraxasterol, between C20 and C21 (Figure 3.1).

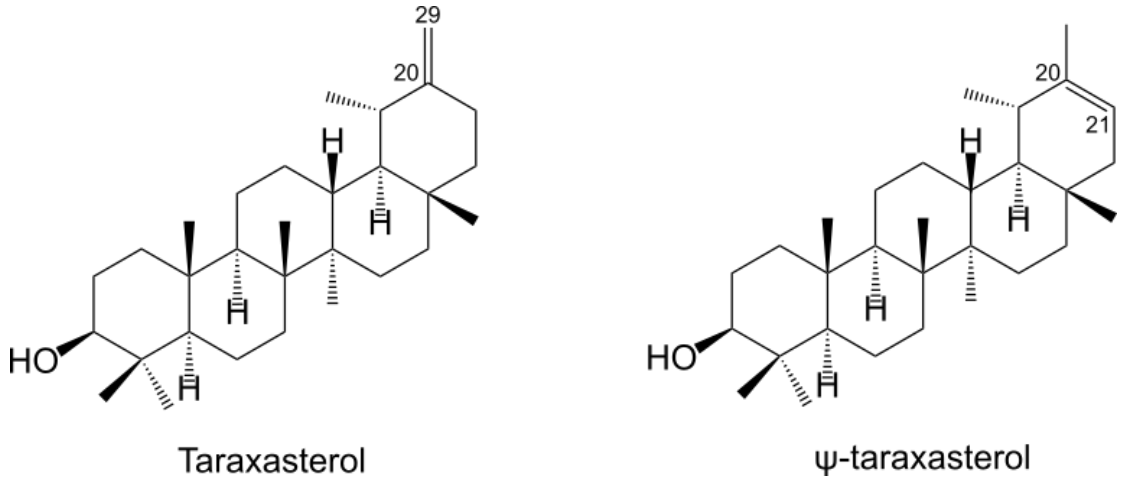


Figure 3.1 The structures of taraxasterol and ψ-taraxasterol. Both taraxasterol and ψ-taraxasterol are pentacyclic triterpenes with a double bond in their fifth ring (E-ring). Taraxasterol has a double bond between C20 and C29 while ψ-taraxasterol has a double bond between C20 and C21.

Taraxasterol was first discovered in the roots of *Taraxacum officinale* (common dandelion) as non-saponifiable compounds (Power and Browning, 1912). Subsequently, Akashi et al. observed that taraxasterol was most abundant in the latex of *T. officinale* (Akashi et al., 1994). Metabolite profiling of diverse plant lineages has identified the presence of taraxasterol and/or its derivatives in species from at least 21 eudicot families, including the early diverging *Nelumbo nucifera* (lotus) (Jiao et al., 2022; Sharma and Zafar, 2015). However, most reports are from species of Asteraceae. Further, taraxasterol/ψ-taraxasterol and their derivatives have been reported to be the most abundant terpenoids in several Asteraceae, including *Calendula officinalis*, *T. officinale* and *Lactuca sativa* (Choi et al., 2020; Yan et al., 2024; Zimmermann and Häberle, 2025).

An oxidosqualene cyclase (OSC) that catalyses the production of, predominantly, taraxasterol was discovered in *T. kok-saghyz* (Russian dandelion), in which the latex is rich in pentacyclic triterpenes (Pütter et al., 2019). This enzyme, TkOSC1 (TkTXSS in this thesis), was characterised by heterologous expression in *Nicotinana benthamiana* and *Saccharomyces cerevisiae*. It was found to be multifunctional, producing small quantities of α-amyrin, β-amyrin and lupeol (lup-19(21)-en-3-ol), as

well as taraxasterol. Consistent with the accumulation of taraxasterol in the latex, *TkOSC1* was found to be highly expressed in laticifer cells.

Subsequent studies in *T. coreanum* (Korean dandelion) and *Lactuca sativa* (lettuce) identified similar enzymes, TcOSC1 and LsOSC1 (TcTXSS and LsTXSS in this thesis). These enzymes were shown to predominantly produce taraxasterol together with smaller quantities of ψ-taraxasterol, α-amyrin, β-amyrin and dammarenediol-II when heterologously expressed in yeast (Choi et al., 2020; Han et al., 2019). Phylogenetic analysis showed that TcOSC1 and LsOSC1 are closely related to TkOSC1 and that these enzymes are most closely related to OSCs from *Panax ginseng* and *Olea europaea* that predominantly produce dammarenediol (Choi et al., 2020). In addition, OSCs (including those from non-Asteraceae species) that predominantly produce other triterpene scaffolds have been shown to produce ψ-taraxasterol and taraxasterol as minor products. For example, two OSCs from *O. europaea* and *Bauhinia forficate* that predominantly produce α-amyrin were reported to produce minor quantities of ψ-taraxasterol and taraxasterol respectively (Alagna et al., 2023; Srisawat et al., 2019).

Taraxasterol is of pharmaceutical interest with numerous previous studies reporting bioactivity, including anti-cancer and antioxidant properties (Jiao et al., 2022). In recent work, Dr Daria Golubova, a colleague in the Patron Lab, demonstrated that ψ-taraxasterol has weak anti-inflammatory bioactivity, suppressing the release of the pro-inflammatory cytokine, interleukin 6 (IL-6), in lipopolysaccharide (LPS)-activated human monocytic (THP-1) cells (Golubova et al., 2025). In contrast, taraxasterol was shown to be pro-inflammatory, increasing IL-6 release, whereas C:16 hydroxylation of either substrate resulted in triterpene diols with very strong anti-inflammatory activity. In the same study, Dr Melissa Salmon, another colleague in the Patron Lab, found that *C. officinalis,* an ancient medicinal herb, accumulates large quantities of ψ-taraxasterol compounds exclusively in its flowers (Figure 3.2). Comparative transcriptomics enabled the identification of an OSC from this species that predominantly produces ψ-taraxasterol (Golubova et al., 2025). This enzyme, named CoTXSS, also produces smaller quantities of taraxasterol, β-amyrin and lupeol. The same species was found to encode an OSC that predominantly produces α-amyrin and β-amyrin with smaller quantities of ψ-taraxasterol and taraxasterol. This was named mixed amyrin synthase (CoMAS). In addition, Dr Melissa Salmon also sequenced the leaf and flower transcriptomes of five other Asteraceae species (*Achillea millefolium*, *Calendula arvensis*, *Eupatorium cannabinum*, *Inula britannica* and *Inula ensifolia*).

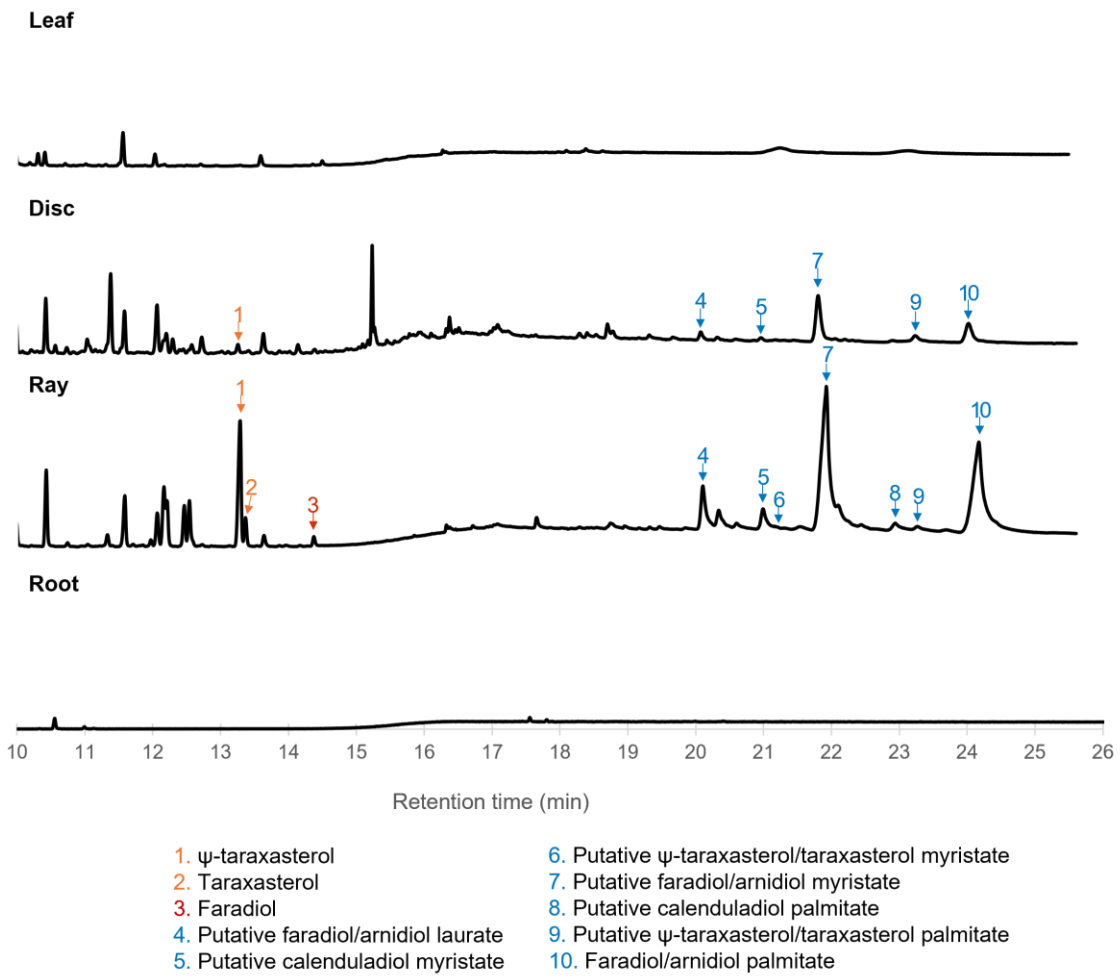


Figure 3.2 Metabolite analysis of *Calendula officinalis* by gas chromatography mass-spectometry (GC-MS). Total Ion Chromatograms (TICs) of leaf, disc floret, ray floret and root extracts. Peaks representing triterpene monols (orange), triterpene diols (red) and triterpene fatty acid esters (blue) derived from ψ-taraxasterol and taraxasterol are labelled. Adapted from Golubova et al., 2025.

At the start of this project, TXSSs had only been described from four Asteraceae species (*T. kok-saghyz*, *T. coreanum*, *L. sativa* and *C. officinalis*), yet their products had been reported to accumulate in different tissues. Furthermore, while these enzymes appear to be closely related, their evolutionary origin as well as the molecular basis of how they cyclise 2,3-oxidosqualene into ψ-taraxasterol and taraxasterol remained unknown. Given the reported bioactivities of ψ-taraxasterol/taraxasterol compounds, investigating their occurrence across plants and their biosynthesis is of immense interest.

Note: Data from this chapter are included in Golubova et al., 2025 (doi: 10.1038/s41467-025-62269-w). Figures 3.5, 3.7, 3.8, 3.9, 3.10, 3.11 and 3.12 are included in that manuscript.

3.2 Aims

In this chapter, I aimed to:

(1) Explore the origin and evolutionary history of TXSSs.
(2) Investigate the molecular basis of ψ-taraxasterol/taraxasterol production.
(3) Investigate the accumulation patterns of ψ-taraxasterol/taraxasterol.

3.3 Contributions by others

All work described in this chapter was done by the author of this thesis except:

Dr Melissa Salmon (Earlham Institute) cloned and expressed TXSSs from *Calendula officinalis*, *Cynara cardunculus*, *Taraxacum kok-saghyz*, *Taraxacum coreanum*, *Lactuca sativa*, *Cichorium endivia*, *Calendula arvensis* and *Helianthus annuus* as well as MAS from *C. officinalis* in *Nicotiana benthamiana*.

Ms Sahr Mian and Prof Ilia Leitch (Royal Botanic Gardens Kew) provided instructions in genome size estimation of *Achilliea millefolium* and *Inula britannica* with flow cytometry. Ms Sahr Mian performed chromosomal counting of *I. britannica*.

3.4 Results

3.4.1 TXSSs are likely specific to Asteraceae

Taraxasterol has been reported to occur in species of many eudicot families but is known to be a minor product of multifunctional OSCs that predominantly produce amyrins. In contrast, TXSSs have only been characterised from Asteraceae species. To investigate whether TXSS is specific to Asteraceae, CoTXSS was used as a BLAST query to identify candidate orthologues in the publicly available genomes and transcriptomes of 88 eudicot species, including 67 Asteraceae species, 10 non-Asteraceae Asterales species and 11 non-Asterales species which have been reported to accumulate taraxasterol (Supplementary Table 1). Sequences with more than 50% query coverage and E-value less than 0.1 were retained and a neighbour-joining phylogenetic tree of the retained sequences and 165 previously characterised OSCs (used in Golubova et al. 2025) was constructed. Sequences in the same monophyletic clade as previously characterised TXSSs were classified as candidate TXSSs. This analysis only detected the presence of genes or transcripts encoding candidate TXSSs in Asteraceae species, indicating that this enzyme is likely specific to the Asteraceae family (Figure 3.3).

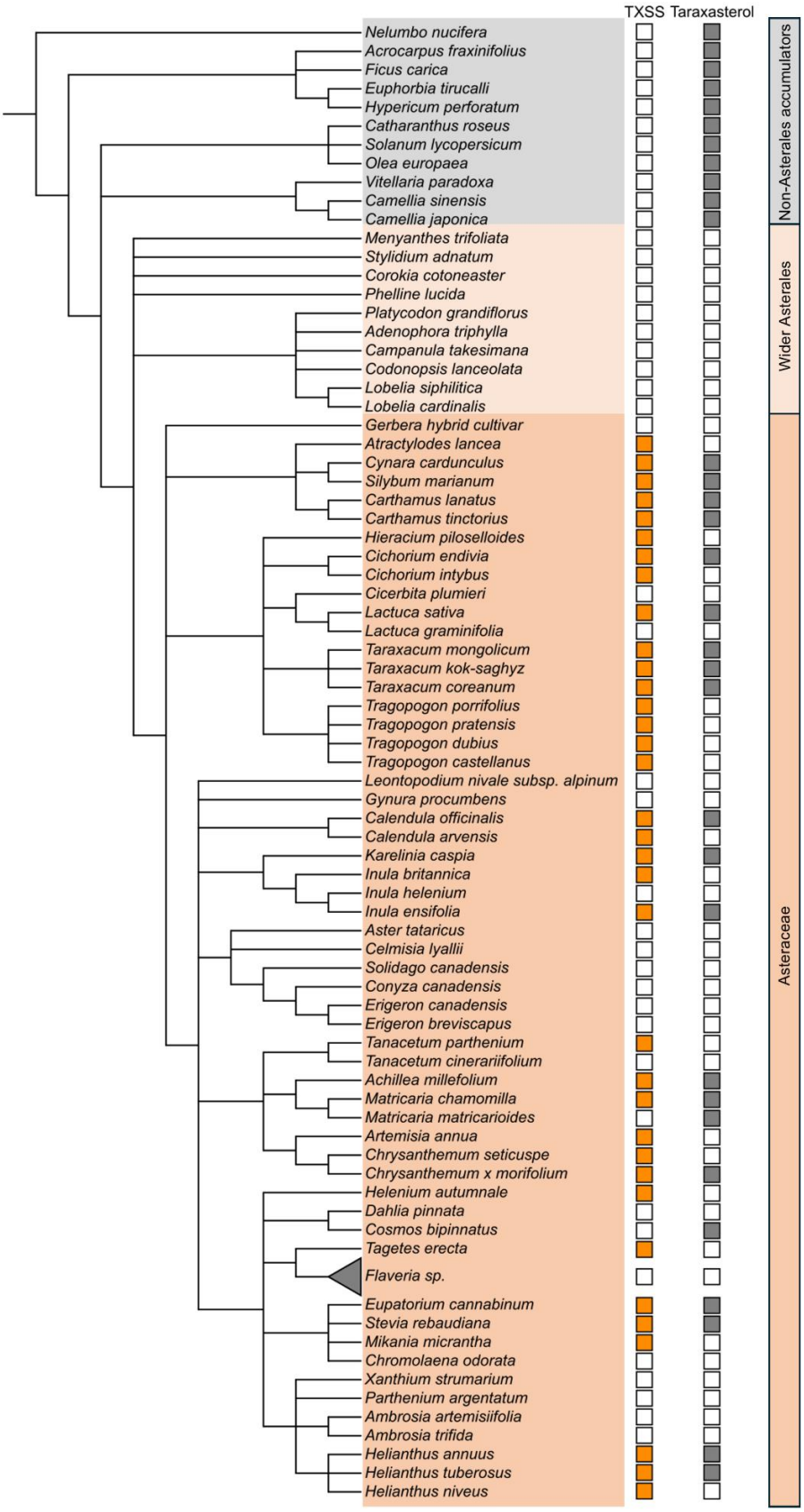


Figure 3.3 Detection of TXSS in publicly available genomes and transcriptomes of eudicots and the presence of taraxasterol in those species. Filled orange boxes indicate that a gene or transcript for a candidate TXSS was detected in the genome or transcriptome (left column). Filled grey boxes indicate that the presence of taraxasterol/ψ-taraxasterol has been reported in this species (right column).

3.4.2 The phylogeny of TXSS is generally consistent with that of the Asteraceae

To explore if the evolutionary history of TXSS is consistent with that of the Asteraceae, 43 candidate TXSSs from the three major Asteraceae subfamilies were used to construct a maximum likelihood phylogenetic tree of the aligned protein sequences together with 12 non-TXSS OSC sequences to serve as an outgroup. These included eudicot mixed amyrin synthases (MASs), dammarenediol synthases (DDSs) and baurenol synthases (BAUSs) (Figure 3.4).

In this analysis, the phylogeny of TXSS was found to be generally consistent with the species phylogeny of Asteraceae (Mandel et al., 2019). TXSSs from most Carduoideae, Cichorioideae and Asteroideae form three monophyletic clades and Carduoideae diverge before the partitioning of Cichorioideae and Asteroideae. TXSSs from all Asteroideae tribes are monophyletic.

There is, however, some incongruency between the TXSS and species phylogenies: Carduoideae *Cynara cardunculus* TXSS falls into the same monophyletic clade with Cichorioideae TXSSs. Calenduleae TXSSs form a sister group to the rest of the Asteroideae TXSSs, rather than forming a monophyletic clade with Anthemideae TXSSs. The deep split between monophyletic supertribe Heliantheae alliance (including Tageteae, Heliantheae, Helenieae and Eupatorieae) from the remaining Asteroideae tribes is not observed in the TXSS tree. Heliantheae alliance TXSSs are paraphyletic, with one monophyletic clade of Tageteae and Heliantheae TXSSs and the other of Helenieae and Eupatorieae TXSSs.

3.4.3 When heterologously expressed in *N. benthamaina*, TXSSs show variation in product specificity

To date, only four TXSSs have been characterised, including CoTXSS from *Calendula officinalis*. To compare the product profiles of TXSSs from different lineages, the coding sequences of the four previously characterised TXSSs from *T. kok-saghyz*, *T. coreanum*, *L. sativa* and *C. officinalis* (TkTXSS, TcTXSS, LsTXSS and CoTXSS) and five uncharacterised TXSSs from *Cynara cardunculus*, *Cichorium endivia*, *Tragopogon dubius*, *C. arvensis* and *Helianthus annuus* (CcTXSS, CeTXSS, TdTXSS, CarTXSS and HaTXSS), which represent three main subfamilies, were cloned into *N. benthamiana* expression vectors (Chapter 2.3.1 and 2.3.2) and transformed into *A. tumefaciens* (Chapter 2.5.1). Constructs expressing TXSSs were co-agroinfiltrated into *N. benthamiana* leaves with constructs expressing the suppressor of silencing p19 from Tomato Bushy Stunt Virus and a truncated HMGR (tHMGR), which enhances flux through the mevalonate (MVA) pathway. Metabolites were extracted in ethyl acetate and analysed by GC-MS (Chapter 2.5.4). This analysis showed that all orthologues of TXSS mainly produce ψ-taraxasterol and taraxasterol, with minor quantities of β-amyrin and lupeol (Figure 3.5).

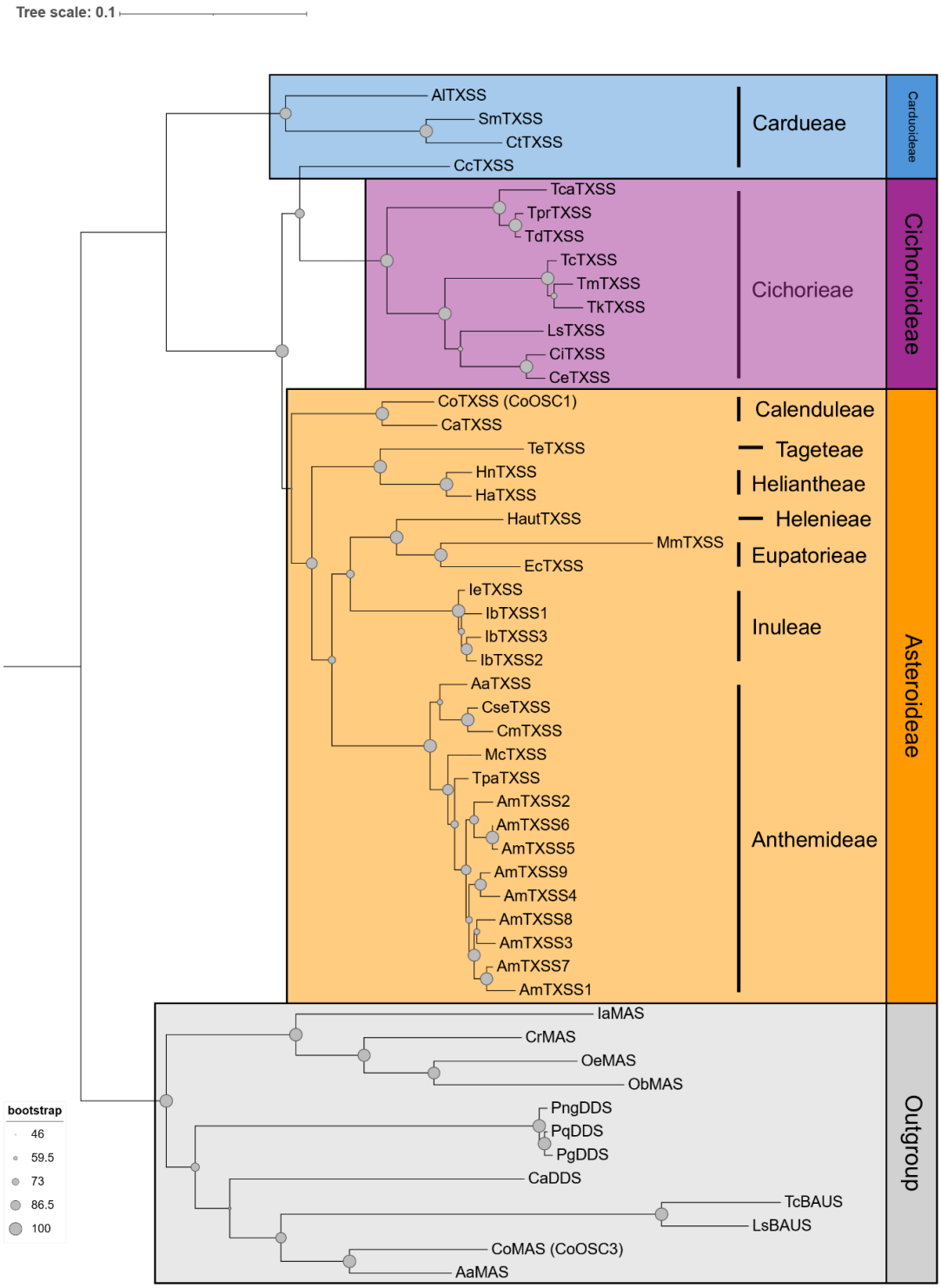


Figure 3.4 Maximum likelihood phylogenetic tree of selected TXSSs and outgroup OSCs. Species abbreviations: Al, *Atractylodes lancea*; Sm, *Silybum marianum*; Ct, *Carthamus tinctorius*; Cc, *Cynara cardunculus*; Tca, *Tragopogon castellanus;* Tpr, *Tr. pratensis*; Td, *Tr. dubius*; Tk, *Taraxacum kok-saghyz*; Tc, *Ta. coreanum*; Ls, *Lactuca sativa*; Ci, *Cichorium intybus*; Ce, *Ci. endivia*; Co, *Calendula officinalis*; Car, *Ca. arvensis*; Te, *Tagetes erecta*; Hn, *Helianthus niveus*; Ha, *H. annuus*; Haut, *Helenium autumnale*; Mm, *Mikania micrantha*; Ec, *Eupatorium cannabinum*; Ie, *Inula ensifolia*; Ib, *I. britannica*; Aa, *Artemisia annua*; Cse, *Chrysanthemum seticuspe*; Cm, *Ch. x morifolium*; Mc, *Matricaria chamomilla*; Tpa, *Tanacetum parthenium*; Am, *Achillea millefolium*; Ia, *Ilex asprella var. asprella*; Cr, *Catharanthus roseus*; Oe, *Olea europaea*; Ob, *Ocimum basilicum*; Ca, *Centella asiatica*; Png, *Panax notoginseng*; Pq, *P. quinquefolius*; Pg, *P. ginseng*. Phylogenetic tree was visualised with iTOL (Letunic & Bork, 2021). The scale bar shows the number of substitutions per site.

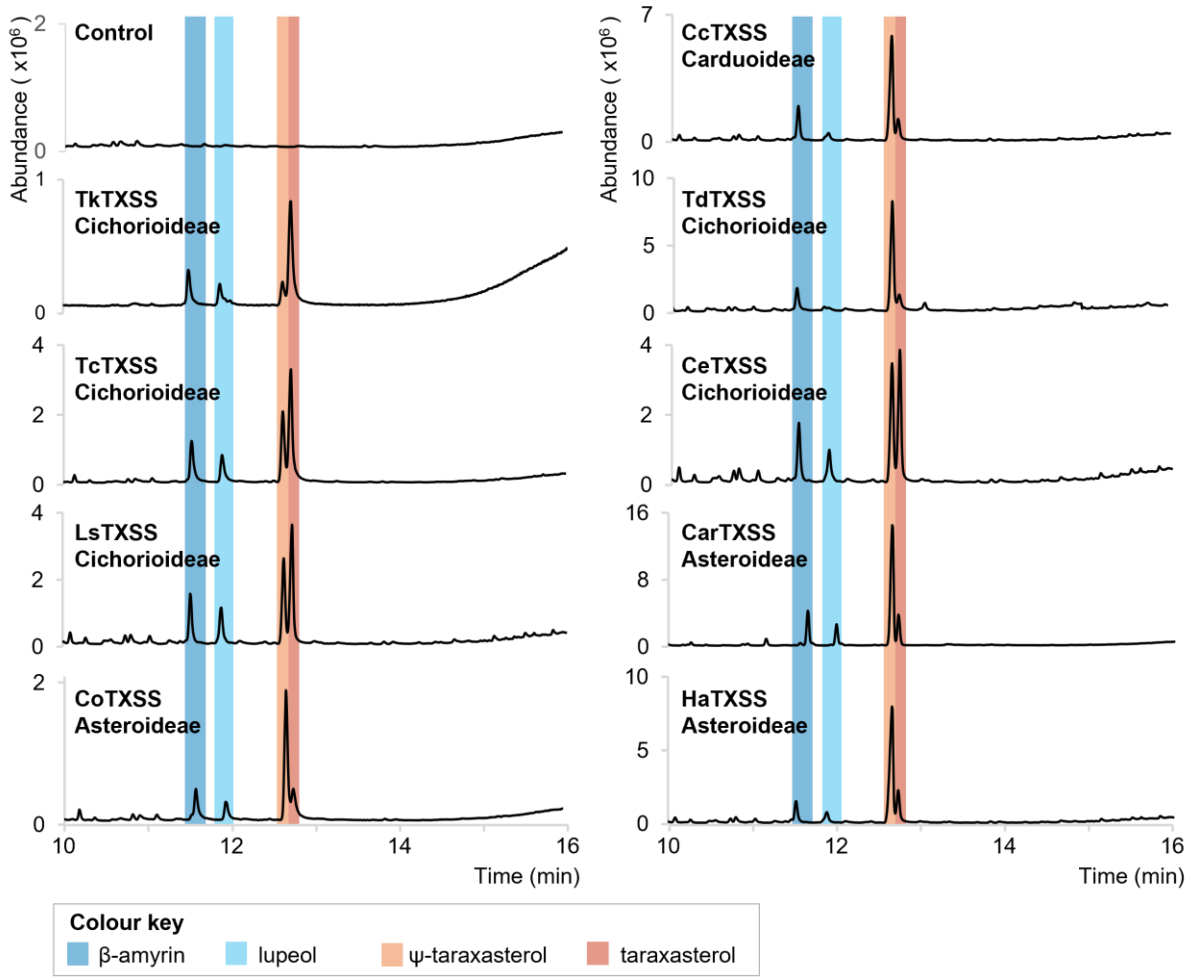


Figure 3.5 GC-MS analysis of *N. benthamiana* leaves transiently expressing TXSS orthologues. Representative total ion chromatograms of ethyl acetate extracts of *N. benthamiana* leaves transiently expressing *TXSS* genes from Asteraceae species. All TXSSs were co-expressed with p19 and tHMGR; the control sample only expresses these constructs. Abbreviations: Cc, *Cynara cardunculus*; Tk, *Taraxacum kok-saghyz*; Td, *Tragopogon dubius*; Tc, *Taraxacum coreanum*; Ce, *Cichorium endivia*; Ls, *Lactuca sativa*; Car, *Calendula arvensis*; Co, *Calendula officinalis*; Ha, *Helianthus annuus*.

Variation was observed between the product profiles of TXSSs from different lineages. The product profiles of the four previously characterised TXSSs were consistent with previous results. All TXSSs from species of Carduoideae and Asteroideae were observed to produce more ψ-taraxasterol than taraxasterol (Figure 3.5). In contrast, those from the Cichorioideae varied with TdTXSS producing predominantly ψ-taraxasterol, TcTXSS, LsTXSS and CeTXSS producing slightly more taraxasterol than ψ-taraxasterol, and TkTXSS producing predominantly taraxasterol (Figure 3.5). *Tragopogon dubius* belongs to a basal subtribe of Cichorieae (the Scorzonerinae), while the other three species are from non-basal lineages (Kilian et al., 2009). Thus, I hypothesised that the divergence in product ratio favouring taraxasterol over ψ-taraxasterol occurred after the divergence of the

Scorzonerinae. The testing of this hypothesis, together with an investigation of genetic basis of TXSS product specificity, is described in Chapter 4.

3.4.4 The copy number of *TXSS* genes increased following gene- and genome-duplication events

All diploid species analysed were found to encode only one *TXSS* gene. In contrast, the tetraploid *C. officinalis* genome encodes one *TXSS* gene per genome: an expressed gene (*CoOSC1*), encoding an active enzyme and a likely pseudogene (*CoOSC17*) (Table 3.1).

Species	Subfamily	Ploidy	Number of *TXSS*
Cynara cardunculus	Carduoideae	Diploid	1
Taraxacum kok-saghyz	Cichorioideae	Diploid	1
Cichorium endivia	Cichorioideae	Diploid	1
Cichorium intybus	Cichorioideae	Diploid	1
Lactuca sativa	Cichorioideae	Diploid	1
Artemisia annua	Asteroideae	Diploid	1
Chrysanthemum seticuspe	Asteroideae	Diploid	1
Helianthus annuus	Asteroideae	Diploid	1
Calendula officinalis	Asteroideae	Tetraploid	2

Table 3.1 Ploidy and number of TXSS in nine Asteraceae species with publicly available genomes.

Although one copy of *TXSS* was identified in the genomes of most species, nine and three copies of *TXSS* were found in the concatenated flower and leaf transcriptomes of *A. millefolium* and *I. britannica*, respectively. The former has been recorded to have intraspecific ploidy variation, with different populations reported to be diploid, tetraploid, hexaploid and octaploid (López-Vinyallonga et al., 2015). To investigate if the gene number observed in these samples was likely to have arisen via genome or gene duplication events, the genome size and ploidy of these two species were estimated using flow cytometry and chromosomal counting (Figure 3.6).

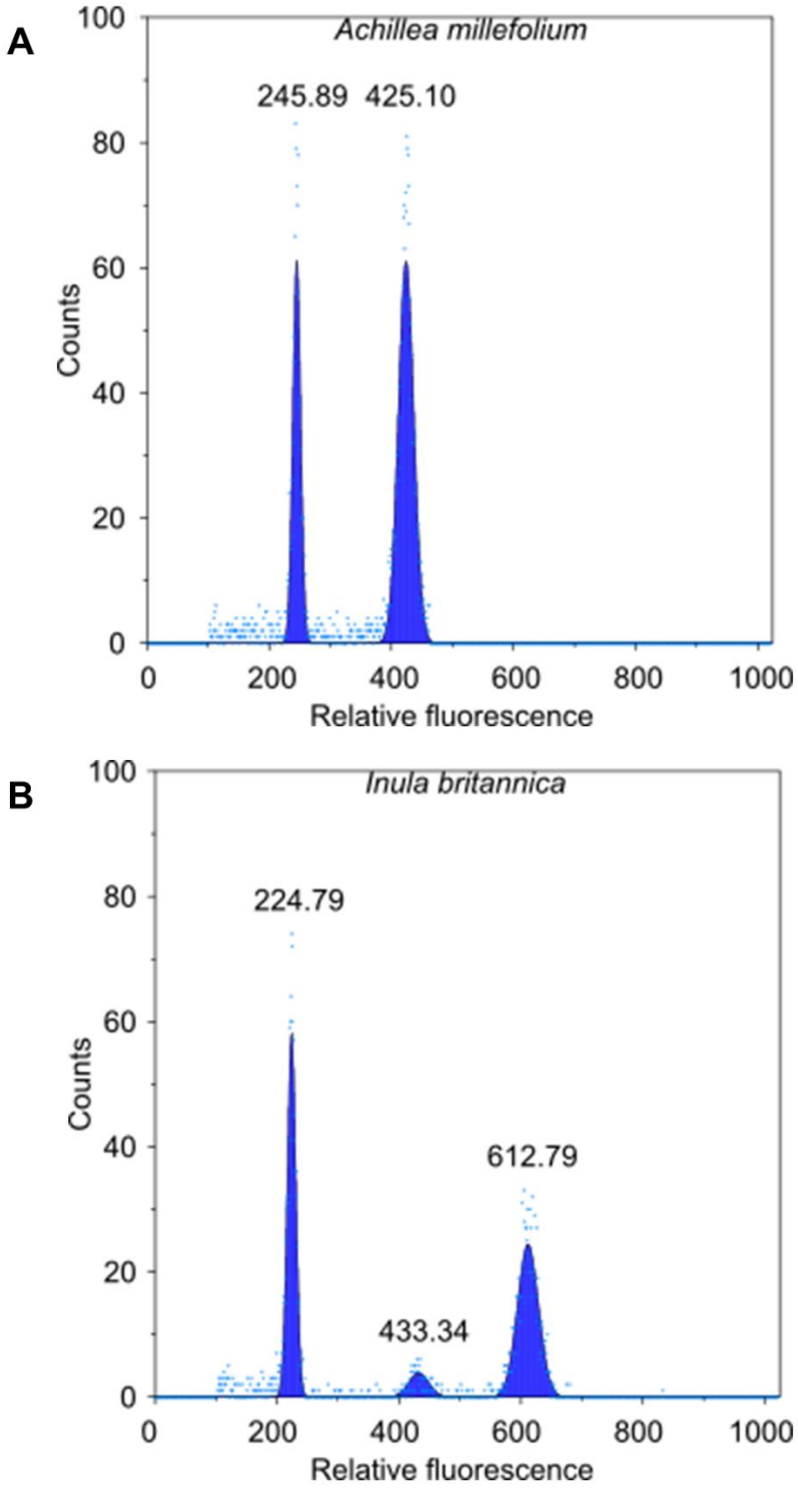


Figure 3.6 Genome size analysis of Asteraceae. Genome size analysis of (A) *A. millefolium* sample with reference to *Pisum sativum* (2C = 9.09 pg) and (B) *I. britannica* sample with reference to *Petroselinum crispum* (2C = 4.5 pg) estimated from flow cytometry. In each plot, the first peak represents nuclei from the reference sample, and the second larger peak represents nuclei from the named species. The peak with mean relative fluorescence of 433.34 in plot (B) represents G2 and M phase nuclei in the reference sample. Mean relative fluorescence of each peak represents the size of the corresponding nuclei.

Flow cytometry estimated the 2C genome sizes of *A. millefolium* and *I. britannica* samples to be 15.717 pg and 12.267 pg respectively. *A. millefolium* has a reference sample in the C-value database (Pellicer and Leitch, 2020) with 2C genome size of 9.75 pg. The genome size of the *A. millefolium* ecotype used in this thesis was found to be 1.612 times larger than the reference. We deduced that the *A. millefolium* reference is tetraploid and the *A. millefolium* sample used in this study is hexaploid. In parallel, chromosome counting by Ms Sahr Mian revealed that the *I. britannica* sample has 32 chromosomes and is likely to be tetraploid. Thus, the *A. millefolium* and *I. britannica* samples all appear to have more than one copy of *TXSS* per genome suggesting the presence of paralogues.

3.4.5 TXSS likely evolved from a mixed amyrin synthase (MAS) following the divergence of Asteraceae

Phylogenetic analysis revealed that the clade containing TXSSs is sister to a clade containing MASs from Asteraceae lineages, which produce predominantly α-amyrin and β-amyrin with a trace amount of ψ-taraxasterol and taraxasterol, as well as clades with enzymes known to catalyse the production of dammenediol and baurenol (Figure 3.7). Basal to this lineage is a clade containing MASs from other plant species. This suggests that TXSSs are most likely to have evolved via gene duplication and neofunctionalisation of an MAS.

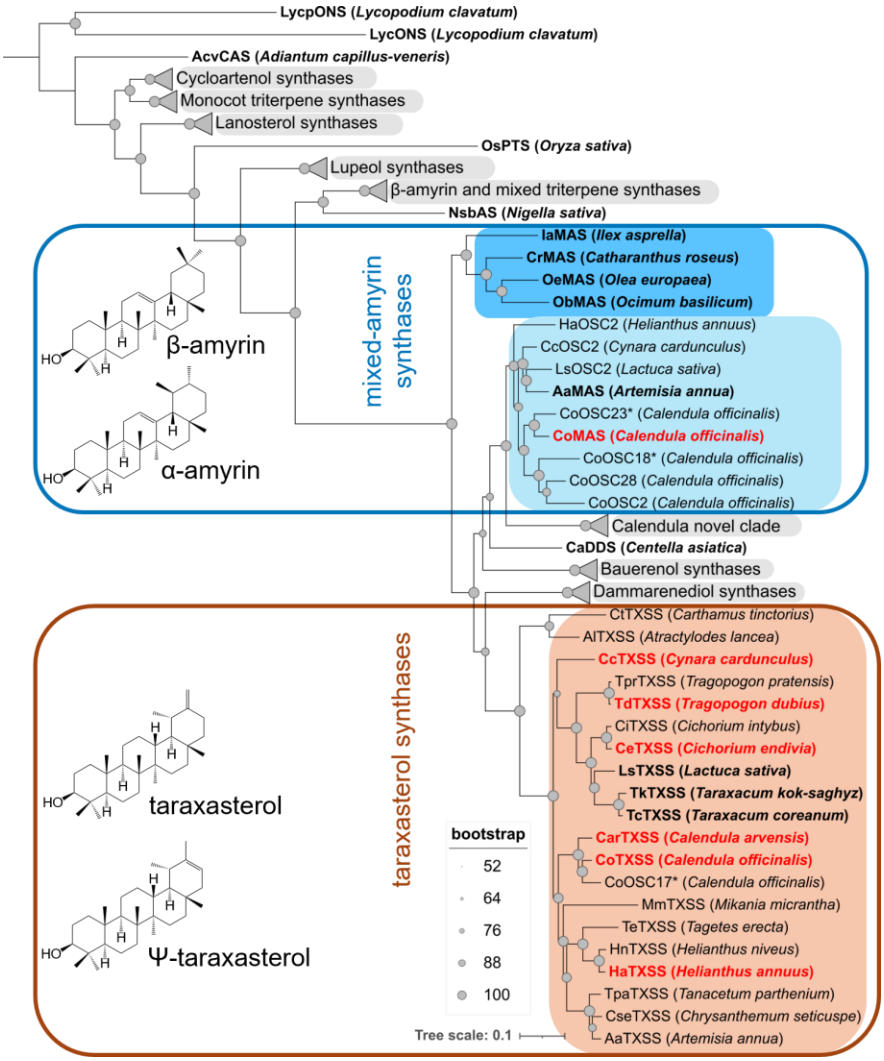


Figure 3.7 Maximum likelihood phylogenetic tree of plant OSCs. TXSSs are highlighted in an orange bubble and MASs are highlighted in a light blue (Asteraceae) and blue (non-Asteraceae Asterids) bubble and annotated with chemical structures of their major products. Characterised OSCs are shown in bold with those characterised in this study highlighted in red. An asterisk (*) indicates likely pseudogenes.

Amyrins and taraxasterols differ in the structure of the third and the fifth ring. Both MASs and TXSSs first fold 2,3-oxidosqualene into the four-ringed dammarenyl cation. MASs predominantly direct ring expansion and closure to create an oleanyl cation (for deprotonation into β-amyrin) or an ursanyl cation (for deprotonation into α-amyrin). In contrast, TXSSs direct a further methyl shift to create a taraxasteryl cation for deprotonation into ψ-taraxasterol or taraxasterol (Figure 3.8).

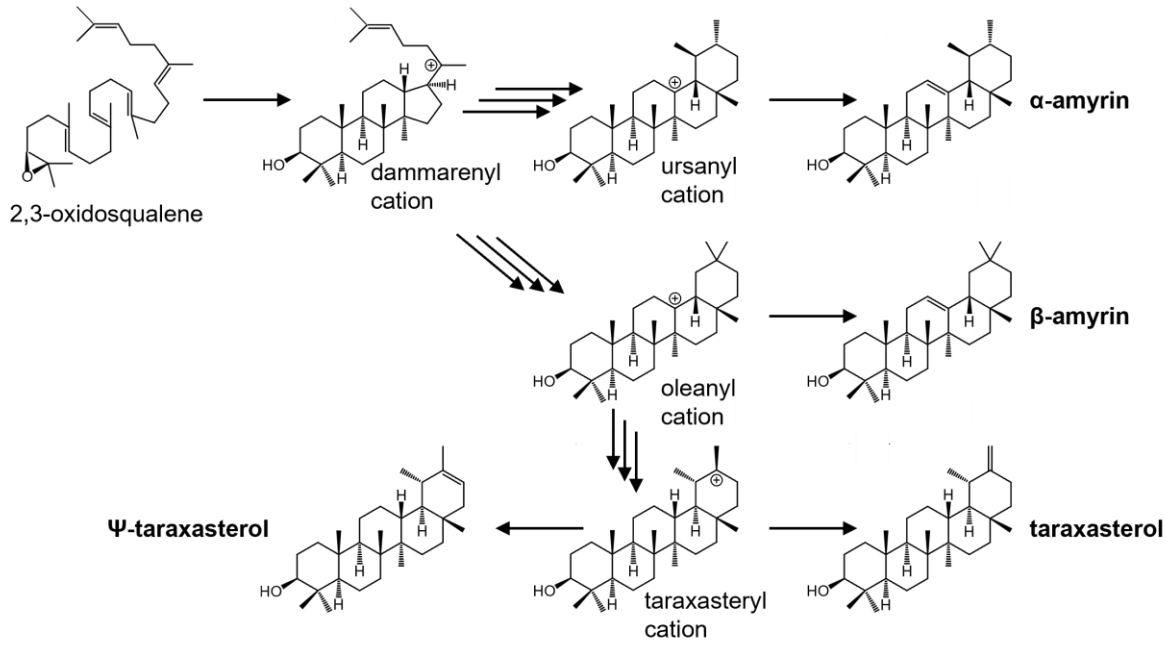


Figure 3.8 The biosynthesis of Ψ-taraxasterol, taraxasterol, α-amyrin and β-amyrin by 2,3-oxidosqualene cyclases.

To identify candidate residues important for ψ-taraxasterol and taraxasterol production, AlphaFold2 was used to create structural models of CoMAS and CoTXSS. These models showed that these enzymes both have high structural similarity (RMSD = 0.806 and = 0.778, respectively) to the crystal structure of human lanosterol synthase (PDB: 1W6K) (Jumper et al., 2021; Thoma et al., 2004). Using these models, the active sites of CoMAS and CoTXSS were predicted from the docked taraxasteryl cation, the position of which was based on the location of lanosterol in 1W6K. Residues within 12 Angstroms of the predicted active sites were proposed to be involved in shaping the conformation of the active site and determining product specificity.

Next, a multiple sequence alignment of all characterised MASs and TXSSs was constructed and two active site residues that differ between MASs and TXSSs were identified: I367 and E371 (in CoMAS numbering; Figure 3.9).

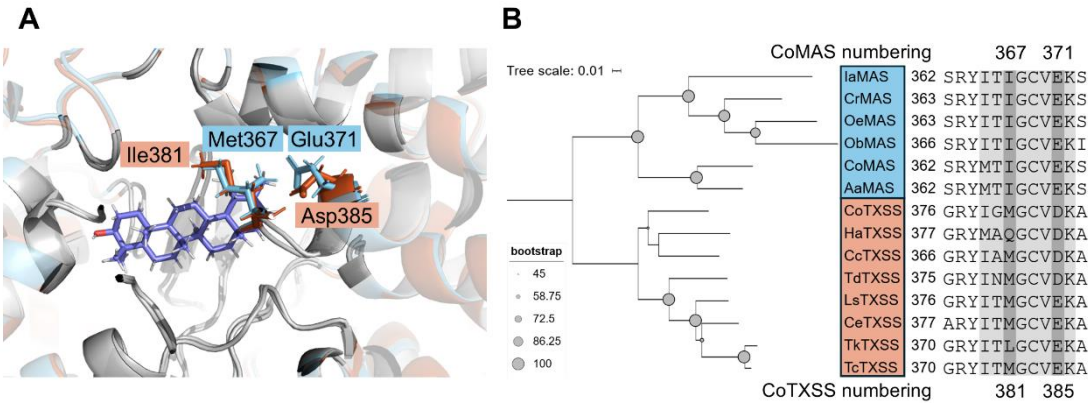


Figure 3.9 The active sites of *Calenduala officinalis* mixed amyrin synthase (CoMAS) and *Calenduala officinalis* taraxaseterol synthase (CoTXSS). (A) The taraxasteryl cation was docked into the structural models of CoMAS (Blue) and CoTXSS (Orange). Residues that differ between them are highlighted and labelled. (B) Phylogenetic tree and sequence alignment of selected residues of MASs and TXSSs with active site residues highlighted in grey and residues that differ in dark grey.

To investigate the contribution of I367 and E371 to the production of taraxasterol, these were mutated individually or in combination in CoMAS. The mutants were cloned into overexpression vectors and transiently co-expressed in *N. benthaminana* with p19 and tHMGR as previously described. The product profile of these mutants was compared to the wild-type CoTXSS and CoMAS using GC-MS, and the amount of each metabolite was quantified by comparison to an internal standard (Figure 3.10).

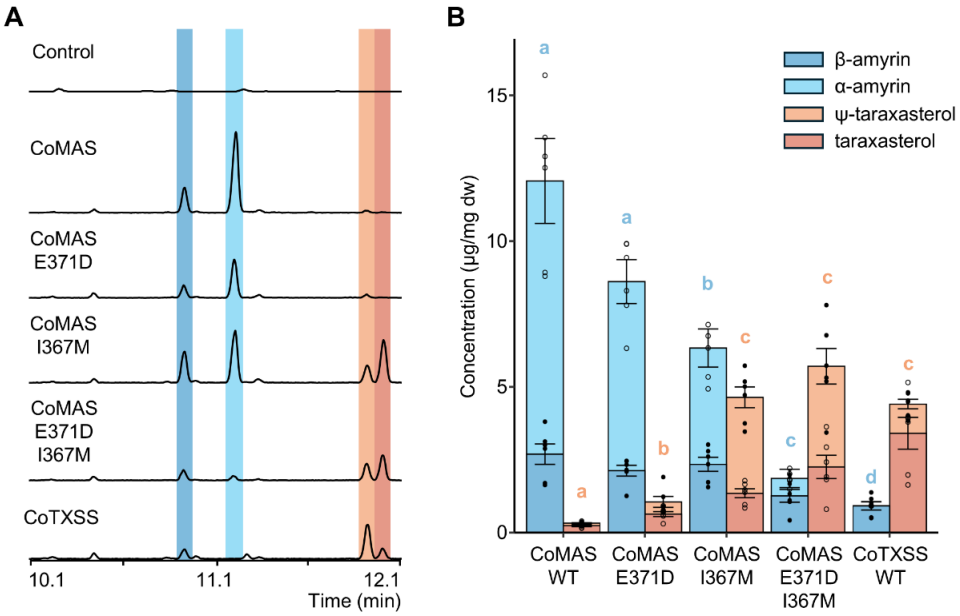


Figure 3.10 GC-MS analysis of *N. benthamiana* leaves transiently expressing CoMAS, CoMAS mutants and CoTXSS (A) Total ion chromatograms of extracts of *N. benthamiana* leaves transiently expressing wild type and mutated CoMAS and CoTXSS. (B) Quantification of triterpenes produced by wild type and mutated CoMAS and CoTXSS relative to an internal standard; n=6; error bars represent standard error. Statistical significance in α/β-amyrin content (blue lowercase letters) and ψ-taraxasterol/taraxasterol content (orange lowercase letters) were analysed using a Kruskal-Wallis test followed by post-hoc Wilcoxon rank sum test with a Benjamini-Hochberg correction. Samples that do not share the same lower-case letter are significantly different from each other ($p<0.05$).

CoMAS^{E371D} produced significantly more ψ-taraxasterol/taraxasterol than CoMAS. However, the production of α/β-amyrin did not change significantly. CoMAS^{I367M} produced significantly less α/β-amyrin and significantly more ψ-taraxasterol/taraxasterol than CoMAS and CoMAS^{E371D}. CoMAS$^{E367M, E371D}$ produced significantly less α/β-amyrin than both CoMAS and single mutants as well as similar amounts of ψ-taraxasterol/taraxasterol to CoTXSS. Consequently, these two active site residues act in concert to determine α/β-amyrin specificity.

3.4.6 TXSSs likely evolved soon after the emergence of the Asteraceae.

The evidence above suggests that TXSSs emerged within the Asteraceae. To investigate this further, Bayesian phylogenetic analysis was performed to estimate the date of the MAS-TXSS split event.

Codons of all the MASs and TXSSs characterised in this chapter were aligned, and a phylogenetic tree was built with a strict molecular clock model, using an Artemisia fossil and a *Cichorium intybus* type fossil as calibrators (Tremetsberger et al., 2013; WANG, 2004) (Figure 3.11). The results suggest that the common ancestor of all Asteraceae MASs and TXSSs emerged 83-121 mya and the split between MAS and TXSS occurred soon afterwards. The estimated MAS-TXSS divergence time overlaps with previously estimated divergence time of Asteraceae which took place 64-91 mya. Thus, it is likely that TXSS diverged soon after the divergence of Asteraceae.

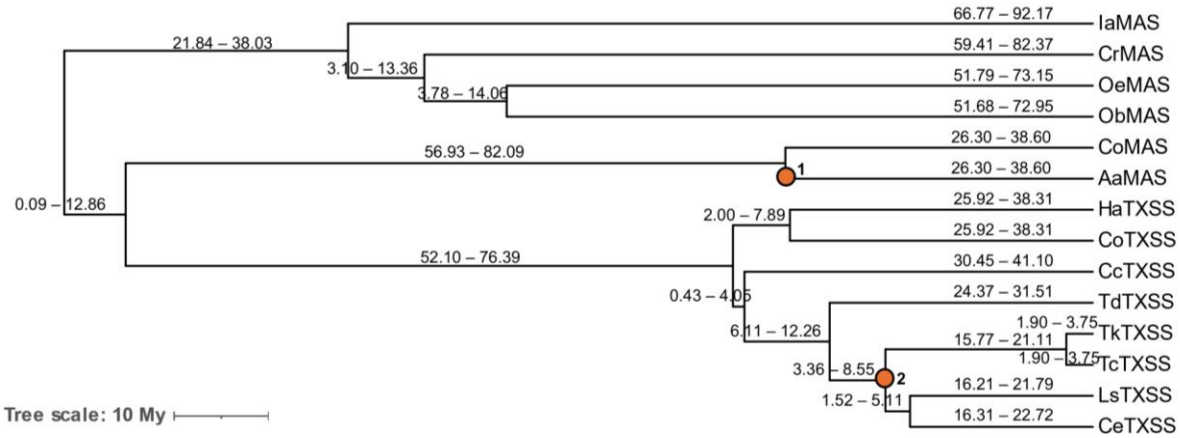
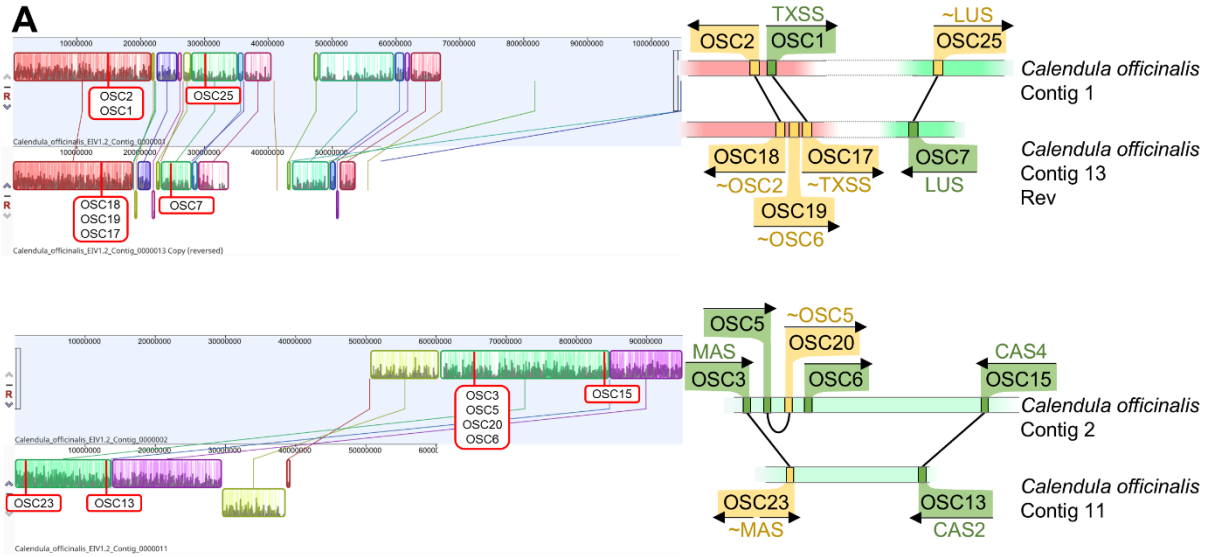
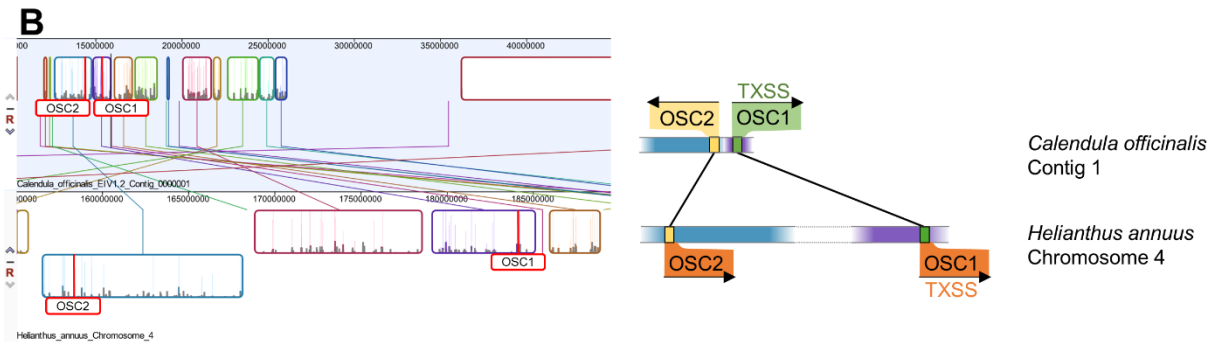


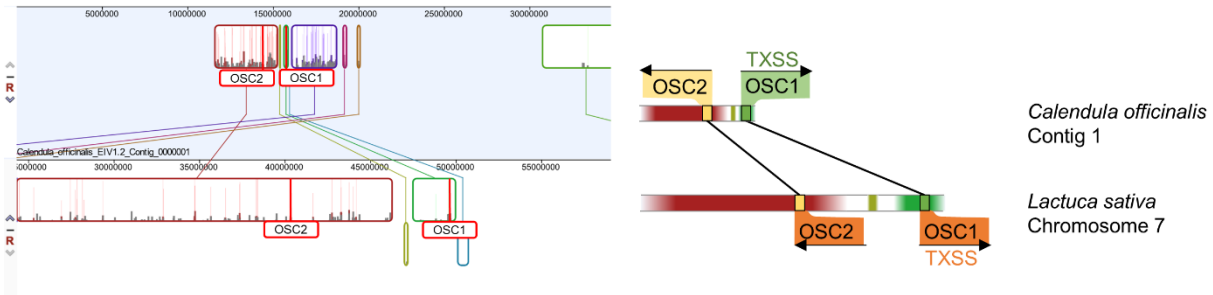
Figure 3.11 Bayesian phylogenetic tree of MAS and TXSS. Bayesian phylogenetic analysis was performed using BEAST2 with a GTR model and an Artemisia fossil (orange circle 1; 1.31 mya) and a *Cichorium intybus* type fossil (orange circle 2; 2.22 mya) as calibrators. Numbers show branch length ranges.

3.4.7 The genomic location of TXSS supports its evolution by duplication and neofunctionalization of an MAS

To investigate if there was evidence of a gene duplication event in the genome, the genomic location of TXSS and MASs were compared. In *C. officinalis*, *CoTXSS* is co-located with the gene encoding a MAS that is phylogenetically related to other Asteracece MASs (*CoOSC2*) (Figure 3.12). This protein is highly similar to the characterised CoMAS (Pairwise identity: 86.8%; BLOSUM62 similarity: 92.4%), however, no transcripts mapping to this *CoOSC2* gene were detected in the leaf and floral transcriptomes. In contrast, *CoMAS* is located on a different chromosome (Figure 3.12). The genomic regions in which *TXSS* is located were found to be syntenic in three other Asteraceae genomes, *H. annuus*, *L. sativa*, and *C. cardunculus*, with *TXSS* being adjacent to a gene which was found to be orthologous to *CoOSC2* (Figure 3.12).


A
OSC2
OSC1
OSC25
OSC18
OSC19
OSC17
OSC7
OSC3
OSC5
OSC20
OSC6
OSC15
OSC23
OSC13
TXSS
~LUS
~OSC2
~TXSS
LUS
~OSC6
~OSC5
MAS
CAS4
~MAS
CAS2
Calendula officinalis
Contig 1
Calendula officinalis
Contig 13
Rev
Calendula officinalis
Contig 2
Calendula officinalis
Contig 11


B
OSC2
OSC1
TXSS
Helianthus annuus
Chromosome 4
Calendula officinalis
Contig 1


OSC2
OSC1
TXSS
Lactuca sativa
Chromosome 7
Calendula officinalis
Contig 1

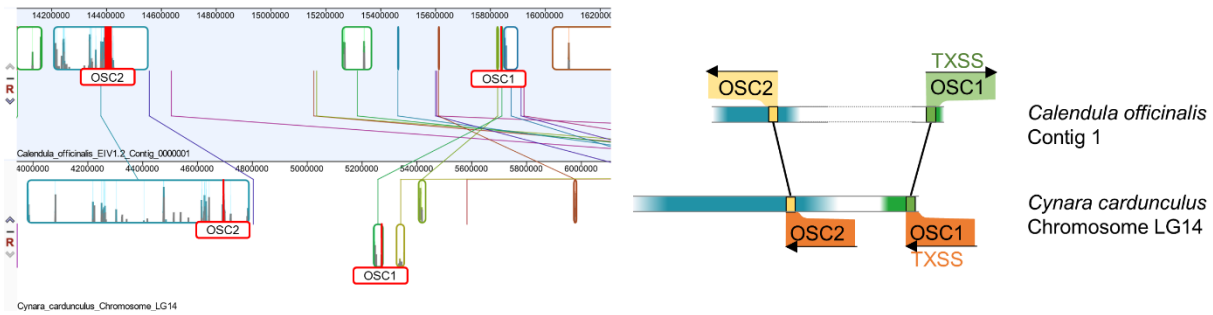

OSC2
OSC1
TXSS
Cynara cardunculus
Chromosome LG14
Calendula officinalis
Contig 1

Figure 3.12 Genomic locations and orientations of Asteraceae *OSCs* (A) The relative locations and orientations of *CoTXSS*, *CoMAS* and *CoOSC2* in the *Calendula officinalis* (pot marigold) genome. Contigs are paired to show synteny between genes found on homeologous contigs. (B) The genomic locations of pot marigold, *Helianthus annuus* (sunflower), *Lactuca sativa* (lettuce) and *Cynara cardunculus* (globe artichoke) genes encoding OSC1 (TXSS) and OSC2**.** A graphical representation of the relative position, orientation and similarity of *OSC* genes. Contigs and chromosomes are paired to show synteny between genes found on homologous regions. Genes that were expressed in the pot marigold transcriptome are shown in green. Those for which expression was not detected are shown in yellow. Homologues in sunflower, lettuce and globe artichoke are shown in orange.

3.4.8 Tissue accumulation pattern of ψ-taraxasterol/taraxasterol and their derivatives varies amongst lineages

In Chapter 3.4.3, characterisation of TXSSs from different Asteraceae revealed that TXSSs from species of Carduoideae and Asteroideae as well as those from the basal lineages of Cichorioideae produce more ψ-taraxasterol than taraxasterol (Figure 3.5). In contrast, those found in species of the non-basal Cichorioideae produce either equal quantities of both products, or predominantly taraxasterol (Figure 3.5).

Interestingly, in *C. officinalis*, ψ-taraxasterol/taraxasterol and their derivatives were only detected in floral tissues with the dominant compounds being derivatives of ψ-taraxasterol (Figure 3.2). In contrast, ψ-taraxasterol/taraxasterol compounds have been previously reported to accumulate in the leaves of *T. coreanum* and *L. sativa*, and in the roots of *T. kok-saghyz*, with the dominant compound being taraxasterol. However, in no species have all tissues been sampled.

To determine if an increase in the production of taraxasterol over ψ-taraxasterol correlates with accumulation in non-floral tissues, different tissues of Asteraceae species from the three main lineages were metabolically profiled. To do this, GC-MS was performed on root, floral and leaf tissue extracts from two species of Carduoideae, five species of Cichorioideae and four species of Asteroideae. The peak areas of all ψ-taraxasterol and taraxasterol derivatives in each tissue were quantified (Figure 3.13). These species were selected because they represent diverse phylogenetic lineages within the Asteraceae and because seeds were available from seedbanks or botanic gardens.

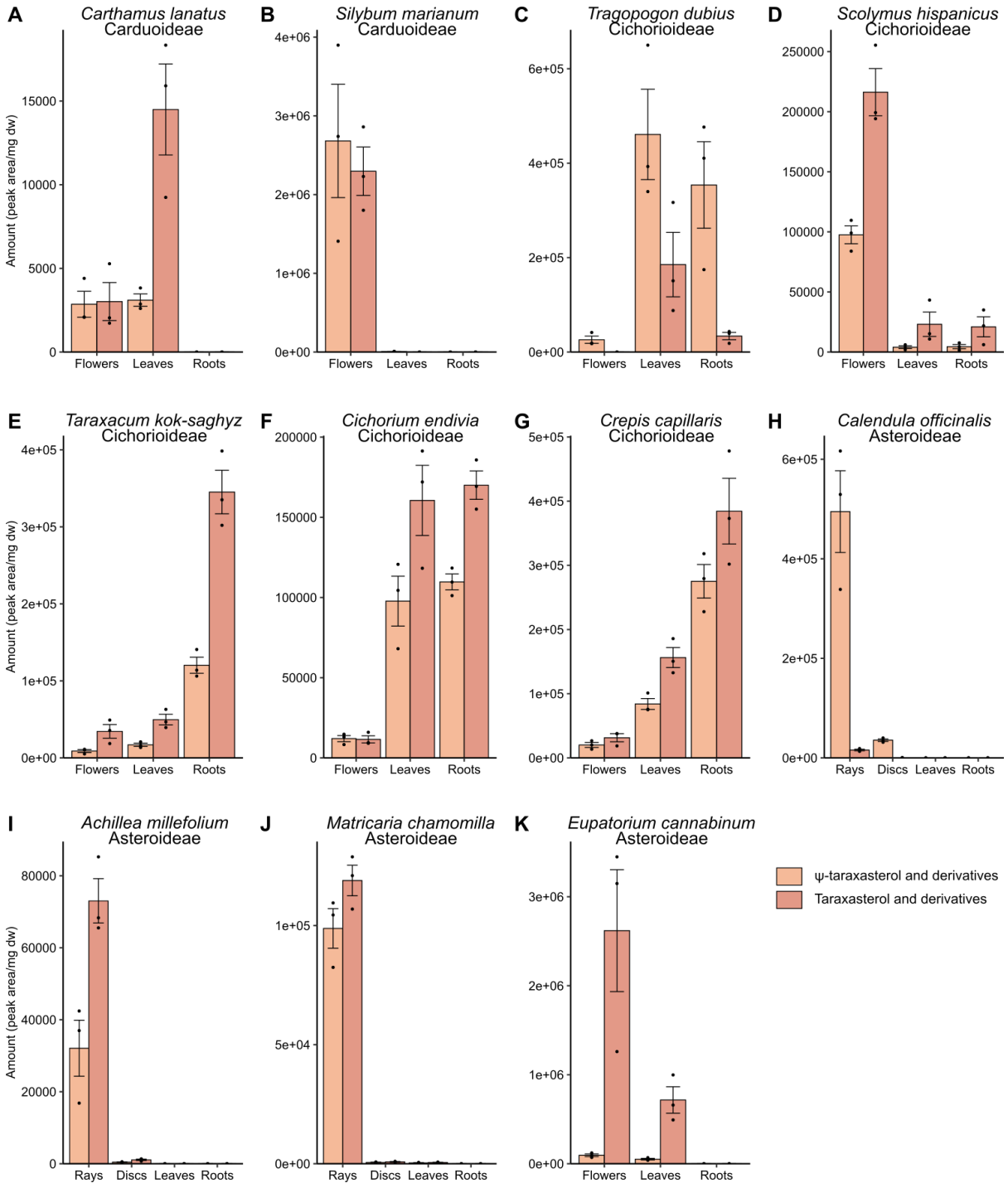


Figure 3.13 Accumulation of ψ-taraxasterol and derivatives versus taraxasterol and derivatives accumulated in different tissues of 11 species from three main Asteraceae subfamilies. N=3. Error bars represent standard error.

Metabolic profiling revealed that, within the Carduoideae, *S. marianum* accumulates slightly more ψ-taraxasterol than taraxasterol compounds in flowers, while *C. lanatus* accumulates similar amounts of both compounds in flowers but more taraxasterol compounds in leaves (Figure 2.13 A, B).

Within the Cichorioideae, the basal lineages differed with *T. dubius* producing mainly ψ-taraxasterol, which accumulated in leaves and roots, while *S. hispanicus* mainly

produced taraxasetrol, which accumulated in the flowers and leaves. The non-basal Cichorioideae lineages all showed the highest concentration of ψ-taraxasterol and taraxasterol compounds in the roots, with more taraxasterol than ψ-taraxasterol compounds (Figure 3.13 C, D, E, F, G).

All species of Asteroideae showed a higher concentration of ψ-taraxasterol and taraxasterol compounds in floral tissues, predominantly in the ray florets (Figure 3.13 H, I, J, K). Interestingly, *A. millfeolium, M. chamomilla* and *E. cannabium,* accumulated more taraxasterol than ψ-taraxasterol compounds, unlike *C. officinalis*.

3.5 Discussion

3.5.1 TXSSs likely emerged in the Asteraceae and have been maintained in all major lineages

TXSSs are OSCs that predominantly produce taraxasterols from 2,3-oxidosqualene. Previous studies have characterised TXSSs from three Asteraceae species: *T. kok-saghyz*, *T. coreanum* and *L. sativa* (Choi et al., 2020; Han et al., 2019; Pütter et al., 2019). Further, in preliminary work, Dr Melissa Salmon characterised a TXSS from *C. officinalis*. I sought to identify if TXSSs are widely distributed across plants and discovered that TXSSs are only present in the genomes of Asteraceae species (Figure 3.3). This suggests that reports of taraxasterol in non-Asteraceae species are likely minor products of non-TXSS OSCs, particularly MASs. This is not unexpected as many triterpene-producing OSCs have been reported to be promiscuous, catalysing the production of more than one triterpene scaffold (Thimmappa et al., 2014).

There are 13 subfamilies of Asteraceae, however, only four subfamilies contain species for which high-quality genome or transcriptome sequence data are available. These are the Mutisioideae, Carduoideae, Cichorioideae and Asteroideae, and TXSSs were only found in the latter three. Therefore, it is not possible to conclude if TXSS evolved in a common ancestor all Asteraceae lineages, or in a common ancestor of the three main subfamilies. However, Bayesian analysis (Figure 3.11) suggested that the common ancestor of Asteraceae MASs and TXSSs likely evolved between 83 and 121 mya, which overlaps with previous estimates of the emergence of Asteraceae of between 64 and 91 mya (Mandel et al., 2019). The analysis dated the MAS-TXSS divergence to slightly earlier than Asteraceae divergence date. This difference may be attributed to fewer lineages being sampled for the Bayesian gene tree than for the species tree described in Mandel et al. With more MASs and TXSSs, the two fossils used in the prior could be placed in more accurate phylogenetic positions, which might enhance the accuracy of divergence date estimation.

In the phylogenetic reconstruction of OSCs, all TXSSs formed a monophyletic clade (Figure 3.4), suggesting that TXSS likely evolved only once. The order of divergence of three main subfamilies in the TXSS and species trees was generally consistent, with Carduoideae diverging first, followed by Cichorioideae and Asteroideae. Some

phylogenetic incongruency was observed, including grouping CcTXSS with Cichorioideae TXSSs and inconsistent Asteroideae intertribal relationships. However, this may be due to the relatively few TXSSs sampled and, potentially, a higher substitution rate in TXSSs compared to those used for the species tree.

Whole genome duplication (WGD) events have been proposed to be important drivers of secondary metabolism innovation, exemplified by the evolution of new glucosinolates in Brassica through WGD-driven expansion of CYP79s and triterpene biosynthesis pathway diversification in the Maleae (apple tribe) (Edger et al., 2015; Su et al., 2021). The emergence of multiple gene copies reduces selection pressure on individual genes, enabling gene copies to accumulate mutations that lead to new functions. Further, intraspecific ploidy has been proposed as a mechanism to alter the abundance of metabolites and to enable rapid adaptation to new environments (Moore et al., 2014).

The presence of TXSS in the genomes of all three of the main Asteraceae subfamilies suggests that is has been widely maintained. Most genomes encoded one TXSS gene. For example, *C. officinalis* is a tetraploid and was found to encode one functional (*CoOSC1*) as well as a non-functional homeologue (*CoOSC17*), likely degraded due to the diploidisation of the genome (Figure 3.12). However, the *A. millefolium* sample used in this study was identified to be hexaploid and would, therefore, be expected to encode three copies of TXSS (one in each genome) from the WGD events that are well-documented in this species (López-Vinyallonga et al., 2015). Given that nine copies were identified in the transcriptome, it is likely further single gene duplication events gave rise to the three additional copies. In the future, it would be interesting to profile the product abundance and gene expression of *TXSS* genes across the ploidy variants of *A. millefolium* to understand the contribution of the multiple gene copies found in these genomes. The *I. britannica* sample was found to be tetraploid. Although WGD has not been documented in *I. britannica*, it has been widely recorded in other Inula species such as *I. racemosa* and *I. grandiflora* (Himshikha et al., 2017; Mudassir Jeelani et al., 2022). The presence of three TXSS copies in *I. britannica* transcriptomes also suggests gene-duplication events. As the genomes of these species have not been sequenced, the relative locations of the *TXSS* genes cannot be investigated.

3.5.2 TXSS likely evolved via the duplication and neofunctionalisation of a MAS in which the active site was reshaped

Phylogenetic reconstruction of OSCs suggests that TXSSs likely evolved from an ancestral MAS during the evolution of Asteraceae. OSCs share high structural similarity and subtle changes in their active sites have been shown to alter their product specificity. For example, *Pisum sativum* cycloartenol synthase was converted into a cucurbitadienol synthase by a Y118L mutation and *Cucurbita pepo* cucurbitadienol synthase was converted into a parkeol synthase by a L125Y mutation (Takase et al., 2015). The substitution of residues within the active site

residues may lead to the perturbation of substrate folding, charge transfer and deprotonation processes, resulting in changes in product specificity.

Structural comparison of CoMAS and CoTXSS revealed differences in two residues within the active site, I367 and E371. Substituting these residues in CoMAS with the corresponding residues from CoTXSS led to the loss of amyrin specificity and an increase in the production of taraxasterol (Figure 3.9).

During the biosynthesis of triterpenes, OSC folds 2,3-oxidosqualene into a chair-chair-chair conformation and protonates the epoxide group, followed by a series of ring expansion and charge transfer events (Thimmappa et al., 2014). During charge transfer, carbocation intermediates are stabilised by π-cation interactions between the positive charge and aromatic amino acids in the active site (Thimmappa et al., 2014). When the positive charge reaches a carbon close to a proton acceptor, the acceptor will subtract a proton to terminate the reaction and release the triterpene (Thimmappa et al., 2014). Only amino acids with a deprotonated R-group located less than 5 Angstroms away from the proton donor can perform proton subtraction (Liang et al., 2022; Thoma et al., 2004; Zhang et al., 2023).

Both I367 and E371 were estimated to be more than 5 Angstroms away from the docked taraxasteryl cation substrate, for which the position can be predicted to be the same as for the structurally similar ursanyl and oleanyl cations. Thus, E371 is unlikely to be able to directly subtract the proton from the C-ring to produce amyrin. It is likely that I367 and E371 may play indirect role in the determination of product specificity by interacting with other residues to shape the active site to favour the transferral of the positive charge into C-ring and deprotonating of the C-ring. The substitution of these residues to those found in CoTXSS likely reshaped the active site to favour the transferral of the positive charge within the E-ring and enable deprotonation.

Epistatic effects are frequently observed when mutating multiple active site residues in OSCs, exemplified by the concerted actions of three active site residues in the interconversion between *Avena strigosa* hopenol B synthase and hop-17(21)-en-3β-ol synthase (Liang et al., 2022). In the CoMAS mutagenesis experiment, it was observed that I367M contributed to an increase in the production of ψ-taraxasterol and a reduction in amyrin biosynthesis; E371D further boosted this effect.

Gene duplication and neofunctionalization has been described as leading to the product diversification of other OSCs. For example, *Chenopodium quinoa* OSCs producing rare B, C-ring-opened triterpenes are likely derived from an ancient β-amyrin synthase, while *Apostichopus japonicus*'s parkeol synthases and lanostadienol synthases are likely derived from an ancient lanosterol synthase which was subsequently lost (Thimmappa et al., 2022; Zhou et al., 2024). As gene duplication provides functional redundancy, these events are thought to enable a greater tolerance of mutations, which paves way for functional divergence and adaptation to new environments (Birchler and Yang, 2022).

MASs that mainly produce α-amyrin and β-amyrin as well as small quantities of ψ-taraxasterol and taraxasterol were found in both non-Asteraceae and Asteraceae lineages (Figure 3.10). It is likely a *MAS* was duplicated during the emergence of Asteraceae. Following duplication, at least one functional *MAS* was maintained, and possible underwent subsequent duplications as multiple closely related *MAS*s were found in some genomes (Figure 3.7), with one copy specialised in the production of taraxasterols.

In plants, tandem gene duplication usually occurs through recombination between displaced repetitive sequences on homologous chromosomes during meiosis, or via the activity of transposable elements (Birchler and Yang, 2022). Tandemly duplicated genes often remain adjacent to their ancestral gene. In the genome of *C. officinalis, CoOSC2*, a predicted *MAS* (based on phylogenetic analysis) was observed to be located upstream of *CoTXSS* in contig1. *CoOSC2* was not characterised in this study as its expression was not detected in the tissues profiled. However, the coding sequence was not mutated, and it may be expressed in other tissue or conditions. Orthologues of *CoOSC2 and CoTXSS* were also adjacent in the genomes of *C. cardunculus*, *L. sativa* and *H. annuus*, which belong to three main Asteraceae subfamilies, supporting a duplication event in a common ancestor.

C. officinalis encodes five candidate *MAS*s, namely *CoOSC2*, *CoOSC3* (*CoMAS*), *CoOSC18*, *CoOSC23* and *CoOSC28*. Of these, two (*CoOSC18* and *CoOSC28*) are likely pseudogenes as they were not expressed and were predicted to encode premature stop codons. The two residues that were found to determine CoMAS product specificity (I367 and E371) are conserved in CoOSC18 and CoOSC23 but were substituted to T and S respectively in CoOSC2 and CoOSC28. However, I367 and E371 are conserved in CcOSC2, LsOSC2 and HaOSC2. Hence, it is likely that Asteraceae MASs with signature residues I367 and E371 arose once during the divergence of Asteraceae. In *C. officinalis*, CoOSC2 and CoOSC28 substitutions at these two residues might have led to a change in product specificity.

Together, phylogenetic relationships, Bayesian molecular clock analysis and genome localisation all indicate that TXSSs likely evolved from duplication and neofunctionalization of a MAS. This event likely happened soon after the divergence of Asteraceae, but more rigorous phylogenetic analyses and genome analyses of species from additional lineages are required to confirm this.

3.5.3 The accumulation patterns of taraxasterol-derived terpenoids differ among species and might play an adaptive role in the Cichorioideae

Metabolic profiling of different tissues of diverse Asteraceae species revealed that while ψ-taraxasterol/taraxasterol and derivatives only accumulate in aerial tissues of Carduoideae and Asteroideae species, they are detected in the roots of Cichorioideae species. Notably, the roots of the non-basal Cichorieae lineages (*T. kok-saghyz*, *C. endivia*, *C. capillaris*) had the higher concentrations of ψ-taraxasterol/taraxasterol compounds than leaves or flowers. Tuberous roots are prevalently observed in Cichorioideae lineages, and they harbour metabolite-

enriched lactiferous canals (Kilian et al., 2009). Thus, it can be hypothesised that tuberous roots provide a new metabolite reservoir for ψ-taraxasterol/taraxasterol and derivatives.

Four of the five Cichorioideae species were found to accumulate mostly taraxasterol compounds, while there was no consistency in taraxasterol vs ψ-taraxasterol in species of the Carduoideae and Asteroideae. For four of these species (*T. dubius*, *T. kok-saghyz*, *C. endivia* and *C. officinalis*), there was a consistency between the abundance of ψ-taraxasterol/taraxasterol compounds and the product specificity of the TXSS, also suggesting that TXSS is the main source of ψ-taraxasterol/taraxasterol in those species. Notably, the multiple TXSSs from *A. millefollium* were not characterised. In the future, it would be interesting to determine if the multiple copies encoded in this genome all show the same product specificity or if different copies are specialised to different products and/or are only expressed in specific tissues. It should also be noted that MASs, and perhaps additional OSCs, produce small quantities of ψ-taraxasterol or taraxasterol, which may contribute to the metabolic profiles of different tissues. Changes in the site of expression (roots) and to product specificity (more taraxasterol) in the Cichorioideae might suggest a potential adaptive role for taraxasterol compounds in interacting with the environment surrounding the tuberous roots. To date, although the bioactivity of taraxasterol on human cells and pathogens has been explored, and anti-inflammation, anti-proliferation, anti-oxidation and anti-pathogen activities have been documented (Jiao et al., 2022), it remains unknown if these compounds affect plant pathogens or pests. Indeed, we still have a limited understanding of the biological function of ψ-taraxasterol/taraxasterol in the Asteraceae. Although taraxasterol and taraxasterone constitute around 50% of the dry weight of metabolites extracted from *T. kok-saghyz* roots, evidence from feeding assays of *Melolontha melolontha* (May cockchafer) on wild-type and *TkOSC1 (TkTXSS)-RNAi* lines suggested that they are unlikely to be anti-herbivory compounds (Böttner et al., 2023; Pütter et al., 2019). Interestingly, the ψ-taraxasterol-derived faradiol palmitate constitutes around 50% of the dry weight of extracted metabolites from *C. officinalis* flowers (Golubova et al., 2025). Given the hydrophobicity of faradiol palmitate it could be hypothesised to have a role in floral cuticle formation.

Alternatively, the adaptive roles of these compounds might be subtle, for example related to reshaping root microbiota. Compared to wild-type, *Arabidopsis thaliana* mutants unable to produce thalianin, arabidin and their triterpene fatty acid ester derivatives showed differences in root microbiota composition, with significantly higher Bacteriodetes abundance and lower Deltaproteobacteria abundance (Huang et al., 2019). Hence, future experiments comparing root microbiota of species which accumulate taraxasterol in the roots and those of species that do not could provide insights into the biological function of taraxasterol.

3.6 Conclusions

Combining evidence from genome and transcriptome mining, molecular clock analysis, genome location analysis and site-directed mutagenesis, it was found that TXSS likely evolved from a MAS soon after the divergence of Asteraceae. TXSS has been maintained in the Asteraceae with most species containing one copy per diploid genome, although it has been expanded in certain lineages through tandem duplication. The main products of TXSS (ψ-taraxasterol and taraxasterol) typically accumulate in higher abundance in the aerial tissues of Carduoideae and Asteroideae, while greater quantities accumulate in tuberous roots of the Cichorioideae.

4. Chapter 4 Identification of residues involved in determining the product specificity of taraxasterol synthases

4.1 Introduction

4.1.1 Variation in the product specificity of OSCs

OSCs cyclise 2,3-oxidosqualene into triterpene scaffolds through a cascade of substrate folding, substrate protonation and ring formation events, followed by rearrangement and deprotonation into the final product (see also Chapter 1.4) (Thimmappa et al., 2014). Some OSCs have precise control over each step of the reaction, allowing them to deprotonate the carbocation intermediate at a specific carbon atom and produce a single, specific product. For instance, AtLAS and AtTHAS from *Arabidopsis thaliana* specialise in cyclising 2,3-oxidosqualene into lanosterol and thalianol respectively (Fazio et al., 2004; Suzuki et al., 2006). However, many OSCs do not precisely control each step or efficiently stabilise the intermediate. In addition, the presence of multiple residues able to act as proton acceptors can lead to the termination of cation transfer reaction by deprotonation at an alternative carbon resulting in more than one product. For instance, most α-amyrin synthases also produce β-amyrin (Thimmappa et al., 2014) and, in an extreme case, AtBARS1 produces baruol and 22 minor products through aberrant cyclisation, rearrangement and deprotonation (Lodeiro et al., 2007).

Despite divergence at the sequence level, OSCs share high structural similarity (Thimmappa et al., 2014). Within the active site, residues involved in core processes such as reaction initiation and substrate folding are highly conserved among OSCs, and their substitution generally leads to a loss of activity. In addition, the active sites of OSCs are enriched in aromatic amino acids, of which the side chains interact and stabilise carbocation intermediates through π-cation interactions. Hence, given the conserved nature of the active site, residues that vary between OSCs tend to contribute to variation in product specificity (Chen et al., 2021). Although some variable residues may directly interact with reaction intermediates to alter product specificity, others are likely to play more subtle roles through active site reshaping.

Variant residues can be identified using active site prediction and multiple sequence alignments. The function of such residues is then experimentally investigated. Site-directed mutagenesis followed by functional characterisation in a heterologous host has been extensively used to investigate the molecular basis of OSC product specificity. For example, Liang et al. observed that three residues in the active site of *Avena strigosa* hopenol B synthase AsHS1 and hop-17(21)-en-3β-ol synthase AsHS2 differ (Liang et al., 2022). Mutating these residues in AsHS1 to the corresponding residues in AsHS2 led to the production of hop-17(21)-en-3β-ol, and reciprocal mutagenesis of AsHS2 led to the production of hopenol B and isomotiol. Similarly, Zhang et al. found that mutating a single active site residue (F726T) was sufficient to convert *Alisma orientale* cycloartenol synthase (AoCAS) into a

protostadienol synthase, but the reciprocal mutation T727F on *A. orientale* protostadienol synthase (AoPDS) did not have the same effect (Zhang et al., 2023).

4.1.2 Molecular dynamics and quantum mechanics/molecular mechanics for inferring mechanisms of enzyme catalysis

The rapid development of molecular modelling and simulation tools is providing new insights into enzyme functions and properties, both complementing and guiding experimental studies such as those described above (van der Kamp et al., 2008). These tools can be applied to protein folding, structural analyses, interactions with small molecules, and to the inference of catalysis mechanisms.

Although protein modelling and ligand docking are useful for predicting residues important to catalysis, dynamic atomic interaction patterns between enzyme active site residues and ligands are rarely captured. Classical molecular dynamics simulations (MD) allow molecular models to move and interact under predefined forcefields within a time frame and provide more information on the dynamic evolution of molecular systems.

Forcefields are sets of functions and parameters that describe the forces between atoms and are used to calculate the potential energy of entire macromolecular systems. Force fields are crucial in MD simulations as they model the interactions between atoms and molecules, allowing predictions of system behaviours under different conditions. Depending on the type of molecules and purpose of the simulation, different forcefields can be applied. For example, AMBER forcefields are typically applied to proteins, and GAFF forcefields are applied to ligands (Tian et al., 2020; Wang et al., 2004). Forcefields consist of both intra- and inter-atomic interaction terms. The former includes bond interactions, angle interactions and dihedral interactions, while the latter includes electrostatic interactions and van der Waal interactions.

When simulating enzyme-ligand complexes, ligands are docked to the active site of enzyme models before solvating them in a water box to imitate their behaviour in solution (or a cellular compartment) and performing energy minimisation. The complexes are heated to 300 K to accelerate atomic movement, followed by pressurisation to 1 atm to equilibrate the system. Finally, a longer MD simulation is performed (Liang et al., 2022). Many features can be extracted during the simulation, including atomic distances between ligands and residues of interest.

Classical MD can only capture interactions between atoms using forcefields, which limits the ability to model chemical reactions involving bond breaking and formation. Hence, methods describing electronic interactions are required to simulate chemical reactions. These can be achieved using quantum mechanics approaches (QM) such as semiempirical methods and density function theory (DFT) (van der Kamp and Mulholland, 2013). In combined quantum mechanics/molecular mechanics (QM/MM) simulation, small regions where bond breaking and formation occur are treated with QM methods, whereas the rest of the system is treated with MM methods to reduce

the computational cost and account for the effect of wider environment on the reaction. To calculate the energy profile of a chemical reaction, QM/MM can be combined with an enhanced sampling method such as umbrella sampling, where a biased potential is applied along the coordinate to overcome free energy barrier (Hirvonen et al., 2022). The resulting data is analysed using the weighted histogram analysis method (WHAM) to reconstruct free energy profile (Kumar et al., 1992). MD and QM/MM techniques have been extensively used to investigate the catalytic mechanisms and guide product specificity engineering of terpene synthases. These include monoterpene synthases such as bacterial linalool synthase and cineole synthase (Leferink et al., 2022); sesquiterpene synthases such as selinadiene synthase (Srivastava et al., 2024) and patchoulol synthase (Srivastava et al., 2023); and OSCs such as α-amyrin synthase (Wu et al., 2020), hopane synthase (Liang et al., 2022) and protostadienol synthase (Zhang et al., 2023).

4.1.3 Positive selection and its detection

Enzymes involved in the production of metabolites that confer advantages to survival or reproduction are predicted to be under selective pressure. Evidence of positive selection can reveal ongoing and past adaptive responses to environmental changes and are indicative of advantageous biological function.

Known as the neutral theory of molecular evolution, Kimura proposed that most evolutionary changes occur at molecular level and most genetic variation is caused by stochastic genetic drift events which fix selective neutral mutations (Kimura, 1968). Despite its focus on neutrality, this theory recognises Darwinian selection: deleterious mutations are rapidly eliminated, and adaptive mutations, although less frequently fixed than neutral ones, may eventually be maintained (Duret, 2008). Positive selection occurs when an adaptive mutation spreads across a population. At molecular level, it is exhibited as a higher ratio of nonsynonymous (dN) to synonymous substitution (dS). The dN/dS ratio represents selection strength (ω), where $\omega>1$ indicates positive selection, $\omega=1$ suggests neutral selection and $\omega<1$ reflects purifying selection.

Goldman and Yang developed a basic codon evolution model (M0), which assumes a constant ω across all branches and sites in the codon alignment and serves as a null model (Goldman and Yang, 1994). Later, they extended it to the branch model, which allows ω to vary among branches and suitable for detecting lineage-specific selections (Yang and Nielsen, 1998).

Further developments also led to several site models where ω varies between sites but stays constant between branches (Nielsen and Yang, 1998; Yang et al., 2000). Model M1a assumes a mixture of purifying and neutral selection across all sites and serves as the null model for both M2a, which allows another class of sites under positive selection, and M3, which assumes three classes of sites with estimated ω values. In comparison, M7 assumes purifying selection across all sites with ω following a beta distribution and serves as the null model for M8, which adds another class of sites with an estimated ω. Positive selection is reported when the

loglikelihood of the alternative model is significantly higher than the null model and sites under selection are recovered using the Bayes Empirical Bayes method (Yang et al., 2005).

Finally, Yang and Nielson proposed a branch-site model to allow ω to vary both among different branches and at different sites (Yang and Nielsen, 2002). Lineages of interest are assigned as foreground, and the remaining lineages constitute the background. In the widely used branch-site model A, four classes of sites are allowed (Table 4.1). The null model forces neutral selection on the foreground while the alternative model allows it to be under positive selection with an estimated ω.

Site class	Background	Foreground
0	Purifying selection ($0<\omega_0<1$)	Purifying selection ($0<\omega_0<1$)
1	Neutral selection ($\omega_1=1$)	Neutral selection ($\omega_1=1$)
2a	Purifying selection ($0<\omega_0<1$)	Neutral or positive selection ($\omega_2\geq1$)
2b	Neutral selection ($\omega_1=1$)	Neutral or positive selection ($\omega_2\geq1$)

Table 4.1 Conditions of ω in four different site classes of the branch site model A.

Positive selection analysis has been employed to understand changes in product specificity of plant terpene synthases. The divergence of germacrene synthases from caryophyllene synthases in Oryza (Chen et al., 2014), parkeol synthases from orysatinol synthases also in Oryza (Xue et al., 2018a) and the rise of lupeol synthases, β-amyrin synthases and dammarenediol synthases in Panax (Yang et al., 2023) were attributed to positive selection. The ability to produce new terpenes potentially contributes to the species adaptation to the environment and confers selective advantage.

As noted above, variations in the product specificity of OSCs are common with many enzymes producing more than one product. In Chapter 3, I observed variation in the products produced by TXSS from different species. However, it remains unclear if the production of different products has been selected or is a product of genetic drift. Further the molecular basis and mechanism of TXSS remain unknown.

Note: Data from this chapter are included in Golubova et al., 2025 (doi: 10.1038/s41467-025-62269-w). Figures 4.1, 4.3 and 4.4 are included in that manuscript as is a variant of figure 4.2.

4.2 Aims

In this chapter, I aimed to:

(1) Explore the molecular basis of TXSS product specificity.
(2) Investigate if TXSSs in which product specificity has changed more recently are under positive selection.
(3) Use computational modelling and simulation to explore the mechanism of TXSS catalysis

4.3 Work contributed by others

All work described in this chapter was done by the author of this thesis.

Advice on positive selection analysis was received from Dr Wilfried Haerty, Dr Dave Wright and Dr Will Nash (Earlham Institute).

Advice on structural modelling was received from Prof Andrew Hemmings (University of East Anglia).

Advice on MD simulation and QM/MM was received from Dr Marc van der Kamp (University of Bristol).

4.4 Results

4.4.1 Identification of residues involved in determining TXSS product specificity

In Chapter 3.4.3, nine TXSSs were characterised using heterologous expression in *N. benthamiana*. Variation in product specificity was observed with TXSSs from *Cynara cardunculus*, *Tragopogon dubius*, *Calendula officinalis*, *Calendula arvensis* and *Helianthus annuus* producing predominantly ψ-taraxasterol and TXSSs from *Taraxacum kok-saghyz*, *Taraxacum coreanum*, *Lactuca sativa* and *Cichorium endivia* producing predominantly taraxasterol (Figure 3.5). Knowledge of which residues contribute to such variation is useful both to understand enzyme mechanisms and to direct enzyme engineering. Thus, differences in the structures of enzymes in these two groups were investigated using CoTXSS and TkTXSS as exemplars.

To achieve this, structural models were built using AlphaFold2 (Jumper et al., 2021) (Figure 4.1). These models showed that both CoTXSS and TkTXSS have high structural similarity (RMSD = 0.806 and = 0.863, respectively) to the crystal structure of human lanosterol synthase (PDB: 1W6K) (Thoma et al., 2004). Using these models, the active sites of CoTXSS and TkTXSS were predicted from the docked taraxasteryl cation, the position of which was based on the location of lanosterol in 1W6K. Residues within 12 Angstroms of the predicted active sites were proposed to be involved in shaping the conformation of the active site and determining product specificity.

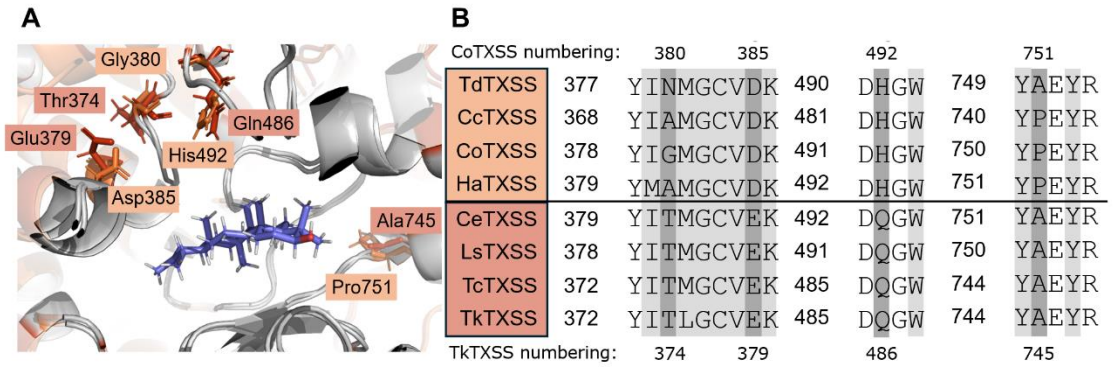


Figure 4.1 The active sites of *Calendula officinalis* taraxasterol synthase (CoTXSS) and *Taraxacum kok-saghyz* taraxasterol synthase (TkTXSS). (A) The taraxasteryl cation was docked into the structural models CoTXSS (Orange) and TkTXSS (Red). Residues that differ between them are highlighted and labelled. (B) Multiple sequence alignment of selected residues with active site residues highlighted in grey and residues that differ in dark grey.

Next, a multiple sequence alignment of all characterised TXSSs was conducted and residues in the active sites of TXSSs that produce more ψ-taraxasterol and those that produced more taraxasterol were compared. Four residues that differ between these two categories were identified, namely G380, D385, H492 and A751 (in CoTXSS numbering; Figure 4.1).

4.4.2 Signatures of positive selection were detected non-basal Cichorieae TXSSs

In Chapter 3.3.8, it was observed that non-basal lineages of the Cichorieae predominantly accumulate taraxasterol compounds and that these are mainly found in the roots. This contrasted with basal Cichorieae lineages as well as species from other subfamilies, most of which accumulate them in aerial tissues and many of which predominantly accumulate ψ-taraxasterol compounds. In support of this, TXSSs found in non-basal Cichorieae lineages produced more taraxasterol than ψ-taraxasterol (Figure 3.5). Given TXSS is likely to produce the majority of the taraxasterol found in these species, I hypothesised that the observed variation in TXSS product specificity in the non-basal lineages of the Cichorieae has been driven by positive selection.

To investigate this, a maximum likelihood tree of 14 representative TXSSs was constructed, and their corresponding coding sequences were aligned. Site model tests were performed by fitting the null model M0 and the site models M1a, M2a, M3 (K=3), M7 and M8 to the phylogeny (Yang, 2007).

Model	Number of parameters	Log likelihood	Estimates of Parameters	Positively selected sites
M0: one-ratio	27	-10123.6803	ω=0.1126	None
M1a: nearly neutral	28	-10030.3391	p0=0.9026, p1=0.0974, ω0=0.0681, ω1=1	Not allowed
M2a: selection	30	-10030.3391	p0=0.9026, p1=0.0974, p2=0, ω0=0.0681, ω1=1	None
M3: discrete (K=3)	31	-9995.6966	p0=0.7362, p1=0.2638, p2=0, ω0=0.0246, ω1=0.3909	None
M7: beta	28	-9992.7206	p=0.2748, q=1.8912	Not allowed
M8: beta and ω	30	-9992.7203	p0=0.9994, p=0.2758, q=1.9064, p1=0.0006, ω=1	None

Table 4.2 Log likelihoods, estimated parameters and sites likely to be under positive selection from the M0 model and site models M1a, M2a, M3, M7 and M8. Abbreviations: p0, proportion of sites subjected to selection pressure of ω0; p1, proportion of sites subjected to selection pressure of ω1; p2, proportion of sites under selection pressure different from ω0 and ω1; p and q in M7, ω ($0 \leq \omega \leq 1$) follows a beta distribution of beta (p, q); p0 in M8, proportion of sites whose ω ($0 \leq \omega \leq 1$) follows a beta distribution of beta (p, q); p1 in M8, proportion of sites with a different ω.

Both the M1a and M2a models are significantly better than M0 (M1a: $2\Delta l=186.68$, p-value<0.001 with df=1; M2a: $2\Delta l=186.68$, p-value<0.001 with df=2). In M1a, 90% of sites are estimated to be under strong purifying selection and the remaining 10% of sites under neutral selection. However, the same estimations were made for M2a with no new sites under positive selection and M2a is not significantly better than M1a ($2\Delta l=0$, p-value=1 with df=0). The M3 model is significantly better than M0 and M1a ($2\Delta l=255.96$, p-value<0.001 with df=4 and $2\Delta l=69.29$, p-value<0.001 with df=3 respectively). However, M3 suggests that all sites are under purifying selection with different strengths. Both the M7 and M8 models are significantly better than M0 (M7: $2\Delta l=261.92$, p-value<0.001 with df=1; M8: $2\Delta l=261.92$, p-value<0.001 with df=3). However, the M8 model is not significantly better than M7 ($2\Delta l=0.006$, p-value=0.499 with df=2), basically suggesting that all sites are under purifying selection with ω following a beta distribution.

Next, TXSSs from non-basal Cichorieae were selected as the foreground with all other TXSSs constituting the background and fitted both a null and alternative branch-site model A to the phylogeny (Figure 4.2; Table 4.3).

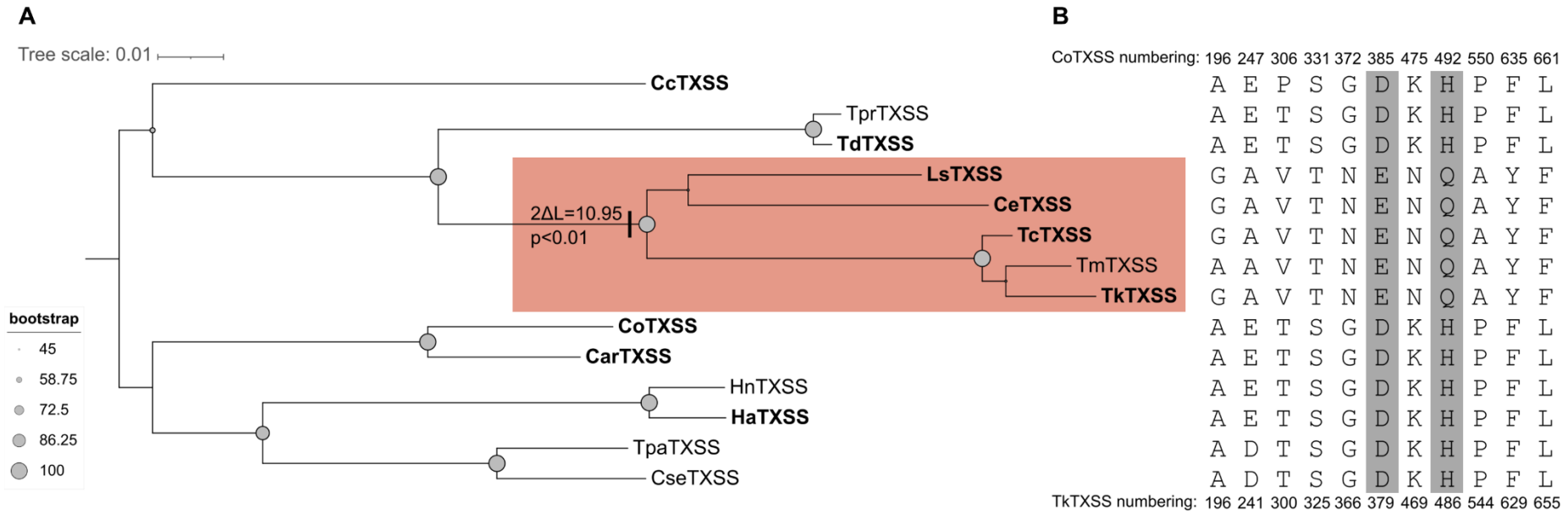


Figure 4.2 Branch-site test for positive selection in TXSSs. (A) The phylogenetic tree of TXSSs used for positive selection analysis. Characterised TXSSs are labelled in bold. The orange box represents the foreground and consists of TXSSs from non-basal Cichorieae that produce more taraxasterol than ψ-taraxasterol. (B) Residues identified as being under positive selection in the foreground were recovered using Bayesian Empirical Bayes (probability > 0.9). Two residues in the predicted active site are shaded in grey.

Model	Number of parameters	Log likelihood	Estimates of Parameters	Positively selected sites
M0: one-ratio	27	-10123.6803	ω=0.1126	None
Null model	29	-10022.0229	p0=0.6092, (p1+p2b)=0.0964, p2a=0.2944, ω0=0.0642, ω1=ω2=1	Not allowed
Alternative model	30	-10016.5456	p0=0.8735, p1=0.0925, (p2a+p2b)=0.0340, ω0=0.0645, ω1=1, ω2=33.1372	Shown in **Figure 4.2 (B)**

Table 4.3 Log likelihoods, estimated parameters and sites likely to be under positive selection from M0, null and alternative branch-site models A. Abbreviations: p0, proportion of sites in site class 0; p1, proportion of sites in site class 1; p2a, proportion of sites in site class 2a; p2b, proportion of sites in site class 2b; ω_0, ω_1 and ω_2 are consistent with Table 4.1.

Both the null and alternative models are significantly better than M0 (null model: $2\Delta l=211.37$, p-value<0.001 with df=2; alternative model: $2\Delta l=219.88$, p-value<0.001 with df=3), and the alternative model is significantly better than the null model ($2\Delta l=10.95$, p-value<0.001 with df=1). Based on the branch model, it is likely that non-basal Cichorieae TXSSs are under positive selection ($2\Delta l=10.95$, p-value<0.01). Based on the site model, all 2a and 2b class sites of non-basal Cichorieae TXSSs discovered in the alternative model are likely under positive selection.

Bayesian Empirical Bayes (BEB) was used to recover 11 residues for which the probability of being under positive selection was higher than 0.9. Two of these residues (385 and 492 in CoTXSS numbering) are located in the predicted active site of the CoTXSS structural model.

4.4.3 Single residue mutagenesis

In total, four active site residues were found to differ between TXSSs that mainly produce ψ-taraxasterol and those that mainly produce taraxasterol. To test the roles of these in determining production specificity, site-directed mutagenesis was performed on CoTXSS and TkTXSS. This was done using mutagenesis PCR on the expression vectors (Chapter 2.3.3 and Supplementary Table 2), which were subsequently used for heterologous expression in *N. benthamiana* leaves (Chapter 2.5.2).

First, reciprocal mutants of CoTXSS and TkTXSS in which a single residue had been mutated were compared. In these mutants, each of the four selected active-site residues was swapped to the corresponding residue (Figure 4.3).

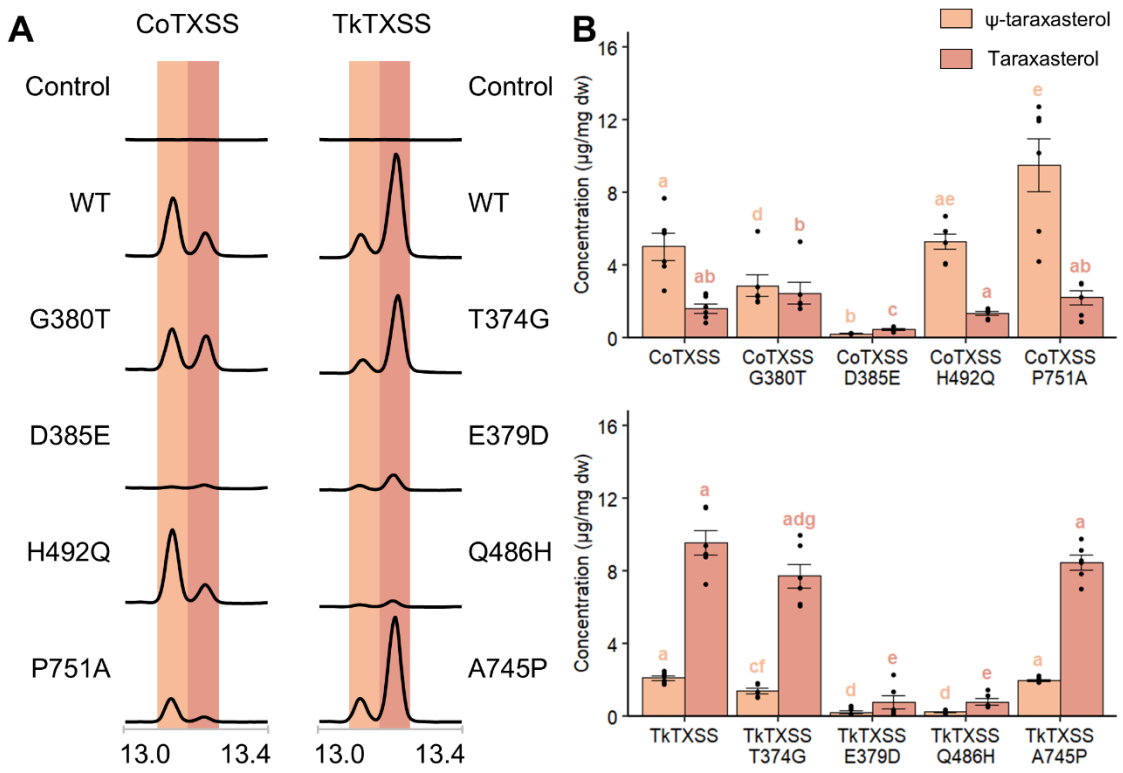


Figure 4.3 GC-MS analysis of *N. benthamiana* leaves transiently expressing CoTXSS and TkTXSS mutants (A) Total ion chromatograms of extracts of *N. benthamiana* leaves transiently expressing wild type and mutated CoTXSS and TkTXSS. (B) Quantification of triterpenes produced by wild type and mutated CoTXSS and TkTXSS relative to an internal standard; n=6; error bars represent standard error. Statistical significance in ψ-taraxasterol content (orange lowercase letters) and taraxasterol content (red lowercase letters) were analysed using a Kruskal-Wallis test followed by post-hoc Wilcoxon rank sum test with a Benjamini-Hochberg correction. Samples that do not share the same lower-case letter are significantly different from each other ($p<0.05$).

Compared to their corresponding wild type enzymes, CoTXSS^{G380T} and TkTXSS^{T374G} both produced significantly less ψ-taraxasterol but there was no significant change to the production of taraxasterol. CoTXSS^{G380T} produced similar amounts of ψ-taraxasterol and taraxasterol with slightly more ψ-taraxasterol, whereas TkTXSS^{T374G} still produced much more taraxasterol. Both CoTXSS^{D385E} and TkTXSS^{E379D} showed a significant reduction in activity, resulting in the reduction of both ψ-taraxasterol and taraxasterol. Despite minimal activity, both produced slightly more taraxasterol than ψ-taraxasterol. CoTXSS^{H492Q} did not show a significant change in activity, while TkTXSS^{Q486H} showed a reduction in activity, producing less ψ-taraxasterol and taraxasterol. CoTXSS^{P751A} unexpectedly displayed a significant increase in the production of ψ-taraxasterol but production of taraxasterol remained unaffected. TkTXSS^{A745P} did not display any significant change in the production of either ψ-taraxasterol or taraxasterol.

In general, single reciprocal mutations did not reverse the product ratio of either CoTXSS or TkTXSS without significant loss in activity. However, CoTXSS^{G380T} changed the ψ-taraxasterol:taraxasterol ratio such that it became more similar to TkTXSS through a reduction in the production of ψ-taraxasterol.

4.4.4 Multiple residue mutagenesis

Next, enzymes with multiple mutations were tested. This included enzymes in which the two active site residues identified to be under positive selection were mutated and enzymes in which all four active site residues that differ between TXSSs producing mainly ψ-taraxasterol and those producing mainly taraxasterol were mutated (Figure 4.4).

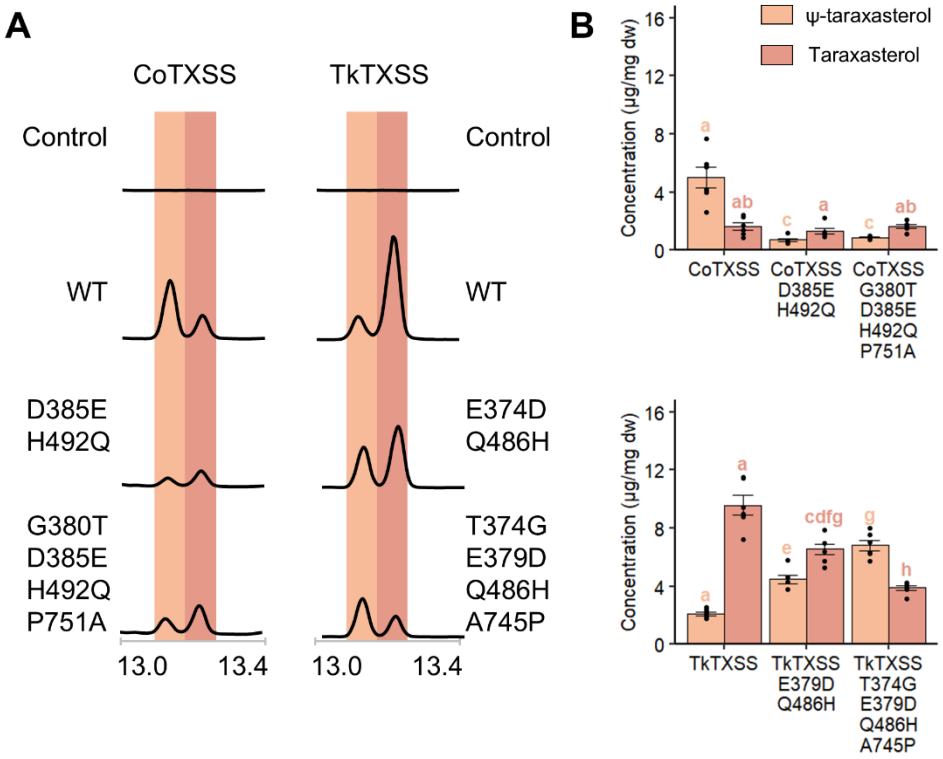


Figure 4.4 GC-MS analysis of *N. benthamiana* leaves transiently expressing CoTXSS and TkTXSS double and quadruple mutants (A) Total ion chromatograms of extracts of *N. benthamiana* leaves transiently expressing wild type and mutated CoTXSS and TkTXSS. (B) Quantification of triterpenes produced by wild type and mutated CoTXSS and TkTXSS relative to an internal standard; n=6; error bars represent standard error. Statistical significance in ψ-taraxasterol content (orange lowercase letters) and taraxasterol content (red lowercase letters) were analysed using a Kruskal-Wallis test followed by post-hoc Wilcoxon rank sum test with a Benjamini-Hochberg correction. Samples that do not share the same lower-case letter are significantly different from each other ($p<0.05$).

Compared to wildtype CoTXSS, CoTXSS$^{D385E,\ H492Q}$ displayed a significant reduction in the production of ψ-taraxasterol but no significant change to the amount of taraxasterol. Thus, there was a reversal of the dominant product, with more taraxasterol than ψ-taraxasterol being produced. TkTXSS$^{E379D,\ Q486H}$ produced significantly more ψ-taraxasterol and significantly less taraxasterol. However, it still produced more taraxasterol than ψ-taraxasterol.

The quadruple mutant, CoTXSS$^{G380T, D385E, H492Q, P751A}$, was similar to the CoTXSS$^{D385E, H492Q}$ double mutant, showing a significant reduction in ψ-taraxasterol production but no significant change in taraxasterol production, which resulted in a reversal of the dominant product. In contrast, TkTXSS$^{T374A, E379D, Q486H, A745P}$ produced significantly more ψ-taraxasterol and significantly less taraxasterol than both the TkTXSS$^{E379D, Q486H}$ double mutant and the TkTXSS wildtype, ultimately resulting in a reverse in the dominant product.

In summary, mutating the two active site residues under positive selection was sufficient to change the dominant product of CoTXSS to taraxasterol, whereas mutating all four variable active site residues was required to alter the dominant product of TkTXSS to ψ-taraxasterol. However, CoTXSS mutants also showed a reduction in activity, while TkTXSS mutants produced similar levels of products overall.

4.4.5 Loss of function mutagenesis

Finally, to explore the importance of the four residues to TXSS activity, non-polar and corresponding residues (CoTXSS G380 and P751; TkTXSS T374 and A745) were mutated to D and polar residues (CoTXSS D385 and H492; TkTXSS E379 and Q486) were mutated to A (Figure 4.5). The activity of these mutants was characterised in *N. benthamiana* as previously described.

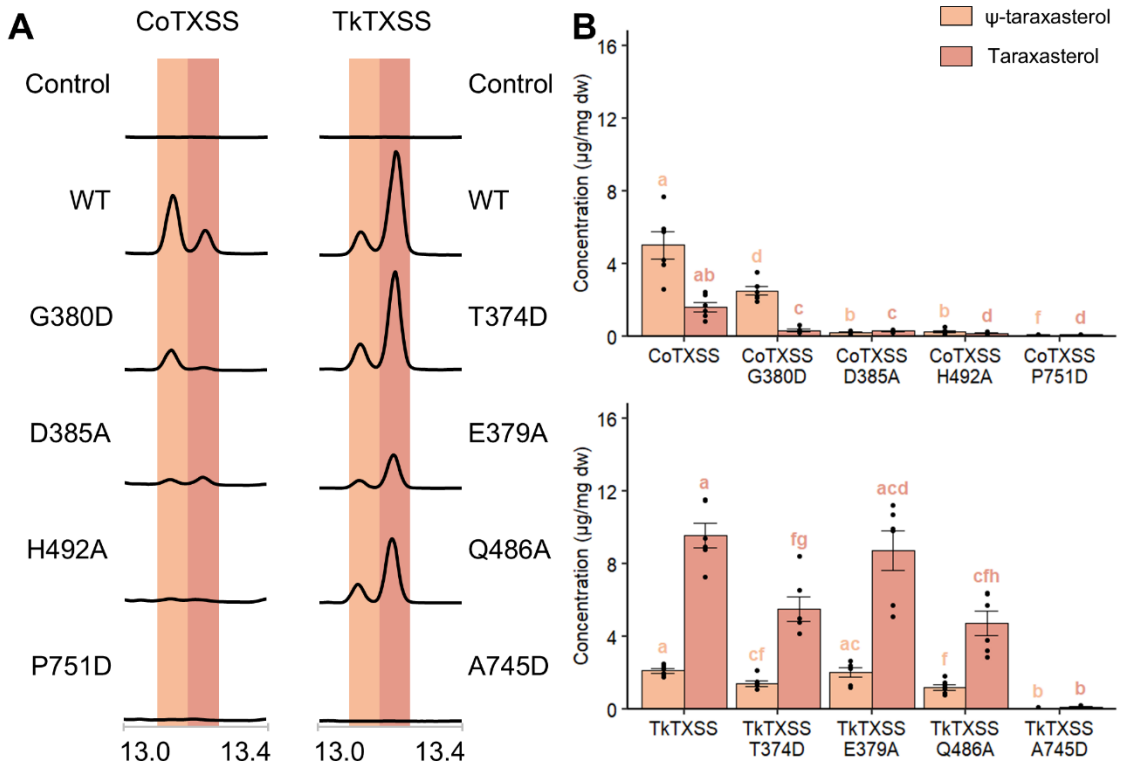


Figure 4.5 GC-MS analysis of *N. benthamiana* leaves transiently expressing CoTXSS and TkTXSS loss-of-function mutants (A) Total ion chromatograms of extracts of *N. benthamiana* leaves transiently expressing wild type and mutated CoTXSS and TkTXSS. (B) Quantification of triterpenes produced by wild type and mutated CoTXSS and TkTXSS relative to an internal standard; n=6; error bars represent standard error. Statistical significance in ψ-taraxasterol content (orange lowercase letters) and taraxasterol content (red lowercase letters) were analysed using a Kruskal-

Wallis test followed by post-hoc Wilcoxon rank sum test with a Benjamini-Hochberg correction. Samples that do not share the same lower-case letter are significantly different from each other ($p<0.05$).

Both CoTXSS^{G380D} and TkTXSS^{T374D} showed a significant reduction in activity, producing less ψ-taraxasterol and taraxasterol. CoTXSS^{D385A} had minimal activity producing only trace amounts of ψ-taraxasterol and taraxasterol, whereas the activity of TkTXSS^{E379A} was not significantly different to the wild type TkTXSS. Similarly, CoTXSS^{H492A} had minimal activity, while TkTXSS^{Q486A} showed a significant reduction in activity. Finally, both CoTXSS^{P751D} and TkTXSS^{A745D} showed minimal activity. In summary, mutating these active site residues to unrelated amino acids resulted in a significant loss of activity in CoTXSS, whereas TkTXSS was more resistant to these mutations.

4.4.6 Exploring the molecular mechanism of TXSS product specificity with molecular simulation

The cyclisation of the linear 2,3-oxidosqualene into ψ-taraxasterol and taraxasterol is a complex reaction. The variation between these two products arises from differential deprotonation of the final intermediate, the taraxasteryl cation (Figure 4.6). Deprotonation of taraxasteryl cation at C21 leads to double bond formation between C20 and C21, giving rise to ψ-taraxasterol, while deprotonation at C29 leads to double bond formation between C20 and C29, giving rise to taraxasterol.

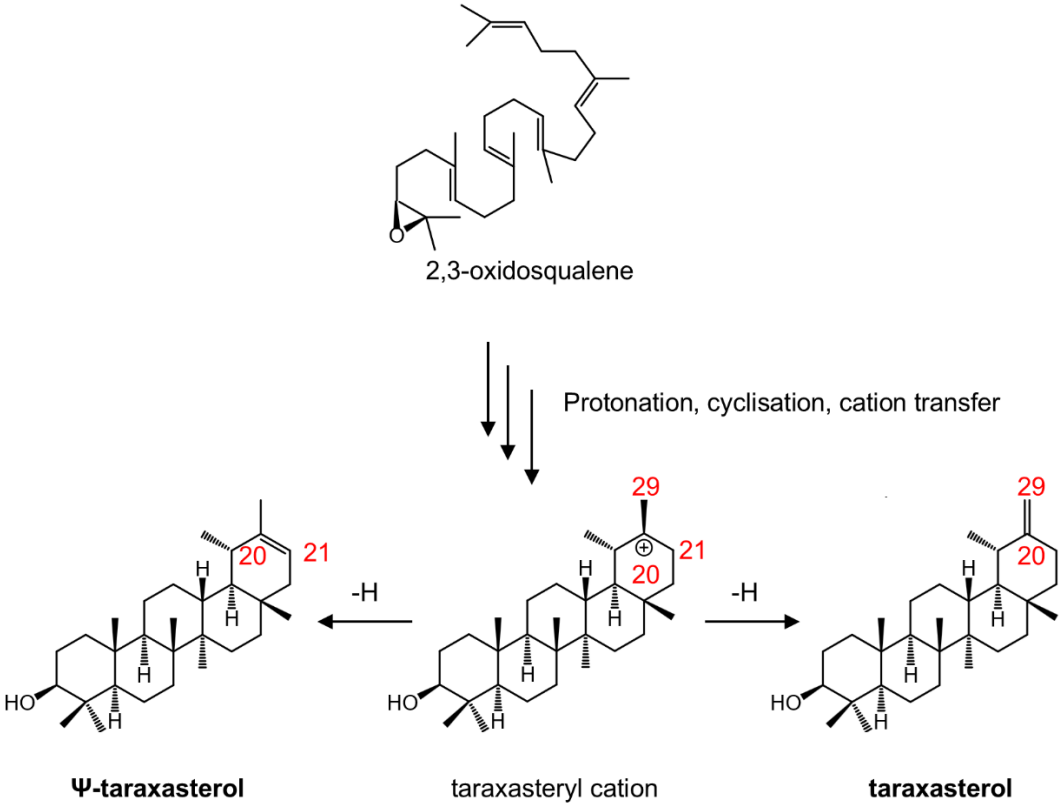


Figure 4.6 The biosynthesis of ψ-taraxasterol and taraxasterol by TXSSs. The taraxasteryl cation acts as the final intermediate, with deprotonation at C21 giving rise to ψ-taraxasterol and deprotonation at C29 giving rise to taraxasterol.

To provide a mechanistic explanation of TXSS product specificity, the identity of the proton acceptor was investigated using molecular dynamic (MD) simulations to explore dynamic interactions between candidate active site residues and taraxasteryl cation. Proton acceptors of terpene synthases are generally polar charged residues. Due to the uncertainty of ligand model docking into structural model, polar charged residues less than 5 Angstroms are predicted as candidate proton acceptors.

In CoTXSS and TkTXSS static enzyme-substrate models, no polar charged residues are found within 5 Angstroms of the deprotonation site. The closest potential proton acceptors are D385 (CoTXSS) and E379 (TkTXSS). Given that mutating these residues contributed to changes in product specificity (Figure 4.3) and/or activity (Figure 4.5), molecular dynamic (MD) simulations were used to explore dynamic interactions between these residues and the taraxasteryl cation.

Throughout the course of MD simulation, the RMSD for all replicates remained below 3 Angstroms, suggesting no significant conformational change occurred during the MD (Figure 4.7). Distance monitoring indicated that the protonatable oxygen atom of CoTXSS D385 was able to move within 5 Angstroms from taraxasteryl cation C21. In contrast, the protonatable oxygen atom of TkTXSS E379 remained greater than 5 Angstroms away from taraxasteryl cation C29. Therefore, CoTXSS D385 is capable of directly deprotonating the taraxasteryl cation but TkTXSS E379 is not.

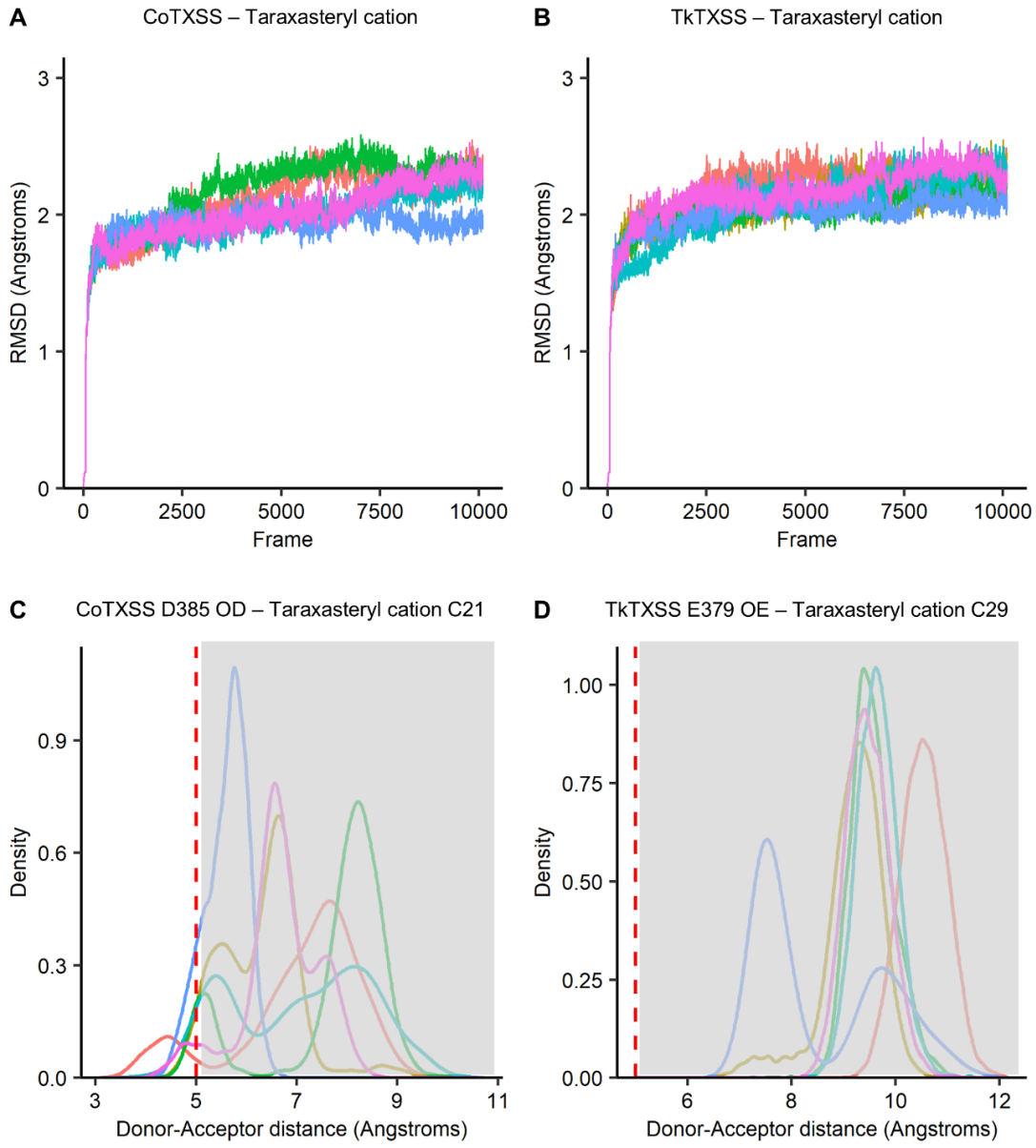


Figure 4.7 Molecular dynamics simulations of CoTXSS-taraxasteryl cation and TkTXSS-taraxasteryl cation complexes. (A) RMSD of CoTXSS-taraxasteryl cation complex throughout the MD simulation. (B) RMSD of TkTXSS-taraxasteryl cation complex throughout the MD simulation. The first frame was used as the reference for both RMSD calculations. (C) Density distribution of distance between CoTXSS D385 protonatable oxygen atom and taraxasteryl cation C21 in the last 70 ns of simulation. (D) Density distribution of distance between TkTXSS E379 protonatable oxygen atom and taraxasteryl cation C29 in the last 70 ns of simulation. Grey box represents frames where the interatomic distance is greater than 5 Angstroms. Six replicates with different seeds were run per sample.

In the CoTXSS-taraxasteryl cation model, Y273 was observed to obstruct direct interactions between CoTXSS D385 and the E-ring of taraxasteryl. The side chain of CoTXSS Y273 stabilises the taraxasteryl cation through π-cation interaction (Figure 4.8). During the MD simulation, CoTXSS Y273 could be displaced over time without losing the π-cation interaction, allowing the side chain of CoTXSS D385 side chain and protonatable oxygen atom to move closer to the E-ring. However, obstruction of

CoTXSS Y273 still dominates, which causes the distance between the D385 protonatable oxygen and taraxasteryl cation C21 to remain mostly at greater than 5 Angstroms throughout the courses of the simulation.

On the other hand, in the TkTXSS-taraxasteryl cation model, in addition to the obstruction of TkTXSS Y267, the side chain of TkTXSS E379 points away from the active site. This positioning was maintained throughout the simulation, preventing the protonatable oxygen of TkTXSS E379 from moving close to taraxasteryl cation C29.

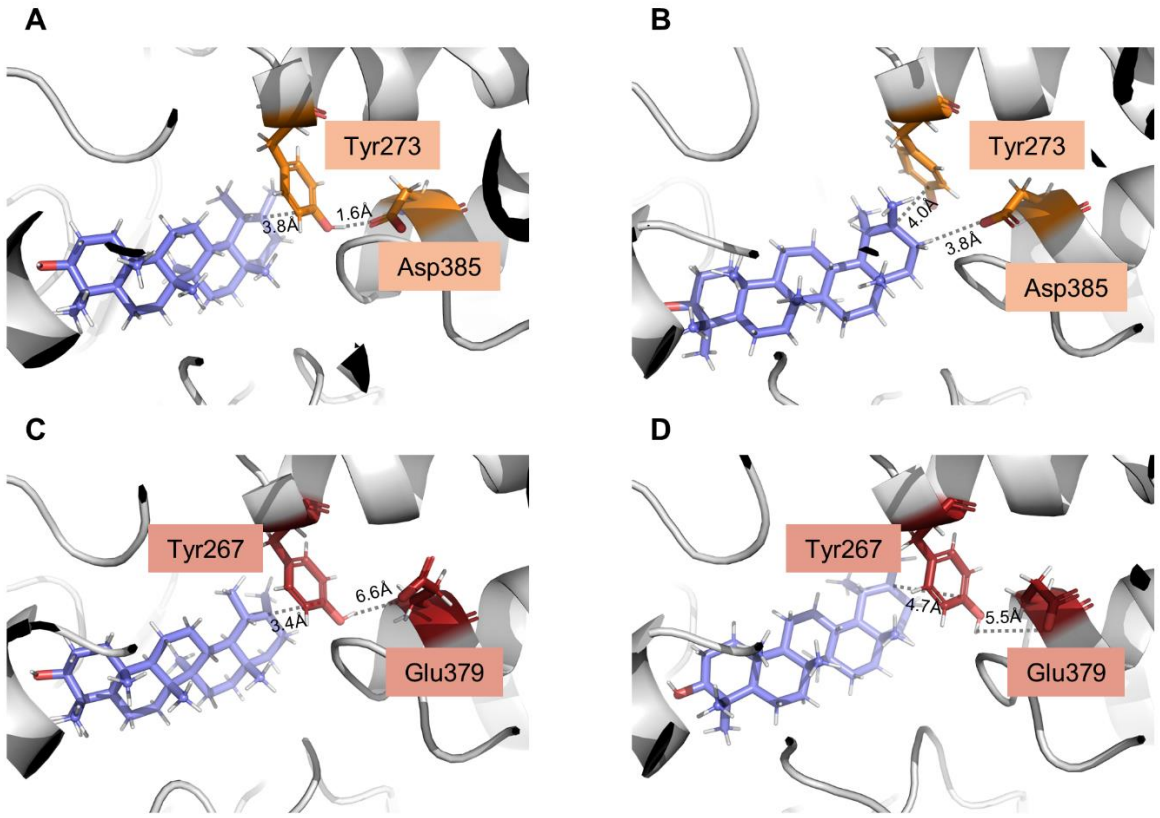


Figure 4.8 Interaction between TXSS and a taraxasteryl cation. (A) Interaction between CoTXSS and a taraxasteryl cation in the static structural model before MD simulation. (B) Interaction between CoTXSS and a taraxasteryl cation in a MD frame where the D385 side chain is close to C21. (C) Interaction between TkTXSS and a taraxasteryl cation in the static structural model before MD simulation. (D) Interaction between TkTXSS and a taraxasteryl cation in a MD frame. Cutoff distance for π-cation interaction was 6 Angstroms.

Given the proximity between the tyrosine and taraxasteryl cation during the simulation and previous reports of C and T as unconventional proton acceptors in OSCs (Liang et al., 2022; Zhang et al., 2023), the possibility of Y273 and Y267 acting as proton acceptor in CoTXSS and TkTXSS was explored by mutating it to F to remove the hydroxyl group and characterising the mutants (Figure 4.9). Neither CoTXSS^{Y273F} nor TkTXSS^{Y267F} showed abolishment of ψ-taraxasterol and taraxasterol production activity, suggesting that tyrosine is unlikely a proton acceptor and is likely only to be involved in intermediate stabilisation.

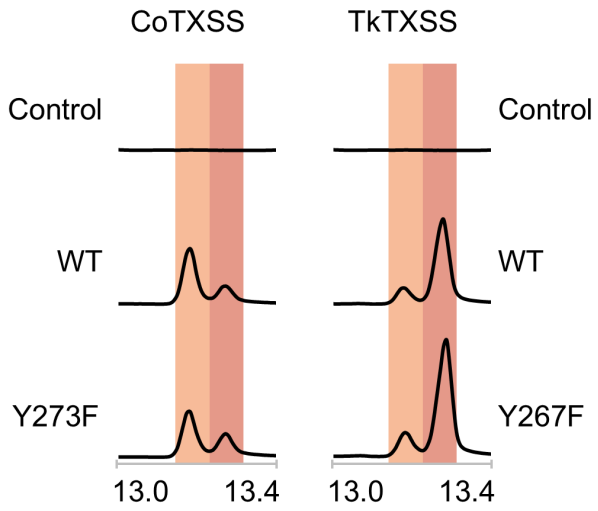


Figure 4.9 GC-MS analysis of *N. benthamiana* leaves transiently expressing CoTXSS and TkTXSS Y-to-F mutants. Total ion chromatograms of extracts of *N. benthamiana* leaves transiently expressing wild type and mutated CoTXSS and TkTXSS.

Classical MD simulation revealed that CoTXSS D385 can act as a potential proton acceptor whereas the corresponding residue in TkTXSS (E379) cannot. To further explore the mechanism of product specificity, QM/MM simulation was performed to calculate the energy requirement to deprotonate taraxasteryl C21 and C29, which lead to ψ-taraxasterol and taraxasterol, respectively.

As the TkTXSS proton acceptor is still unknown, the simulation could only be performed on CoTXSS. To do this, two frames were selected from independent MD runs where the protonatable oxygen atom CoTXSS D385 was within 5 Angstroms of taraxasteryl C21 and C29. From these starting points, the deprotonation process was simulated with umbrella sampling and the free energy of each step was calculated using WHAM (Figure 4.9).

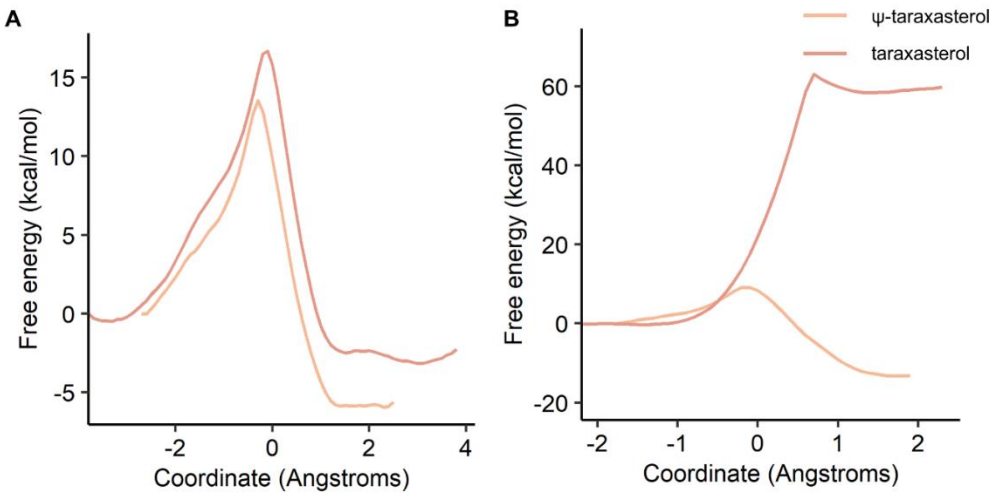


Figure 4.10 Free energy profiles of the deprotonation of taraxasteryl cations into ψ-taraxasterol and taraxasterol by CoTXSS during QM/MM simulation. (A) Free energy profile of the reactions in frame 1. (B) Free energy profile of the reactions in frame 2.

In the first frame, 13.53 kcal/mol is required to overcome the free energy barrier for C21 deprotonation and ψ-taraxasterol formation, compared to 16.66 kcal/mol for C29 deprotonation and taraxasterol formation. Given the lower free energy barrier for C21 deprotonation, proton transfer from C21 to CoTXSS D385 is more favourable than from C29, which is consistent with the experimental observation of preferential ψ-taraxasterol production.

In the second frame, 9.12 kcal/mol is required to overcome the free energy barrier for C21 deprotonation and ψ-taraxasterol formation, in contrast to C29 deprotonation and taraxasterol formation which is thermodynamically infeasible. Closer inspection suggests that before the protonatable oxygen atom moves closer to C29, the proton from C21 has already been transferred to CoTXSS D385.

4.5 Discussion

4.5.1 Positive selection is likely to have changed the product specificity of TXSS in non-basal Cichorieae

In Chapter 3, it was observed that TXSSs from non-basal Cichorieae species predominantly produce taraxasterol. In contrast, TXSSs from other lineages predominantly produce ψ-taraxasterol. The difference between ψ-taraxasterol and taraxasterol is subtle, which is the position of the double bond in the E-ring. This raises the question of whether the change in specificity is simply chemical noise, or if it is an adaptation that confers a selective advantage. A previous study reported that selective pressure on *Oryza sativa* (rice) orysatinol synthases changed their specificity to produce parkeol, which likely plays a role in resistance to insects and pathogens (Xue et al., 2018a).

A site model was fit to a phylogeny of characterised TXSSs to investigate whether any site, regardless of lineage, was likely to be under positive selection. This approach, which ignored any variation in selection pressures amongst lineages, found that none of the positive selection models were significantly better than the corresponding null model. This suggests that it is unlikely that there was selection on any site of all TXSSs. However, it is important to note that as the focus of this study was comparison of TXSSs that produce different dominant products, a balanced number of TXSSs from non-basal Cichorieae species and from other species were included; to gain insights on how selection acted on TXSSs as a family, a greater number of taxa should be characterised and included in the phylogeny.

To address the question of whether TXSSs from non-basal Cichorieae are under positive selection, a branch site model was fit to the TXSS phylogeny with these taxa as foreground. This revealed that these TXSSs are likely under positive selection and additionally recovered 11 residues that are likely under positive selection (Figure 4.2). Therefore, TXSSs from non-basal Cichorieae likely evolved more rapidly than TXSSs in the background, with 11 residues experiencing higher substitution rates. This finding suggests that the switch to the production of taraxasterol as the dominant product is unlikely to be chemical noise and it is likely to have conferred a

selective advantage. As noted in Chapter 3, this change in product specificity is accompanied by a change in the site of accumulation from aerial tissues to the roots, which also supports a change in biological function. Although the function of both products remains unknown, accumulation in the roots, which in this lineage tend to be tuberous roots, suggests a role in the defence of soil-borne pathogens or in shaping the root microbiome. Interaction between root-secreted triterpenes and root microbiomes has been recorded in the literature, including thalianols regulating the abundance of Bacteriodetes and Deltaproteobacteria (Huang et al., 2019), soyasaponin Bb enhancing growth of potential plant growth-promoting bacteria *Novosphingobium* (Fujimatsu et al., 2020) and cucurbitacin B enriching *Enterobacter* and *Bacillus*, which in turns inhibits pathogenic *Fusarium oxysporum* (Zhong et al., 2022).

TXSSs are likely the main contributors of taraxasterol-derived triterpenoid accumulation. Hence, difference in taraxasterol-derived triterpenoid tissue specificity is likely attributed to changes in regulatory sequences which lead to differential expression of *TXSS*. In *C. officinalis*, *CoTXSS* regulatory sequences likely contain binding sites for flower specific transcriptional activators to enable their flower specific expression in, whereas in *T. kok-saghyz TkTXSS* regulatory sequences likely contain binding sites for the root specific counterparts. Differential expression of triterpenoid biosynthesis pathway genes has been demonstrated in the Cucurbitaceae family, in which cultivated *Cucumis sativus*, *C. melo* and *Citrullus lanatus* show lower expression of cucurbitacin biosynthesis genes and lower cucurbitacin accumulation in fruits than wild counterparts, as the result of reduced expression of a fruit-specific bHLH regulatory transcription factor (Zhou et al., 2016).

4.5.2 Structure guided mutations of CoTXSS and TkTXSS revealed residues important for product specificity

Consistent with the observation in the crystal structure of human lanosterol synthase (Thoma et al., 2004), the active site of TXSS is primarily hydrophobic and enriched in aromatic amino acids (F, Y and W), effectively stabilising reaction intermediates through extensive π-cation interactions. It also has the characteristic DCTAE motif for reaction initiation. Those highly conserved active site features are also observed in other OSCs with diverse product specificities, including *Euphorbia tirucalli* β-amyrin synthase, *A. strigosa* hopenol synthases and *Oryza sativa* parkeol synthase (Hoshino et al., 2017; Liang et al., 2022; Xue et al., 2018a).

From a structure-guided multiple sequence alignment, residues predicted to be important for the product specificity of TXSSs were selected. Four residues within 12 Angstroms of the active site were identified that differ between TXSSs that predominantly produce ψ-taraxasterol versus taraxasterol. However, these are not in direct contact with the substrate (Figure 4.1). Hence, it is likely that, in addition to CoTXSS D385, which the MD simulations indicated as a likely proton acceptor, all residues play an indirect role in determining product specificity through active site reshaping. However, in most other OSC mutagenesis studies, the target residues

are closer to the substrate and likely directly interact with them. For example, *A. strigosa* hopenol synthase (AsHS) specificity towards hopenol B/ hop-17(21)-en-3β-ol is determined by three residues 4 Angstroms away from the substrate (Liang et al., 2022) and *C. quinoa* quinoxide synthase specificity towards B,C-ring-opened triterpenes/ β-amyrin is determined by four residues 6 Angstroms away from the substrate (Zhou et al., 2024). Nevertheless, mutagenesis studies on *Barbarea vulgaris* amyrin synthases demonstrated that residues further away from the active sites act in concert with active site residues to determine α-amyrin/β-amyrin specificity (Günther et al., 2021).

Single reciprocal mutations of CoTXSS or TkTXSS were unable to alter product specificity without affecting activity. The activity and product specificity of CoTXSS were more susceptible to mutations, whereas those of TkTXSS were more resistant to mutations. The reduction in the production of ψ-taraxasterol by CoTXSS^{G380T} indicates the importance of G380 in this enzyme. Interestingly, however, the identity of residue 380 (CoTXSS numbering) is G, A and N in different TXSSs that mainly produce ψ-taraxasterol (Figure 4.1). In contrast, the maintenance of taraxasterol production in the reciprocal TkTXSS^{T374G} mutant suggests that T374 play a less important role (Figure 4.3). As the predominant production of taraxasterol by TkTXSS was likely driven by positive selection, it is less likely to be affected by changes in the identities of residues that are not under selection. In both CoTXSS^{G380D} and TkTXSS^{T374D} activity was reduced but product specificity was maintained, thus a D at this site is suboptimal (Figure 4.5).

CoTXSS^{D385E} and TkTXSS^{E379D} both showed a large reduction in activity (Figure 4.3). This was unexpected as both are acidic polar amino acids. This suggests that D and E may have different roles in CoTXSS and TkTXSS. Compared to D, E has an extra methylene group. The shorter side chain of D means it is more rigid, which may be preferable within an active site. In CoTXSS, the side chain of D385 points towards the taraxasteryl substrate, in contrast to the side chain of TkTXSS E379 which points away. MD simulation suggested that D385 could act as a proton acceptor while E379 could not. Consistently, CoTXSS^{D385A} showed minimal activity while the activity of TkTXSS^{E379A} was similar to the wild type (Figure 4.5).

The activity of CoTXSS^{H492Q} was unchanged, whereas the reciprocal mutation of TkTXSS^{Q486H} resulted in a loss of activity (Figure 4.3). In contrast, CoTXSS^{H492A} lost activity, whereas TkTXSS^{Q486A} did not (Figure 4.5). This suggests that this residue plays distinct structural and mechanistic roles in both enzymes dependent on the microenvironment: a polar residue is needed to maintain CoTXSS activity and a non-charged residue to maintain TkTXSS activity.

Surprisingly, CoTXSS^{P751A} showed a significant increase in ψ-taraxasterol production without significant reduction in taraxasterol production (Figure 4.3). This residue is therefore a target for engineering an increase in ψ-taraxasterol production. Evidence that this may work in other TXSSs comes from TdTXSS, which has an A at this position and produced large quantities of ψ-taraxasterol (Figure 3.5). The activity of

TkTXSS^{A745P} was not significantly changed compared to the wild type (Figure 4.3), therefore, the P/A identity at position 751 may play a less important role in determining product specificity. However, activity was depleted in both CoTXSS^{P751D} and TkTXSS^{A745D}, suggesting that a non-polar residue is essential for activity.

The dominant product of both CoTXSS$^{D385E,\ H492Q}$ and CoTXSS$^{G380T,\ D385E,\ H492Q,\ P751A}$ was reversed compared to the wild type. This was achieved through a reduction in the production of ψ-taraxasterol production as taraxasterol production did not increase (Figure 4.4). In contrast, ψ-taraxasterol production increased and taraxasterol production decreased in TkTXSS$^{E379D,\ Q486H}$ (Figure 4.4). Further, both effects increased in the quadruple mutant TkTXSS$^{T374G,\ E379D,\ Q486H,\ A745P}$ reversing the major product (Figure 4.4). Interestingly, although the activity of TkTXSS^{E379D} and TkTXSS^{Q486H} was minimal, TkTXSS$^{E379D,\ Q486H}$ was active. Thus, there is evidence that E379 and Q486, which are both under positive selection, act in concert and may have co-evolved. Overall, selection to increase the production of taraxasterol appears to have made TkTXSS more resistant to mutations. This supports the hypothesis that preferential production of taraxasterol may be contributed to the adaptation of non-basal Cichorieae species to their environment.

Consistent with the CoMAS mutagenesis study described in Chapter 3.4.5, mutation epistasis, where simultaneous mutation of multiple residues reinforces the effect of each other, was observed in the mutation of CoTXSS and TkTXSS. Single mutations either failed to reverse product specificity or exerted detrimental effects to enzyme activity, whereas double and quadruple mutations were able to reverse product ratios of CoTXSS and TkTXSS respectively. Additive effects of individual mutations and interactions between residues were consistent with the previously reported AsHS mutagenesis study (Liang et al., 2022). Single reciprocal mutations could not reverse AsHS1 specificity towards hop-17(21)-en-3β-ol or AsHS2 specificity towards hopenol B, whereas triple reciprocal mutations could achieve the swap in specificity.

Positive selection has been prevalently observed in the evolutionary history of OSCs (Xue et al., 2012). It has been described as leading to the functional diversification of cycloartenol synthases in the plant kingdom, the emergence of parkeol synthases in *Oryza* (Xue et al., 2018a) and the divergence of lupeol synthases, β-amyrin synthases and dammarenediol synthases in *Panax* (Yang et al., 2023). However, previous studies on OSCs have not experimentally characterised residues identified to be under positive selection. Nevertheless, as only residues within the active site were studied, further work on residues outside the active site could be conducted to assess the effect of selection in different regions of TXSS and their contribution to product specificity.

4.5.3 Computational modelling and MD simulation provide a mechanistic explanation for the dominant products of CoTXSS and TkTXSS

Classical MD simulation and QM/MM have been used to identify deprotonation residues of OSCs and explain differences in product specificity. For example, this approach identified an unusual deprotonation base (C366) in AsHS1, which is 3

Angstroms away from the proton donor carbon of hopenol B carbocation intermediate (Liang et al., 2022). Furthermore, QM/MM calculation revealed that the dominant product (hopenol B) of AsHS1 is the result of a lower energy barrier along the reaction pathway. It also demonstrated that a change in product specificity in a mutant, AsHS1 Y410A, could be attributed to a relief in steric hindrance of the aromatic side chain of Y410 within 4 Angstroms from the intermediate. This led to a reduction in energy required for further hydride shift from hopenol B carbocation intermediate and enabled production of isomer hop-17(21)-en-3β-ol with the double bond at a different position from hopenol B.

Similarly, it was discovered that AoPDS protostadienol synthase uses the unusual residue, T727, as its deprotonation base (Zhang et al., 2023). T727 is 3.1 Angstroms away from the proton donor carbon of protostadienol carbocation intermediate. QM/MM calculation revealed that the introduction of F726T mutation into a cycloartenol synthase, AoCAS, which changed the dominant product to protostadienol, was attributable to a reduced energy barrier. AoCAS F726 corresponds to AoPDS T727 and is located around 5 Angstroms away from the carbocation intermediates, allowing π-cation interaction. A F726T mutation changed the reaction energy profile of protostadienol production in AoCAS, making it similar to that of wild-type AoPDS.

In this study, MD simulation provided a dynamic image of the interactions between TXSS and the taraxasteryl cation, revealing transient conformational changes that could not be captured by the static enzyme-substrate model. In the CoTXSS-taraxasteryl cation structural model, Y273 creates a barrier between D385 and the taraxasteryl cation and prevents the protonatable atom of D385 from approaching the cation. However, during MD simulation, periods of gradual Y273 displacement without the loss of the π-cation interaction were observed, allowing D385 to serve as a proton acceptor in CoTXSS (Figure 4.8). The role of D385 was supported by loss of function of the CoTXSS^{D385A} mutant (Figure 4.6). Although non-canonical proton acceptors were reported in AsHS and AoPDS (C and T) (Liang et al., 2022; Zhang et al., 2023), Y273 does not act as proton acceptor in CoTXSS and its only known role is stabilising intermediates through π-cation interaction (Figure 4.9).

From energy barrier calculations during the QM/MM simulation it was observed that CoTXSS needs to overcome a lower free energy barrier to deprotonate C21 than it does to deprotonate C29, consistent with the experimentally observed preferential production of ψ-taraxasterol (Figure 4.10). Nevertheless, it is possible that additional residues contribute to deprotonation, and water molecules may act as intermediaries that relay protons to D385. Intriguingly, MD simulation of the TkTXSS-taraxasteryl cation model and mutagenesis experiments both suggest that the corresponding residue in TkTXSS (E379) is unlikely to be involved in deprotonation (Figure 4.6, 4.8). Y267, which corresponds to CoTXSS Y273, also does not involve in deprotonation (Figure 4.9). As residues with protonatable side chains are not observed in the vicinity of the taraxasteryl cation in TkTXSS, an active site water

molecule is most likely the proton acceptor. In this model, the proton is predicted to transfer through a water chain accommodated inside TkTXSS to a protonatable residue away from the active site. Proton transfer to water has been documented in *Alicyclobacillus acidocaldarius* squalene hopene synthase (AaSHS) and in some sesquiterpene synthases (Srivastava et al., 2024, 2023; Wendt et al., 1999).

TXSSs that mainly produce ψ-taraxasterol via deprotonation of taraxasteryl C21 by D385 are ancestral to the taraxasterol-producing TXSSs found in the non-basal Cichorieae. The substitution of D385E would have driven the necessity for an alternate proton acceptor to increase the production of taraxasterol. Thus, the likely selective pression on TkTXSS may have reshaped the active site, rendering it more resistant to mutations.

Due to the paucity of experimentally determined enzyme structures with product intermediates, structural models of TXSSs with a rationally docked taraxasteryl cation model were used as inputs for MD simulation. In addition, given the size of the molecular system and computational cost, I used semi-empirical method PM6 in the QM region. These uncertainties may have contributed to inaccuracies in the simulation results. Nevertheless, results from computational simulation and experimental results are consistent for the role of CoTXSS D385. Thus, molecular modelling and simulation is a useful tool to generate hypotheses and to further investigate mechanisms suggested from experimental data.

4.6 Conclusions

Using structure-informed multiple sequence alignment and mutational studies, four residues were identified that influence the product specificity of TXSSs. Evolutionary studies revealed that the production of taraxasterol by non-basal Cichorieae likely conferred a selected advantage and identified two residues in the active site of TkTXSS that are likely to be under positive selection. Interestingly, mutating these residues individually was detrimental to enzyme activity, whereas mutating both was not, supporting co-evolution. Together with the difference in the site of accumulation, observed in Chapter 3, these findings support a distinct biological function for taraxasterol compared to ψ-taraxasterol. Although the proton acceptor and the molecular mechanism for closure of the E ring by TkTXSS remains unknown, computational modelling and simulation revealed that CoTXSS D385 is a likely proton acceptor, preferentially deprotonating C21 of the taraxasteryl cation, leading to the dominant production of ψ-taraxasterol. In contrast, the corresponding residue in TkTXSS (E379) is unlikely to be involved in deprotonation. This new knowledge of the structural determinants of TXSS product specificity and underlying mechanisms could guide future engineering of TXSS for more specific and efficient production of ψ-taraxasterol or taraxasterol.

5. Chapter 5 The regulation of faradiol fatty acid ester biosynthesis pathway in *Calendula officinalis* flowers

5.1 Introduction

Plant gene expression is regulated at multiple levels, with transcriptional (mRNA production) and post-transcriptional (mRNA processing, stability, and translation into protein) regulation being key mechanisms, together with epigenetic regulation (changes in chromatin structure) and post-translational modification of proteins.

Interactions between DNA-binding proteins, known as transcription factors (TFs), and regulatory sequences, particularly cis-regulatory elements (CREs), play an important role in regulating transcription (Strader et al., 2022). CREs are regions of non-coding DNA that bind TFs and other proteins to activate or repress transcription and are broadly classified into promoters, enhancers, silencers and insulators (Spitz and Furlong, 2012; Swinnen et al., 2016). TFs either directly bind to specific cognate DNA motifs known transcription factor binding sites (TFBSs) or bind indirectly (via another TF) to activate or repress transcription by enabling or blocking the recruitment of RNA polymerase complexes. CREs contain TFBSs for multiple TFs and, reciprocally, individual TFs interact with multiple CREs to control the expression of multiple genes simultaneously.

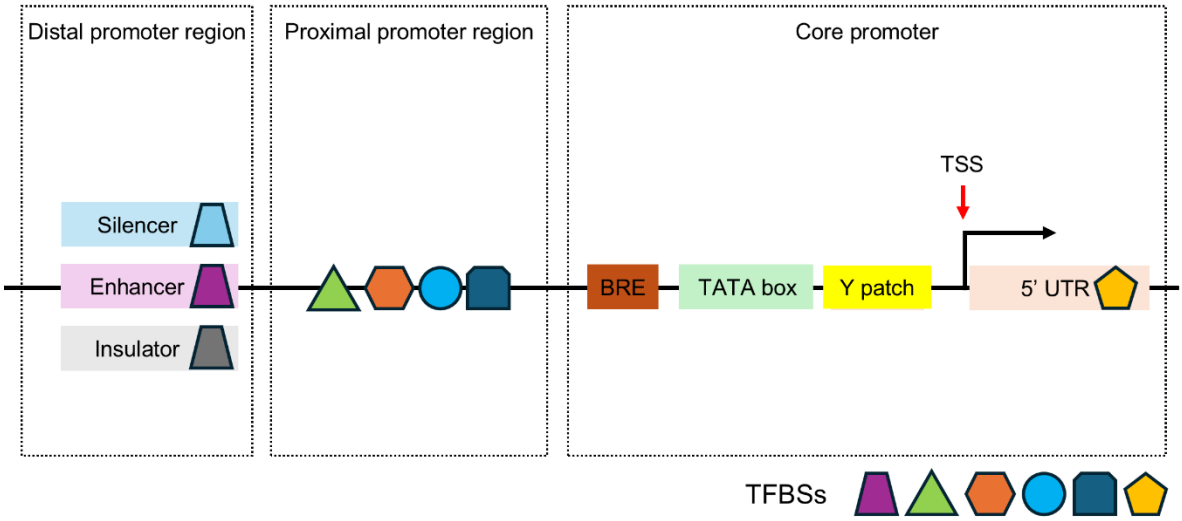


Figure 5.1 Generalised architecture of the promoter of a protein-coding plant gene. Transcription factors contribute to the regulation of expression by binding to cognate motifs in all regions. The TATA box is not a prerequisite for transcription and is absent from many plant promoters.

Promoters are located directly upstream of the transcript. They regulate and control basal levels of transcription via the recruitment of TFs and RNA polymerases (Yasmeen et al., 2023). The promoters of protein coding genes are transcribed by RNA polymerase II (RNAPolII) and typically include a core promoter region, a proximal region, and one more distal enhancer, suppressor or insulator (Figure 5.1)

(Yasmeen et al., 2023). The core promoter is the minimal sequence required to initiate transcription and typically spans −/+50 bp from the transcription start site (TSS). Core promoters of plant protein-coding genes generally contain a recognition element for transcription factor II B (a crucial protein involved in recruitment of RNApolII) but may or may not contain a TATA box (a landmark to initiate transcription) or a Y patch (for interaction with the transcriptional machinery, particularly in the absence of a TATA box). Protein-coding genes also contain untranslated regions, which are transcribed, non-coding sequences before the start codon (5′ UTRs) and after the stop codon (3′ UTRs). 5′ UTRs often contain structural motifs and open reading frames important for regulation (Jia et al., 2020). Some plant 5′ UTRs contain a pyrimidine-rich region that accelerates the rate of transcription. The structure of the 3′ UTRs and any introns also contribute to mRNA stability and influence protein levels. Enhancers recruit TFs and other proteins that interact with promoters (and the proteins bound to them) to increase transcription. As well as being located within or close to the promoter region, they can also be located at varying distances and up- and down-stream of their target transcript(s) (Schmitz et al., 2022). In contrast, silencers interact with promoters to reduce their activity, also in a distance and orientation independent manner (Schmitz et al., 2022). Insulators are located between CREs, including between genes, isolating promoters from the effects of enhancers and repressors but are relatively poorly characterised in plants (Kurbidaeva and Purugganan, 2021).

These regulatory sequences define the spatiotemporal expression patterns of genes, restricting expression to individual tissues or even cell types. Expression of a given gene is determined both by the cellular concentrations of the specific TFs that interact with its CREs, and by the chromatin accessibility of the regulatory sequences (Klemm et al., 2019). Chromatin accessibility is controlled by composition and organisation of nucleosomes and non-histone proteins that interact with chromatin.

Based on their expression patterns, plant promoters can be classified into two types. Constitutive promoters drive gene expression in most cells and tissues throughout growth and development, whereas variable promoters are only active in specific/tissues, at specific developmental stages, or in response to environmental stimuli such as biotic and abiotic stressors (Brooks et al., 2023). As secondary metabolites are not essential for normal growth and development, their promoters are often variable, limiting accumulation to the locations and conditions required. The expression of such genes may be regulated by changes to chromatin states. For example, the production of camelexins in *Arabidopsis thaliana* is regulated by histone modifications in response to pathogen attacks (Zhao et al., 2021).

Genes that function in the same biosynthetic pathway are often co-regulated. For example, in *Medicago truncatula*, two bHLH-type TFs, TSAR1 and TSAR2, bind to and activate the promoters of genes encoding 3-hydroxy-3-methylglutaryl-CoA

reductase 1 (HMGR), cytochrome P450s, and UDP-dependent glycosyltransferases involved in biosynthesis of oleanane-type triterpene saponins (Mertens et al., 2016). The genes of some plant biosynthetic pathways, including those of terpenoids, are physically co-located or co-located in the genome. For example, the pathways for thalianol and marneral in *A. thaliana*, the pathway for avenacin A-1 in *Avena strigosa*, and the pathways for cucurbitacins in *Citrullus* (Nützmann and Osbourn, 2014; Zhou et al., 2016). Gene clusters have been observed to be regulated both by TFs and at the chromatin level (Nützmann and Osbourn, 2015).

TFs from multiple families have been documented to regulate triterpenoid biosynthesis. In the outer tissues of *A. thaliana* root tips, methyl jasmonate-induced expression of thalianol and marneral is activated by two redundant bHLH-type TFs while expression in inner root tissues is repressed by a DOF-type TF (Nguyen et al., 2023). Members in R2R3-MYB family have also been implicated in regulating triterpenoid synthesis, including MdMYB66, which was shown to activate the expression of a lupeol synthase in *Malus domestica* (Falginella et al., 2021). In *Citrus sinensis,* CiMYB42 activates expression of *CiOSC*, which encodes an OSC involved in limonoid biosynthesis (Zhang et al., 2020).

The identification of TFs that regulate biosynthetic genes is desirable because it provides insights into the conditions required for gene expression. In addition, expression levels of TF genes can be manipulated to affect the expression of the whole pathway. Some studies have identified TFs regulating triterpenoid biosynthesis by searching transcriptomics datasets for TF genes with expression profiles that most closely match those of the pathway genes. For example, this approach was used to discover a bHLH-type TFs activating oleanane-type triterpenoid biosynthesis in *Medicago truncatula* (Mertens et al., 2016). However, this ideally requires multiple transcriptomes across time or space and assumes that the pathway genes are tightly co-expressed.

Several sequencing techniques enable the identification of the specific binding motifs recognised by DNA binding proteins. For example, in chromatin immunoprecipitation sequencing (ChIP-seq), DNA is crosslinked to a TF of interest before the DNA is sheared and antibodies specific to the TF are used to pull down the protein-bound fragments. The DNA is then purified and sequenced, identifying the footprints of the TF (Kidder et al., 2011). In another method, known as DNA affinity purification sequencing (DAP-seq), TFs of interest are recombinantly expressed with an affinity tag and used to capture fragmented genomic DNA, which is subsequently sequenced (O'Malley et al., 2016). The experimental identification of TFBSs allows the inference of positional frequency matrices (PFMs), which represent the frequency of each nucleotide at each position within the TFBS. These PFMs can then be used to predict the presence of TFBSs in other DNA sequences. This has made it possible to predict which TFs can bind to a given promoter sequence.

The DNA-binding domains of TFs are highly conserved (Zenker et al., 2025). Thus, PFMs from orthologous TFs from related species can be used to predict TFBS in species for which ChIP-seq/DAP-seq data is missing. Recently, this has demonstrated at the genome scale across ten species spanning 150 million years of flowering plant evolution, where TF orthologues from distant species were found to share preference towards the same TFBSs in DAP-seq experiments (Baumgart et al., 2025).

Once candidate TFs have been identified, interactions with target promoters can be experimentally determined using several methods. In yeast-one-hybrid assays, the coding sequence of TFs of interest are fused with a transcriptional activation domain and co-expressed in yeast with a construct encoding the test promoter fused to a reporter (Reece-Hoyes and Walhout, 2012). Promoter-TF interactions are identified by detection of the reporter signal. Physical protein DNA interactions can be tested using electrophoretic mobility shift assays (EMSAs), in which DNA is electrophoresed with and without recombinantly expressed TF protein. Detection of binding is based on the principle that free DNA migrates faster than protein-bound DNA during electrophoresis (Hellman and Fried, 2007). A microplate-based variation of EMSA, called relative quantification of TF interactions with DNA probes (qTFD), quantifies TF-DNA binding by measuring the amount of reporter-tagged TF that binds to DNA bound to a microtiter plate (Cai et al., 2023).

The ability of a candidate TFs to activate their proposed target genes can be assessed *in vivo* using the Transient Assay Reporting Genome-wide Effects of Transcription factors (TARGET) method (Bargmann et al., 2013). In TARGET, protoplasts are transfected with a construct encoding a translational fusion of the TF coding sequence and a glucocorticoid receptor tag, which enables translocation to the nucleus only upon dexamethasone treatment allowing the timing of delivery to be controlled. RNA is extracted from protoplasts, and the expression levels of target genes are evaluated by either by qPCR or transcriptome sequencing. Protoplasts can also be differentially treated with cycloheximide to block further protein synthesis, enabling direct and indirect effects to be distinguished. In an alternative method known as the transactivation assay, a construct coding for the constitutive expression of the TF of interest is co-expressed (typically in protoplasts) with a construct encoding a reporter gene controlled by the target promoter (Wehner et al., 2011). An increase or reduction in the reporter signal in the presence of TF relative to its absence indicates activation or repression, respectively.

In *Calendula officinalis*, faradiol fatty acid esters, and their intermediates, ψ-taraxasterol and faradiol, only accumulate in flowers (rays and discs) (Chapter 3.1). Differential gene expression analysis conducted by colleagues revealed that all five genes involved in this biosynthesis pathway (*CoTXSS*, *CoCYP1*, *CoCYP2*, *CoACT1* and *CoACT2*) show floral specific expression but that their temporal

expression through development is not synchronous. *CoTXSS*, *CoACT1* and *CoACT2* are most highly expressed in developing buds (S1) then decline, while *CoCYP1*, *CoCYP2* are most highly expressed in mature buds (S3) (Figure 5.2) (Golubova et al., 2025). Analysis of the genome revealed that these genes are not physically clustered (Golubova et al., 2025).

Figure 5.2 The developmental stages of *C. officinalis* flowers. Ray petals from the six developmental stages, from young buds (S1), developing buds (S2), mature buds (S3), junior flowers (S4), mature flowers (S5) to senescence flowers (S6), were sampled for RNA extraction. Figure adapted from Golubova et al., 2025.

To identify candidate regulators of *CoTXSS*, it was hypothesised that its floral-specific expression might be activated by the recruitment of the same set of TFs as *CoACT1* and *CoACT2* and potentially *CoCYP1* and *CoCYP2*.

5.2 Aims

In this chapter, I aimed to:
(1) Identify candidate TFs of genes involved in the biosynthesis of faradiol fatty acid esters in pot marigold flowers.
(2) Determine if these candidate TFs can activate the expression of *CoTXSS* and other faradiol fatty acid ester biosynthesis genes.

Samples of cDNA synthesised from RNA extracted from *C. officinalis* flowers were provided by Dr Daria Golubova (Earlham Institute). The transactivation assays in *Lactuca sativa* (lettuce) protoplasts were performed together with Dr Juanjuan Wang (University of Cambridge).

5.3 Contributions by others

All work described in this chapter was done by the author of this thesis.

5.4 Results

5.4.1 Identification of candidate TFBSs in the promoters of faradiol palmitate biosynthesis pathway genes.

To pursue the hypothesis that all pathway genes are regulated by a shared TF, the DNA sequences of the target genes were analysed to determine if there were any TFBSs common to their promoter regions. As chromatin accessibility data is lacking for *C. officinalis*, an arbitrary region of 1000 bp upstream of the TSSs plus the 5'UTR of each gene was analysed. These regions were named *pCoTXSS*, *pCoCYP1*, *pCoCYP2*, *pCoACT1* and *pCoACT2*.

The software package Find Individual Motif Occurrences (FIMO) was used to predict candidate TFBSs in the selected regions using PFMs of all known plant TFs documented in the TF binding profile database JASPAR 2024 (Grant et al., 2011; Rauluseviciute et al., 2024). This analysis predicted the occurrence of binding sites for R2R3-MYB TFs in all promoters (Figure 5.3).

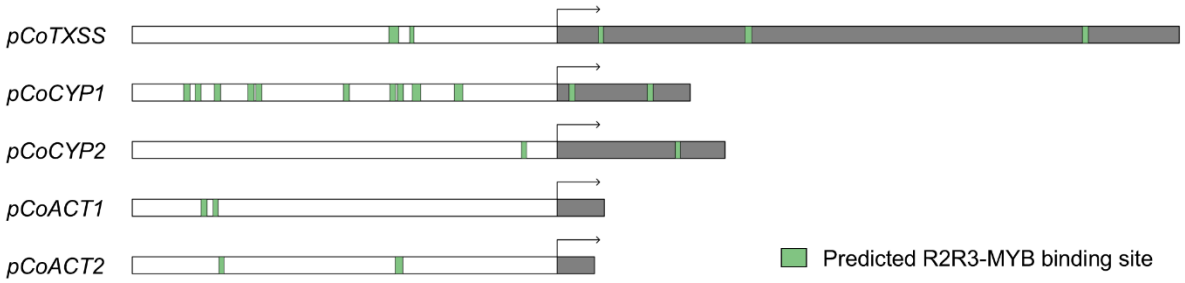


Figure 5.3 Predicted R2R3-MYB binding sites in *pCoTXSS*, *pCoCYP1*, *pCoCYP2*, *pCoACT1* and *pCoACT2*. Arrow represents transcription start site (TSS). White box represents sequences 1000 bp upstream of TSS and grey box represents 5' UTR.

To investigate if any *C. officinalis* R2R3-MYBs show similar expression patterns as the five target genes, co-expression analysis of the five faradiol fatty acid ester pathway genes was performed. Pearson's correlation analysis indicated that the expression levels of the pathway genes across different tissues are not correlated (Table 5.1). Thus, identification of specific candidate R2R3-MYB TFs via this approach was unfeasible.

	CoTXSS	*CoCYP1*	*CoCYP2*	*CoACT1*	*CoACT2*
CoTXSS	1	0.610273	0.662066	0.394915	0.685227
CoCYP1	0.610273	1	0.974653	0.779697	0.973304
CoCYP2	0.662066	0.974653	1	0.729096	0.988389
CoACT1	0.394915	0.779697	0.729096	1	0.712346
CoACT2	0.685227	0.973304	0.988389	0.712346	1

Table 5.1 Pearson's correlation of log2 normalised reads between different pathway genes.

As an alternative, R2R3-MYB TFs that were more highly expressed in *C. officinalis* flowers compared to leaves were identified. Genes that were found to have significantly higher expression in rays or disc tissue than in leave tissue (adjusted p-value<0.05, log2FC>1) were designated as flower-specific genes.

The translated peptide sequences of these genes were used to construct a BLASTp database. This database was queried using the DNA-binding domains of all *Arabidopsis thaliana* R2R3-MYB (AtMYB) peptide sequences. Sequences with an e-value lower than 0.01 were retained and designated as flower-specific *C. officinalis* R2R3-MYBs (CoMYBs). In total, 86 flower-specific CoMYBs were identified. CoMYBs with ray-to-leaf and disc-to-leaf log2FC values greater than 7 in the differential gene expression analysis were selected for further evaluation (Table 5.2).

Transcript	Ray-to-Leaf log2FC	Disc-to-Leaf log2FC	Reason for elimination
0503340.11	21.360	20.576	Low absolute expression
0549710.1	14.295	12.564	
0813200.1	13.782	11.653	
0872580.1	12.968	13.039	Fragmented
0423330.1	12.382	12.828	
0679380.1	12.243	12.427	
0684640.1	12.216	12.278	Shorter than 130 residues
0813330.1	11.733	12.323	
0589950.3	8.962	8.476	
0610460.1	8.707	8.854	Shorter than 130 residues
0380470.1	8.698	7.311	
0452040.1	7.654	4.457	No motif found in promoters
0663610.1	7.171	5.861	Motif not conserved

Table 5.2 *Calendula officinalis* R2R3-MYB genes *(CoMYBs)* identified as being more highly expressed in flowers. Grey boxes represent *CoMYB*s that were eliminated. Orange shading indicates sequences selected for experimental analysis.

Four flower-specific CoMYBs were eliminated due to low raw expression values (raw value less than 100), sequence completeness (incomplete open reading frame) and sequence length (shorter than 130 amino acids).

To predict the cognate TFBSs of the remaining ten flower-specific CoMYBs, the first 130 amino acids were predicted as DNA binding domains, as PROSITE indicates this as the DNA-binding domain of R2R3-MYBs (Sigrist et al., 2013). The putative DNA-binding domains were aligned with the DNA binding domains of AtMYBs present in the JASPAR 2024 database. This alignment was used to construct maximum-likelihood phylogenetic trees (Figure 5.4, 5.6, 5.8, 5.9). MYBs in the same

cluster have higher similarity in their MYB DNA-binding domains and are, therefore, likely bind to similar TFBSs.

CoMYBs encoded by transcripts 0548710.1, 0813200.1, 0423330.1, 0679380.1 and 0813330.1 were found to cluster with AtMYB24 and AtMYB57. Their DNA binding domains were found to be more similar to AtMYB24 than AtMYB57 (94.231% to 95.192% vs 82.692% to 83.654%) (Figure 5.4). Hence, those CoMYBs are likely orthologues of AtMYB24 and to bind to similar TFBSs as AtMYB24.

To determine if these CoMYBS are likely to regulate the target genes, I investigated if the promoters contained TFBSs that were likely to be bound by these proteins. This was done by using the AtMYB24 TFBS PFM (NNRTTHGGY), obtained from the JASPAR 24 database, to search for candidate motifs in the promoters of all five biosynthetic genes in Benchling software. Candidate binding sites were found in all five promoters (Figure 5.5). The CoMYB encoded by transcript 0549710.1 had the highest ray-to-leaf log2FC and was selected for further analysis and named CoMYB24.

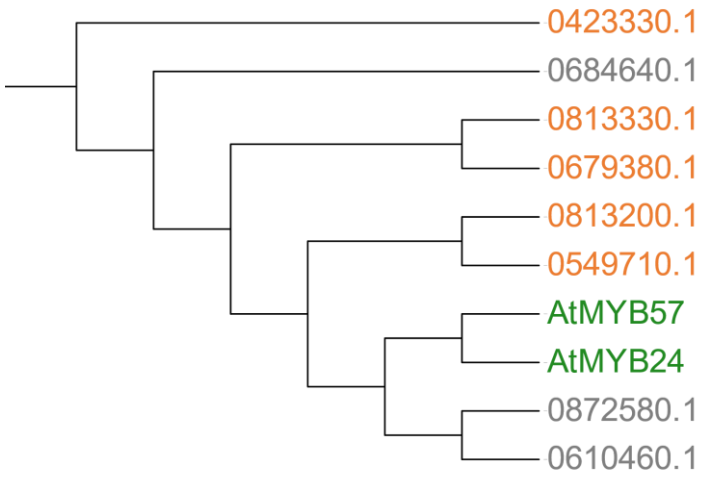


Figure 5.4 Phylogenetic analysis of the MYB domains of AtMYB24 and its putative *C. officinalis* orthologues. *A. thaliana* MYBs with known TFBS PFM are labelled in green. CoMYBs with ray-to-leaf log2FC above 7 are labelled in orange. CoMYBs eliminated due to short or fragmented sequences are labelled in grey.

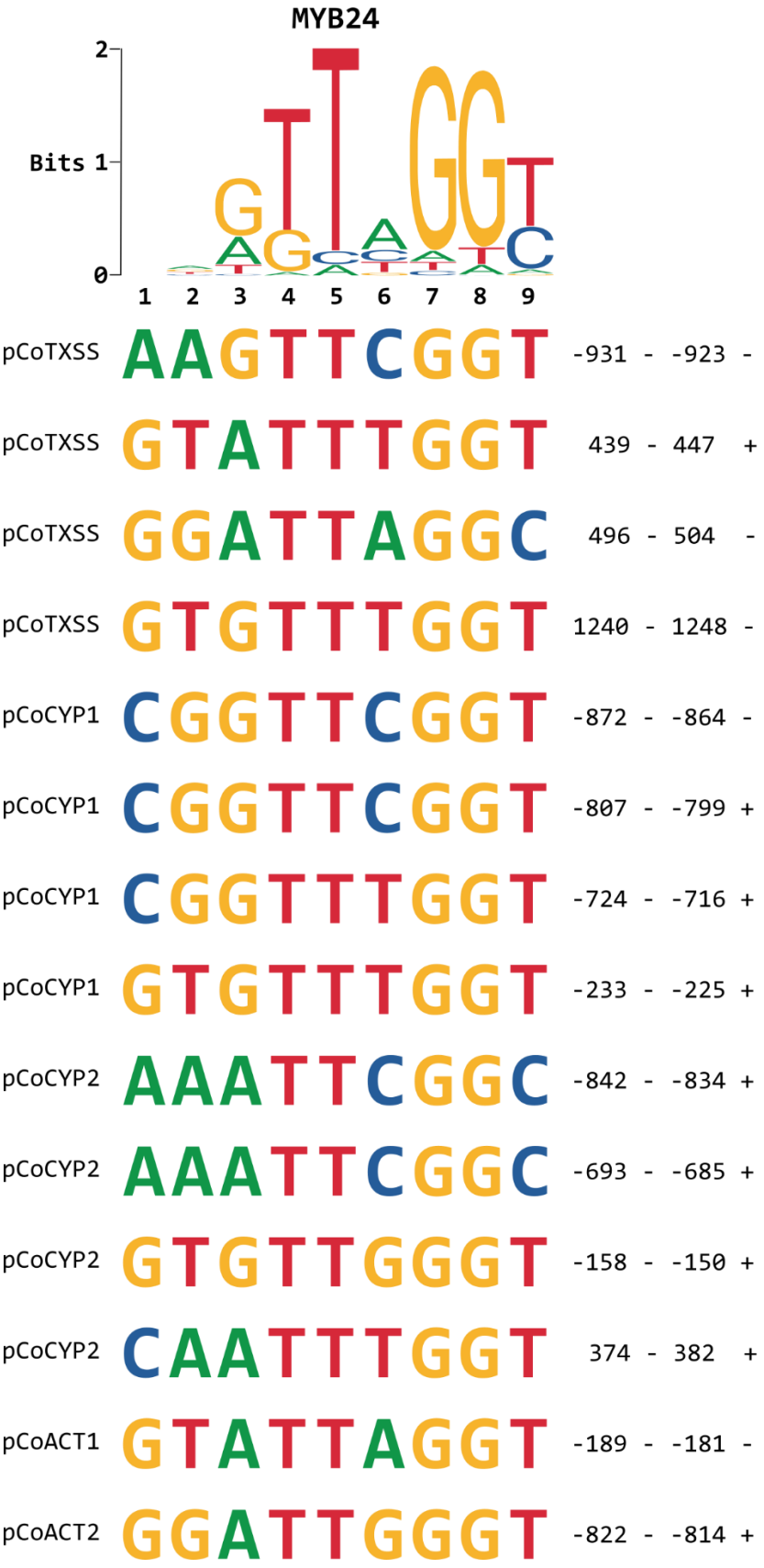


Figure 5.5 Logo plot of the position frequency matrix (PFM) of the AtMYB24 binding sites and potential binding sites in *pCoTXSS*, *pCoCYP1*, *pCoCYP2*, *pCoACT1* and *pCoACT2*. CoMYB24 is predicted to bind to the same motif. Number ranges after the binding motifs are the positions of the motifs on those predicted promoters.

The CoMYB encoded by transcript 0589950.3 was found to cluster with AtMYB3 and AtMYB4. Its DNA binding domain is more similar to AtMYB4 than AtMYB3 (97.155% vs 91.346%) (Figure 5.6). Sequence motifs similar to the AtMYB4 TFBS PFM (YAMCWAMY) were found in all promoters except *pCoACT2* (Figure 5.7). Thus, the CoMYB encoded by transcript 0589950.3 was selected for further analysis and named CoMYB4.

Similarly, the CoMYB encoded by transcript 0380470.1 clusters with AtMYB111, and its DNA binding domain shares 83.654% similarity with AtMYB111 (Figure 5.6). Sequence motifs similar to the AtMYB111 TFBS PFM (YAMCWAMY) were found in all promoters except *pCoACT2.* Thus, the CoMYB encoded by transcript 0380470.1 was selected for further analysis and named CoMYB111.

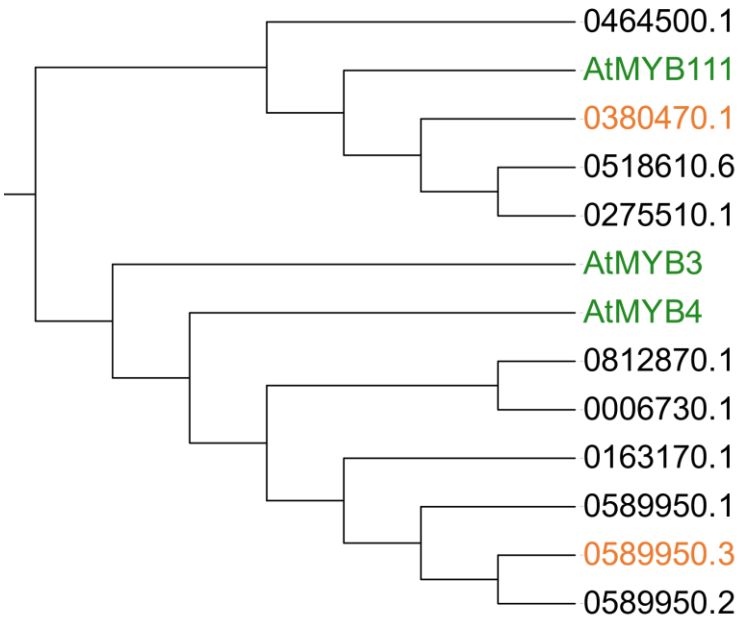


Figure 5.6 Phylogenetic analysis of the MYB domains of AtMYB4 and AtMYB111 and their putative *C. officinalis* orthologues. *A. thaliana* MYBs with known TFBS PFM are labelled in green. CoMYBs with ray-to-leaf log2FC above 7 are labelled in orange. Other CoMYBs are labelled in black.

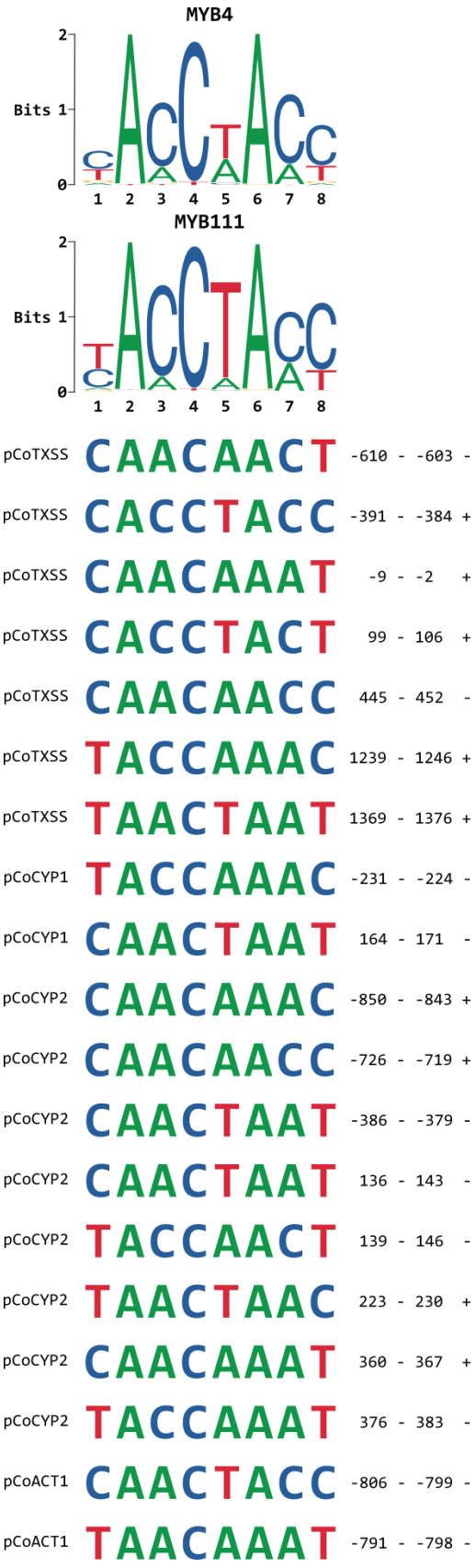


Figure 5.7 Logo plots of the position frequency matrix (PFM) of AtMYB4 and AtMYB111 binding sites and their potential binding sites in *pCoTXSS*, *pCoCYP1*, *pCoCYP2* and *pCoACT1*. CoMYB4 and CoMYB111 are predicted to bind to the same motifs as AtMYB4 and AtMYB111. Number ranges after the binding motifs are the positions of the motifs on those predicted promoters.

The CoMYB encoded by transcript 0452040.1 was found to cluster with AtMYB27 and its DNA binding domain shares 68.269% similarity with AtMYB27 (Figure 5.8). However, sequence motifs similar to the AtMYB27 TFBS PFM (NNNWWUTTAGGTNNN) were not found in any of the five promoters. Therefore, this CoMYB is unlikely to be involved in the regulation of faradiol fatty acid ester biosynthesis.

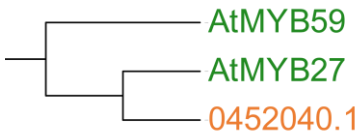


Figure 5.8 Phylogenetic analysis of the MYB domains of AtMYB27 and its putative *C. officinalis* orthologues. *A. thaliana* MYBs with known TFBS PFM are labelled in green. CoMYBs with ray-to-leaf log2FC above 7 are labelled in orange.

The CoMYB encoded by 0663610.1 clusters with AtMYB80 and AtRAX3. However, its DNA binding domain shared only 45.370% sequence similarity with AtMYB80 and 47.706% to AtRAX3 (Figure 5.9). Hence, it is unlikely that they are able to bind to the same TFBSs, and it was not possible to identify candidate binding sites in the target gene promoters.

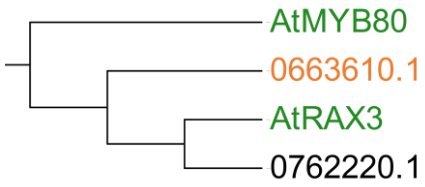


Figure 5.9 Phylogenetic analysis of the MYB domains of AtMYB80 and AtRAX3 and itheir putative *C. officinalis* orthologues. *A. thaliana* MYBs with known TFBS PFM are labelled in green. CoMYBs with ray-to-leaf log2FC above 7 are labelled in orange. Other CoMYBs are labelled in black.

To summarise, by combining differential gene expression analysis, phylogenetic analysis and motif searching, three R2R3-MYB TFs were identified as candidate regulators of *CoTXSS* expression and faradiol fatty acid ester biosynthesis. Of these, CoMYB24 was predicted to regulate all five genes, while CoMYB4 and CoMYB111 were predicted to regulate all genes except *CoACT2*.

5.4.2 Validation of the flower specific expression of candidate CoMYBs

Genes involved in faradiol fatty acid ester biosynthesis were identified through differential gene expression between flower and leaf transcriptomes in plants at mature flower (S5) stage (Figure 5.2). Subsequently, gene expression analysis of these genes through floral development using qPCR found that *CoTXSS*, *CoACT1* and *CoACT2* are most highly expressed in young buds (S1) and *CoCYP1* and *CoCYP2* are most highly expressed in mature buds (S3) and moderately expressed

during S1 (Golubova et al., 2025). The expression levels of all five genes were found to have declined by S5.

Therefore, I next confirmed that the genes encoding candidate CoMYBs are significantly more highly expressed in S1 stage relative to leaves. To investigate this, qPCR primers for *CoMYB24*, *CoMYB4* and *CoMYB111* were designed and evaluated for efficiency. Primers CoSAND2 PF and CoSAND2 PR for housekeeping gene *CoSAND* were used to calculate relative expression levels (Chapter 2.3.9).

*CoMYB*s and *CoSAND* amplicons were amplified from S1 ray total cDNA, and purified amplicons were used in a dilution series from 10^{-3} to 10^{-8} ng/µL to determine primer efficiency (Figure 5.10).

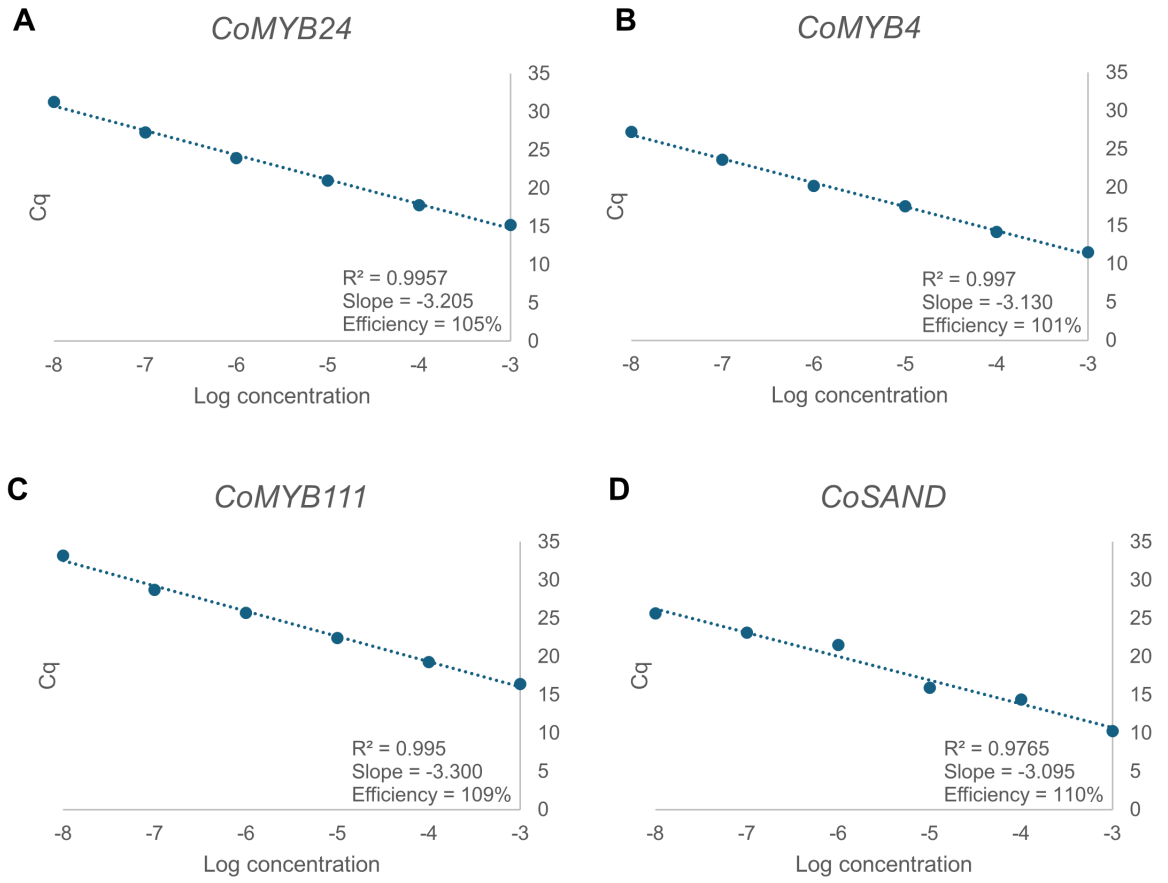


Figure 5.10 Correlation between Cq values and concentrations of the DNA segments during the qPCR reaction for different *C. officinalis* genes. Cq values relative to log concentrations of DNA segments from *CoMYB24* (A), *CoMYB4* (B), *CoMYB111* (C) and *CoSAND* (D) were plotted. Best fitted lines were plotted between Cq values and logarithmic of the concentrations and R square, slopes and primer efficiencies were calculated.

The efficiencies of all primer pairs were between 100% and 110%, which indicates that they are moderately efficient and differences in their amplification efficiencies are negligible. These primer pairs were used to evaluate the expression levels of *CoMYB24*, *CoMYB4* and *CoMYB111* in S1 rays, S1 discs and leaves (Figure 5.11).

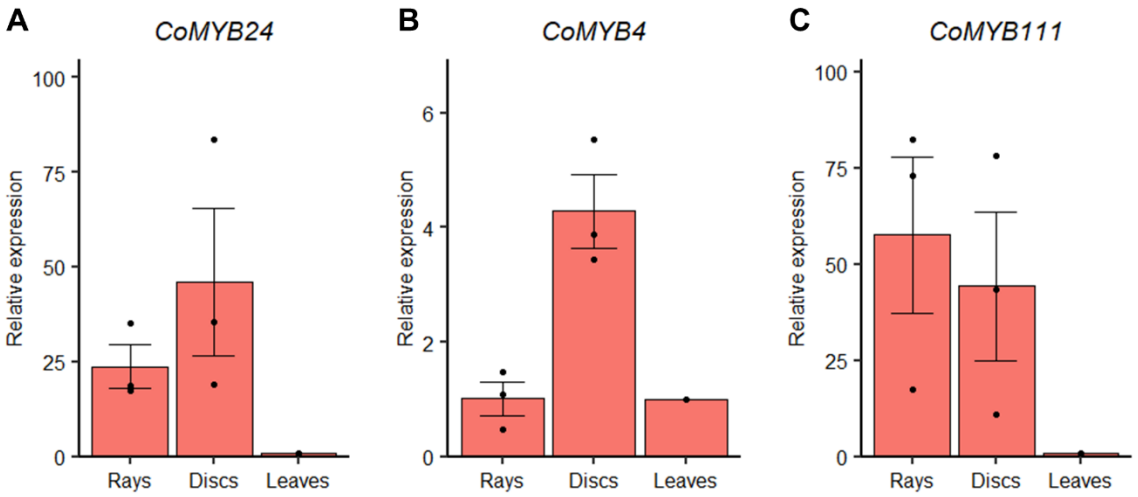


Figure 5.11 Relative expression levels of *CoMYB24* (A), *CoMYB4* (B) and *CoMYB111* (C) in S1 rays, S1 discs and leaves. The expression levels of *CoMYB24*, *CoMYB4* and *CoMYB111* were quantified relative to housekeeping gene *CoSAND* and normalised against leaves.

The expression of *CoMYB24* was only detected in flowers, with the highest expression in discs. Similarly, *CoMYB111* was only expressed in flowers, but with similar expression in rays and discs. Both CoMYB24 and CoMYB111 were selected for further evaluation.

In contrast, although the expression of *CoMYB4* was highest in discs, it showed similar levels of expression level in rays and leaves, which was inconsistent with the expression pattern of the target genes.

5.4.3 Transactivation assays in *L. sativa* protoplasts

To investigate if CoMYB24 and CoMYB111 were able to activate expression of *pCoTXSS*, *pCoCYP1*, *pCoCYP2*, *pCoACT1* and *pCoACT2,* transactivation assays were performed. First, CoMYB24 and CoMYB111 were cloned into plant expression vectors under the control of the strong constitutive promoter CaMV35s (Chapter 2.3.2). The selected promoter regions were amplified from genomic DNA and used to assemble expression constructs with the coding sequences of the nanoluciferase reporter (LucN) (Chapter 2.3.2). One exception was *pCoCYP1* for which I was unable to amplify a sequence for the correct allele.

Transactivation assays were performed in protoplasts of *L. sativa*, an Asteraceae species for which protoplast isolation and transfection are well-established (Chupeau et al., 1989). Aliquots of 1 x 10^5 protoplasts were transfected with three constructs (Chapter 2.5.5): a construct constitutively expressing CoMYB (or NLS:YFP as a negative control), a construct with the promoter of interest fused to LucN, and a construct constitutively expressing a calibrator (LucF) (Figure 5.12). Transfected protoplasts were harvested 24 hours after transfection (Chapter 2.5.5). Protoplasts expressing NLS:YFP were visualised using epifluorescence microscopy to assess transfection efficiency (cutoff: 70%). Total proteins were extracted from protoplasts

and LucF and LucN luminescence was quantified using the NanoGlo dual luciferase kit.

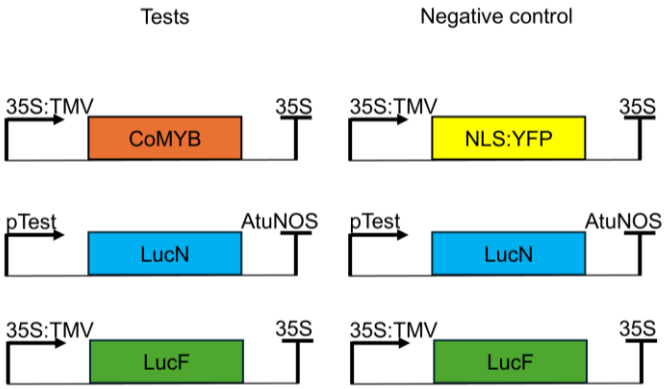


Figure 5.12 Genetic constructs used in *L. sativa* protoplast transactivation assay.

To determine if the CoMYB being tested effected the expression of the target promoter, the ratio of LucN/LucF in protoplasts in which *CoMYB* was co-expressed was compared to the negative control in which YFP was co-expressed. From the presence of TFBSs identified in the previous section, CoMYB24 was expected to increase the expression of *pCoTXSS, pCoCYP2, pCoACT1* and pCoACT2, and CoMYB111 was predicted to increase expression of *pCoTXSS, pCoCYP2* and *pCoACT1.*

Co-expression of CoMYB24 significantly increased the expression of *pCoTXSS*:LucN, *pCoCYP2*:LucN and *pCoACT1*:LucN. In contrast, although the expression of *pCoACT2*:LucN also increased, this was not significant (Figure 5.13). CoMYB111 significantly increased expression of *pCoACT1*:LucN but did not significantly increase the expression of *pCoTXSS*:LucN, *pCoCYP2*:LucN or *pCoACT2*:LucN (Figure 5.14).

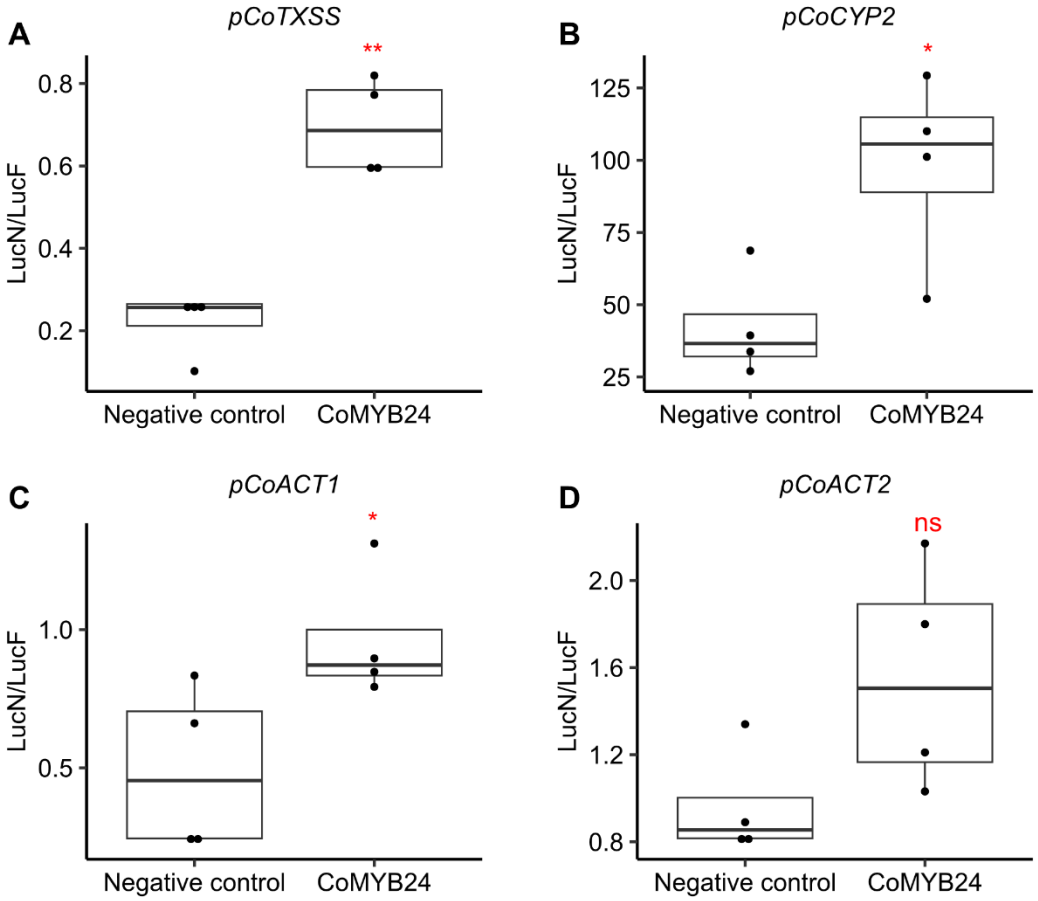


Figure 5.13 Protoplast transactivation assays with CoMYB24. The ratio of LucN/LucF was measured in protoplasts co-transfected with CaMV35s:LucF, either CaMV35s:YFP (control) or CaMV35s:CoMYB24, and *pCoTXSS*:LucN (A), *pCoCYP2*:LucN (B), *pCoACT1*:LucN (C) and *pCoACT2*:LucN (D). Statistical significance in LucN/LucF ratio was analysed using a Dunnett's test; ns, $p>0.05$; *, $p<0.05$; **, $p<0.01$; n=4.

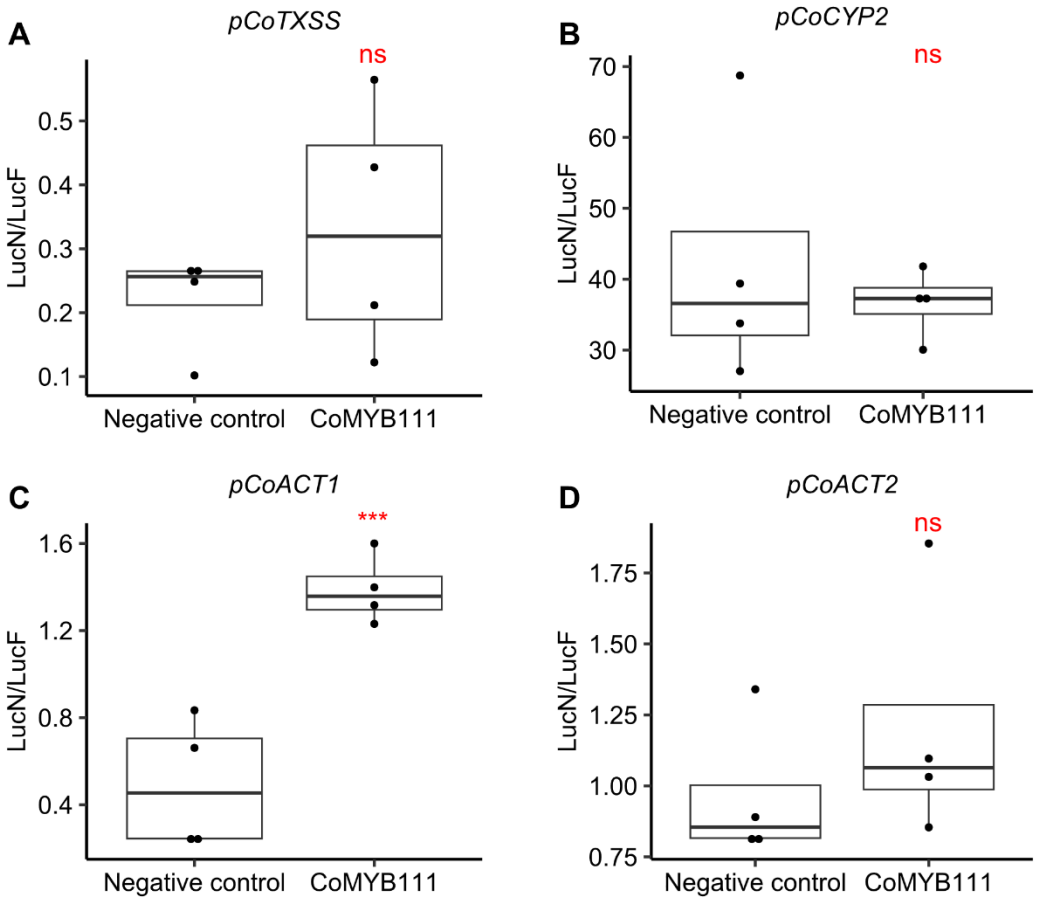


Figure 5.14 Protoplast transactivation assays with CoMYB111 The ratio of LucN/LucF was measured in protoplasts co-transfected with CaMV35s:LucF, either CaMV35s:YFP or CaMV35s:CoMYB111, and *pCoTXSS*:LucN (A), *pCoCYP2*:LucN (B), *pCoACT1*:LucN (C) and *pCoACT2*:LucN (D). Statistical significance in LucN/LucF ratio was analysed using a Dunnett's test; ns, p>0.05; *, p<0.05; **, p<0.01; n=4.

5.5 Discussion

5.5.1 Promoters of genes involved in the biosynthesis of faradiol fatty acid ester biosynthesis contain candidate binding sites for R2R3-MYB-family transcription factors

In *C. officinalis,* the five genes involved in the biosynthesis of faradiol fatty acid ester biosynthesis are only expressed in flowers. Previous analysis of the genome has found that these genes are not clustered (Golubova et al., 2025). I hypothesised that the floral expression of some or all of these genes is controlled by the same set of TFs. Studies to identify *cis*-regulatory elements and TFBSs have not been conducted in this species. In previous studies conducted in other species, co-expression analyses have been used to identify TFs regulating triterpenoid biosynthesis pathway genes, exemplified by discovery of bHLH-type TFs activating oleanane-type triterpenoid biosynthesis in *Medicago truncatula* (Mertens et al., 2016). This

approach was not employed for faradiol fatty acid ester biosynthesis as the pathway genes are not co-expressed over time (Golubova et al., 2025) and across tissues (Table 5.1). This may mean that different TFs contribute to the regulation of expression through flower development, however, as expression is limited to floral tissues and is higher in rays than discs, they may have common cis regulatory elements.

As DNA binding domains of homologous TFs from different species have been shown to be highly conserved (Zenker et al., 2025), it is possible to infer candidate binding sites for TFs using data from other species. For example, Bian et al. used Arabidopsis TFBS data to identify binding sites for tomato orthologues of TFs involved in nitrate responses (Bian et al., 2025). Using a combination of motif identification, gene expression analysis and identification of Arabidopsis orthologues, candidate TFBSs for R2R3-MYBs were identified in the promoter regions of *C. officinalis* genes involved in faradiol fatty acid ester biosynthesis. R2R3-MYB TFs have been shown to play important roles in plant secondary metabolism regulation, modulating the biosynthesis of benzenoids, phenylpropanoids, terpenoids, and glucosinolates (Wu et al., 2022). In the case of terpenoids, R2R3-MYBs were observed to enhance the expression of monoterpene synthase genes in *Lilium* 'Siberia' flowers and sesquiterpene synthase genes in *A. thaliana* flowers, which are involved in producing volatile terpenoids for pollinator attraction (Guo et al., 2023; Reeves et al., 2012). R2R3-MYBs were also described to activate expression of OSC genes in *M. domestica* and *C. sinensis* (Falginella et al., 2021; Zhang et al., 2020).

5.5.2 CoMYB24 may positively regulate the biosynthesis of faradiol fatty acid esters

A protoplast transactivation assay indicated that the floral-specific R2R3-MYB-family transcription factor, CoMYB24, can increase expression of *pCoTXSS*, *pCoCYP2* and *pCoACT1* but not *pCoACT2* (Figure 5.13). In contrast, although candidate binding sites for CoMYB111 were identified in *pCoTXSS*, *pCoCYP2* and *pCoACT1*, it was only able to increase the expression of *pCoACT1* in transactivation assays.

Biosynthesis of faradiol fatty acid esters in *N. benthamiana* was achieved by co-expression of CoTXSS, plus either CoCYP1 or CoCYP2, and either CoACT1 or CoACT2 (Golubova et al., 2025). Thus, CoMYB24 is a potential positive regulator of the pathway. As the expression of these genes is not tightly co-regulated through floral development, it is likely that additional TFs contribute to the expression of each gene. For practical reasons, this analysis was limited to an arbitrary 1 kB region; chromosomal occupancy information acquired from techniques such as Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq) could guide the identification of more candidate TFBSs (Grandi et al., 2022).

Inconsistency between the presence of a predicted TFBS and the transactivation assay experimental results can be attributed to multiple factors. A simple reason is that CoMYBs might need to interact with other TFs that were not present in the lettuce protoplasts. For example, MYBs are known to interact with bHLH and WD40 TFs to form regulatory complexes and activate transcription (Chezem and Clay, 2016). In *Vitis vinifera*, MYB24 was found to interact with MYC2 to mediate methyl jasmonate-induced upregulation of terpene synthase genes (Zhang et al., 2025).

The limitation of transactivation assays is that they are unable to provide information about which sites a TF is interacting with. It is possible that the activation is indirect with the TF activating expression of another regulator, which in turn regulates the expression of the target promoter. Conducting transactivation assays in heterologous species, reduces this probability, however, it cannot be ruled out. In the future, direct interactions could be investigated using electrophoretic mobility shift assay (EMSA) or relative quantification of TF interactions with DNA probes (qTFD) to validate physical interactions between TFs and their candidate TFBSs (Cai et al., 2023; Hellman and Fried, 2007). Another method to confirm direct activation of target genes is the TARGET assay on *C. officinalis* protoplasts (Bargmann et al., 2013). However, this would require the ability to prepare and transfect *C. officinalis* protoplasts. While the development of protoplast protocols for leaf tissues is reasonably straightforward, the absence of expression in leaf tissues means that they may be regulated at the chromatin level. Thus, TARGET assays may need to be conducted in protoplasts extracted from *C. officinalis* ray petals, which is technically challenging.

A conclusive but also technically challenging route to validating the role of CoMYBs in metabolite accumulation would be to knock down, mutate (knock out), or overexpress the *CoMYB* genes in *C. officinalis*. These methods require the development of transgenesis protocols or robust and systemic virus-induced gene silencing method for *C. officinalis*, which are currently lacking for this species.

5.6 Conclusion

In conclusion, using sequence analysis, phylogenetic clustering with orthologous genes, and gene expression analysis, two R2R3-MYB-family TFs, CoMYB24 and CoMYB111, were predicted to regulate the expression of faradiol fatty acid ester biosynthesis genes in *C. officinalis*. Experimental analyses in *L. sativa* protoplasts found that these TFs were able to activate expression of at least one pathway gene with CoMYB24 able to increase the expression of genes involved in all pathway steps. Further work is required to determine the contribution of these TF to their overall expression levels of pathway genes *C. officinalis*.

6. Chapter 6 Towards membrane dissociated plant OSCs

6.1 Introduction

Synthetic biology and metabolic engineering have the potential to enable large scale production of industrially and pharmaceutically valuable molecules in heterologous hosts (Nielsen and Keasling, 2016). To date, most metabolic engineering effort has been focussed on microbes, exemplified by the successful production of butanediol (precursor of the synthetic elastic fibre, spandex) in *E. coli* and artemisinic acid (precursor of the anti-malarial drug, artemisinin) in *Saccharomyces cerevisiae* (Ro et al., 2006; Yim et al., 2011). However, metabolic engineering in microbes faces several challenges: For example, the solubility and function of membrane associated enzymes, such as OSCs and cytochrome P450s, involved in the biosynthesis of valuable compounds in *E. coli* remains challenging (Schlegel et al., 2010). Moreover, questions on the sustainability of some feedstocks used for microbial cultivation have been raised (Lips, 2021).

Photosynthetic chassis, including microalgae and plants, provide potential alternatives. These have the advantages of precursors derived from photosynthesis, different subcellular compartments, and the ability to fold and express numerous classes of enzymes found in plants, animals and fungi (Golubova et al., 2024). *Nicotiana benthamiana*, a close relative of tobacco (*Nicotiana tabacum*), is indigenous to Australia and is increasingly being used as a chassis for pathway elucidation and small molecule production (Golubova et al., 2024). The widely used 'LAB' strain is an extremophile containing a loss-of-function mutation in RNA-dependent RNA polymerase 1 gene (*Rdr1*), rendering it susceptible to viruses and infection by *Agrobacterium tumefaciens* (Bally et al., 2015). Agroinfiltration of *N. benthamiana* leaves has enabled milligram- to gram-scale production of the triterpene β-amyrin and its derivatives (Reed et al., 2017; Stephenson et al., 2018). Despite some success in the development of *N. benthamiana* as a production chassis for therapeutic proteins (Lomonossoff and D'Aoust, 2016), it has yet to be adopted for the commercial production of small molecules. However, it has potential for producing molecules that are difficult to achieve in microbial chassis. These include triterpenoids, for which biosynthesis often requires multiple membrane-associated enzymes, as discussed in previous chapters.

To boost the yield of target molecules, many studies have over-expressed enzymes involved in the production of precursors. For example, overexpression of a feedback insensitive 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase (HMGR) enhanced the yield of β-amyrin (Reed et al., 2017). Similarly, overexpression of deoxyxylulose 5-phosphate synthase and hydroxymethylbutenyl diphosphate reductase (HDR) increased flux towards the MEP pathway and improved the yields of diterpenes casbene and jolkinol C (Forestier et al., 2021). One problem that has been frequently

reported when producing small molecules in *N. benthamiana* is derivatisation by endogenous enzymes. For example, when expressing the pathway for parthenolide biosynthesis, derivatives conjugated with cysteine and glutathione were detected (Liu et al., 2014). In other studies, unwanted glycosylation of geraniol, artemisinic acid, faradiol and arnidiol (Dong et al., 2016; Golubova et al., 2025; van Herpen et al., 2010), and hyperglycosylation of saponins (Khakimov et al., 2015) were observed following the expression of their biosynthesis pathways in *N. benthamiana* leaves. One of the most reported derivatisations is the addition of single or multiple hexose sugars, predicted to be catalysed by promiscuous UDP-glycosyltransferases (UGTs) (Lu et al., 2023). Dudley et al. demonstrated that the engineering of *N. benthamiana* lines with loss-of-function mutations in UGTs reduced the accumulation of nepetalactol and geraniol derivatives (Dudley et al., 2022).

A potential approach to avoid derivatisation by UGTs, which are mainly cytosolic, is to relocate metabolic pathways to chloroplasts. Chloroplasts are cyanobacteria-derived organelles where photosynthesis takes place. Chloroplasts also produce the basic precursors of terpenoid biosynthesis (isopentenyl diphosphate (IPP) and dimethyallyl diphosphate (DMAPP)) via the methylerythritol phosphate (MEP) pathway and are the site of mono- and di-terpenoid production in plants. In a metabolic engineering study, when dhurrin biosynthesis was relocated to chloroplasts, a decrease in the accumulation of derivatives and a five-fold increase in yield was observed (Henriques de Jesus et al., 2017).

As previously discussed, the first committed step of triterpene biosynthesis is the cyclisation of 2,3-oxidosqualene into triterpene scaffolds by OSCs. OSCs are generally monotopic integral membrane proteins that contain exposed hydrophobic segments, which locate them to the cytosolic face of the endoplasmic reticulum (ER). To date, no crystal structures of plant OSCs have been obtained. However, the crystal structure of human lanosterol synthase (1W6K), a homologue of plant OSCs, indicates that three helices are bound to the nonpolar octyl tails of detergent β-OG, which allows the enzyme to be inserted into the membrane (Thoma et al., 2004) (Figure 6.1). Similarly, a cryo-EM structure of *Tripterygium wilfordii* friedelin synthase (8J5Z) indicates that one loop and two helices are inserted into the membrane (Luo et al., 2023) (Figure 6.1).

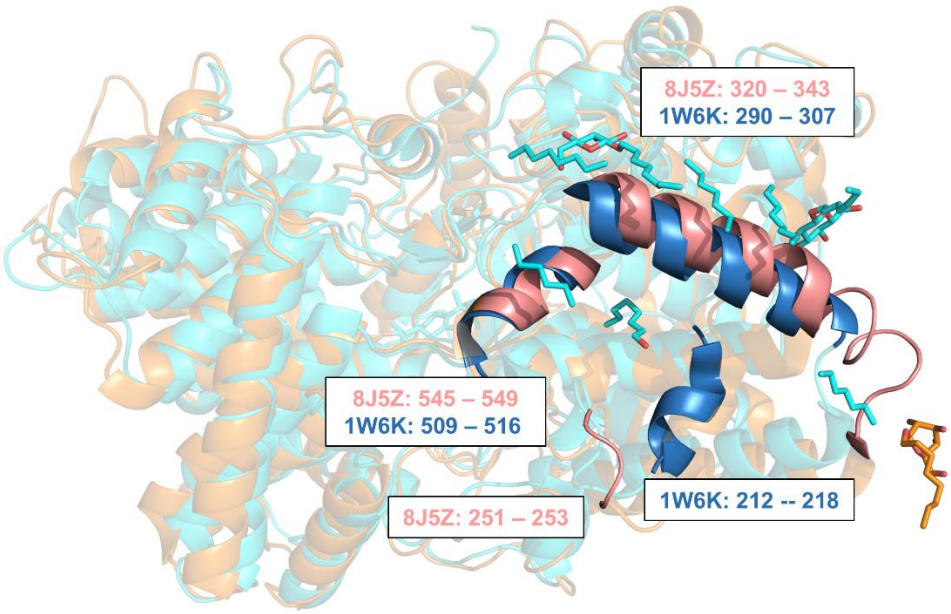


Figure 6.1 Structures of human lanosterol synthase (1W6K) and *Tripterygium wilfordii* friedelin synthase (8J5Z). Human lanosterol synthase is in cyan and *T. wilfordii* friedelin synthase is in orange. Their membrane embedded helicies are coloured dark blue and rose respectively. Detergent molecules used in resolving their structures are shown as cyan and orange sticks, respectively. Structural alignment RMSD = 1.646.

Expressing OSCs in prokaryotes such as *E. coli* or in prokaryote-derived eukaryotic compartments such as chloroplasts typically results in insolubility and/or instability due to insertion into membranes with different lipid compositions (Carpenter et al., 2008). Previously, unsuccessful attempts have been made to purify plant OSCs following heterologous expression in both *E. coli* and yeasts (Dokarry, 2010; Pfalzgraf, 2022). Interestingly, however, *Methylococcus capsulatus* lanosterol synthase (McLS), was found to be soluble. This OSC was characterised and purified following heterologous expression in *E. coli* (Lamb et al., 2007). McLS is structurally similar to human lanosterol (RMSD = 0.931) and has been proposed to interact and associate with a membrane-bound squalene monooxygenase in order to access the 2,3-oxidosqualene substrate. The discovery of McLS indicates that the structural topology of OSCs is amenable to membrane dissociation. Thus, it can be hypothesised that it may be possible to engineer soluble versions of membrane-bound OSCs. This might facilitate the production or triterpenes in prokaryotic production chassis such as *E. coli* or in plant chloroplasts.

Rational protein engineering, where knowledge of chemistry guides protein engineering, has been employed to alter the product specificity and enhance the thermostability of terpene synthases (Diaz et al., 2011; Johnson and Allemann, 2025). There are also rational design rules to enhance protein solubility. When designing oligomeric α-helical coiled coils, polar amino acids E, Q and K were often put on the solvent exposed surface to enhance solubility without breaking the helices (Woolfson, 2023). In recent years, deep learning-based methods with impressive

success rates have revolutionised the field of protein design and engineering. Dauparas et al. developed ProteinMPNN, a robust deep learning-based method which rescued previously failed designs and improved the success rate of designing functional proteins (Dauparas et al., 2022). ProteinMPNN has been employed to enhance the stability and solubility of diverse protein candidates, including myoglobin, tobacco etch virus protease and non-haem iron enzyme (King et al., 2025; Sumida et al., 2024). A variation of ProteinMPNN was made by retraining it with datasets composed of exclusively soluble proteins and named as SolubleMPNN (Goverde et al., 2024). This was able to generate soluble homologues of several membrane proteins, including claudin, rhomboid and G protein-coupled receptor.

In this chapter, I work towards the goal of expressing OSCs in bacteria and chloroplasts. I do this by exploring the use of rational and SolubleMPNN-guided protein design methods to engineer a soluble (membrane detached) *Calendula officinalis* taraxasterol synthase (CoTXSS).

6.2 Aims

In this chapter, I aim to:
(1) Use rational and computational protein design methods to enhance the solubility of CoTXSS.
(2) Examine the subcellular localisation and function of redesigned CoTXSS candidates in *N. benthamiana*.
(3) Examine the solubility of redesigned CoTXSS candidates in *E. coli*.

6.3 Work contributed by others

All work described in this chapter was done by the author of this thesis. Advice on rational and computational protein design was received from Prof Derek N. Woolfson and Mr Rokas Petrenas (University of Bristol). Dr Connor Tansley and Dr Facundo Romani (University of Cambridge) provided advice and training in the selection of appropriate settings for confocal microscopy.

6.4 Results

6.4.1 CoTXSS localises to the ER in *N. benthamiana*

Although the notion that plant OSCs are monotopic integral membrane proteins located at the ER membrane is widely accepted, few studies have experimentally confirmed their subcellular localisation. Thus, to confirm the subcellular localisation of CoTXSS, the coding sequence was translationally fused to a linker and a YFP and assembled into a plant expression vector regulated by the CaMV35s promoter and terminator (Chapter 2.3.2). This construct was agroinfiltrated into *N. benthamiana*

together with strains expressing the p19 suppressor of silencing as well as mCherry translationally fused to signal peptide OsRAmy3A (MGKQMAALCGFLLVALLWLTPDVASG) and an ER retention signal (HDEL), hereafter referred to as mCherryER (Figure 6.2A) (Chapter 2.5.2). OsRAmy3A is an *Oryza sativa* α-amylase containing an N-terminal signal peptide that directs it into the secretory pathway (Sutliff et al., 1991). The function of this signal peptide has previously been characterised in *N. benthamiana* (Engler et al., 2014). The HDEL C-terminal extension was previously characterised as being sufficient for ER retention of secretory proteins in *N. benthamiana* (Lee et al., 2018). Hence, mCherryER was used as an ER marker.

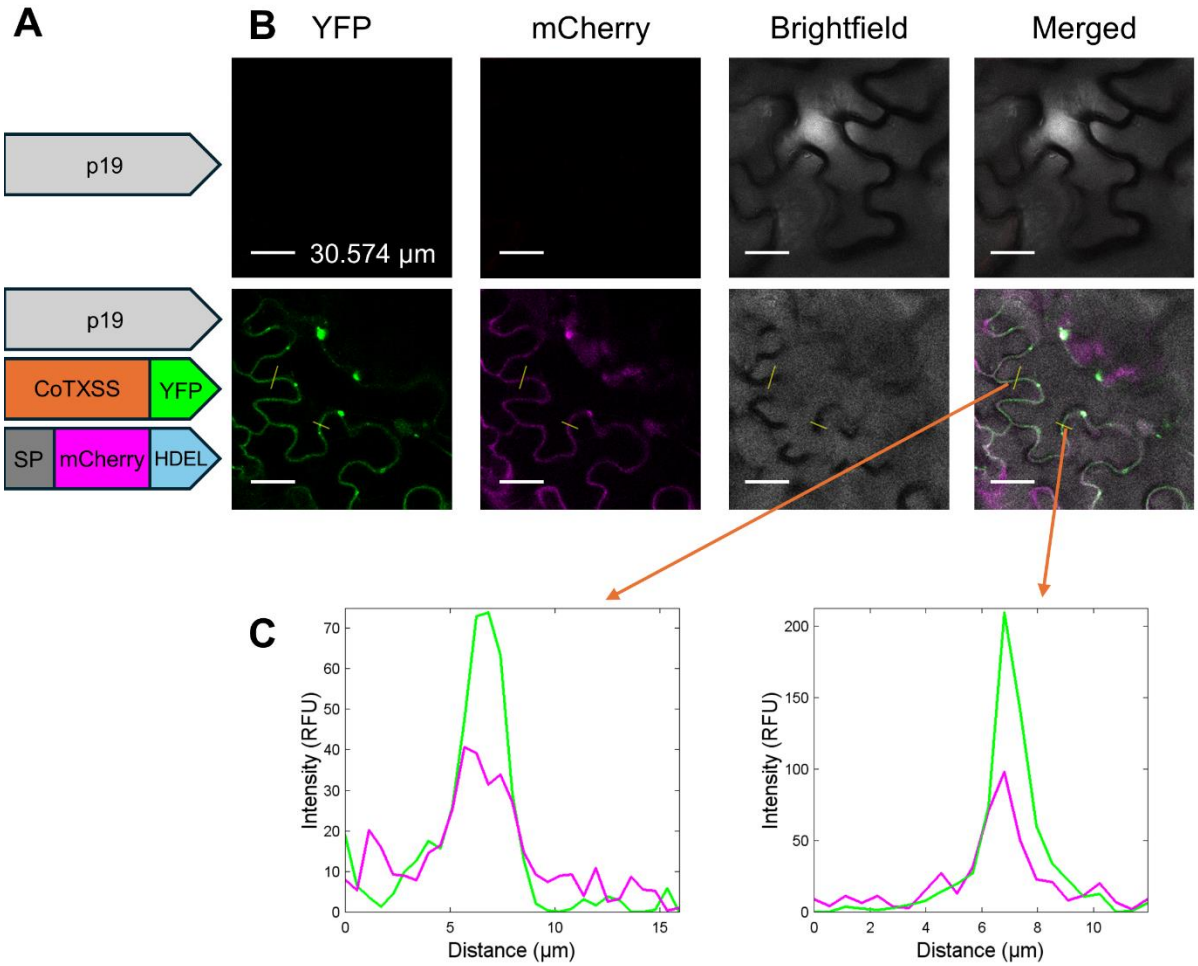


Figure 6.2 Confocal micrographs showing co-localisation of CoTXSS:YFP and mCherryER (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

Five days post-infiltration, infiltrated leaves were visualised using confocal microscopy (Figure 6.2B). This revealed the co-localisation of the YFP and mCherry signals. As the ER and cytosol are difficult to resolve visually, pixel intensity profiles were performed along two cross-sections (Figure 6.2C). The peaks of the YFP and

mCherry channels overlap, indicating co-localisation. This experiment was performed in triplicate and similar patterns were observed in all replicates. Thus, as expected, CoTXSS is likely localised to the ER.

6.4.2 McLS localises to the cytosol in *N. benthamiana*

Next, the subcellular localisation of McLS, which is soluble when expressed in *E. coli*, was investigated in *N. benthamiana*. To do this, the McLS coding sequence was cloned in a translational fusion with YFP into a plant expression vector regulated by the CaMV35s promoter and terminator (Figure 6.3A). This construct was co-infiltrated into in *N. benthamiana* leaves together with constructs expressing mCherryER and p19. Five days post-infiltration, the leaves were examined by confocal microscopy (Figure 6.3B).

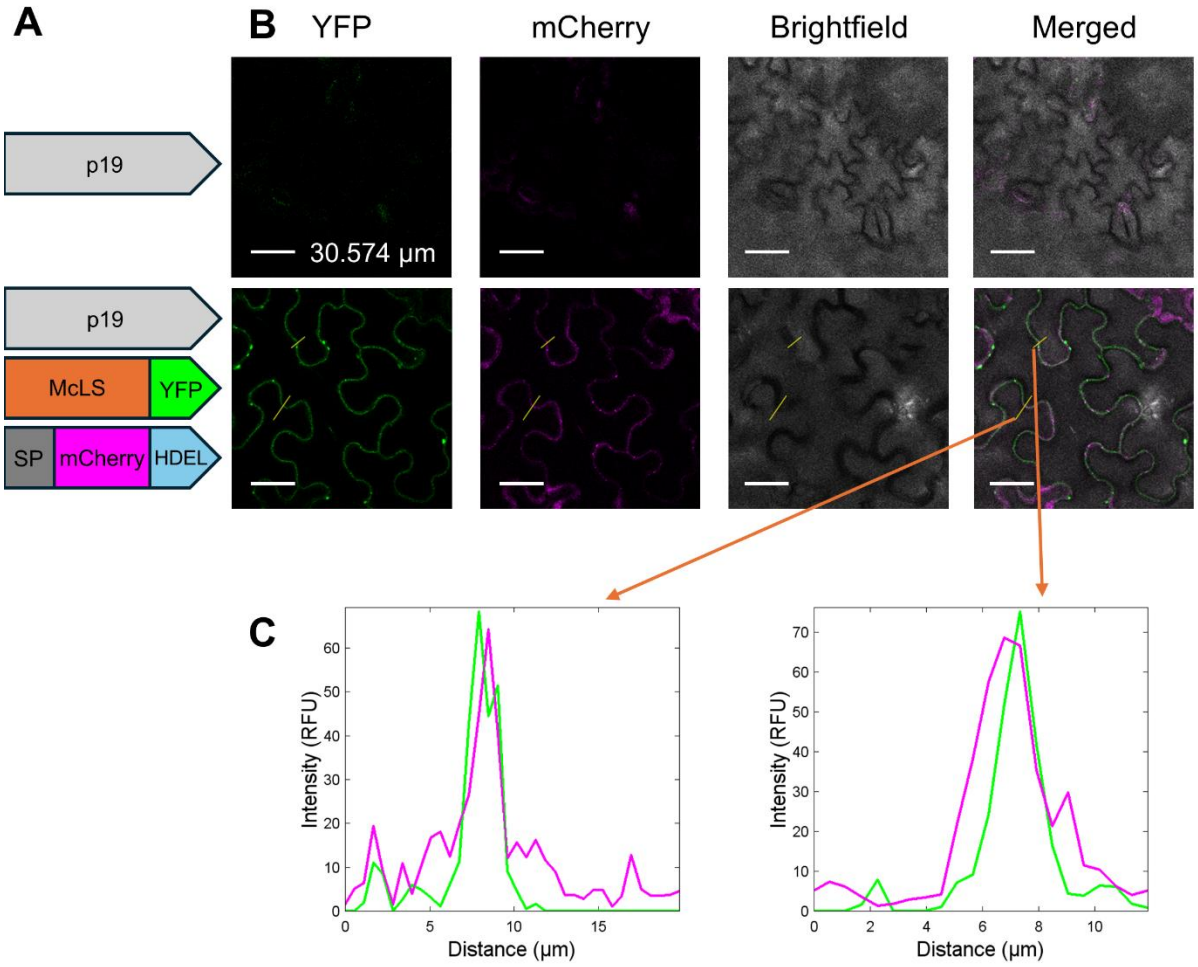


Figure 6.3 Confocal micrographs of McLS:YFP in *N. benthamiana* leaves. (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

The YFP and mCherry signals were observed to be adjacent rather than overlapping. As previously, this is difficult to visualise and pixel intensity profiles of YFP and mCherry were performed along two cross-sections (Figure 6.3C). In these profiles, the YFP and mCherry peaks are offset, which indicates that McLS is not localised to the ER. Thus, McLS is likely to be soluble and located in the cytosol when expressed in *N. benthamiana*. This experiment was performed in triplicate and similar patterns were observed in all replicates.

6.4.3 Computational redesign of membrane-embedded helices

In human lanosterol synthase, three helices are embedded into the ER membrane. These helices are enriched in hydrophobic residues (Figure 6.1). As CoTXSS is a homologue of human lanosterol synthase and shares high structural similarity (RMSD = 0.778), I predicted that the equivalent helices in CoTXSS are also membrane-embedded (Figure 6.3).

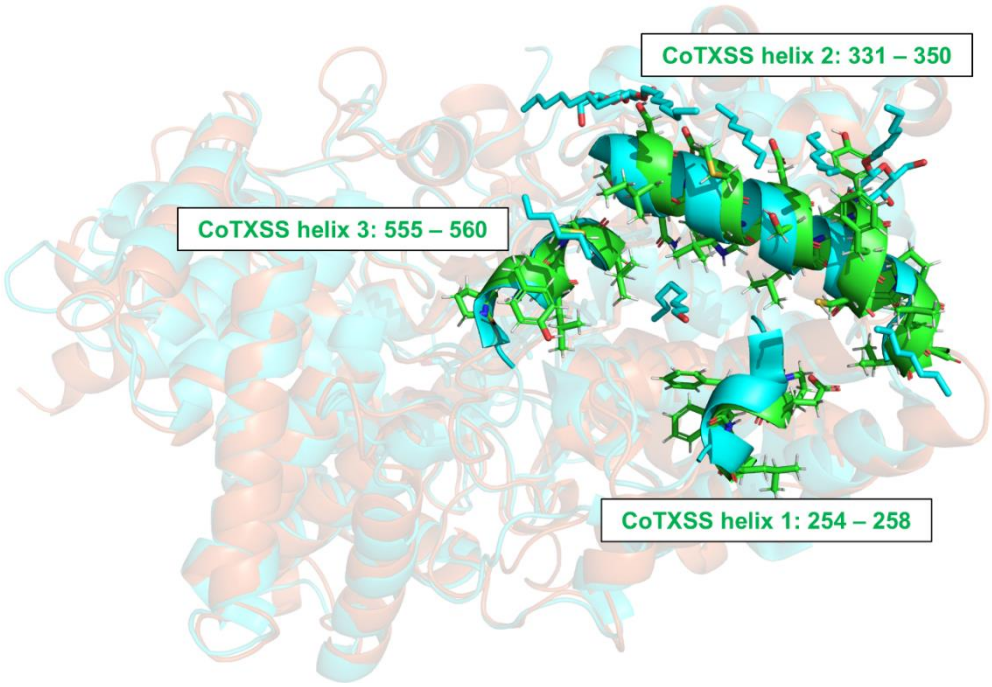


Figure 6.4 Structural model of CoTXSS aligned to the crystal structure of human lanosterol synthase (HsLS) (1W6K). The three membrane-embedded helices are highlighted with CoTXSS shown in green and HsLS synthase in cyan. CoTXSS side chains are shown in green and the detergent molecules used to resolve the HsLS structure are shown in cyan.

To explore the hypothesis that solubilising the three membrane-embedded helices would dissociate the enzyme from the ER, rational and computational protein design approaches were applied. Rational design approaches were applied to helices 1 (254-258) and 3 (555-560). These contain hydrophobic residues on the surface of the enzyme: L258 in helix 1; Y556 and M559 in helix 3. These three residues were substituted with the polar acidic residue E as it does not break the helix but is

expected to repel the protein from the negatively charged phospholipid heads of the ER membrane (Woolfson, 2023).

Helix 2 is the longest membrane-embedded helix and is predicted to be the main source of hydrophobicity. To solubilise it, three protein design approaches were implemented: In the first approach, all exposed residues were redesigned using SolubleMPNN (Goverde et al., 2024). Three non-polar residues (L337, L341 and V349) were fixed because their side chains extend into the interior of the enzyme and may interact with the substrate. In the second approach, the entire helix was redesigned using SolubleMPNN. In the third approach, all exposed hydrophobic residues (I333, M336, Y343, F344 and L348) were substituted to either E or Q, which are polar residues predicted not to break the helix (Woolfson, 2023).

In total, 18 variants of helix 2 were designed, six from each approach. These were combined with the redesigned 1 and 3 helices to produce 18 CoTXSS variants. All 18 variants were modelled using AlphaFold2. The quality was checked (pLDDT and RMSD to wild type) and the variants were ranked based on their hydrophobic solvent accessible surface area (hydrophobic SASA) (Fleishman et al., 2011; Jumper et al., 2021) (Table 6.1).

Name	Helix 1 (254 - 258)	Helix 2 (331 - 350)	Helix 3 (555 - 560)	Helix 1 pLDDT	Helix 2 pLDDT	Helix 3 pLDDT	Hydrophobic SASA	RMSD
CoTXSS	**PEFWL**	**SSIQDMLWDSLHYFCEPLVK**	**PYLQML**	95.96	92.88	90.59	4949.42	0.00
CoTXSS_SME1	**PEFWE**	**TEELIELEDYLYNEVEPLVN**	**PELQEL**	95.73	95.33	91.11	4740.06	1.40
CoTXSS_SME2	**PEFWE**	**TDELLALQDQLYNEVEPLVE**	**PELQEL**	95.69	94.80	89.85	4842.19	0.89
CoTXSS_SME6	**PEFWE**	**TDELIALEDNLYNNVVPLVE**	**PELQEL**	95.14	94.30	90.96	4889.08	0.84
CoTXSS_SME4	**PEFWE**	**TDELLALYDTLFNEVEPLVE**	**PELQEL**	95.63	94.49	91.02	5026.96	1.46
CoTXSS_SME5	**PEFWE**	**TEELLALYDYLYNEVEPLVE**	**PELQEL**	95.74	95.38	90.84	5028.80	1.63
CoTXSS_SME3	**PEFWE**	**TEELIALEDYLYNEVEPLVN**	**PELQEL**	95.60	95.34	90.71	5048.27	1.51
CoTXSS_SMH3	**PEFWE**	**TEEQIELENYLFNEVEPLLN**	**PELQEL**	95.68	94.87	89.57	4847.51	1.40
CoTXSS_SMH6	**PEFWE**	**SQELLDLYDYLYNEVEPLLN**	**PELQEL**	95.93	94.97	91.64	4874.86	1.49
CoTXSS_SMH2	**PEFWE**	**TEEQIALEDYLYNEVEPLLE**	**PELQEL**	95.69	94.95	89.72	4907.47	1.44
CoTXSS_SMH1	**PEFWE**	**TQELLDYYDYLFNEVEPLLE**	**PELQEL**	95.69	95.29	89.70	4925.98	1.41
CoTXSS_SMH5	**PEFWE**	**TEELLALYDTLYNEVEPLLE**	**PELQEL**	95.41	94.78	90.54	4967.93	1.47
CoTXSS_SMH4	**PEFWE**	**TDELIALEDFLYNVVEPLLN**	**PELQEL**	95.59	95.57	88.02	5027.45	1.60
CoTXSS_RDH2	**PEFWE**	**SSQQDQLWDSLHQQCEPQVK**	**PELQEL**	94.98	92.20	87.77	4554.67	0.22
CoTXSS_RDH4	**PEFWE**	**SSQQDELWDSLHQECEPQVK**	**PELQEL**	94.98	92.41	87.15	4568.63	0.30
CoTXSS_RDH3	**PEFWE**	**SSEQDQLWDSLHEQCEPEVK**	**PELQEL**	95.01	92.84	88.38	4586.99	0.23
CoTXSS_RDH5	**PEFWE**	**SSEQDQLWDSLHQECEPQVK**	**PELQEL**	95.10	92.63	87.66	4693.21	1.46
CoTXSS_RDH6	**PEFWE**	**SSEQDQLWDSLHEQCEPQVK**	**PELQEL**	94.02	92.63	85.64	4788.64	1.46
CoTXSS_RDH1	**PEFWE**	**SSEQDELWDSLHEECEPEVK**	**PELQEL**	94.86	92.11	87.28	4840.17	1.66

Table 6.1 Summary of sequences and scores of CoTXSS and its variants generated through computational protein design. Sequences and AlphaFold pLDDT scores of membrane-embedded helices, hydrophobic solvent accessible surface areas (hydrophobic SASA) of the variant and RMSDs to wild-type CoTXSS are displayed. CoTXSS_SME sequences were created through designing exposed residues on helix 2 with SolubleMPNN (Approach 1). CoTXSS_SMH sequences were created through designing entire helix 2 with SolubleMPNN (Approach 2). CoTXSS_RDH sequences were created through substituting exposed hydrophobic residues on helix 2 with E or Q (Approach 3). Sequences shaded in orange were chosen for subsequent experimentation. Sequences shaded in grey were eliminated due to higher hydrophobic SASA than wild type CoTXSS or RMSD > 1.50. pLDDT scores of all redesigned helices are higher than 85, indicating high prediction confidence.

6.4.4 Subcellular localisation of redesigned CoTXSSs

Six candidates (two candidates designed using each approach) with the lowest hydrophobic SASA scores were selected for heterologous expression: SME1 and SME2 were designed by running SolubleMPNN on exposed residues of helix 2; SMH3 and SMH6 were designed by running SolubleMPNN on the entire helix 2; and RDH2 and RDH4 were designed by substituting exposed hydrophobic residues on helix 2 with E or Q.

Constructs containing the coding sequence of each CoTXSS variant were created via iterative rounds of site-directed mutagenesis (Chapter 2.3.3). The variant coding sequences were then assembled into plant expression vectors translationally fused to YFP and regulated by the CaMV35s promoter and terminator. These were co-expressed in *N. benthamiana* leaves with constructs expressing mCherryER and p19 (Figure 6.5A).

The signals from CoTXSS_SMH3:YFP and mCherryER did not overlap (Figure 6.5B) and pixel intensity profiles along two cross-sections indicate that the YFP and mCherry peaks are offset suggesting that CoTXSS_SMH3 does not localise to the ER. This experiment was performed in triplicated and similar patterns were observed in all replicates.

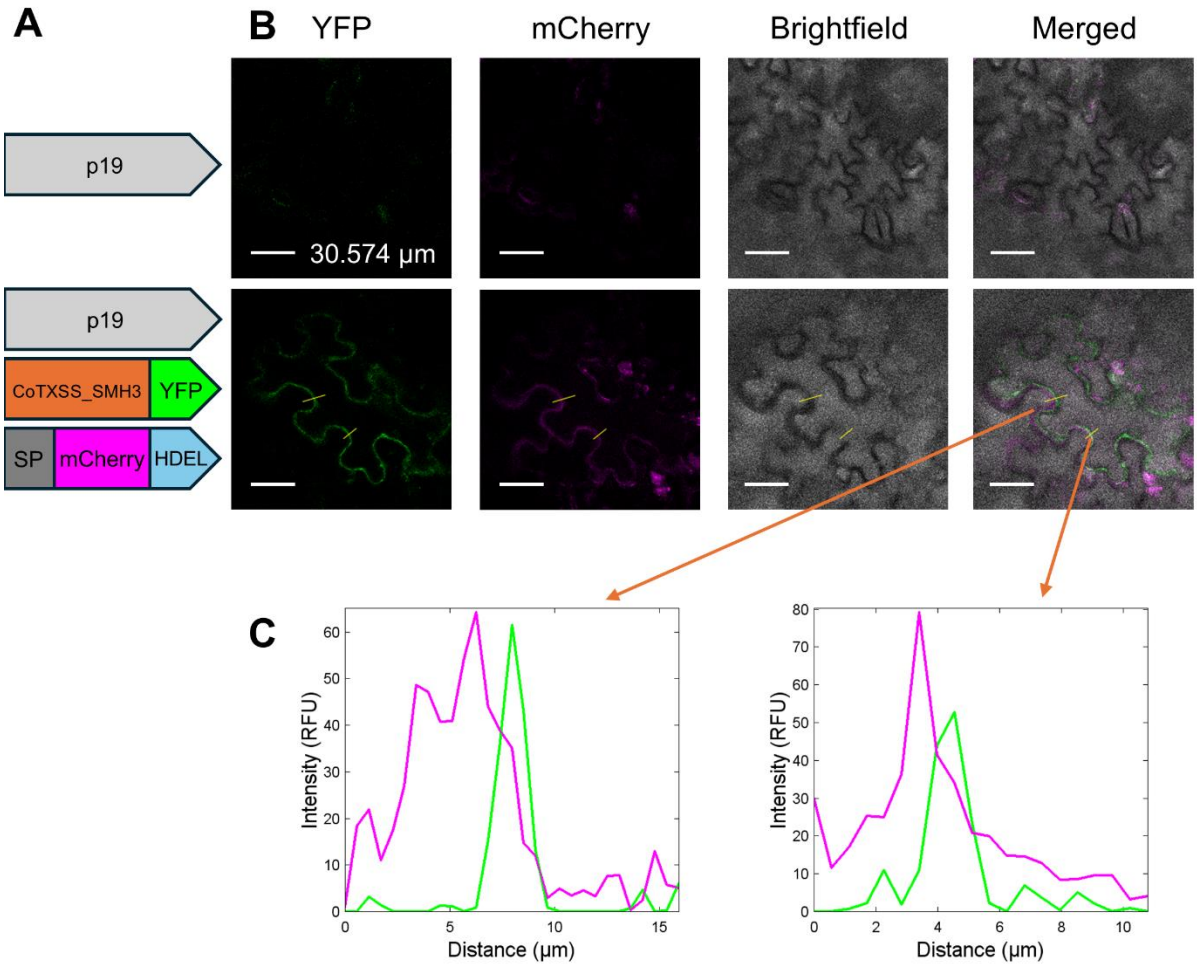


Figure 6.5 Confocal micrographs of CoTXSS_SMH3:YFP and mCherryER in *N. benthamiana* leaves (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

In contrast, all five of the other CoTXSS variants (CoTXSS_SME1, CoTXSS_SME2, CoTXSS_SMH6, CoTXSS_RDH2, CoTXSS_RDH4) appeared to localise to the ER and pixel intensity profiles along two cross-sections showed overlapping peaks (Figures 6.6 – 6.10). In all cases, experiments were performed in triplicate and similar patterns were observed across replicates.

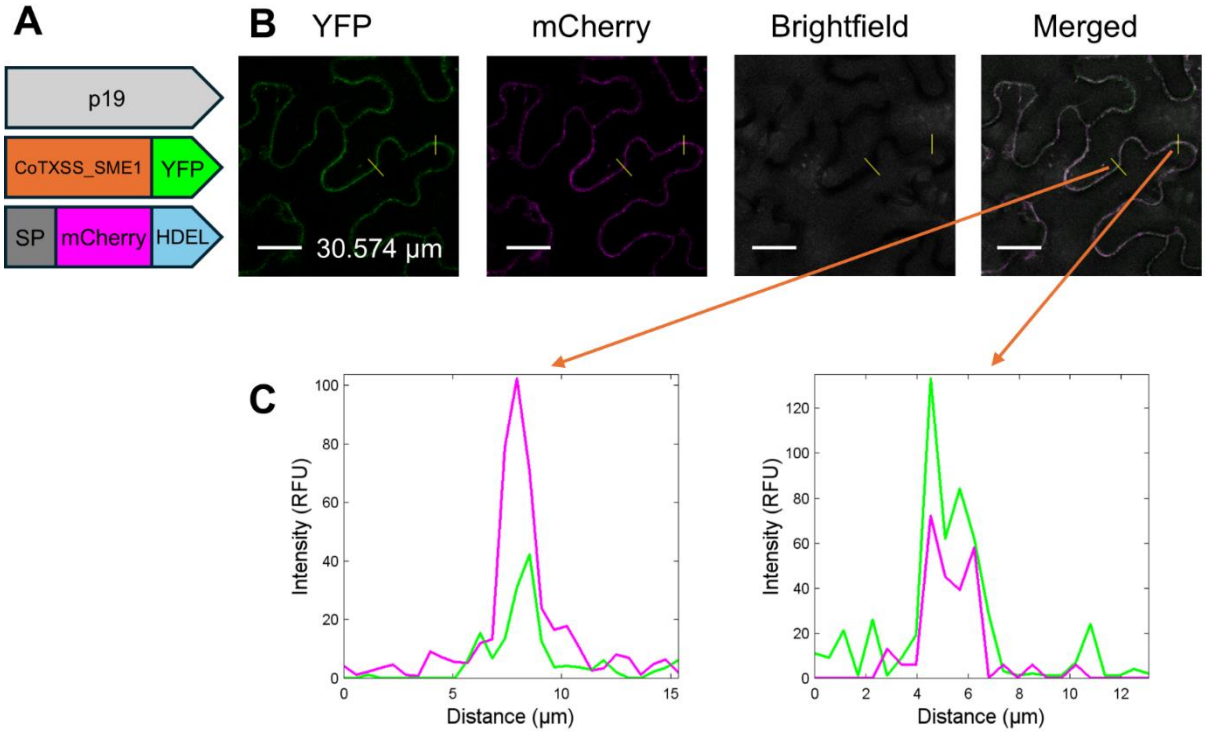


Figure 6.6 Confocal micrographs of CoTXSS_ ME1:YFP and mCherryER in *N. benthamiana* leaves (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

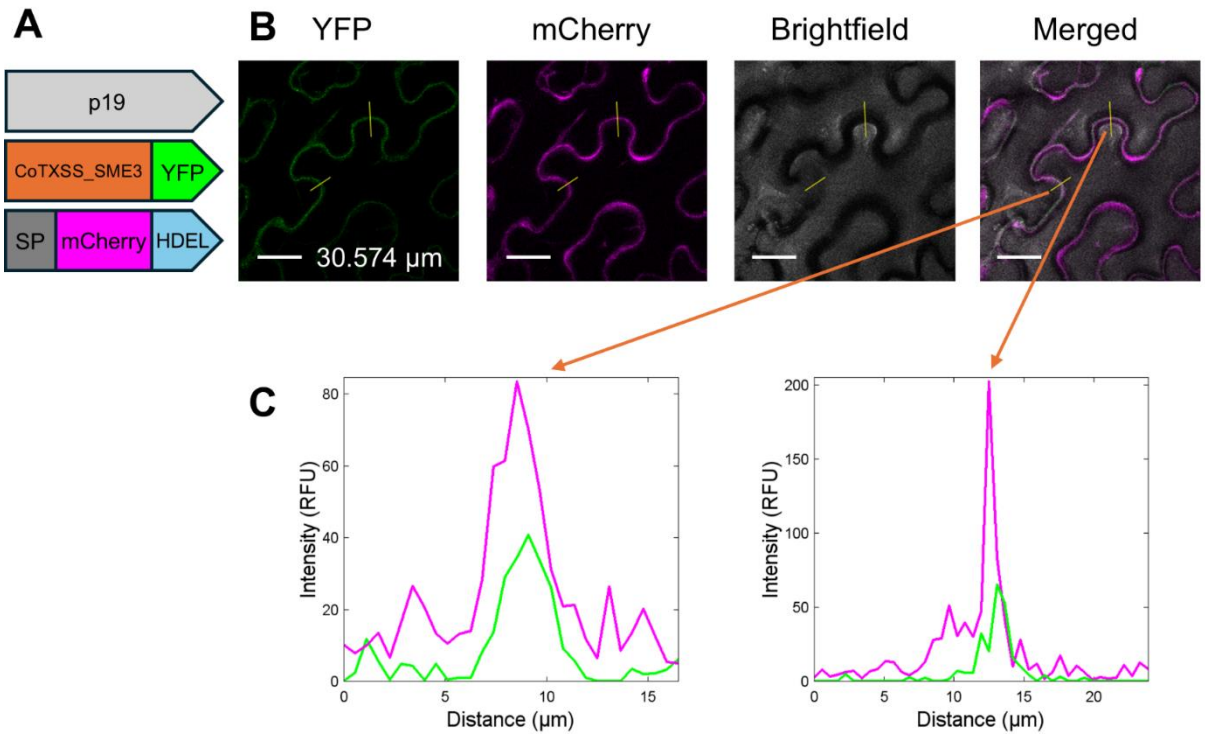


Figure 6.7 Confocal micrographs of CoTXSS_SME3:YFP and mcherryER in *N. benthamiana* leaves (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

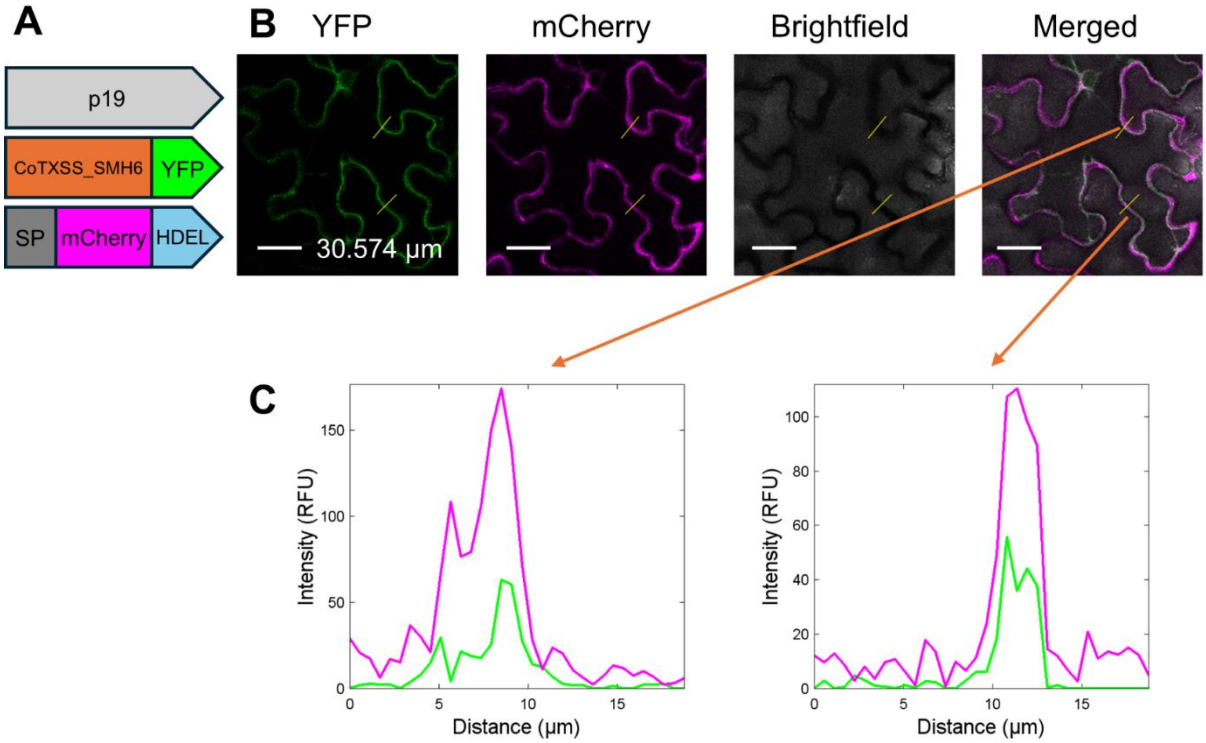


Figure 6.8 Confocal micrographs of CoTXSS_SMH6:YFP and mCherryER in *N. benthamiana* leaves (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and

merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

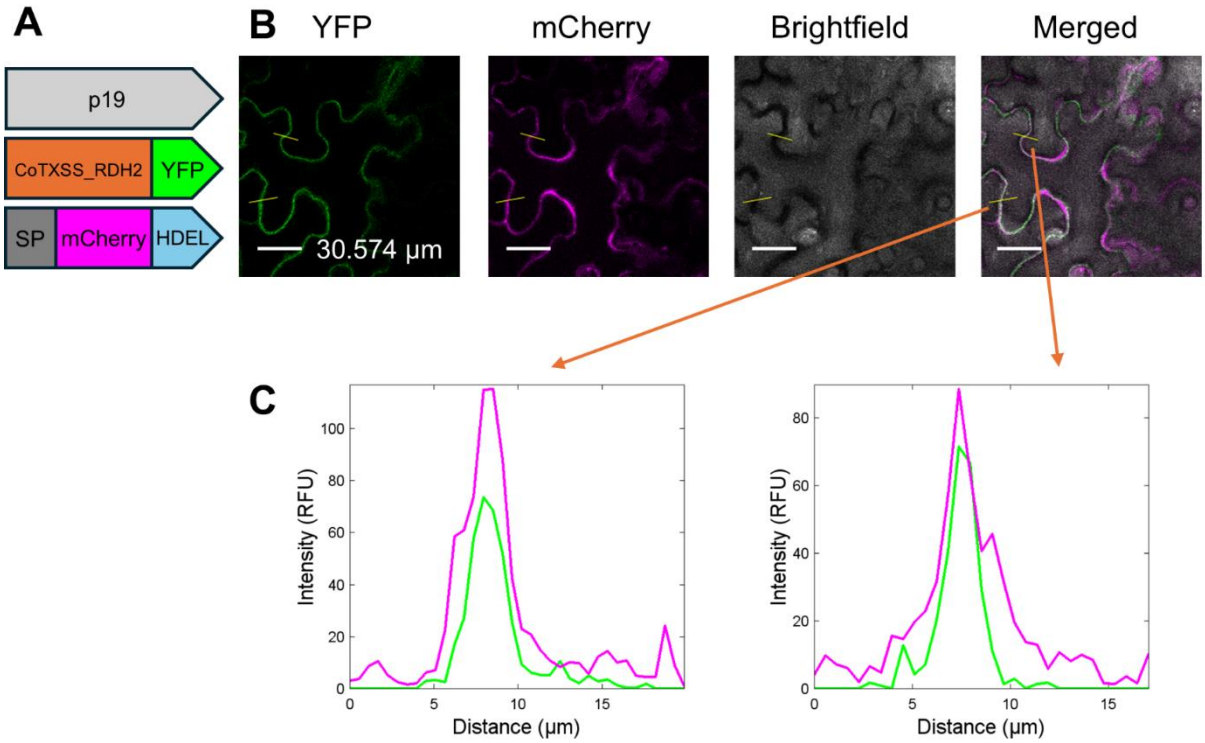


Figure 6.9 Confocal micrographs of CoTXSS_RDH2:YFP and mcherryER in *N. benthamiana* leaves (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

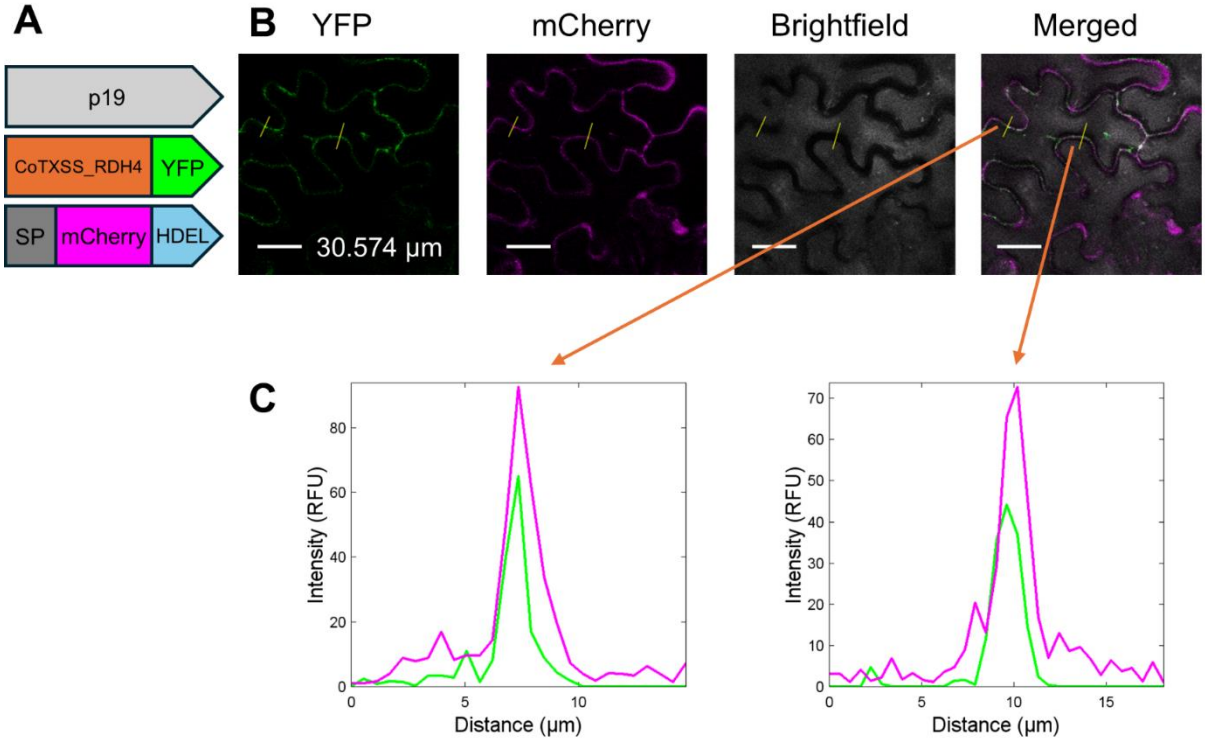


Figure 6.10 Confocal micrographs of CoTXSS_RDH4:YFP and mCherryER in *N. benthamiana* leaves (A) Schematic representation of the synthetic coding sequences transiently over expressed in *N. benthamiana*. SP: OsRAmy3A signal peptide. HDEL: ER-retention signal. p19: suppressor of silencing. (B) Confocal micrographs of *N. benthamiana* showing YFP, mCherry, brightfield and merged channels. Scale bar = 30.574 µm. (C) Pixel intensity profiles of YFP and mCherry along two cross-sections.

6.4.5 Activity of CoTXSS variants in *N. benthamiana*

To assess if the CoTXSS variants were catalytically active, their coding sequences were cloned into plant expression vectors without any tags. These were assembled into plant expression vectors between the CaMV35s promoter and terminator and transiently expressed in *N. benthamiana* leaves together with p19 and a truncated HMGR (tHMGR), which has been previously shown to enhance flux through the mevalonate (MVA) pathway (Reed et al., 2017).

Metabolites were extracted from samples of infiltrated leaves and analysed using GC-MS (Chapter 2.5.4). Constructs expressing the wild type CoTXSS and McLS sequences were also infiltrated (Figure 6.11).

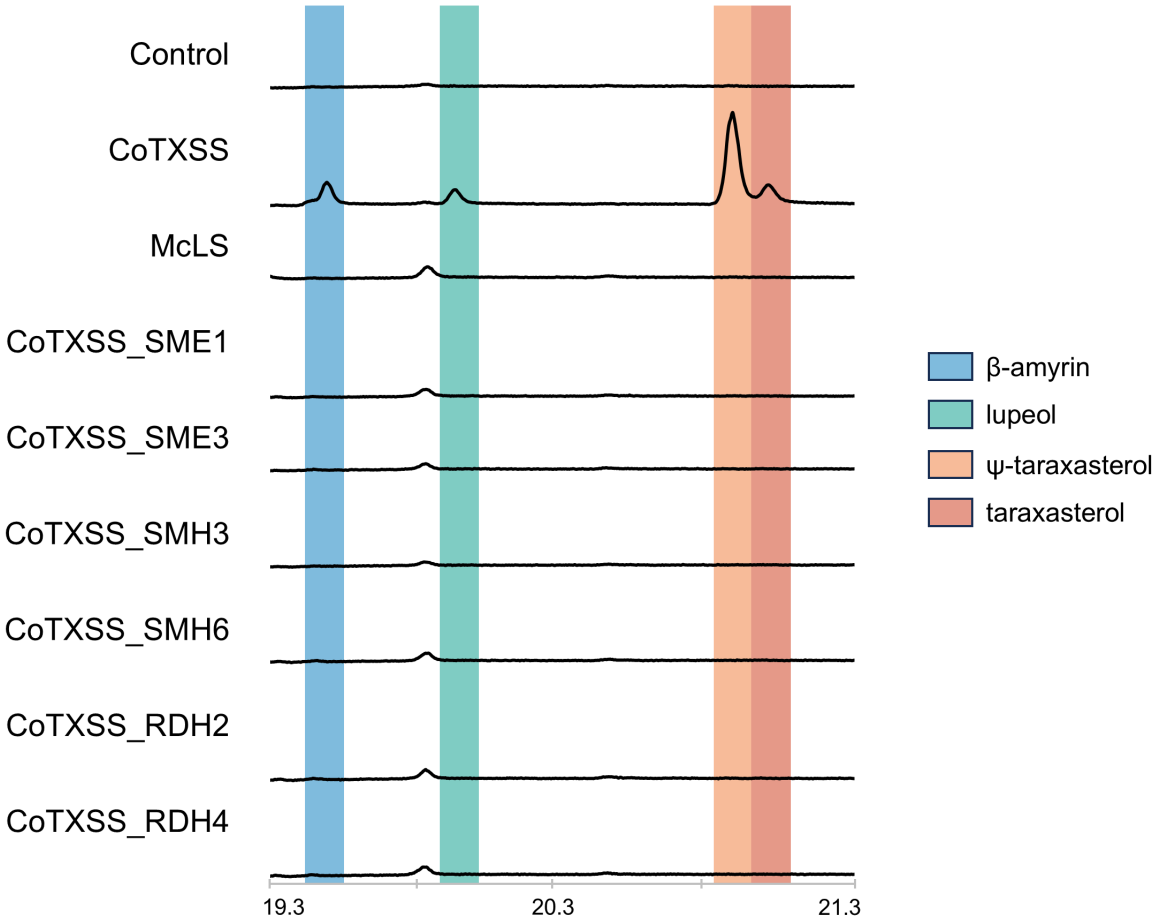


Figure 6.11 GC-MS analysis of *N. benthamiana* leaves transiently expressing CoTXSS, CoTXSS variants and McLS. Representative total ion chromatograms of ethyl acetate extracts of *N. benthamiana* leaves co-infiltrated with constructs overexpressing each OSC the p19 suppressor of silencing and tHMGR; the control sample only expresses p19 and tHMGR. Lanosterol was expected to elute at a time point between lupeol and ψ-taraxasterol.

As previously observed, CoTXSS produced ψ-taraxasterol, taraxasterol, β-amyrin and lupeol. However, these compounds were not detected in leaves infiltrated with engineered CoTXSS suggesting that they have either lost activity or that the substrate is no longer accessible to the enzyme. Interestingly, lanosterol was also not detected in leaves transiently expressing McLS.

It is possible that when no longer properly embedded in the ER, OSCs are unable to access the 2,3-oxidosqualene substrate, which is produced by ER-bound squalene epoxidases (Laranjeira et al., 2015). In previous work, McLS was found to produce lanosterol from fed 2,3-oxidosqualene *in vitro* (Lamb et al., 2007). Therefore, total protein was extracted from five *N. benthamiana* leaf discs (approximate mass =120 mg) transiently expressing CoTXSS variants, CoTXSS or McLS. This was fed with 10 µg 2,3-oxidosqualene and the reaction analysed using GC-MS (Figure 6.12).

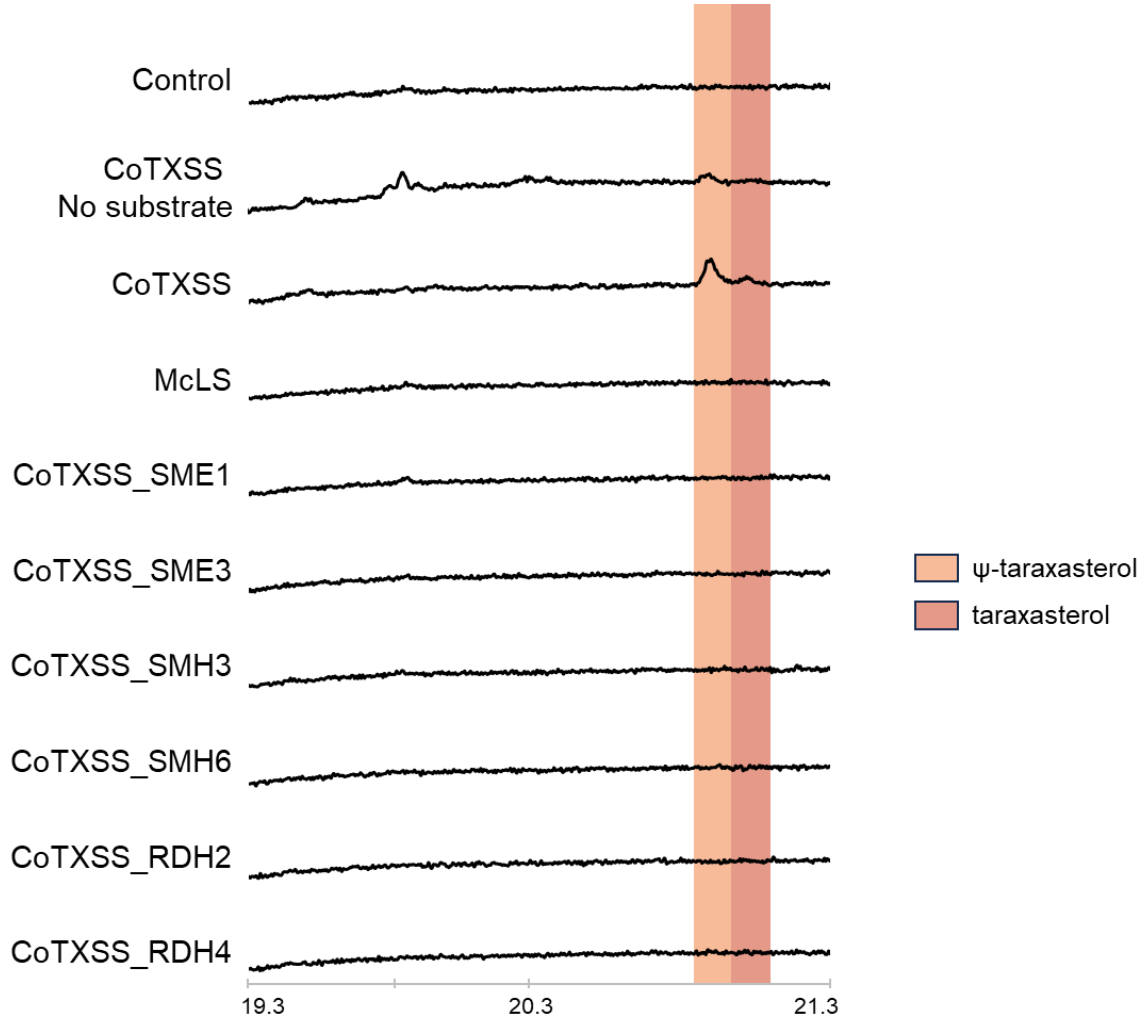


Figure 6.12 GC-MS analysis of *in vitro* bioassays of total proteins extracted from *N. benthamiana* leaves transiently expressing CoTXSS variants and McLS incubated with 2,3-oxidosqualene substrate. Representative total ion chromatograms of ethyl acetate extracts of *in vitro* bioassays. All OSCs were co-expressed with p19; the control sample only expresses p19. Lanosterol is expected to be eluted at a time point before ψ-taraxasterol.

A trace amount of ψ-taraxasterol and taraxasterol was detected in samples of total protein from leaves expressing wild type CoTXSS fed with 2,3-oxidosqualene, with amount higher than samples not fed with 2,3-oxidosqualene. However, no target products were detected in samples of total protein from leaves expressing CoTXSS variants or McLS. Further work to determine if the enzymes were inactive in the leaves of if activity was compromised by extraction could not be completed due to lack of time.

6.4.6 Expression of CoTXSS variants in *E. coli*

Previous reports indicate that McLS was soluble in *E. coli* (Lamb et al., 2007). To investigate if the CoTXSS variants are soluble in *E. coli*, McLS, wild type CoTXSS and variants were cloned into an *E. coli* expression vector with N-terminal 9x His-tag (Chapter 2.3.4). Constructs were transformed into *E. coli* SHuffle T7 express. Single colonies were grown on selection at 37 degrees until OD600 reaches 0.6 to 0.8. To identify optimal growth conditions for protein expression, following induction of expression with 1mM IPTG, cells were incubated at three different temperatures for 16-20 hours. Cells were subsequently harvested through centrifugation and lysed in lysis buffer containing lysozymes and β-mercaptoethanol (Chapter 2.3.7).

Soluble and insoluble fractions of total protein extracts were separated on a 4–12% NuPAGE Bis-Tris (Chapter 2.3.8). Western blot analysis was performed, and target proteins were detected using anti-His antibodies (Chapter 2.3.8) (Figure 6.13).

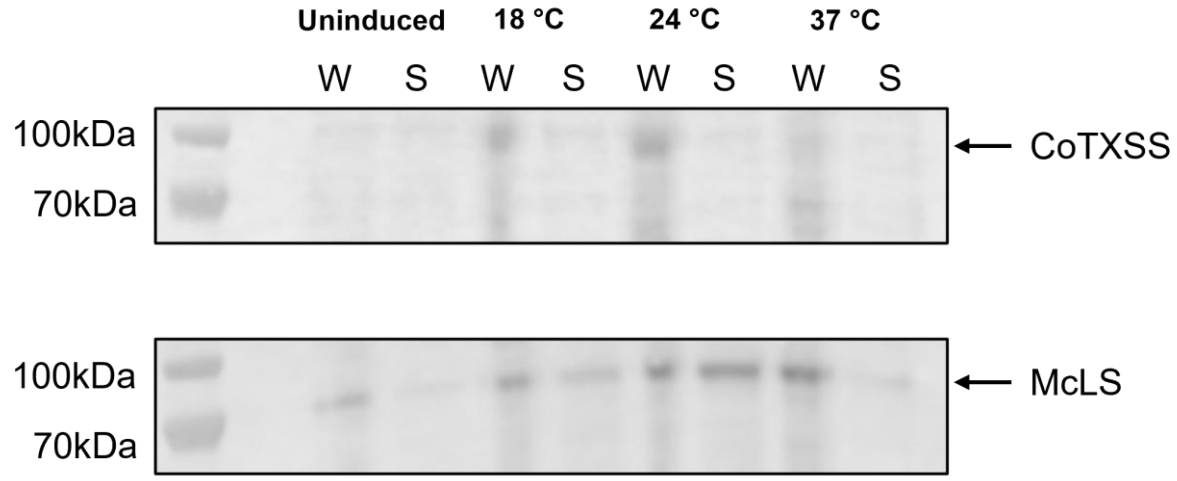


Figure 6.13 Western blot analysis of total protein extracted from *E. coli* expressing recombinant McLS and CoTXSS. Membranes were probed with and anti-His antibody tagged with HRP (1:5000) and visualised with ECL. Prior to extraction, cultures were diluted to OD600. W: Whole cell lysates. S: Supernatant.

The predicted molecular weight of His-tagged CoTXSS is 92.07 kDa. Bands corresponding to this predicted weight were observed in the whole cell lysates of *E. coli* cultures expressing this protein that had been incubated at 18°C and 24 °C post-induction.

The predicted molecular weight of His-tagged McLS is 77.99 kDa. Bands corresponding to this predicted weight were observed in the whole cell lysates and supernatants of *E. coli* cultures expressing this protein that had been incubated at 18°C and 24 °C post-induction. More protein was detected in samples incubated at 24 °C and this was, therefore, selected for future experiments.

E. coli strains expressing recombinant CoTXSS and CoTXSS variants were induced with 1mM IPTG and incubated for 16-20 hours at 24°C. Cells were subsequently harvested, lysed and analysed as above (Figure 6.14).

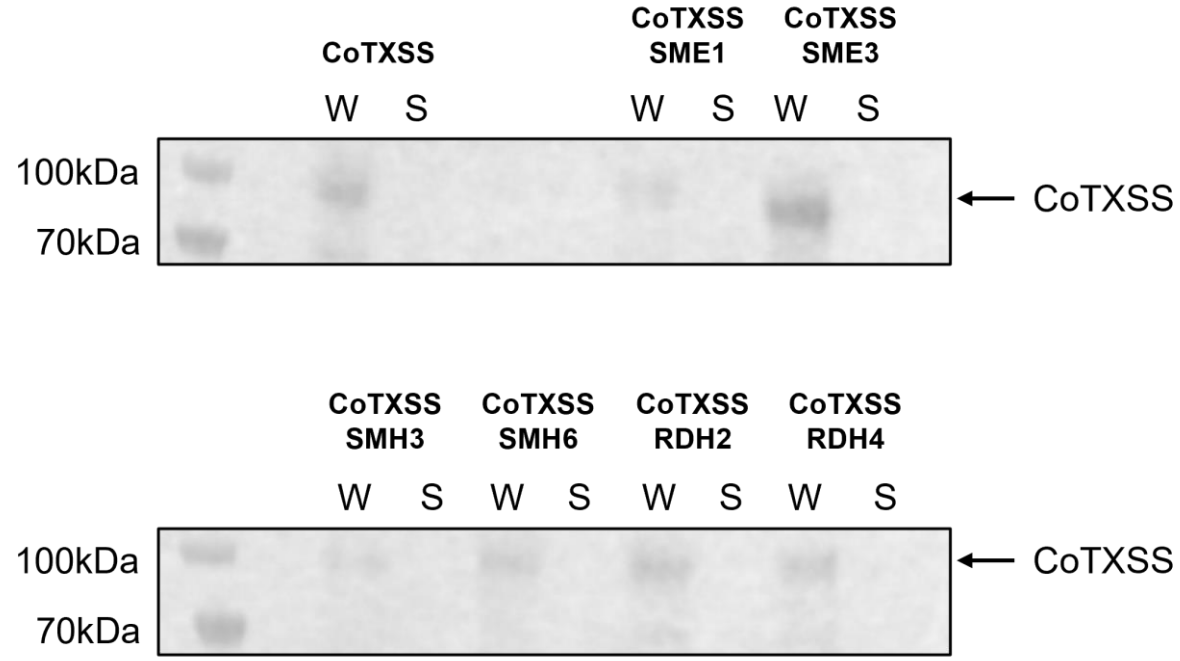


Figure 6.14 Western blot analysis of total protein extracted from *E. coli* expressing recombinant CoTXSS and CoTXSS variants. Membranes were probed with anti-His antibody tagged with HRP (1:5000) and visualised with ECL reagents. Prior to extraction, *E. coli* cultures were diluted to OD600. W: Whole cell lysates. S: Supernatant.

For *E. coli* strains expressing wild-type CoTXSS or CoTXSS variants, protein bands corresponding to the predicted size of His-tagged CoTXSS were observed in the whole cell lysates but the not supernatants. This indicates that the recombinant proteins are insoluble and/or present within inclusion bodies. Due to lack of time, this was not explored further.

6.5 Discussion

6.5.1 CoTXSS localises to the ER membrane in *N. benthamiana*

Through purification and structural elucidation with X-ray crystallography, human lanosterol synthase was characterised as a monomeric monotopic protein located at the outer ER membrane (Thoma et al., 2004). As plant OSCs are homologous to

human lanosterol synthase and share high structural similarity, plant OSCs it is generally assumed in the literature that they localise to the ER (Thimmappa et al., 2014). However, the experimental evidence for the subcellular localisation of plant OSCs is surprisingly sparse. Luo et al. demonstrated that *T. wilfordii* friedelin synthase is located to the ER membrane (Luo et al., 2023). In addition, Dong et al. concluded that *Iberis amarala* and *Citrullus lanatus* of cucurbitadienol synthases are located to the ER and Almeida et al. reported that *Ononis spinosa* α-onocerin synthase is present in the cytosol using confocal microscopy imaging (Almeida et al., 2018; Dong et al., 2021). However, these latter studies did not include compartmental marker controls, which makes them less robust.

In this study, by co-expressing CoTXSS translationally fused to YFP with an ER-localised mCherry, I observed that CoTXSS localises to the ER membrane (Figure 6.2). As plant squalene epoxidases, which produce 2,3-oxidosqualene, also localise to the ER (Laranjeira et al., 2015; Unland et al., 2018), co-localisation enables efficient substrate channelling.

6.5.2 *M. capsulatus* lanosterol synthase localises to the cytosol expression in plants and is soluble in *E. coli*

McLS is one of the few OSCs that have been reported to be soluble (Lamb et al., 2007). It shows high structural similarity to eukaryotic OSCs and the structural model has high similarity to human lanosterol synthase crystal structure (RMSD = 0.931) as well as to the structural model of CoTXSS (RMSD = 0.883). Consistent with previous reports, recombinant McLS was soluble in *E. coli* (Figure 6.13). In addition, when expressed in *N. benthamina*, McLS did not colocalise with the ER marker but was adjacent, which indicates a likely cytosolic location (Figure 6.4). This experiment should ideally be repeated with a cytosolic marker to confirm the subcellular location of McLS.

However, in this study, McLS did not produce lanosterol when expressed in *N. benthamiana* (Figure 6.11). This might be explained by the reduced accessibility of ER-bound 2,3-oxidosqualene substrate in plant cells, but further investigation is needed. Lanosterol production was also not detected in an *in vitro* assay in which 2,3-oxidosqualene was incubated with protein extracts of *N. benthamiana* leaves expressing McLS. This contradicts the original study, which demonstrated its activity *in vitro* when purified from *E. coli* (Lamb et al., 2007). Further investigations are required to investigate if the extraction process led to loss in activity. Assay optimisation will also be attempted.

An alternative approach to assessing activity would be to target McLS to *N. benthamiana* chloroplasts by including an N-terminal transit peptide. Co-expression with chloroplast targeted squalene synthase and squalene epoxidase would convert

the chloroplast pool of farnesyl pyrophosphate to 2,3-oxidosqualene. Synthetic chloroplast targeting peptides have been shown to robustly target heterologous protein to this compartment (Engler et al., 2014). Successful squalene production in plant chloroplasts has been demonstrated with the best average yield of 0.302 mg/g fresh weight (Bibik et al., 2022). However, engineering 2,3-oxidosqualene biosynthesis in chloroplasts has yet to be achieved. In addition, the sub-compartmental localisation of 2,3-oxidosqualene within chloroplasts is difficult to predict. In the native *M. capsulatus*, McLS was proposed to be tethered to the membrane-bound squalene epoxidase through protein-protein interaction to gain access to membrane-bound 2,3-oxidosqualene substrate (Lamb et al., 2007). Therefore, if the substrate remains largely insoluble in chloroplasts (e.g. in the thylakoid or inner organellar membrane), the soluble McLS could be tethered to a chloroplast-targeted squalene epoxidase to mimic the *M. capsulatus* protein complex and enable substrate access.

6.5.3 An engineered variant of CoTXSS is not localised to the ER

Of six CoTXSS variants, only CoTXSS_SMH3 exhibited enhanced solubility and likely became detached from the ER (Figure 6.5). Helix 2 of this variant was designed using SolubleMPNN. Compared to helix 2 of the other five variants, CoTXSS_SMH3 does not have any sequence or structural feature which makes it stand-out (Table 6.1). Hence, the detachment from ER might be attributed to a unique synergistic effect of the residues in redesigned helix 2.

The low success rate of solubilisation could be attributed to the misidentification of exposed residues due to inaccurate side chain packing in the model. Benchmarking studies have recently demonstrated that AlphaFold2 sometimes struggles to accurately predict membrane-embedded regions of proteins and packing side chains (Gut and Lemmin, 2025; Hegedűs et al., 2022). Thus, some residues predicted to be exposed by in fact be buried within the structure and, conversely, residues identified as buried might be exposed.

Furthermore, inaccurate side chain packaging could have led to the misidentification of residues required for CoTXSS activity. For example, residues important for interacting with the 2,3-oxidosqualene substrate and maintaining active site hydrophobicity may have been altered. Given that all variants may be inactive, it is possible that L258, Y556 and M559 on helix 1 and helix 3, which were mutated in all variants, might be important for CoTXSS activity. Alternatively, given that activity of a previously characterised soluble McLS was not, the assay may not be sufficiently robust to determine if the redesigned CoTXSS variants are active.

6.5.4 CoTXSS variants are insoluble in *E. coli*

In previous work, Pfalzgraf attempted to solubilise *Avena strigosa* β-amyrin synthase by engineering helix 2 through substituting to corresponding residues in McLS. However, the engineered variants remained insoluble (Pfalzgraf, 2022). In this work, despite engineering three helices, recombinant CoTXSS variants were not detected in the soluble fractions of *E. coli* (Figure 6.14).

One possibility is that additional regions beyond the three identified helices, contribute to insolubility or membrane association. However, proteins can also be found in the insoluble fractions of extracts due to misfolding. Misfolding could lead to the exposure of hydrophobic regions, which interact with similar proteins and lead to formation of aggregates and inclusion bodies (Burgess, 2009). Given McLS is soluble in *E. coli*, the insolubility of CoTXSS and variants might be attributable to unsuitable residue composition, leading to misfolding and forming inclusion bodies, rather than a less favourable topology. In the future, protein purification from inclusion bodies and refolding could be attempted, as inclusion bodies might retain native-like secondary structures and purification of functional enzymes, such as β-lactamase, β-galactosidase, and human granulocyte-colony stimulating factor, from inclusion bodies has been demonstrated (Singh et al., 2015). To improve solubility expression of CoTXSS in *E. coli*, additional regions could be redesigned. A successful example of this is the redesign of non-conserved regions outside active site of tobacco etch virus protease using ProteinMPNN. This enabled soluble and stable expression in *E. coli* with enhanced cleavage activity (Sumida et al., 2024). In general, to engineer enzyme solubility while preserving its activity, a comprehensive knowledge of how enzyme functions, key residues involved in proper folding and regions contributing to protein hydrophobicity is needed. This has been accelerated by advancement of sequence searching and alignment methods such as HHblits and deep-learning driven protein modelling and design methods (Dauparas et al., 2022; Jumper et al., 2021; Remmert et al., 2012).

6.6 Conclusions

A soluble CoTXSS variant could pave way for enabling production of ψ-taraxasterol compounds in *E. coli*, which would facilitate scalable production. It might also enable triterpene production in chloroplasts, which contain fewer enzymes known to derivatise heterologously produced small molecules. Using protein modelling and structural comparison, regions contributing to CoTXSS hydrophobicity were identified. Using rational design and computational design, a variant of CoTXSS that is unlikely to be attached to the ER of plant cells was engineered. However, the functionality of this protein could not be verified, and the protein remained insoluble in *E. coli.*

7. Chapter 7 Discussion

7.1 Introduction

Plants produce diverse chemical compounds, many of which are of industrial and pharmaceutical interest. However, within their native species, such compounds are often present in complex mixtures and/or produced in a limited number of cell types. This can make purification complicated, expensive and inefficient, limiting accessibility (Golubova et al., 2024). Although some species are suitable for cultivation, other species have been harvested from the wild due to challenges with cultivation. In some cases, this has led to unsustainable practices: for example, paclitaxel is a compound used in chemotherapeutics that used to be harvested from the bark of the Pacific yew (*Taxus brevifolia*). Overharvesting eventually led to a 30% decline in population and the listing of the species as 'near threatened' in 2012 (Howes et al., 2020). In recent years, cell culture has provided promising alternatives to extraction from mature trees and the yield from cell cultures has been improved by optimising induction and culturing conditions (Zhong, 2002).

An alternative to obtaining natural products by extraction from native species is to identify the genes involved in their biosynthesis, enabling reconstruction in heterologous hosts (Golubova et al., 2024). Successful examples of this include the anti-malarial drug precursor artemisinic acid and the flavour compound vanillin (Gallage et al., 2014; Paddon et al., 2013).

Pathways are typically identified by first predicting a series of reactions (and therefore types of enzymes) that lead to the final product using so-called 'chemical logic' (Han et al., 2024). Rapid advances in DNA sequencing technology together with bioinformatics tools and databases such as Pfam domains have facilitated the identification of genes that encode enzymes from specific families (Mistry et al., 2021). However, as plants often encode large families of biosynthetic enzymes and it remains difficult to predict substrate and/or product specificity from the primary sequence, identifying specific enzymes can be challenging. In some cases, pathway elucidation can be accelerated when biosynthetic genes are co-located in gene clusters, for example, the triterpenoid avenacin A-1 in *Avena strigosa*, the steroidal glycoalkaloid α-tomatine in *Solanum lycopersicum*, and the alkaloid noscapine in *Papaver somniferum* (Nützmann and Osbourn, 2014). More often, however, plant genes are not clustered and other methods are required. Common approaches to narrow down candidate enzymes involved in a specific pathway are to compare transcriptome or proteome datasets from multiple tissues or different conditions to identify enzymes for which expression correlates with the presence of the molecule (Swamidatta and Lichman, 2024). However, it is often necessary to test many candidates. Recent advances in plant single-cell transcriptomics have been used to further narrow down candidate genes and have enabled the pathway elucidation of metabolites that show cell-type specific accumulation such as hyperforin in *Hypericum perforatum* (Wu et al., 2024). Nevertheless, the pathway elucidation remains challenging, particularly when biosynthetic steps involve previously

undescribed reactions catalysed by unconventional enzymes. For example, after more than 30 years of research, the biosynthesis pathway of paclitaxel was only fully elucidated after the application of single-cell omics led to the discovery of a protein from a family with no previously known roles in metabolite biosynthesis (Liang et al., 2025; McClune et al., 2025).

Once pathways have been elucidated, they can be reconstructed in heterologous hosts, also called production chassis. *E. coli* and *Saccharomyces cerevisiae* are commonly used chassis due to their well-characterised metabolism, the availability of tools and methods for genetic manipulation, rapid growth rates, and the availability of infrastructure and expertise for large-scale growth (Zhu et al., 2021). However, even in these well-characterised hosts, reconstructing complex pathways can still be challenging (Zhu et al., 2021). For example, expression of functional plant cytochrome P450s in *E. coli* and *S. cerevisiae* typically requires co-expression of a corresponding cytochrome P450 reductase and is further complicated by the membrane-bound nature of these enzyme families (Hausjell et al., 2018). In addition, concerns about the sustainability of some feedstocks used for microbial cultivation have been raised (Lips, 2021). Plant chassis such as *Nicotiana benthamiana* provides an alternative with several advantages including the presence of precursors for many classes of secondary metabolites and the ability to express and fold most enzymes (Golubova et al., 2024).

Understanding the relationship between structure and product specificity has the potential to assist in pathway discovery and to enable the rational engineering of enzymes for the efficient and specific production of compounds of interest. This can be facilitated by integrating molecular modelling and simulation with evolutionary analysis. A successful example of molecular simulation guided engineering is the application of QM/MM to *Streptomyces pristinaespiralis* selenadiene synthase. This successfully predicted that a G305E mutation would enable hydroxylation and production of selin-7(11)-en-4-ol (Srivastava et al., 2024). Similarly, MD guided introduction of an Y525F mutation in *Pogostemon cablin* patchoulol synthase resulted in an enzyme that was able to abolish water capture and produce β-caryophyllene and α-bulnesene (Srivastava et al., 2023). Evolutionary analysis was used to identify a residue in the active site of *Datura wrightii* salicylic acid and to predict that a M156H mutation would lead to equal preferences for salicylic acid and benzoic acid substrates (Barkman et al., 2007). Similarly, the identification of sites under positive selection was used to guide the replacement of five residues in the active site of *Pinus tabuliformis* glutathione S-transferase to broaden its substrate (Lan et al., 2013).

In addition, metabolite yield could be enhanced through engineering the heterologous host to funnel substrates efficiently and reduce competing pathways. In plants, the presence of multiple sub-cellular compartments such as chloroplasts, the ER and the cytosol, can be used to facilitate the efficient use of metabolic pools and optimise pathway partitioning to increase yields. For example, in *N. benthamiana,*

locating the early biosynthetic steps of strictosidine biosynthesis in the chloroplast with the later steps in cytosol led to increased yield (Dudley et al., 2022). Similarly, overexpression of enzymes involved in precursor production together with chloroplast expression of a synthetic fusion of a truncated cytochrome P450 and corresponding reductase improved yields of taxadiene-5α-ol (Li et al., 2019).

Another approach to increasing yields is to manipulate the genetics of the native plant. This was successfully achieved in *Beta vulgaris*, where overexpression of a regulator of betalain biosynthesis (BvMYB1) in white beet enhanced betalain content (Hatlestad et al., 2015). Similarly, overexpression of ANT1, a R2R3-MYB in *S. lycopersicum*, boosted anthocyanin content more than 500-fold (Mathews et al., 2003). Thus, knowledge of the transcriptional regulators of specific pathways can be used to inform approaches for manipulating gene expression levels to increase yields.

This thesis focusses on taraxasterol synthase (TXSS). In *C. officinalis*, this enzyme catalyses the first committed step of the production of faradiol fatty acid esters, which colleagues recently demonstrated to be responsible for the anti-inflammatory activity of floral extracts (Golubova et al., 2025). In Chapter 3, the evolutionary origin of this enzyme is uncovered. In Chapters 3 and 4, combining evolutionary analyses with structural modelling and computational simulation, uncovered the relationship between the structure and product specificity of TXSS, and identified a likely mechanism of catalysis. In Chapter 5, a flower-specific R2R3-MYB was found to contribute to regulating the expression of *CoTXSS* and two downstream genes involved in faradiol fatty acid ester biosynthesis. In Chapter 6, rational and computational methods were used to engineer a variant of CoTXSS that is likely to have lost its association with the ER. Together, these findings shed light on the evolutionary history of taraxasterol biosynthesis, key TXSS residues to engineer to change product specificity and subcellular localisation, and transcription factor to overexpress to enhance anti-inflammatory faradiol fatty acid ester yield in native plants.

7.2 Structural modelling provides insights into the specificity and subcellular localisation of OSCs

The 2024 Nobel Prize for Chemistry was jointly award to scientists that advanced protein structure prediction and computational protein design. This highlights the growing importance of computational structural biology. Protein structural modelling has been particularly useful for understanding the chemistry of plant oxidosqualene cyclases (OSCs), particularly as no X-ray crystal structure have been acquired. The most closely related is human lanosterol synthase (1W6K) (Thoma et al., 2004). The difficulty in acquiring crystal structures of plant OSC is due to the monotopic integral membrane nature of these proteins. These difficulties are exemplified by unsuccessful attempts to purify soluble proteins of *Avena strigosa* β-amyrin synthase, *Euphorbia tirucalli* β-amyrin synthase and, in this thesis, CoTXSS (Dokarry, 2010; Pfalzgraf, 2022).

In Chapter 3 and Chapter 4, the structures of CoMAS, CoTXSS and TkTXSS were modelled using AlphaFold2 and reaction intermediates were docked into their active sites. Despite sequence divergence, these enzymes share high structural similarity with human lanosterol synthase. The βγ fold, characteristic of class II terpene synthase, was observed in all structural models (Thimmappa et al., 2014). In addition, conserved motifs such as DCTAE, which initiates cyclisation, and MWCYCR, which stabilises the D/E ring, were observed in the active sites of all structural models. The active sites were also observed to be enriched in hydrophobic residues and aromatic amino acids for intermediate stabilisation with π-cation interactions. Active site residues that differ among different OSCs play a role in shaping active site geometry, resulting in differences in product specificity. Based on the information gained from these structural models, two residues that determine amyrin/taraxasterol specificity in CoMAS, and four residues that determine ψ-taraxasterol/taraxasterol specificity in CoTXSS and TkTXSS were identified. Subsequent site-directed mutagenesis experiments confirmed that a CoMAS double reciprocal mutant preferentially produced taraxasterols instead of amyrins (Figure 3.10), and both CoTXSS and TkTXSS quadruple reciprocal mutants experienced change in product specificity and preferentially produced taraxasterol and ψ-taraxasterol respectively (Figure 4.4).

Similar methods have been employed to investigate the structural basis of product specificity in other plant OSCs, such as *A. strigosa* hopenol synthases, *Chenopodium quinoa* quinoxide synthase, *Astragalus membranaceus* β-amyrin and cycloartenol synthase and *Oryza sativa* orysatinol and parkeol synthases (Chen et al., 2022; Liang et al., 2022; Xue et al., 2018a; Zhou et al., 2024). For *A. strigosa* hopenol synthases, reciprocal mutations of three active site residues interconverted the product specificity of a hopenol B synthase and a hop-17(21)-en-3β-ol synthase (Liang et al., 2022). Similarly, reciprocal mutations of four active site residues converted a *C. quinoa* quinoxide synthase into a β-amyrin synthase, whereas conversion of a *C. quinoa* β-amyrin synthase into a quinoxide synthase required seven reciprocal mutations (Zhou et al., 2024). Consistent with this study, OSCs with divergent sequences and product specificity were found to share high structural similarity and key active site motifs were found to be conserved. However, a larger active site was defined in this study (12 Angstroms from the centre of the intermediate), implying that residues important for determining product specificity might not be in direct contact with the intermediate.

Although structural modelling allows the identification of key residues determining product specificity, the catalytic mechanisms can remain elusive as models only provide a snapshot of ligand-enzyme complexes. In Chapter 4, proton acceptors that terminate the proton transfer reaction leading to the final product could not be identified. Hence, molecular dynamic simulations (MD) were performed to gain dynamic insights into the interactions between enzymes and ligands. Using this approach, CoTXSS D385 was found to be involved in deprotonation but no TkTXSS amino acid was identified. Site-directed mutagenesis supported the role of D385,

and QM/MM simulation provided a mechanistic explanation of the specificity of CoTXSS for ψ-taraxasterol. MD and QM/MM were previously conducted to investigate the product specificity of *A. strigosa* hopenol synthases, *Alisma orientale* prostadienol and cycloartenol synthases and *Tripterygium wilfordii* friedelin synthase and rationally redesign product specificity (Liang et al., 2022; Luo et al., 2023; Zhang et al., 2023). However, MD and QM/MM potentially suffer from several drawbacks. In the study of TXSSs described in Chapter 4, potential inaccuracies arising from docking a reaction intermediate model to enzyme models led to the use of a less stringent (5 Angstrom) criterion between the proton donor and acceptor. In addition, the blockage by Y273 limited D385 accessibility to intermediate. Therefore, the availability of a crystal might have improved the accuracy of the MD simulation. The quality of the enzyme-ligand model remains the most important factor in successful identification of cryptic interactions between enzymes and ligands during computational simulations.

Finally, in chapter 6, computational modelling enabled the identification of residues that contribute to the ER-localisation of OSCs, which only few studies to date have addressed. Previously, only an X-ray crystallography study of human lanosterol synthase and a cryo-EM study of *Tripterygium wilfordii* friedelin synthase have provided direct experimental evidence of ER-localisation (Luo et al., 2023; Thoma et al., 2004). With the goal of perturbing the ER-localisation of CoTXSS, the inverse folding protein design method SolubleMPNN was applied to the main hydrophobic element. Deep learning-driven protein design enabled the design of a protein (CoTXSS_SMH3) that is likely to be soluble (detached from the ER membrane and present in the cytosol). However, more work is required to establish assays to determine the function of soluble OSCs.

ProteinMPNN and SolubleMPNN have previously been used to find the optimal and probable amino acids of a given backbone, based on what they have learnt from the examples in the Protein Data Bank (PDB) or (for SolubleMPNN) soluble structures in PDB (Dauparas et al., 2022; Goverde et al., 2024). By combining rational design and ProteinMPNN, new α-helical barrel proteins were designed and structurally characterised with a high success rate (>60% soluble) (Albanese et al., 2024). In addition, ProteinMPNN and SolubleMPNN were successful at rescuing previously failed designs, enhancing protein solubility and thermostability, and creating soluble analogues of membrane proteins, suggesting that they have broad applicability and strong accuracy (Dauparas et al., 2022; Goverde et al., 2024; Sumida et al., 2024). Deep learning driven protein design methods are relatively recent inventions and have yet to be applied to plant metabolic engineering. However, the combination of computational protein design and plant synthetic biology is being applied to other plant proteins, leading to useful products. For example, rational engineering of abscisic acid receptor enabled it to act as an organophosphate sensor in *Arabidopsis thaliana* (Park et al., 2024). Additionally, deep learning driven protein design tools generated *de novo* binders against *Magnaporthe oryzae* effectors to enhance plant immunity (Bucknell et al., 2025).

In summary, computational modelling of enzyme-ligand complexes are facilitating our understanding of the molecular determinants of enzyme activity, specificity and subcellular localisation. Simulation methods such as MD and QM/MM complement computational modelling by revealing molecular interactions in a dynamic context. In the context of engineering biology, knowledge from computational modelling is guiding the rational design of proteins with novel activities, specificities and locations.

7.3 Phylogenetic analyses reveal the evolutionary history of TXSS

The distribution pattern of individual plant secondary metabolites varies across species. While some metabolites such as β-amyrin are found in multiple monocot and eudicot lineages, others are only found within a single lineage. An example of the latter is paclitaxel, which has only been detected in the *Taxus* genus (Xie et al., 2025). Hence, the question of when and how certain plants evolved the ability to make new secondary metabolites are of great interest. Phylogenetic tools, including maximum likelihood phylogenetic analysis, positive selection analysis and molecular clock analysis, are particularly useful for exploring the evolutionary history of secondary metabolites and the enzymes that make them.

Maximum likelihood phylogenetic analysis of enzyme sequences can be used to identify gene duplication and neofunctionalisation events. In this study, although taraxasterols had previously been identified from diverse dicot lineages, genome/transcriptome mining together with phylogenic analyses revealed that, outside of the Asteraceae, taraxasterols are most likely to be minor product of mixed amyrin synthases (MASs). TXSSs, of which taraxasterols are the major product, are restricted to the Asteraceae (Chapter 3). Maximum likelihood phylogenetic analysis revealed that TXSSs are closely related to MASs and most likely arose from MASs via duplication and neofunctionalisation. This phylogenetic result was supported by the colocation of CoTXSS and a putative CoMAS in the genome. Further, the introduction of a few mutations into CoMAS led to a change in the dominant product from amyrins to taraxasterols. Similar approaches were previously used to identify a likely origin of *C. quinoa* quinoxide synthase by duplication and neofunctionalization of a β-amyrin synthase (Zhou et al., 2024).

Positive selection analysis is a powerful tool for inferring functional divergence. Positive selection is represented by high dN/dS ratio in specific lineages (where dN is the rate of non-synonymous substitutions and dS is the rate of synonymous substitutions) suggesting positive selection, potentially driving the evolution of new functions. During its expansion in plants, the OSC family experienced varied selection pressure with a mixture of purifying selection, neutral selection and positive selection on different lineages (Xue et al., 2012). However, high dN/dS ratios and signals of positive selection are observed in most eudicot and monocot OSCs, including OSCs evolved from cycloartenol synthases and *O. sativa* parkeol synthase (Xue et al., 2018a, 2012). On the other hand, all *Triticum aestivum* OSCs were found to be under purifying selection and to retain their original functions (Guo et al., 2022).

In this study, TXSSs encoded by non-basal Cichorieae were found to be likely to be under positive selection. This is likely to have contributed to a functional divergence of TXSS in this lineage, leading to the production of taraxasterol as the primary product rather than ψ-taraxasterol (Chapter 4). This is consistent with previous observations of positive selection of OSCs. Although the specific biological functions of taraxasterol and its derivatives remain unknown, it was found to mainly accumulate in the roots of Cichorioideae species (Figure 3.13). In other species, triterpenoids found in the roots have been found to be involved in shaping the microbiome of the rhizosphere and in pathogen defence (Fujimatsu et al., 2020; Huang et al., 2019; Zhong et al., 2022).

Positive selection analysis has been employed to explain changes in the substrate and product specificity of a number of plant secondary metabolic enzymes, such as iridoid synthases in three *Nepeta* species, which were shown to diverge from progesterone 5β-reductases (Lichman et al., 2020). Similarly, positive selection likely drove DODAα and CYP76AD1α found in the Caryophyllales to develop specificity for betalain biosynthesis (Brockington et al., 2015).

Bayesian phylogenetic analysis using a molecular clock and fossil calibrations can help to establish an absolute timeline for enzyme functional divergence. In this study (Chapter 3), two fossil calibrators were used to infer the divergence of TXSS from MAS, which was identified to have occurred 83 to 121 Mya, consistent with the proposed origin of Asteraceae (64 to 91 Mya) and corroborating the hypothesis that TXSSs arose in Asteraceae (Figure 3.11). Molecular clocks have previously been used to investigate the co-evolution of pathway enzymes: for example, iridoid synthases and nepetalactol-related short-chain reductase/dehydrogenases in three *Nepeta* species, which act as partners to control iridoid stereo-isomerisation, were found to evolve catalytic activities at a similar time from their corresponding ancestors (Lichman et al., 2020). The ability to infer timing therefore enables direct comparisons between different evolutionary events.

Although not used in this study, ancestral sequence construction is a useful tool to study enzyme evolutionary history and direct the rational engineering of enzymes. Ancestral sequence construction has been successfully used to explore the origins of enzymes involved in plant secondary metabolism, including lupeol and poaceatapetol synthases found in species of *Oryza*, iridoid synthases found in species of *Nepeta,* and betalain biosynthesis enzymes found in Caryophyllales (Lichman et al., 2020; Ma et al., 2024; Walker-Hale et al., 2025). In the future, it would be interesting to characterise the predicted ancestor of all TXSSs as well as the predicted ancestor of MASs and TXSSs.

Even though the biosynthetic pathways of many industrially and pharmaceutically valuable plant secondary metabolites have been discovered, far fewer studies have identified their biological function in the plant species in which they evolved. Triterpenoids have been documented to play roles in plant growth and development processes, including acting as plant hormones, controlling grain sizes, forming pollen

coats and cuticles and regulating seed germination (Dong and Qi, 2025). They are also involved in defence against pathogens and herbivores and regulate root microbiomes (Dong and Qi, 2025). *C. officinalis* is likely to have evolved in the Southwest Mediterranean (Zournatzis et al., 2025) and is thought to have spread to Northern Europe and other continents partly due to its recorded use as a medicinal herb since antiquity (Riley, 1855). As for many plants for which the range and habitats have been extensively expanded by humans, we lack knowledge of the ecology in which *C. officinalis* evolved. It is, therefore, difficult to make hypotheses about the biological functions of specific fates. Hence, the biological functions were not explicitly addressed in this thesis. However, the differences in the site of accumulation of taraxasterol compounds in the roots of non-basal Cichorieae and the floral accumulation of ψ-taraxasterol in *C. officinalis* (Figure 3.13; Figure 4.2), suggest different functions for these molecules, which is discussed further in the next section. In the future, field work in the native habitat of *Calendula sp.* might provide insights into the abiotic and biotic challenges that these species face and suggest biological functions for these metabolites.

In summary, the evolutionary history of enzymes involved in the production of plant secondary metabolites complements structural studies and furthers our understanding of structure-function relationships.

7.4 Promoter sequence analysis identifies regulators of taraxasterol production

Plant secondary metabolic pathways tend to localise to specific tissues, cells or specialised structures such as trichomes and laticifers. As noted above, genes involved in the biosynthesis of some triterpenoids, for example avenacin A-1 in *Avena strigosa* and thalianol in *Arabidopsis thaliana*, are clustered within the genome (Nützmann and Osbourn, 2014). Gene clusters are characterised by distinct chromatin modifications, including histone 3 lysine trimethylation (H3K27me3) associated with repression and histone 2 variant H2A.Z linked to activation. This chromatin landscape allows for coordinated expression of clustered genes and is proposed to enable fine tuning of cluster expression, as chromatin modifications can influence access of transcription factors (Yu et al., 2016). The expression of the five genes involved in the production of faradiol fatty acid esters (*CoTXSS*, *CoCYP1, CoCYP2*, *CoACT1* and *CoACT2*) in *C. officinalis* was found to be limited to floral tissues. However, the genes are not clustered in the genome (Golubova et al., 2025). Unclustered genes are common in many biosynthesis pathways, including those involved in the production of saponins in *Prunella vulgaris* and *Barbarea vulgaris* (Khakimov et al., 2015; Zhang et al., 2024) and may still be co-regulated, with their expression controlled by common transcription factors.

In the case of faradiol fatty acid esters, although the genes showed similar spatial expression patterns, they were not temporally co-regulated: expression of *CoTXSS, CoACT*1 and *CoACT2* all peaked early in bud development with relatively low expression at later stages, while expression of *CoCYP1* and *CoCYP2* peaked in

emerging flowers (Golubova et al., 2025). This suggests that there may be multiple transcription factors controlling the expression of the different pathway genes.

By identifying sequence motifs common to pathway genes and selecting TFs with cognate DNA-binding domains and similar expression profiles, CoMYB24 was identified as a candidate regulator of pathway genes. Functional analysis suggests that this transcription factor may be positive regulator of *CoTXSS, CoCYP2* and *CoACT1* (Figure 5.13). The ability to manipulate the expression of *CoMYB24* in *C. officinalis* would help to confirm this and could enable the future development of high-producing plant lines. However, methods for genetic manipulation of this species have yet to be developed.

A large number of Asteraceae species were identified to encode at least one *TXSS* gene (Figure 3.3). Although expression data from multiple tissues is lacking for most species, the tissues in which taxarasterol-derived compounds were found to accumulate differ across lineages (Figure 3.13). Strikingly, taraxasterol-derived compounds mainly accumulate in roots of Cichorioideae species, which is concurrent with a change in product specificity in favour of taraxasterol over ψ-taraxasterol. Hence, it is likely that changes in regulatory elements such as promoters controlling *TXSS* expression in the Cichoroideae enabled them to interact with TFs highly expressed in roots. To investigate this, well-annotated genomes and tissue-level transcriptomes of Asteraceae species that accumulate taraxasterol in roots are needed. Those resources might enable the identification of TFs that are highly expressed in roots and interact with *TXSS* promoters.

TF-driven changes in triterpenoid accumulation patterns have been previously observed. For example, a reduction in the accumulation of cucurbitacins in the fruits of cultivated *Cucumis sativus*, *C. melo* and *Citrullus lanatus* compared to wild relatives was caused by a reduction in the expression of a fruit-specific bHLH TF (Zhou et al., 2016). Variation in the accumulation patterns of taraxasterol-derived metabolites across different Asteraceae species may be a result of adaptation. The ψ-taraxasterol fatty acid compounds found in aerial tissues may contribute to cuticular wax, as previously reported for triterpenoid mixture in *Sorghum bicolor* (Busta et al., 2021). In contrast, taraxasterol compounds found in roots may potentially be involved in reshaping the soil microbiome or guarding against pathogenic microbes, as has previously been reported for root-secreted triterpenoids in *Arabidopsis thaliana*, *Glycine max* and *Cucumis melo* (Fujimatsu et al., 2020; Huang et al., 2019; Zhong et al., 2022). The observed variation in the accumulation of taraxasterol/ψ-taraxasterol in the Asteraceae is most likely to have resulted from changes in the *TXSS* regulatory sequences and, thus, differences in their interactions with TFs.

7.5 Conclusions

Plants produce a diversity of chemical compounds, many of which have industrial and pharmaceutical value. With the advancement of genome sequencing technology and bioinformatics tools, the biosynthesis pathways of many compounds have been elucidated, allowing them to be reconstructed in heterologous chassis organisms. Nevertheless, the evolutionary history and regulation of those pathways, as well as the product specificity and subcellular locations of the enzymes involved remain elusive. In this thesis, by integrating evolutionary analysis, molecular modelling and simulation and protein design with experimental validation, insights into the evolutionary history, structure-function relationship and regulation of taraxasterol synthase (TXSS) were acquired.

Exclusively found in Asteraceae, TXSSs were likely derived via gene duplication and neofunctionalization of a mixed amyrin synthase soon after the divergence of that family (Chapter 3). Within TXSSs, variation in product specificity was likely driven by positive selection acting on several Cichorioideae lineages (Chapter 4). Four residues within the active site of TXSSs likely determine TXSS product specificity (ψ-taraxasterol vs taraxasterol), with one residue contributing to reaction termination in *Calendula officinalis* TXSS (CoTXSS) (Chapter 4). CoMYB24 was found to regulate expression of *CoTXSS* and other pathway enzymes (Chapter 5). Finally, using computational protein design methods, an engineered CoTXSS variant was detached from the ER setting the scene for pathway engineering in chloroplasts (Chapter 6). Together, these findings pave the way for engineering the product specificity and subcellular localisation of TXSSs for efficient and specific heterologous production of ψ-taraxasterol and taraxasterol and enhancing the yield in native plants by manipulating transcriptional regulation.

References

Akashi, T., Furuno, T., Takahashi, T., Ayabe, S.-I., 1994. Biosynthesis of triterpenoids in cultured cells, and regenerated and wild plant organs of *Taraxacum officinale*. Phytochemistry, The International Journal of Plant Biochemistry 36, 303–308. https://doi.org/10.1016/S0031-9422(00)97065-1

Alagna, F., Reed, J., Calderini, O., Thimmappa, R., Cultrera, N.G.M., Cattivelli, A., Tagliazucchi, D., Mousavi, S., Mariotti, R., Osbourn, A., Baldoni, L., 2023. OeBAS and CYP716C67 catalyze the biosynthesis of health-beneficial triterpenoids in olive (Olea europaea) fruits. New Phytol 238, 2047–2063. https://doi.org/10.1111/nph.18863

Albanese, K.I., Petrenas, R., Pirro, F., Naudin, E.A., Borucu, U., Dawson, W.M., Scott, D.A., Leggett, G.J., Weiner, O.D., Oliver, T.A.A., Woolfson, D.N., 2024. Rationally seeded computational protein design of α-helical barrels. Nat Chem Biol 20, 991–999. https://doi.org/10.1038/s41589-024-01642-0

Almeida, A., Dong, L., Khakimov, B., Bassard, J.-E., Moses, T., Lota, F., Goossens, A., Appendino, G., Bak, S., 2018. A Single Oxidosqualene Cyclase Produces the Seco-Triterpenoid α-Onocerin. Plant Physiol 176, 1469–1484. https://doi.org/10.1104/pp.17.01369

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Brière, C., Owens, G.L., Carrère, S., Mayjonade, B., Legrand, L., Gill, N., Kane, N.C., Bowers, J.E., Hubner, S., Bellec, A., Bérard, A., Bergès, H., Blanchet, N., Boniface, M.-C., Brunel, D., Catrice, O., Chaidir, N., Claudel, C., Donnadieu, C., Faraut, T., Fievet, G., Helmstetter, N., King, M., Knapp, S.J., Lai, Z., Le Paslier, M.-C., Lippi, Y., Lorenzon, L., Mandel, J.R., Marage, G., Marchand, G., Marquand, E., Bret-Mestries, E., Morien, E., Nambeesan, S., Nguyen, T., Pegot-Espagnet, P., Pouilly, N., Raftis, F., Sallet, E., Schiex, T., Thomas, J., Vandecasteele, C., Varès, D., Vear, F., Vautrin, S., Crespi, M., Mangin, B., Burke, J.M., Salse, J., Muños, S., Vincourt, P., Rieseberg, L.H., Langlade, N.B., 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546, 148–152. https://doi.org/10.1038/nature22380

Bally, J., Nakasugi, K., Jia, F., Jung, H., Ho, S.Y.W., Wong, M., Paul, C.M., Naim, F., Wood, C.C., Crowhurst, R.N., Hellens, R.P., Dale, J.L., Waterhouse, P.M., 2015. The extremophile Nicotiana benthamiana has traded viral defence for early vigour. Nat Plants 1, 15165. https://doi.org/10.1038/nplants.2015.165

Barba, F.J., Nikmaram, N., Roohinejad, S., Khelfa, A., Zhu, Z., Koubaa, M., 2016. Bioavailability of Glucosinolates and Their Breakdown Products: Impact of Processing. Front Nutr 3, 24. https://doi.org/10.3389/fnut.2016.00024

Barco, B., Clay, N.K., 2019. Evolution of Glucosinolate Diversity via Whole-Genome Duplications, Gene Rearrangements, and Substrate Promiscuity. Annu Rev Plant Biol 70, 585–604. https://doi.org/10.1146/annurev-arplant-050718-100152

Bargmann, B.O.R., Marshall-Colon, A., Efroni, I., Ruffel, S., Birnbaum, K.D., Coruzzi, G.M., Krouk, G., 2013. TARGET: A Transient Transformation System for Genome-Wide Transcription Factor Target Discovery. Mol Plant 6, 978–980. https://doi.org/10.1093/mp/sst010

Barkman, T.J., Martins, T.R., Sutton, E., Stout, J.T., 2007. Positive Selection for Single Amino Acid Change Promotes Substrate Discrimination of a Plant

Volatile-Producing Enzyme. Mol Biol Evol 24, 1320–1329. https://doi.org/10.1093/molbev/msm053
Barreda, V.D., Palazzesi, L., Tellería, M.C., Olivero, E.B., Raine, J.I., Forest, F., 2015. Early evolution of the angiosperm clade Asteraceae in the Cretaceous of Antarctica. Proceedings of the National Academy of Sciences 112, 10989–10994. https://doi.org/10.1073/pnas.1423653112
Baumgart, L.A., Greenblum, S.I., Morales-Cruz, A., Wang, P., Zhang, Y., Yang, L., Chen, C., Dilworth, D.J., Garretson, A.C., Grosjean, N., He, G., Savage, E., Yoshinaga, Y., Blaby, I.K., Daum, C.G., O'Malley, R.C., 2025. Recruitment, rewiring and deep conservation in flowering plant gene regulation. Nat Plants 11, 1514–1527. https://doi.org/10.1038/s41477-025-02047-0
Bian, C., Demirer, G.S., Oz, M.T., Cai, Y.-M., Witham, S., Mason, G.A., Di, Z., Deligne, F., Zhang, P., Shen, R., Gaudinier, A., Brady, S.M., Patron, N.J., 2025. Conservation and divergence of regulatory architecture in nitrate-responsive plant gene circuits. Plant Cell 37, koaf124. https://doi.org/10.1093/plcell/koaf124
Bibik, J.D., Weraduwage, S.M., Banerjee, A., Robertson, K., Espinoza-Corral, R., Sharkey, T.D., Lundquist, P.K., Hamberger, B.R., 2022. Pathway Engineering, Re-targeting, and Synthetic Scaffolding Improve the Production of Squalene in Plants. ACS Synth. Biol. 11, 2121–2133. https://doi.org/10.1021/acssynbio.2c00051
Birchler, J.A., Yang, H., 2022. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Plant Cell 34, 2466–2474. https://doi.org/10.1093/plcell/koac076
Böttner, L., Malacrinò, A., Schulze Gronover, C., van Deenen, N., Müller, B., Xu, S., Gershenzon, J., Prüfer, D., Huber, M., 2023. Natural rubber reduces herbivory and alters the microbiome below ground. New Phytologist 239, 1475–1489. https://doi.org/10.1111/nph.18709
Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10, e1003537. https://doi.org/10.1371/journal.pcbi.1003537
Bremer, K., 1996. Major clades and grades of the Asteraceae. Compositae: systematics 1, 1–7.
Bremer, K., Jansen, R.K., 1992. A New Subfamily of the Asteraceae. Annals of the Missouri Botanical Garden 79, 414. https://doi.org/10.2307/2399777
Brockington, S.F., Yang, Y., Gandia-Herrero, F., Covshoff, S., Hibberd, J.M., Sage, R.F., Wong, G.K.S., Moore, M.J., Smith, S.A., 2015. Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. New Phytologist 207, 1170–1180. https://doi.org/10.1111/nph.13441
Brooks, E.G., Elorriaga, E., Liu, Y., Duduit, J.R., Yuan, G., Tsai, C.-J., Tuskan, G.A., Ranney, T.G., Yang, X., Liu, W., 2023. Plant Promoters and Terminators for High-Precision Bioengineering. Biodesign Research 5, 0013. https://doi.org/10.34133/bdr.0013
Bucknell, A.H., Xi, Y., Knight, G., Gentle, A., Ryder, L.S., Yan, X., Watson, J.L., Emmrich, P.M.F., Talbot, N.J., Banfield, M.J., Bentham, A.R., 2025. Engineering new-to-nature immune sensors for novel recognition of Magnaporthe oryzae. https://doi.org/10.5281/zenodo.15880524

Burgess, R.R., 2009. Refolding solubilized inclusion body proteins. Methods Enzymol 463, 259–282. https://doi.org/10.1016/S0076-6879(09)63017-2
Burke, C.C., Wildung, M.R., Croteau, R., 1999. Geranyl diphosphate synthase: Cloning, expression, and characterization of this prenyltransferase as a heterodimer. Proceedings of the National Academy of Sciences 96, 13062–13067. https://doi.org/10.1073/pnas.96.23.13062
Busta, L., Schmitz, E., Kosma, D.K., Schnable, J.C., Cahoon, E.B., 2021. A co-opted steroid synthesis gene, maintained in sorghum but not maize, is associated with a divergence in leaf wax chemistry. Proc Natl Acad Sci U S A 118, e2022982118. https://doi.org/10.1073/pnas.2022982118
Cai, Y.-M., Witham, S., Patron, N.J., 2023. Tuning Plant Promoters Using a Simple Split Luciferase Method to Assess Transcription Factor-DNA Interactions. ACS Synth. Biol. 12, 3482–3486. https://doi.org/10.1021/acssynbio.3c00094
Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348
Carlquist, S., 1976. Tribal Interrelationships and Phylogeny of the Asteraceae. Aliso: A Journal of Systematic and Floristic Botany 8, 465–492.
Carpenter, E.P., Beis, K., Cameron, A.D., Iwata, S., 2008. Overcoming the challenges of membrane protein crystallography. Curr Opin Struct Biol 18, 581–586. https://doi.org/10.1016/j.sbi.2008.07.001
Case, D., Aktulga, H.M., Belfon, K., Ben-Shalom, I., Brozell, S., Cerutti, D., Cheatham, T., Cisneros, G.A., Cruzeiro, V., Darden, T., Duke, R., Giambasu, G., Gilson, M., Gohlke, H., Götz, A., Harris, R., Izadi, S., Izmailov, S., Kollman, P., 2022. Amber 2022. https://doi.org/10.13140/RG.2.2.31337.77924
Chappell, J., 2002. The genetics and molecular genetics of terpene and sterol origami. Current Opinion in Plant Biology 5, 151–157. https://doi.org/10.1016/S1369-5266(02)00241-8
Chen, H., Li, G., Köllner, T.G., Jia, Q., Gershenzon, J., Chen, F., 2014. Positive Darwinian selection is a driving force for the diversification of terpenoid biosynthesis in the genus Oryza. BMC Plant Biol 14, 239. https://doi.org/10.1186/s12870-014-0239-x
Chen, K., Zhang, M., Xu, L., Yi, Y., Wang, L., Wang, H., Wang, Z., Xing, J., Li, P., Zhang, X., Shi, X., Ye, M., Osbourn, A., Qiao, X., 2022. Identification of oxidosqualene cyclases associated with saponin biosynthesis from Astragalus membranaceus reveals a conserved motif important for catalytic function. J Adv Res 43, 247–257. https://doi.org/10.1016/j.jare.2022.03.014
Chen, K., Zhang, M., Ye, M., Qiao, X., 2021. Site-directed mutagenesis and substrate compatibility to reveal the structure–function relationships of plant oxidosqualene cyclases. Nat. Prod. Rep. 38, 2261–2275. https://doi.org/10.1039/D1NP00015B
Chezem, W.R., Clay, N.K., 2016. Regulation of Plant Secondary Metabolism and Associated Specialized Cell Development by MYBs and bHLHs. Phytochemistry 131, 26–43. https://doi.org/10.1016/j.phytochem.2016.08.006
Choi, H.S., Han, J.Y., Choi, Y.E., 2020. Identification of triterpenes and functional characterization of oxidosqualene cyclases involved in triterpene biosynthesis in lettuce (*Lactuca sativa*). Plant Science 301, 110656. https://doi.org/10.1016/j.plantsci.2020.110656
Chupeau, M.-C., Bellini, C., Guerche, P., Maisonneuve, B., Vastra, G., Chupeau, Y., 1989. Transgenic Plants of Lettuce (Lactuca sativa) Obtained Through

Electroporation of Protoplasts. Nat Biotechnol 7, 503–508. https://doi.org/10.1038/nbt0589-503
Colombo, E., Sangiovanni, E., D'Ambrosio, M., Bosisio, E., Ciocarlan, A., Fumagalli, M., Guerriero, A., Harghel, P., Dell'Agli, M., 2015. A Bio-Guided Fractionation to Assess the Inhibitory Activity of Calendula officinalis L. on the NF-κB Driven Transcription in Human Gastric Epithelial Cells. Evidence-Based Complementary and Alternative Medicine 2015, 727342. https://doi.org/10.1155/2015/727342
Darling, A.E., Mau, B., Perna, N.T., 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE 5, e11147. https://doi.org/10.1371/journal.pone.0011147
Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Courbet, A., de Haas, R.J., Bethel, N., Leung, P.J.Y., Huddy, T.F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A.K., King, N.P., Baker, D., 2022. Robust deep learning–based protein sequence design using ProteinMPNN. Science 378, 49–56. https://doi.org/10.1126/science.add2187
Diaz, J.E., Lin, C.-S., Kunishiro, K., Feld, B.K., Avrantinis, S.K., Bronson, J., Greaves, J., Saven, J.G., Weiss, G.A., 2011. Computational design and selections for an engineered, thermostable terpene synthase. Protein Sci 20, 1597–1606. https://doi.org/10.1002/pro.691
Dokarry, M., 2010. β-amyrin synthase: Investigating the structure and function of an oxidosqualene cyclase involved in disease resistance in oats (doctoral). University of East Anglia. School of Biological Sciences.
Doležel, J., Greilhuber, J., Suda, J., 2007. Estimation of nuclear DNA content in plants using flow cytometry. Nat Protoc 2, 2233–2244. https://doi.org/10.1038/nprot.2007.310
Dong, H., Qi, X., 2025. Biosynthesis of triterpenoids in plants: Pathways, regulation, and biological functions. Current Opinion in Plant Biology 85, 102701. https://doi.org/10.1016/j.pbi.2025.102701
Dong, L., Almeida, A., Pollier, J., Khakimov, B., Bassard, J.-E., Miettinen, K., Stærk, D., Mehran, R., Olsen, C.E., Motawia, M.S., Goossens, A., Bak, S., 2021. An Independent Evolutionary Origin for Insect Deterrent Cucurbitacins in Iberis amara. Mol Biol Evol 38, 4659–4673. https://doi.org/10.1093/molbev/msab213
Dong, L., Jongedijk, E., Bouwmeester, H., Van Der Krol, A., 2016. Monoterpene biosynthesis potential of plant subcellular compartments. New Phytologist 209, 679–690. https://doi.org/10.1111/nph.13629
D. Rudolf, J., Chang, C.-Y., 2020. Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases. Natural Product Reports 37, 425–463. https://doi.org/10.1039/C9NP00051H
Dudley, Q.M., Jo, S., Guerrero, D.A.S., Chhetry, M., Smedley, M.A., Harwood, W.A., Sherden, N.H., O'Connor, S.E., Caputi, L., Patron, N.J., 2022. Reconstitution of monoterpene indole alkaloid biosynthesis in genome engineered Nicotiana benthamiana. Commun Biol 5, 949. https://doi.org/10.1038/s42003-022-03904-w
Duret, L., 2008. Neutral Theory: The Null Hypothesis of Molecular Evolution. Nature Education 1, 218.
Edger, P.P., Heidel-Fischer, H.M., Bekaert, M., Rota, J., Glöckner, G., Platts, A.E., Heckel, D.G., Der, J.P., Wafula, E.K., Tang, M., Hofberger, J.A., Smithson, A., Hall, J.C., Blanchette, M., Bureau, T.E., Wright, S.I., dePamphilis, C.W., Eric

Schranz, M., Barker, M.S., Conant, G.C., Wahlberg, N., Vogel, H., Pires, J.C., Wheat, C.W., 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proceedings of the National Academy of Sciences 112, 8362–8366. https://doi.org/10.1073/pnas.1503926112
Engler, C., Youles, M., Gruetzner, R., Ehnert, T.-M., Werner, S., Jones, J.D.G., Patron, N.J., Marillonnet, S., 2014. A Golden Gate Modular Cloning Toolbox for Plants. ACS Synth. Biol. 3, 839–843. https://doi.org/10.1021/sb4001504
Falginella, L., Andre, C.M., Legay, S., Lin-Wang, K., Dare, A.P., Deng, C., Rebstock, R., Plunkett, B.J., Guo, L., Cipriani, G., Espley, R.V., 2021. Differential regulation of triterpene biosynthesis induced by an early failure in cuticle formation in apple. Horticulture Research 8, 75. https://doi.org/10.1038/s41438-021-00511-4
Fawcett, J.A., Maere, S., Van de Peer, Y., 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. Proceedings of the National Academy of Sciences 106, 5737–5742. https://doi.org/10.1073/pnas.0900906106
Fazio, G.C., Xu, R., Matsuda, S.P.T., 2004. Genome Mining To Identify New Plant Triterpenoids. J. Am. Chem. Soc. 126, 5678–5679. https://doi.org/10.1021/ja0318784
Fleishman, S.J., Leaver-Fay, A., Corn, J.E., Strauch, E.-M., Khare, S.D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., Baker, D., 2011. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. PLOS ONE 6, e20161. https://doi.org/10.1371/journal.pone.0020161
Forestier, E.C.F., Czechowski, T., Cording, A.C., Gilday, A.D., King, A.J., Brown, G.D., Graham, I.A., 2021. Developing a Nicotiana benthamiana transgenic platform for high-value diterpene production and candidate gene evaluation. Plant Biotechnology Journal 19, 1614–1623. https://doi.org/10.1111/pbi.13574
Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grün, S., Winklmair, A., Eisenreich, W., Bacher, A., Meeley, R.B., Briggs, S.P., Simcox, K., Gierl, A., 1997. Analysis of a Chemical Plant Defense Mechanism in Grasses. Science 277, 696–699. https://doi.org/10.1126/science.277.5326.696
Frisch, M. ea, Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, Ma., Cheeseman, J.R., Scalmani, G., Barone, V., Petersson, G.A., Nakatsuji, H., 2016. Gaussian 16. Gaussian, Inc. Wallingford, CT.
Fujimatsu, T., Endo, K., Yazaki, K., Sugiyama, A., 2020. Secretion dynamics of soyasaponins in soybean roots and effects to modify the bacterial composition. Plant Direct 4, e00259. https://doi.org/10.1002/pld3.259
Gallage, N.J., Hansen, E.H., Kannangara, R., Olsen, C.E., Motawia, M.S., Jørgensen, K., Holme, I., Hebelstrup, K., Grisoni, M., Møller, B.L., 2014. Vanillin formation from ferulic acid in Vanilla planifolia is catalysed by a single enzyme. Nat Commun 5, 4037. https://doi.org/10.1038/ncomms5037
Geisler, K., Hughes, R.K., Sainsbury, F., Lomonossoff, G.P., Rejzek, M., Fairhurst, S., Olsen, C.-E., Motawia, M.S., Melton, R.E., Hemmings, A.M., Bak, S., Osbourn, A., 2013. Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants. Proc Natl Acad Sci U S A 110, E3360-3367. https://doi.org/10.1073/pnas.1309157110
Glover, B.J., Martin, C., 2012. Anthocyanins. Current Biology 22, R147–R150. https://doi.org/10.1016/j.cub.2012.01.021

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11, 725–736. https://doi.org/10.1093/oxfordjournals.molbev.a040153
Golubova, D., Salmon, M., Su, H., Tansley, C., Kaithakottil, G.G., Linsmith, G., Schudoma, C., Swarbreck, D., O'Connell, M.A., Patron, N.J., 2025. Biosynthesis and bioactivity of anti-inflammatory triterpenoids in Calendula officinalis. Nat Commun 16, 6941. https://doi.org/10.1038/s41467-025-62269-w
Golubova, D., Tansley, C., Su, H., Patron, N.J., 2024. Engineering *Nicotiana benthamiana* as a platform for natural product biosynthesis. Current Opinion in Plant Biology 81, 102611. https://doi.org/10.1016/j.pbi.2024.102611
Goverde, C.A., Pacesa, M., Goldbach, N., Dornfeld, L.J., Balbi, P.E.M., Georgeon, S., Rosset, S., Kapoor, S., Choudhury, J., Dauparas, J., Schellhaas, C., Kozlov, S., Baker, D., Ovchinnikov, S., Vecchio, A.J., Correia, B.E., 2024. Computational design of soluble and functional membrane protein analogues. Nature 631, 449–458. https://doi.org/10.1038/s41586-024-07601-y
Grandi, F.C., Modi, H., Kampman, L., Corces, M.R., 2022. Chromatin accessibility profiling by ATAC-seq. Nat Protoc 17, 1518–1552. https://doi.org/10.1038/s41596-022-00692-9
Grant, C.E., Bailey, T.L., Noble, W.S., 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064
Günther, J., Erthmann, P.Ø., Khakimov, B., Bak, S., 2021. Reciprocal mutations of two multifunctional β-amyrin synthases from Barbarea vulgaris shift α/β-amyrin ratios. Plant Physiol 188, 1483–1495. https://doi.org/10.1093/plphys/kiab545
Guo, C.-F., Xiong, X.-C., Dong, H., Qi, X.-Q., 2022. Genome-wide investigation and transcriptional profiling of the oxidosqualene cyclase (OSC) genes in wheat (Triticum aestivum). Journal of Systematics and Evolution 60, 1378–1392. https://doi.org/10.1111/jse.12738
Guo, Y., Guo, Z., Zhong, J., Liang, Y., Feng, Y., Zhang, P., Zhang, Q., Sun, M., 2023. Positive regulatory role of R2R3 MYBs in terpene biosynthesis in *Lilium* 'Siberia.' Horticultural Plant Journal 9, 1024–1038. https://doi.org/10.1016/j.hpj.2023.05.004
Gut, J.A., Lemmin, T., 2025. Dissecting AlphaFold2's capabilities with limited sequence information. Bioinform Adv 5, vbae187. https://doi.org/10.1093/bioadv/vbae187
Han, J., Miller, E.P., Li, S., 2024. Cutting-edge plant natural product pathway elucidation. Current Opinion in Biotechnology 87, 103137. https://doi.org/10.1016/j.copbio.2024.103137
Han, J.Y., Jo, H.-J., Kwon, E.K., Choi, Y.E., 2019. Cloning and Characterization of Oxidosqualene Cyclases Involved in Taraxasterol, Taraxerol and Bauerenol Triterpene Biosynthesis in Taraxacum coreanum. Plant Cell Physiol 60, 1595–1603. https://doi.org/10.1093/pcp/pcz062
Handelman, C., Kohn, J.R., 2014. Hummingbird color preference within a natural hybrid population of Mimulus aurantiacus (Phrymaceae). Plant Species Biology 29, 65–72. https://doi.org/10.1111/j.1442-1984.2012.00393.x
Hatlestad, G.J., Akhavan, N.A., Sunnadeniya, R.M., Elam, L., Cargile, S., Hembd, A., Gonzalez, A., McGrath, J.M., Lloyd, A.M., 2015. The beet Y locus encodes

an anthocyanin MYB-like protein that activates the betalain red pigment pathway. Nat Genet 47, 92–96. https://doi.org/10.1038/ng.3163
Hausjell, J., Halbwirth, H., Spadiut, O., 2018. Recombinant production of eukaryotic cytochrome P450s in microbial cell factories. Biosci Rep 38, BSR20171290. https://doi.org/10.1042/BSR20171290
Hegedűs, T., Geisler, M., Lukács, G.L., Farkas, B., 2022. Ins and outs of AlphaFold2 transmembrane protein structure predictions. Cell. Mol. Life Sci. 79, 73. https://doi.org/10.1007/s00018-021-04112-1
Hellman, L.M., Fried, M.G., 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. Nat Protoc 2, 1849–1861. https://doi.org/10.1038/nprot.2007.249
Henriques de Jesus, M.P.R., Zygadlo Nielsen, A., Busck Mellor, S., Matthes, A., Burow, M., Robinson, C., Erik Jensen, P., 2017. Tat proteins as novel thylakoid membrane anchors organize a biosynthetic pathway in chloroplasts and increase product yield 5-fold. Metab Eng 44, 108–116. https://doi.org/10.1016/j.ymben.2017.09.014
Himshikha, Gupta, R.C., Kumar, R., Singhal, V.K., 2017. Cytomixis and Intraspecific Polyploidy (2x, 4x) in Inula grandiflora Willd. from Malana Valley, Kullu District, Himachal Pradesh. CYTOLOGIA 82, 273–278. https://doi.org/10.1508/cytologia.82.273
Hooker, J.J., Collinson, M.E., Sille, N.P., 2004. Eocene–Oligocene mammalian faunal turnover in the Hampshire Basin, UK: calibration to the global time scale and the major cooling event. Journal of the Geological Society 161, 161–172. https://doi.org/10.1144/0016-764903-091
Hoshino, T., Nakagawa, K., Aiba, Y., Itoh, D., Nakada, C., Masukawa, Y., 2017. Euphorbia tirucalli β-Amyrin Synthase: Critical Roles of Steric Sizes at Val483 and Met729 and the CH–π Interaction between Val483 and Trp534 for Catalytic Action. ChemBioChem 18, 2145–2155. https://doi.org/10.1002/cbic.201700368
Howes, M.-J.R., Quave, C.L., Collemare, J., Tatsis, E.C., Twilley, D., Lulekal, E., Farlow, A., Li, L., Cazar, M.-E., Leaman, D.J., Prescott, T.A.K., Milliken, W., Martin, C., De Canha, M.N., Lall, N., Qin, H., Walker, B.E., Vásquez-Londoño, C., Allkin, B., Rivers, M., Simmonds, M.S.J., Bell, E., Battison, A., Felix, J., Forest, F., Leon, C., Williams, C., Nic Lughadha, E., 2020. Molecules from nature: Reconciling biodiversity conservation and global healthcare imperatives for sustainable use of medicinal plants and fungi. PLANTS, PEOPLE, PLANET 2, 463–481. https://doi.org/10.1002/ppp3.10138
Huang, A.C., Jiang, T., Liu, Y.-X., Bai, Y.-C., Reed, J., Qu, B., Goossens, A., Nützmann, H.-W., Bai, Y., Osbourn, A., 2019. A specialized metabolic network selectively modulates Arabidopsis root microbiota. Science 364, eaau6389. https://doi.org/10.1126/science.aau6389
Huang, C.-H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., Ma, H., 2016. Multiple Polyploidization Events across Asteraceae with Two Nested Events in the Early History Revealed by Nuclear Phylogenomics. Mol Biol Evol 33, 2820–2835. https://doi.org/10.1093/molbev/msw157
Huang, T., Jander, G., de Vos, M., 2011. Non-protein amino acids in plant defense against insect herbivores: representative cases and opportunities for further functional analysis. Phytochemistry 72, 1531–1537. https://doi.org/10.1016/j.phytochem.2011.03.019

Jansen, R.K., Palmer, J.D., 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proceedings of the National Academy of Sciences 84, 5818–5822. https://doi.org/10.1073/pnas.84.16.5818
Jia, L., Mao, Y., Ji, Q., Dersh, D., Yewdell, J.W., Qian, S.-B., 2020. Decoding mRNA translatability and stability from the 5' UTR. Nat Struct Mol Biol 27, 814–821. https://doi.org/10.1038/s41594-020-0465-x
Jian, W., Cao, H., Yuan, S., Liu, Y., Lu, J., Lu, W., Li, N., Wang, J., Zou, J., Tang, N., Xu, C., Cheng, Y., Gao, Y., Xi, W., Bouzayen, M., Li, Z., 2019. SlMYB75, an MYB-type transcription factor, promotes anthocyanin accumulation and enhances volatile aroma production in tomato fruits. Hortic Res 6, 22. https://doi.org/10.1038/s41438-018-0098-y
Jiao, F., Tan, Z., Yu, Z., Zhou, B., Meng, L., Shi, X., 2022. The phytochemical and pharmacological profile of taraxasterol. Front. Pharmacol. 13. https://doi.org/10.3389/fphar.2022.927365
Johnson, L.A., Allemann, R.K., 2025. Engineering terpene synthases and their substrates for the biocatalytic production of terpene natural products and analogues. Chem. Commun. 61, 2468–2483. https://doi.org/10.1039/D4CC05785F
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2
Katinas, L., Pruski, J., Sancho, G., Tellería, M.C., 2008. The Subfamily Mutisioideae (Asteraceae). Bot. Rev 74, 469–716. https://doi.org/10.1007/s12229-008-9016-6
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772–780. https://doi.org/10.1093/molbev/mst010
Khakimov, B., Kuzina, V., Erthmann, P.Ø., Fukushima, E.O., Augustin, J.M., Olsen, C.E., Scholtalbers, J., Volpin, H., Andersen, S.B., Hauser, T.P., Muranaka, T., Bak, S., 2015. Identification and genome organization of saponin pathway genes from a wild crucifer, and their use for transient production of saponins in Nicotiana benthamiana. Plant J 84, 478–490. https://doi.org/10.1111/tpj.13012
Khoo, H.E., Azlan, A., Tang, S.T., Lim, S.M., 2017. Anthocyanidins and anthocyanins: colored pigments as food, pharmaceutical ingredients, and the potential health benefits. Food Nutr Res 61, 1361779. https://doi.org/10.1080/16546628.2017.1361779
Kidder, B.L., Hu, G., Zhao, K., 2011. ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol 12, 918–922. https://doi.org/10.1038/ni.2117
Kilian, N., Gemeinholzer, B., Lack, H., 2009. Cichorieae, in: Systematics, Evolution, and Biogeography of Compositae. pp. 343–383.

Kim, K.J., Jansen, R.K., 1995. ndhF sequence evolution and the major clades in the sunflower family. Proceedings of the National Academy of Sciences 92, 10379–10383. https://doi.org/10.1073/pnas.92.22.10379
Kimura, M., 1968. Evolutionary Rate at the Molecular Level. Nature 217, 624–626. https://doi.org/10.1038/217624a0
King, B.R., Sumida, K.H., Caruso, J.L., Baker, D., Zalatan, J.G., 2025. Computational Stabilization of a Non-Heme Iron Enzyme Enables Efficient Evolution of New Function. Angew Chem Int Ed Engl 64, e202414705. https://doi.org/10.1002/anie.202414705
Klemm, S.L., Shipony, Z., Greenleaf, W.J., 2019. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 20, 207–220. https://doi.org/10.1038/s41576-018-0089-8
Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., Kollman, P.A., 1992. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. Journal of Computational Chemistry 13, 1011–1021. https://doi.org/10.1002/jcc.540130812
Kurbidaeva, A., Purugganan, M., 2021. Insulators in Plants: Progress and Open Questions. Genes (Basel) 12, 1422. https://doi.org/10.3390/genes12091422
Lamb, D.C., Jackson, C.J., Warrilow, A.G.S., Manning, N.J., Kelly, D.E., Kelly, S.L., 2007. Lanosterol biosynthesis in the prokaryote Methylococcus capsulatus: insight into the evolution of sterol biosynthesis. Mol Biol Evol 24, 1714–1721. https://doi.org/10.1093/molbev/msm090
Lan, T., Wang, X.-R., Zeng, Q.-Y., 2013. Structural and Functional Evolution of Positively Selected Sites in Pine Glutathione S-Transferase Enzyme Family. J Biol Chem 288, 24441–24451. https://doi.org/10.1074/jbc.M113.456863
Laranjeira, S., Amorim-Silva, V., Esteban, A., Arró, M., Ferrer, A., Tavares, R.M., Botella, M.A., Rosado, A., Azevedo, H., 2015. *Arabidopsis* Squalene Epoxidase 3 (SQE3) Complements SQE1 and Is Important for Embryo Development and Bulk Squalene Epoxidase Activity. Molecular Plant 8, 1090–1102. https://doi.org/10.1016/j.molp.2015.02.007
Lee, J.W., Heo, W., Lee, J., Jin, N., Yoon, S.M., Park, K.Y., Kim, E.Y., Kim, W.T., Kim, J.Y., 2018. The B cell death function of obinutuzumab-HDEL produced in plant (Nicotiana benthamiana L.) is equivalent to obinutuzumab produced in CHO cells. PLOS ONE 13, e0191075. https://doi.org/10.1371/journal.pone.0191075
Leferink, N.G.H., Escorcia, A.M., Ouwersloot, B.R., Johanissen, L.O., Hay, S., van der Kamp, M.W., Scrutton, N.S., 2022. Molecular Determinants of Carbocation Cyclisation in Bacterial Monoterpene Synthases. ChemBioChem 23, e202100688. https://doi.org/10.1002/cbic.202100688
Letunic, I., Bork, P., 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res 52, W78–W82. https://doi.org/10.1093/nar/gkae268
Lewis Smith, R.I., Richardson, M., 2011. Fuegian plants in Antarctica: natural or anthropogenically assisted immigrants? Biological Invasions 13, 1–5. https://doi.org/10.1007/s10530-010-9784-x
Li, J., Mutanda, I., Wang, K., Yang, L., Wang, J., Wang, Y., 2019. Chloroplastic metabolic engineering coupled with isoprenoid pool enhancement for committed taxanes biosynthesis in Nicotiana benthamiana. Nat Commun 10, 4850. https://doi.org/10.1038/s41467-019-12879-y

Liang, F., Xie, Y., Zhang, C., Zhao, Y., Motawia, M.S., Kampranis, S.C., 2025. Elucidation of the final steps in Taxol biosynthesis and its biotechnological production. Nat. Synth 1–11. https://doi.org/10.1038/s44160-025-00800-z
Liang, M., Zhang, F., Xu, J., Wang, X., Wu, R., Xue, Z., 2022. A conserved mechanism affecting hydride shifting and deprotonation in the synthesis of hopane triterpenes as compositions of wax in oat. Proceedings of the National Academy of Sciences 119, e2118709119. https://doi.org/10.1073/pnas.2118709119
Lichman, B.R., Godden, G.T., Hamilton, J.P., Palmer, L., Kamileen, M.O., Zhao, D., Vaillancourt, B., Wood, J.C., Sun, M., Kinser, T.J., Henry, L.K., Rodriguez-Lopez, C., Dudareva, N., Soltis, D.E., Soltis, P.S., Buell, C.R., O'Connor, S.E., 2020. The evolutionary origins of the cat attractant nepetalactone in catnip. Sci Adv 6, eaba0721. https://doi.org/10.1126/sciadv.aba0721
Lips, D., 2021. Practical considerations for delivering on the sustainability promise of fermentation-based biomanufacturing. Emerg Top Life Sci 5, 711–715. https://doi.org/10.1042/ETLS20210129
Liu, H., Naismith, J.H., 2008. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. BMC Biotechnol 8, 91. https://doi.org/10.1186/1472-6750-8-91
Liu, Q., Manzano, D., Tanić, N., Pesic, M., Bankovic, J., Pateraki, I., Ricard, L., Ferrer, A., de Vos, R., van de Krol, S., Bouwmeester, H., 2014. Elucidation and in planta reconstitution of the parthenolide biosynthetic pathway. Metab Eng 23, 145–153. https://doi.org/10.1016/j.ymben.2014.03.005
Lodeiro, S., Xiong, Q., Wilson, W.K., Kolesnikova, M.D., Onak, C.S., Matsuda, S.P.T., 2007. An Oxidosqualene Cyclase Makes Numerous Products by Diverse Mechanisms: A Challenge to Prevailing Concepts of Triterpene Biosynthesis. J. Am. Chem. Soc. 129, 11213–11222. https://doi.org/10.1021/ja073133u
Lomonossoff, G.P., D'Aoust, M.-A., 2016. Plant-produced biopharmaceuticals: A case of technical developments driving clinical deployment. Science 353, 1237–1240. https://doi.org/10.1126/science.aaf6638
López-Vinyallonga, S., Soriano, I., Susanna, A., Montserra, J.M., Roquet, C., Garcia-Jacas, N., 2015. The Polyploid Series of the Achillea millefolium Aggregate in the Iberian Peninsula Investigated Using Microsatellites. PLoS One 10, e0129861. https://doi.org/10.1371/journal.pone.0129861
Lu, X., Huang, L., Scheller, H.V., Keasling, J.D., 2023. Medicinal terpenoid UDP-glycosyltransferases in plants: recent advances and research strategies. J Exp Bot 74, 1343–1357. https://doi.org/10.1093/jxb/erac505
Luo, Y., Ma, X., Qiu, Y., Lu, Y., Shen, S., Li, Y., Gao, H., Chen, K., Zhou, J., Hu, T., Tu, L., Zhao, H., Li, D., Leng, F., Gao, W., Jiang, T., Liu, C., Huang, L., Wu, R., Tong, Y., 2023. Structural and Catalytic Insight into the Unique Pentacyclic Triterpene Synthase TwOSC. Angew Chem Int Ed Engl 62, e202313429. https://doi.org/10.1002/anie.202313429
Ma, A., Sun, J., Feng, L., Xue, Z., Wu, W., Song, B., Xiong, X., Wang, X., Han, B., Osbourn, A., Qi, X., 2024. Functional diversity of oxidosqualene cyclases in genus Oryza. New Phytologist 244, 2430–2441. https://doi.org/10.1111/nph.20175
Mandel, J.R., Dikow, R.B., Siniscalchi, C.M., Thapa, R., Watson, L.E., Funk, V.A., 2019. A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. Proceedings

of the National Academy of Sciences 116, 14083–14088. https://doi.org/10.1073/pnas.1903871116
Mathews, H., Clendennen, S.K., Caldwell, C.G., Liu, X.L., Connors, K., Matheis, N., Schuster, D.K., Menasco, D.J., Wagoner, W., Lightner, J., Wagner, D.R., 2003. Activation tagging in tomato identifies a transcriptional regulator of anthocyanin biosynthesis, modification, and transport. Plant Cell 15, 1689–1703. https://doi.org/10.1105/tpc.012963
McClune, C.J., Liu, J.C.-T., Wick, C., De La Peña, R., Lange, B.M., Fordyce, P.M., Sattely, E.S., 2025. Discovery of FoTO1 and Taxol genes enables biosynthesis of baccatin III. Nature 643, 582–592. https://doi.org/10.1038/s41586-025-09090-z
Mertens, J., Pollier, J., Vanden Bossche, R., Lopez-Vidriero, I., Franco-Zorrilla, J.M., Goossens, A., 2016. The bHLH Transcription Factors TSAR1 and TSAR2 Regulate Triterpene Saponin Biosynthesis in Medicago truncatula1[OPEN]. Plant Physiol 170, 194–210. https://doi.org/10.1104/pp.15.01645
Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30, 1188–1195. https://doi.org/10.1093/molbev/mst024
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015
Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Res 49, D412–D419. https://doi.org/10.1093/nar/gkaa913
Moore, B.D., Andrew, R.L., Külheim, C., Foley, W.J., 2014. Explaining intraspecific diversity in plant secondary metabolites in an ecological context. New Phytologist 201, 733–750. https://doi.org/10.1111/nph.12526
Mudassir Jeelani, S., Shahnawaz, Mohd., Prakash Gupta, A., Lattoo, S.K., 2022. Phytochemical Diversity in Relation to Cytogenetic Variability in Inula racemosa Hook.f., an Endangered Medicinal Plant of Himalayas. Chemistry & Biodiversity 19, e202200486. https://doi.org/10.1002/cbdv.202200486
Nguyen, T.H., Thiers, L., Van Moerkercke, A., Bai, Y., Fernández-Calvo, P., Minne, M., Depuydt, T., Colinas, M., Verstaen, K., Van Isterdael, G., Nützmann, H.-W., Osbourn, A., Saeys, Y., De Rybel, B., Vandepoele, K., Ritter, A., Goossens, A., 2023. A redundant transcription factor network steers spatiotemporal Arabidopsis triterpene synthesis. Nature Plants 9, 926–937. https://doi.org/10.1038/s41477-023-01419-8
Nielsen, J., Keasling, J.D., 2016. Engineering Cellular Metabolism. Cell 164, 1185–1197. https://doi.org/10.1016/j.cell.2016.02.004
Nielsen, R., Yang, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148, 929–936. https://doi.org/10.1093/genetics/148.3.929
Niżyński, B., Alsoufi, A.S.M., Pączkowski, C., Długosz, M., Szakiel, A., 2015. The content of free and esterified triterpenoids of the native marigold (Calendula officinalis) plant and its modifications in in vitro cultures. Phytochemistry Letters 11, 410–417. https://doi.org/10.1016/j.phytol.2014.12.017

Nützmann, H.-W., Osbourn, A., 2015. Regulation of metabolic gene clusters in Arabidopsis thaliana. New Phytol 205, 503–510. https://doi.org/10.1111/nph.13189
Nützmann, H.-W., Osbourn, A., 2014. Gene clustering in plant specialized metabolism. Current Opinion in Biotechnology, Food biotechnology ⬢ Plant biotechnology 26, 91–99. https://doi.org/10.1016/j.copbio.2013.10.009
O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., Ecker, J.R., 2016. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell 165, 1280–1292. https://doi.org/10.1016/j.cell.2016.04.038
Paddon, C.J., Westfall, P.J., Pitera, D.J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M.D., Tai, A., Main, A., Eng, D., Polichuk, D.R., Teoh, K.H., Reed, D.W., Treynor, T., Lenihan, J., Jiang, H., Fleck, M., Bajad, S., Dang, G., Dengrove, D., Diola, D., Dorin, G., Ellens, K.W., Fickes, S., Galazzo, J., Gaucher, S.P., Geistlinger, T., Henry, R., Hepp, M., Horning, T., Iqbal, T., Kizer, L., Lieu, B., Melis, D., Moss, N., Regentin, R., Secrest, S., Tsuruta, H., Vazquez, R., Westblade, L.F., Xu, L., Yu, M., Zhang, Y., Zhao, L., Lievense, J., Covello, P.S., Keasling, J.D., Reiling, K.K., Renninger, N.S., Newman, J.D., 2013. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature 496, 528–532. https://doi.org/10.1038/nature12051
Panero, J.L., Crozier, B.S., 2016. Macroevolutionary dynamics in the early diversification of Asteraceae. Molecular Phylogenetics and Evolution 99, 116–132. https://doi.org/10.1016/j.ympev.2016.03.007
Panero, J.L., Funk, V.A., 2008. The value of sampling anomalous taxa in phylogenetic studies: Major clades of the Asteraceae revealed. Molecular Phylogenetics and Evolution 47, 757–782. https://doi.org/10.1016/j.ympev.2008.02.011
Papadopoulou, K., Melton, R.E., Leggett, M., Daniels, M.J., Osbourn, A.E., 1999. Compromised disease resistance in saponin-deficient plants. Proceedings of the National Academy of Sciences 96, 12923–12928. https://doi.org/10.1073/pnas.96.22.12923
Park, S.-Y., Qiu, J., Wei, S., Peterson, F.C., Beltrán, J., Medina-Cucurella, A.V., Vaidya, A.S., Xing, Z., Volkman, B.F., Nusinow, D.A., Whitehead, T.A., Wheeldon, I., Cutler, S.R., 2024. An orthogonalized PYR1-based CID module with reprogrammable ligand-binding specificity. Nat Chem Biol 20, 103–110. https://doi.org/10.1038/s41589-023-01447-7
Patron, N.J., Orzaez, D., Marillonnet, S., Warzecha, H., Matthewman, C., Youles, M., Raitskin, O., Leveau, A., Farré, G., Rogers, C., Smith, A., Hibberd, J., Webb, A.A.R., Locke, J., Schornack, S., Ajioka, J., Baulcombe, D.C., Zipfel, C., Kamoun, S., Jones, J.D.G., Kuhn, H., Robatzek, S., Van Esse, H.P., Sanders, D., Oldroyd, G., Martin, C., Field, R., O'Connor, S., Fox, S., Wulff, B., Miller, B., Breakspear, A., Radhakrishnan, G., Delaux, P.-M., Loqué, D., Granell, A., Tissier, A., Shih, P., Brutnell, T.P., Quick, W.P., Rischer, H., Fraser, P.D., Aharoni, A., Raines, C., South, P.F., Ané, J.-M., Hamberger, B.R., Langdale, J., Stougaard, J., Bouwmeester, H., Udvardi, M., Murray, J.A.H., Ntoukakis, V., Schäfer, P., Denby, K., Edwards, K.J., Osbourn, A., Haseloff, J., 2015. Standards for plant synthetic biology: a common syntax for exchange of DNA parts. New Phytologist 208, 13–19. https://doi.org/10.1111/nph.13532

Pellicer, J., Leitch, I.J., 2020. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytologist 226, 301–305. https://doi.org/10.1111/nph.16261
Pfalzgraf, H., 2022. Towards the structure-informed engineering of enzymes in the avenacin biosynthesis pathway (doctoral). University of East Anglia. School of Biological Sciences.
Pichersky, E., Gang, D.R., 2000. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci 5, 439–445. https://doi.org/10.1016/s1360-1385(00)01741-6
Power, F.B., Browning, H., 1912. CCLII.—The constituents of taraxacum root. J. Chem. Soc., Trans. 101, 2411–2429. https://doi.org/10.1039/CT9120102411
Pütter, K.M., van Deenen, N., Müller, B., Fuchs, L., Vorwerk, K., Unland, K., Bröker, J.N., Scherer, E., Huber, C., Eisenreich, W., Prüfer, D., Schulze Gronover, C., 2019. The enzymes OSC1 and CYP716A263 produce a high variety of triterpenoids in the latex of Taraxacum koksaghyz. Sci Rep 9, 5942. https://doi.org/10.1038/s41598-019-42381-w
Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol 67, 901–904. https://doi.org/10.1093/sysbio/syy032
Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., Castro-Mondragon, J.A., Ferenc, K., Kumar, V., Lemma, R.B., Lucas, J., Chèneby, J., Baranasic, D., Khan, A., Fornes, O., Gundersen, S., Johansen, M., Hovig, E., Lenhard, B., Sandelin, A., Wasserman, W.W., Parcy, F., Mathelier, A., 2024. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. Nucleic Acids Research 52, D174–D182. https://doi.org/10.1093/nar/gkad1059
Reece-Hoyes, J.S., Walhout, A.J.M., 2018. Gateway Recombinational Cloning. Cold Spring Harb Protoc 2018, pdb.top094912. https://doi.org/10.1101/pdb.top094912
Reece-Hoyes, J.S., Walhout, A.J.M., 2012. YEAST ONE-HYBRID ASSAYS: A HISTORICAL AND TECHNICAL PERSPECTIVE. Methods 57, 441–447. https://doi.org/10.1016/j.ymeth.2012.07.027
Reed, J., Stephenson, M.J., Miettinen, K., Brouwer, B., Leveau, A., Brett, P., Goss, R.J.M., Goossens, A., O'Connell, M.A., Osbourn, A., 2017. A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules. Metabolic Engineering 42, 185–193. https://doi.org/10.1016/j.ymben.2017.06.012
Reeves, P.H., Ellis, C.M., Ploense, S.E., Wu, M.-F., Yadav, V., Tholl, D., Chételat, A., Haupt, I., Kennerley, B.J., Hodgens, C., Farmer, E.E., Nagpal, P., Reed, J.W., 2012. A Regulatory Network for Coordinated Flower Maturation. PLOS Genetics 8, e1002506. https://doi.org/10.1371/journal.pgen.1002506
Reiser, L., Bakker, E., Subramaniam, S., Chen, X., Sawant, S., Khosa, K., Prithvi, T., Berardini, T.Z., 2024. The Arabidopsis Information Resource in 2024. Genetics 227, iyae027. https://doi.org/10.1093/genetics/iyae027
Remmert, M., Biegert, A., Hauser, A., Söding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9, 173–175. https://doi.org/10.1038/nmeth.1818
Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D.O., Truco, M.-J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B.C., Michelmore, R.W., 2017. Genome

assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat Commun 8, 14953. https://doi.org/10.1038/ncomms14953
Riley, H.T., 1855. The Natural History of Pliny, A Translation of: Naturalis Historia. Taylor and Francis, Woking, Surrey.
Ro, D.-K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., Chang, M.C.Y., Withers, S.T., Shiba, Y., Sarpong, R., Keasling, J.D., 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440, 940–943. https://doi.org/10.1038/nature04640
Rodríguez-Concepción, M., Boronat, A., 2002. Elucidation of the Methylerythritol Phosphate Pathway for Isoprenoid Biosynthesis in Bacteria and Plastids. A Metabolic Milestone Achieved through Genomics. Plant Physiol 130, 1079–1089. https://doi.org/10.1104/pp.007138
Salmon, M., Thimmappa, R.B., Minto, R.E., Melton, R.E., Hughes, R.K., O'Maille, P.E., Hemmings, A.M., Osbourn, A., 2016. A conserved amino acid residue critical for product and substrate specificity in plant triterpene synthases. Proceedings of the National Academy of Sciences 113, E4407–E4414. https://doi.org/10.1073/pnas.1605509113
Sauret-Güeto, S., Frangedakis, E., Silvestri, L., Rebmann, M., Tomaselli, M., Markel, K., Delmans, M., West, A., Patron, N.J., Haseloff, J., 2020. Systematic Tools for Reprogramming Plant Gene Expression in a Simple Model, Marchantia polymorpha. ACS Synth. Biol. 9, 864–882. https://doi.org/10.1021/acssynbio.9b00511
Scaglione, D., Reyes-Chin-Wo, S., Acquadro, A., Froenicke, L., Portis, E., Beitel, C., Tirone, M., Mauro, R., Monaco, A.L., Mauromicale, G., Faccioli, P., Cattivelli, L., Rieseberg, L., Michelmore, R., Lanteri, S., 2016. The genome sequence of the outbreeding globe artichoke constructed de novo incorporating a phase-aware low-pass sequencing strategy of F1 progeny. Scientific Reports 6, 19427. https://doi.org/10.1038/srep19427
Schandry, N., Becker, C., 2020. Allelopathic Plants: Models for Studying Plant-Interkingdom Interactions. Trends Plant Sci 25, 176–185. https://doi.org/10.1016/j.tplants.2019.11.004
Schlegel, S., Klepsch, M., Gialama, D., Wickström, D., Slotboom, D.J., De Gier, J., 2010. Revolutionizing membrane protein overexpression in bacteria. Microb Biotechnol 3, 403–411. https://doi.org/10.1111/j.1751-7915.2009.00148.x
Schmitz, R.J., Grotewold, E., Stam, M., 2022. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell 34, 718–741. https://doi.org/10.1093/plcell/koab281
Sharma, K., Zafar, R., 2015. Occurrence of taraxerol and taraxasterol in medicinal plants. Pharmacogn Rev 9, 19–23. https://doi.org/10.4103/0973-7847.156317
Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I., 2013. New and continuing developments at PROSITE. Nucleic Acids Res 41, D344–D347. https://doi.org/10.1093/nar/gks1067
Singh, A., Upadhyay, V., Upadhyay, A.K., Singh, S.M., Panda, A.K., 2015. Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. Microbial Cell Factories 14, 41. https://doi.org/10.1186/s12934-015-0222-8

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13, 613–626. https://doi.org/10.1038/nrg3207
Srisawat, P., Fukushima, E.O., Yasumoto, S., Robertlee, J., Suzuki, H., Seki, H., Muranaka, T., 2019. Identification of oxidosqualene cyclases from the medicinal legume tree Bauhinia forficata: a step toward discovering preponderant α-amyrin-producing activity. New Phytologist 224, 352–366. https://doi.org/10.1111/nph.16013
Srivastava, P.L., Johns, S.T., Voice, A., Morley, K., Escorcia, A.M., Miller, D.J., Allemann, R.K., van der Kamp, M.W., 2024. Simulation-Guided Engineering Enables a Functional Switch in Selinadiene Synthase toward Hydroxylation. ACS Catal. 14, 11034–11043. https://doi.org/10.1021/acscatal.4c02032
Srivastava, P.L., Johns, S.T., Walters, R., Miller, D.J., Van der Kamp, M.W., Allemann, R.K., 2023. Active Site Loop Engineering Abolishes Water Capture in Hydroxylating Sesquiterpene Synthases. ACS Catal. 13, 14199–14204. https://doi.org/10.1021/acscatal.3c03920
Stephenson, M.J., Reed, J., Brouwer, B., Osbourn, A., 2018. Transient Expression in Nicotiana Benthamiana Leaves for Triterpene Production at a Preparative Scale. J Vis Exp 58169. https://doi.org/10.3791/58169
Stewart, J.J.P., 2007. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J Mol Model 13, 1173–1213. https://doi.org/10.1007/s00894-007-0233-4
Stracke, R., Werber, M., Weisshaar, B., 2001. The *R2R3-MYB* gene family in *Arabidopsis thaliana*. Current Opinion in Plant Biology 4, 447–456. https://doi.org/10.1016/S1369-5266(00)00199-0
Strader, L., Weijers, D., Wagner, D., 2022. Plant transcription factors - being in the right place with the right company. Curr Opin Plant Biol 65, 102136. https://doi.org/10.1016/j.pbi.2021.102136
Su, W., Jing, Y., Lin, Shoukai, Yue, Z., Yang, X., Xu, J., Wu, J., Zhang, Z., Xia, R., Zhu, J., An, N., Chen, H., Hong, Y., Yuan, Y., Long, T., Zhang, L., Jiang, Y., Liu, Zongli, Zhang, H., Gao, Y., Liu, Y., Lin, H., Wang, H., Yant, L., Lin, Shunquan, Liu, Zhenhua, 2021. Polyploidy underlies co-option and diversification of biosynthetic triterpene pathways in the apple tribe. Proceedings of the National Academy of Sciences 118, e2101767118. https://doi.org/10.1073/pnas.2101767118
Sumida, K.H., Núñez-Franco, R., Kalvet, I., Pellock, S.J., Wicky, B.I.M., Milles, L.F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., Kang, A., De La Cruz, J., Sankaran, B., Bera, A.K., Jiménez-Osés, G., Baker, D., 2024. Improving Protein Expression, Stability, and Function with ProteinMPNN. J. Am. Chem. Soc. 146, 2054–2061. https://doi.org/10.1021/jacs.3c10941
Sutliff, T.D., Huang, N., Litts, J.C., Rodriguez, R.L., 1991. Characterization of an alpha-amylase multigene cluster in rice. Plant Mol Biol 16, 579–591. https://doi.org/10.1007/BF00023423
Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34, W609–W612. https://doi.org/10.1093/nar/gkl315
Suzuki, M., Xiang, T., Ohyama, K., Seki, H., Saito, K., Muranaka, T., Hayashi, H., Katsube, Y., Kushiro, T., Shibuya, M., Ebizuka, Y., 2006. Lanosterol synthase in dicotyledonous plants. Plant Cell Physiol 47, 565–571. https://doi.org/10.1093/pcp/pcj031

Swamidatta, S.H., Lichman, B.R., 2024. Beyond co-expression: pathway discovery for plant pharmaceuticals. Current Opinion in Biotechnology 88, 103147. https://doi.org/10.1016/j.copbio.2024.103147
Swinnen, G., Goossens, A., Pauwels, L., 2016. Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement. Trends in Plant Science 21, 506–515. https://doi.org/10.1016/j.tplants.2016.01.014
Takase, S., Saga, Y., Kurihara, N., Naraki, S., Kuze, K., Nakata, G., Araki, T., Kushiro, T., 2015. Control of the 1,2-rearrangement process by oxidosqualene cyclases during triterpene biosynthesis. Org. Biomol. Chem. 13, 7331–7336. https://doi.org/10.1039/C5OB00714C
Tene, M., Ndontsa, B., Tane, P., De Dieu Tamokou, J., Kuiate, J.-R., 2009. Antimicrobial diterpenoids and triterpenoids from the stem bark of Croton macrostachys. International Journal of Biological and Chemical Sciences 3. https://doi.org/10.4314/ijbcs.v3i3.45331
Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., Osbourn, A., 2014. Triterpene Biosynthesis in Plants. Annual Review of Plant Biology 65, 225–257. https://doi.org/10.1146/annurev-arplant-050312-120229
Thimmappa, R., Wang, S., Zheng, M., Misra, R.C., Huang, A.C., Saalbach, G., Chang, Y., Zhou, Z., Hinman, V., Bao, Z., Osbourn, A., 2022. Biosynthesis of saponin defensive compounds in sea cucumbers. Nat Chem Biol 18, 774–781. https://doi.org/10.1038/s41589-022-01054-y
Thoma, R., Schulz-Gasch, T., D'Arcy, B., Benz, J., Aebi, J., Dehmlow, H., Hennig, M., Stihle, M., Ruf, A., 2004. Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. Nature 432, 118–122. https://doi.org/10.1038/nature02993
Tian, C., Kasavajhala, K., Belfon, K.A.A., Raguette, L., Huang, H., Migues, A.N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., Simmerling, C., 2020. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. J. Chem. Theory Comput. 16, 528–552. https://doi.org/10.1021/acs.jctc.9b00591
Tremetsberger, K., Gemeinholzer, B., Zetzsche, H., Blackmore, S., Kilian, N., Talavera, S., 2013. Divergence time estimation in Cichorieae (Asteraceae) using a fossil-calibrated relaxed molecular clock. Org Divers Evol 13, 1–13. https://doi.org/10.1007/s13127-012-0094-2
Unland, K., Pütter, K.M., Vorwerk, K., van Deenen, N., Twyman, R.M., Prüfer, D., Schulze Gronover, C., 2018. Functional characterization of squalene synthase and squalene epoxidase in Taraxacum koksaghyz. Plant Direct 2, e00063. https://doi.org/10.1002/pld3.63
van der Kamp, M.W., Mulholland, A.J., 2013. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. Biochemistry 52, 2708–2728. https://doi.org/10.1021/bi400215w
van der Kamp, M.W., Shaw, K.E., Woods, C.J., Mulholland, A.J., 2008. Biomolecular simulation and modelling: status, progress and prospects. Journal of The Royal Society Interface 5, 173–190. https://doi.org/10.1098/rsif.2008.0105.focus
van Herpen, T.W.J.M., Cankar, K., Nogueira, M., Bosch, D., Bouwmeester, H.J., Beekwilder, J., 2010. Nicotiana benthamiana as a production platform for artemisinin precursors. PLoS One 5, e14222. https://doi.org/10.1371/journal.pone.0014222

VanEtten, H.D., Mansfield, J.W., Bailey, J.A., Farmer, E.E., 1994. Two Classes of Plant Antibiotics: Phytoalexins versus “Phytoanticipins.” Plant Cell 6, 1191–1192. https://doi.org/10.1105/tpc.6.9.1191
Vetter, J., 2000. Plant cyanogenic glycosides. Toxicon 38, 11–36. https://doi.org/10.1016/s0041-0101(99)00128-2
Vienne, D.M. de, 2016. Lifemap: Exploring the Entire Tree of Life. PLOS Biology 14, e2001624. https://doi.org/10.1371/journal.pbio.2001624
Wagenitz, G., 1976. Systematics and phylogeny of theCompositae (Asteraceae). Pl Syst Evol 125, 29–46. https://doi.org/10.1007/BF00986129
Walker-Hale, N., Guerrero-Rubio, M.A., Brockington, S.F., 2025. Multiple transitions to high l-DOPA 4,5-dioxygenase activity reveal molecular pathways to convergent betalain pigmentation in Caryophyllales. New Phytologist 247, 341–357. https://doi.org/10.1111/nph.70177
Wang, B., Luo, Q., Li, Y., Yin, L., Zhou, N., Li, X., Gan, J., Dong, A., 2020. Structural insights into target DNA recognition by R2R3-MYB transcription factors. Nucleic Acids Res 48, 460–471. https://doi.org/10.1093/nar/gkz1081
Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A., 2004. Development and testing of a general amber force field. J Comput Chem 25, 1157–1174. https://doi.org/10.1002/jcc.20035
Wang, K.-W., 2007. A new fatty acid ester of triterpenoid from Celastrus rosthornianus with anti-tumor activitives. Natural Product Research 21, 669–674. https://doi.org/10.1080/14786410701371447
WANG, W.-M., 2004. On the origin and development of Artemisia (Asteraceae) in the geological past. Bot J Linn Soc 145, 331–336. https://doi.org/10.1111/j.1095-8339.2004.00287.x
Wehner, N., Hartmann, L., Ehlert, A., Böttner, S., Oñate-Sánchez, L., Dröge-Laser, W., 2011. High-throughput protoplast transactivation (PTA) system for the analysis of Arabidopsis transcription factor function. The Plant Journal 68, 560–569. https://doi.org/10.1111/j.1365-313X.2011.04704.x
Wendt, K.U., Lenhart, A., Schulz, G.E., 1999. The structure of the membrane protein squalene-hopene cyclase at 2.0 å resolution 1. Journal of Molecular Biology 286, 175–187. https://doi.org/10.1006/jmbi.1998.2470
Wendt, K.U., Poralla, K., Schulz, G.E., 1997. Structure and function of a squalene cyclase. Science 277, 1811–1815. https://doi.org/10.1126/science.277.5333.1811
Wenzell, K.E., Neequaye, M., Paajanen, P., Hill, L., Brett, P., Byers, K.J.R.P., 2025. Within-species floral evolution reveals convergence in adaptive walks during incipient pollinator shift. Nat Commun 16, 2721. https://doi.org/10.1038/s41467-025-57639-3
Weston, L.A., Mathesius, U., 2013. Flavonoids: their structure, biosynthesis and role in the rhizosphere, including allelopathy. J Chem Ecol 39, 283–297. https://doi.org/10.1007/s10886-013-0248-5
Wink, M., 2016. Secondary Metabolites, the Role in Plant Diversification of, in: Kliman, R.M. (Ed.), Encyclopedia of Evolutionary Biology. Academic Press, Oxford, pp. 1–9. https://doi.org/10.1016/B978-0-12-800049-6.00263-8
Woolfson, D.N., 2023. Understanding a protein fold: The physics, chemistry, and biology of α-helical coiled coils. J Biol Chem 299, 104579. https://doi.org/10.1016/j.jbc.2023.104579
Wu, S., Morotti, A.L.M., Yang, J., Wang, E., Tatsis, E.C., 2024. Single-cell RNA sequencing facilitates the elucidation of the complete biosynthesis of the

antidepressant hyperforin in St. John's wort. Molecular Plant 17, 1439–1457. https://doi.org/10.1016/j.molp.2024.08.003
Wu, S., Zhang, F., Xiong, W., Molnár, I., Liang, J., Ji, A., Li, Y., Wang, C., Wang, S., Liu, Z., Wu, R., Duan, L., 2020. An Unexpected Oxidosqualene Cyclase Active Site Architecture in the Iris tectorum Multifunctional α-Amyrin Synthase. ACS Catal. 10, 9515–9520. https://doi.org/10.1021/acscatal.0c03231
Wu, Y., Wen, J., Xia, Y., Zhang, L., Du, H., 2022. Evolution and functional diversification of R2R3-MYB transcription factors in plants. Hortic Res 9, uhac058. https://doi.org/10.1093/hr/uhac058
Xie, X., Zhai, Y., Cheng, H., Wei, W.-H., Ren, M., 2025. From *Taxus* to paclitaxel: Opportunities and challenges for urban agriculture to promote human health. Plant Physiology and Biochemistry 220, 109502. https://doi.org/10.1016/j.plaphy.2025.109502
Xue, Z., Duan, L., Liu, D., Guo, J., Ge, S., Dicks, J., ÓMáille, P., Osbourn, A., Qi, X., 2012. Divergent evolution of oxidosqualene cyclases in plants. New Phytologist 193, 1022–1038. https://doi.org/10.1111/j.1469-8137.2011.03997.x
Xue, Z., Tan, Z., Huang, A., Zhou, Y., Sun, J., Wang, X., Thimmappa, R.B., Stephenson, M.J., Osbourn, A., Qi, X., 2018a. Identification of key amino acid residues determining product specificity of 2,3-oxidosqualene cyclase in Oryza species. New Phytol 218, 1076–1088. https://doi.org/10.1111/nph.15080
Xue, Z., Xu, X., Zhou, Y., Wang, X., Zhang, Y., Liu, D., Zhao, B., Duan, L., Qi, X., 2018b. Deficiency of a triterpene pathway results in humidity-sensitive genic male sterility in rice. Nat Commun 9, 604. https://doi.org/10.1038/s41467-018-03048-8
Yan, Q., Xing, Q., Liu, Z., Zou, Y., Liu, X., Xia, H., 2024. The phytochemical and pharmacological profile of dandelion. Biomedicine & Pharmacotherapy 179, 117334. https://doi.org/10.1016/j.biopha.2024.117334
Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24, 1586–1591. https://doi.org/10.1093/molbev/msm088
Yang, Z., Li, X., Yang, L., Peng, S., Song, W., Lin, Y., Xiang, G., Li, Y., Ye, S., Ma, C., Miao, J., Zhang, G., Chen, W., Yang, S., Dong, Y., 2023. Comparative genomics reveals the diversification of triterpenoid biosynthesis and origin of ocotillol-type triterpenes in Panax. Plant Comm 4. https://doi.org/10.1016/j.xplc.2023.100591
Yang, Z., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19, 908–917. https://doi.org/10.1093/oxfordjournals.molbev.a004148
Yang, Z., Nielsen, R., 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46, 409–418. https://doi.org/10.1007/PL00006320
Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.-M.K., 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155, 431–449. https://doi.org/10.1093/genetics/155.1.431
Yang, Z., Wong, W.S.W., Nielsen, R., 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol Biol Evol 22, 1107–1118. https://doi.org/10.1093/molbev/msi097

Yasmeen, E., Wang, J., Riaz, M., Zhang, L., Zuo, K., 2023. Designing artificial synthetic promoters for accurate, smart, and versatile gene expression in plants. Plant Commun 4, 100558. https://doi.org/10.1016/j.xplc.2023.100558
Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H.B., Andrae, S., Yang, T.H., Lee, S.Y., Burk, M.J., Van Dien, S., 2011. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat Chem Biol 7, 445–452. https://doi.org/10.1038/nchembio.580
Yonekura-Sakakibara, K., Saito, K., 2009. Functional genomics for plant natural product biosynthesis. Nat. Prod. Rep. 26, 1466–1487. https://doi.org/10.1039/B817077K
Yoo, S.-D., Cho, Y.-H., Sheen, J., 2007. Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat Protoc 2, 1565–1572. https://doi.org/10.1038/nprot.2007.199
Yu, N., Nützmann, H.-W., MacDonald, J.T., Moore, B., Field, B., Berriri, S., Trick, M., Rosser, S.J., Kumar, S.V., Freemont, P.S., Osbourn, A., 2016. Delineation of metabolic gene clusters in plant genomes by chromatin signatures. Nucleic Acids Res 44, 2255–2265. https://doi.org/10.1093/nar/gkw100
Zenker, S., Wulf, D., Meierhenrich, A., Viehöver, P., Becker, S., Eisenhut, M., Stracke, R., Weisshaar, B., Bräutigam, A., 2025. Many transcription factor families have evolutionarily conserved binding motifs in plants. Plant Physiol 198, kiaf205. https://doi.org/10.1093/plphys/kiaf205
Zhang, C., Echeverria, J., Foresti, C., Santiago, A., Moubtassim, H.E.I., Amato, A., Sonego, P., Pindo, M., Zenoni, S., Moretto, M., Matus, J.T., 2025. A Grapevine MYC2-MYB24 Regulatory Module Activates Terpenoid Biosynthesis Upon Methyl Jasmonate Elicitation. https://doi.org/10.1101/2025.01.27.635049
Zhang, F., Wang, Y., Yue, J., Zhang, R., Hu, Y.-E., Huang, R., Ji, A.-J., Hess, B.A., Liu, Z., Duan, L., Wu, R., 2023. Discovering a uniform functional trade-off of the CBC-type 2,3-oxidosqualene cyclases and deciphering its chemical logic. Sci Adv 9, eadh1418. https://doi.org/10.1126/sciadv.adh1418
Zhang, P., Liu, X., Yu, X., Wang, F., Long, J., Shen, W., Jiang, D., Zhao, X., 2020. The MYB transcription factor CiMYB42 regulates limonoids biosynthesis in citrus. BMC Plant Biology 20, 254. https://doi.org/10.1186/s12870-020-02475-4
Zhang, S., Meng, F., Pan, X., Qiu, X., Li, C., Lu, S., 2024. Chromosome-level genome assembly of Prunella vulgaris L. provides insights into pentacyclic triterpenoid biosynthesis. The Plant Journal 118, 731–752. https://doi.org/10.1111/tpj.16629
Zhao, K., Kong, D., Jin, B., Smolke, C.D., Rhee, S.Y., 2021. A novel bivalent chromatin associates with rapid induction of camalexin biosynthesis genes in response to a pathogen signal in Arabidopsis. eLife 10, e69508. https://doi.org/10.7554/eLife.69508
Zhong, J.-J., 2002. Plant cell culture for production of paclitaxel and other taxanes. Journal of Bioscience and Bioengineering 94, 591–599. https://doi.org/10.1016/S1389-1723(02)80200-6
Zhong, Y., Xun, W., Wang, X., Tian, S., Zhang, Y., Li, D., Zhou, Y., Qin, Y., Zhang, B., Zhao, G., Cheng, X., Liu, Y., Chen, H., Li, L., Osbourn, A., Lucas, W.J., Huang, S., Ma, Y., Shang, Y., 2022. Root-secreted bitter triterpene modulates

the rhizosphere microbiota to improve plant fitness. Nat. Plants 8, 887–896. https://doi.org/10.1038/s41477-022-01201-2
Zhou, Q., Sun, P., Xiong, H.-M., Xie, J., Zhu, G.-Y., Tantillo, D.J., Huang, A.C., 2024. Insight into neofunctionalization of 2,3-oxidosqualene cyclases in B,C-ring-opened triterpene biosynthesis in quinoa. New Phytologist 241, 764–778. https://doi.org/10.1111/nph.19345
Zhou, Y., Ma, Y., Zeng, J., Duan, L., Xue, X., Wang, H., Lin, T., Liu, Z., Zeng, K., Zhong, Y., Zhang, S., Hu, Q., Liu, M., Zhang, H., Reed, J., Moses, T., Liu, Xinyan, Huang, P., Qing, Z., Liu, Xiubin, Tu, P., Kuang, H., Zhang, Z., Osbourn, A., Ro, D.-K., Shang, Y., Huang, S., 2016. Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. Nature Plants 2, 16183. https://doi.org/10.1038/nplants.2016.183
Zhu, X., Liu, X., Liu, T., Wang, Y., Ahmed, N., Li, Z., Jiang, H., 2021. Synthetic biology of plant natural products: From pathway elucidation to engineered biosynthesis in plant cells. Plant Comm 2. https://doi.org/10.1016/j.xplc.2021.100229
Zimmermann, B.F., Häberle, E., 2025. New insights into the analysis of faradiol esters and related compounds in Calendula. Journal of Pharmaceutical and Biomedical Analysis 260, 116792. https://doi.org/10.1016/j.jpba.2025.116792
Zournatzis, I., Liakos, V., Papadopoulos, S., Wogiatzi, E., 2025. *Calendula officinalis* - A comprehensive review. Pharmacological Research - Natural Products 6, 100140. https://doi.org/10.1016/j.prenap.2024.100140

Supplementary information

Species	Family	Accession prefix	Database	Tissue
Taraxacum coreanum	Asteraceae	GHDY01	TSA	Whole plant
Taraxacum kok-saghyz	Asteraceae	GFJE01, GFZW01	TSA	Root
Chrysanthemum seticuspe	Asteraceae	BPTQ01	WGS	N/A
Lactuca sativa	Asteraceae	NBSK01	WGS	N/A
Cynara cardunculus	Asteraceae	LEKV01	WGS	Leaf
Helianthus annuus	Asteraceae	MNCJ02	WGS	N/A
Atractylodes lancea	Asteraceae	GEGA01	TSA	Leaves
Helianthus niveus	Asteraceae	GEWS01	TSA	Leaf
Cichorium endivia	Asteraceae	GGQM01	TSA	apices, stems, leaves and roots
Stevia rebaudiana	Asteraceae	GISI01, GANE01	TSA	Leaf
Artemisia annua	Asteraceae	PPK01	WGS	Leaf
Mikania micrantha	Asteraceae	SZYD01	WGS	Leaf, stem, root
Hieracium piloselloides	Asteraceae	GEEH01	TSA	N/A
Karelinia caspia	Asteraceae	GANI01	TSA	Plant tissue
Tagetes erecta	Asteraceae	GGGQ01	TSA	N/A
Cichorium intybus	Asteraceae	GGQG01	TSA	apices, stems, leaves and roots
Helianthus tuberosus	Asteraceae	GHBN01	TSA	Mixed samples
Chrysanthemum x morifolium	Asteraceae	IABW01	TSA	N/A
Silybum marianum	Asteraceae	LMWD01	WGS	Leaf
Carthamus tinctorius	Asteraceae	LUCG01	WGS	Leaf
Erigeron breviscapus	Asteraceae	GDQF01	TSA	N/A
Chromolaena odorata	Asteraceae	GACH01	TSA	Stem
Dahlia pinnata	Asteraceae	GBDN01	TSA	N/A
Ambrosia trifida	Asteraceae	GEOH01, GIMT01	TSA	Leaf

Cosmos bipinnatus	Asteraceae	GEZQ01	TSA	Roots
Parthenium argentatum	Asteraceae	GFTW01	TSA	N/A
Gynura procumbens	Asteraceae	GGYE01	TSA	Leaves
Celmisia lyallii	Asteraceae	GHOT01	TSA	Leaf and root
Erigeron canadensis	Asteraceae	JWSR01	WGS	Leaf
Gerbera hybrid cultivar	Asteraceae	GDBL01, GACN01	TSA	Ray floret
Tanacetum cinerariifolium	Asteraceae	BKCJ01	WGS	N/A
Ambrosia artemisiifolia	Asteraceae	HAAJ01, GEZL01, GFWB01, GFWS01	TSA	Female flower, pollen
Taraxacum mongolicum	Asteraceae	GWHBCHG00000000	GWH	N/A
Matricaria chamomilla	Asteraceae	CAWUDT01	WGS	Leaf
Aster tataricus	Asteraceae	XRCX	1kp	leaves
Carthamus lanatus	Asteraceae	CFRN	1kp	leaves
Cicerbita plumieri	Asteraceae	JNKW	1kp	young leaves
Conyza canadensis	Asteraceae	WSYE, NUSE	1kp	glyphosate suseptible and resistant young leaves and meristems
Flaveria angustifolia	Asteraceae	EXHX,UQAZ	1kp	juvenile leaf,mature leaf
Flaveria bidentis	Asteraceae	QBGG,ZNZC	1kp	mature leaf,juvenile leaf
Flaveria brownii	Asteraceae	IPUI,LMXP	1kp	mature leaf,juvenile leaf
Flaveria cronquistii	Asteraceae	KJBH,DSWR	1kp	mature leaf,juvenile leaf
Flaveria kochiana	Asteraceae	QXWF	1kp	mature leaf
Flaveria palmeri	Asteraceae	NVSO	1kp	juvenile leaf
Flaveria pringlei	Asteraceae	RZSW,GFVW	1kp	mature leaf,juvenile leaf

Flaveria pubescens	Asteraceae	JZHZ,HXDV	1kp	juvenile leaf,mature leaf
Flaveria sonorensis	Asteraceae	UYED	1kp	juvenile leaf
Flaveria trinervia	Asteraceae	RLCS,HRVY	1kp	juvenile leaf,mature leaf
Flaveria vaginata	Asteraceae	HXCD	1kp	juvenile leaf
Helenium autumnale	Asteraceae	DUNJ	1kp	leaves and flower buds
Inula helenium	Asteraceae	AFQQ	1kp	leaves and root
Lactuca graminifolia	Asteraceae	ZGDS,MMJI	1kp	young leaves
Leontopodium nivale subsp. alpinum	Asteraceae	DOVJ	1kp	leaf
Matricaria matricarioides	Asteraceae	OAGK	1kp	shoot with flowers
Solidago canadensis	Asteraceae	TEZA	1kp	shoot with flowers
Tanacetum parthenium	Asteraceae	DUQG	1kp	flower, floral buds, leaves and stem
Tragopogon castellanus	Asteraceae	KFZY	1kp	young leaves
Tragopogon dubius	Asteraceae	DDRL	1kp	N/A
Tragopogon porrifolius	Asteraceae	KGJF	1kp	young leaves
Tragopogon pratensis	Asteraceae	FUPX	1kp	young leaves
Xanthium strumarium	Asteraceae	UBLN	1kp	leaves and stems
Calendula officinalis	Asteraceae	CAXWYG01	WGS	Flower and leaf
Calendula arvensis	Asteraceae	This study, Transcriptome		Flower and leaf
Achillea millefolium	Asteraceae	This study, Transcriptome		Flower and leaf
Eupatorium cannabinum	Asteraceae	This study, Transcriptome		Flower and leaf
Inula britannica	Asteraceae	This study, Transcriptome		Flower and leaf
Inula ensifolia	Asteraceae	This study, Transcriptome		Flower and leaf

Nelumbo nucifera	Nelumbonaceae	AQOG01	WGS	leaves
Acrocarpus fraxinifolius	Fabaceae	N/A	tropiTree	germinated seeds
Ficus carica	Moraceae	VYVB01	WGS	leaves
Euphorbia tirucalli	Euphorbiaceae	GETW01	TSA	Stem, latex
Hypericum perforatum	Hypericaceae	JAZHPW01	WGS	Leaf
Catharanthus roseus	Apocynaceae	GKIJ01	TSA	Leaf
Solanum lycopersicum	Solanaceae	AEKE04	WGS	N/A
Olea europaea	Oleaceae	JBKEJA01	WGS	Leaf
Vitellaria paradoxa	Sapotaceae	N/A	orcAE	N/A
Camellia sinensis	Theaceae	JBNOCF01	WGS	Leaf
Camellia japonica	Theaceae	JAPPVO01	WGS	Young leaf
Menyanthes trifoliata	Menyanthaceae	IXVJ	1kp	leaves
Stylidium adnatum	Stylidiaceae	FXGI	1kp	N/A
Corokia cotoneaster	Argophyllaceae	GIOY	1kp	leaves, veg buds
Phelline lucida	Phellinaceae	AUIP	1kp	N/A
Platycodon grandiflorus	Campanulaceae	IHPC	1kp	leaf
Adenophora triphylla	Campanulaceae	JAFEMH01	WGS	leaf
Campanula takesimana	Campanulaceae	JAOWBW01	WGS	leaf
Codonopsis lanceolata	Campanulaceae	JABEVN02	WGS	leaf
Lobelia siphilitica	Campanulaceae	IZLO	1kp	shoot
Lobelia cardinalis	Campanulaceae	JAOWAN01	WGS	leaf

Supplementary Table 1 Species used for genome and transcriptome mining

Chapter 3 and 4	Site-directed mutagenesis on CoMAS, CoTXSS and TkTXSS	Sequence
HSN001	CoMAS I367M PF	GACTATGGGATGTGTTGAAAAGAGCTTGCAAATGATGTGTTGG
HSN002	CoMAS E371D PF	GACTATTGGATGTGTTGATAAGAGCTTGCAAATGATGTGTTGG
HSN003	CoMAS I367M E371D PF	GACTATGGGATGTGTTGATAAGAGCTTGCAAATGATGTGTTGG
HSN004	CoMAS I367M PR	TCTTTTCAACACATCCCATAGTCATATATCTGCTTTGTTGAGCAC
HSN005	CoMAS E371D PR	TCTTATCAACACATCCAATAGTCATATATCTGCTTTGTTGAGCAC
HSN006	CoMAS I367M E371D PR	TCTTATCAACACATCCCATAGTCATATATCTGCTTTGTTGAGCAC
HSN007	CoTXSS G380T PF	TACATAACCATGGGTTGTGTTGACAAGGCTTTAC
HSN008	CoTXSS D385E PF	GTGTTGAAAAGGCTTTACAAATGATGTGTTTTTATGCCG
HSN009	CoTXSS H492Q PF	CAAGACCAAGGATGGGTTGTATCAGATTGCACTGCAG
HSN010	CoTXSS P751A PF	TTACATTATGCAGAATATAGGAACACTTTTCCGTTATGGGC
HSN011	CoTXSS G380T D385E PF (on G380T)	TACATAACCATGGGTTGTGTTGAAAAGGCTTTAC
HSN012	CoTXSS G380T D385E PF (on D385E)	GTGTTGAAAAGGCTTTACAAATGATGTGTTTTTATGCCG
HSN013	CoTXSS G380T PR	ACCCATGGTTATGTAACGTCCTTCTTCGGATCC
HSN014	CoTXSS D385E PR	AGCCTTTTCAACACAACCCATGCCTATGTAACGTC
HSN015	CoTXSS H492Q PR	CCATCCTTGGTCTTGATCAGAGAAAGTCCATGCCC
HSN016	CoTXSS P751A PR	CCTATATTCTGCATAATGTAACATGCAGTTTTTCATGTACACTC
HSN017	CoTXSS G380T D385E PR (on G380T)	ACCCATGGTTATGTAACGTCCTTCTTCGGATCC
HSN018	CoTXSS G380T D385E PR (on D385E)	AGCCTTTTCAACACAACCCATGGTTATGTAACGTC
HSN019	TkTXSS T374G PF	ATACATAGGATTGGGGTGTGTCGAGAAATCCTTACAAATG
HSN020	TkTXSS E379D PF	TGTGTCGATAAATCCTTACAAATGATGTGCTTCTCAGC
HSN021	TkTXSS Q486H PF	AGGACCATGGCTGGGTTGTTAGTGATTGCACTG
HSN022	TkTXSS A745P PF	CACTACCCCGAATATCGAAACACCTTCCCTTTATGG

HSN023	TkTXSS T374G E379D PF (on T374G)	ATACATAGGATTGGGGTGTGTCGATAAATCCTTACAAATG
HSN024	TkTXSS T374G E379D PF (on E379D)	TGTGTCGATAAATCCTTACAAATGATGTGCTTCTCAGC
HSN025	TkTXSS T374G PR	ACCCCAATCCTATGTATCTACCTTCTTCGGCGTTATATTG
HSN026	TkTXSS E379D PR	AGGATTTATCGACACACCCCAATGTTATGTATCTACCTTC
HSN027	TkTXSS Q486H PR	CAGCCATGGTCCTGGTCGCTAAATGTCCAAGC
HSN028	TkTXSS A745P PR	ATATTCGGGGTAGTGCAGCATGCAATTCTTCATATAAACC
HSN029	TkTXSS T374G E379D PR (on T374G)	ACCCCAATCCTATGTATCTACCTTCTTCGGCGTTATATTG
HSN030	TkTXSS T374G E379D PR (on E379D)	AGGATTTATCGACACACCCCAATCCTATGTATCTACCTTC
HSN031	CoTXSS Y273F PF	TGGTGTttcTGTAGAACGACATACATGCCGATGTC
HSN032	CoTXSS Y273F PR	TCTACAgaaACACCACATTTTTGCTGGATGATAAGGC
HSN033	TkTXSS Y267F PF	tggtgtTTTgtcgaactacctacatgccgatgag
HSN034	TkTXSS Y267F PR	gacaAAAacaccacatcttagctgggtggtagggg
Chapter 5	Promoter cloning	Sequence
HSN035	pCoTXSS PF	ggGCTCTTCgtctcCGGAGAAGACTATCCACGTTTTGGAAACTACAACTTCGTTC
HSN036	pCoTXSS PR	ggGCTCTTCgtctcCCATTTTATATTTTCCCCCCCAGTTATGTGTTGTTGTGATC
HSN039	pCoCYP2 PF	GGGCTCTTCGTCTCCGGAGACACGCAGAAGATGCAATTAGGGAGC
HSN040	pCoCYP2 PR	GGGCTCTTCGTCTCCCATTTTGAATGTTGTACCAAATTGTGTACGATTTGTTGATG
HSN041	pCoACT1 PF	GGGAAGACGGCTCAGGAGCCTTGTCTTTTTCACCAATACCACTAGGTTAG
HSN042	pCoACT1 PR internal (BpiI removal)	ggGAAGACaaAgACTTGCAGTACTTAGGTACTATCTTAGTGAAGG
HSN043	pCoACT1 PF internal (BpiI removal)	ggGAAGACaaGTcTCTTCAAGACTCAATGTAATATTTGTTAAATACACATTG
HSN044	pCoACT1 PR	GGGAAGACGGCTCGCATTATCGATCTTGCTAGGTTTGATCTCTG

HSN045	pCoACT2 PF	GGGAAGACGGCTCAGGAGCATTTAATTAGGCGACTTGTAGAACAAGTCGAAACTC
HSN046	pCoACT2 PR	GGGAAGACGGCTCGCATTATCGATCCTTCTAGATTTGATCTCTGTCTCTCTCAC
Chapter 5	CoMYB qPCR	Sequence
HSC024	CoMYB4 PR 287-271	AATCTTCCGGCGATCAG
HSC025	CoMYB4 PF 191-210	ATCTCAAACGTGGCAATTTT
HSC027	CoMYB24 PF 178-195	TTGAACTATCTCCGGCCA
HSC028	CoMYB24 PR 343-324	GAATCCTTGTCCTCCAGTAA
HSC031	CoMYB111 PF 244-262	TTAAGAGGGGACTTAAAGC
HSC032	CoMYB111 PR 362-345	GTTCGTCCAGGTAAATGG
From Melissa	CoSAND2_F	TCTTTCAGTTGGAACCCTGCA
From Melissa	CoSAND2_R	CTGCAATATAGCACCAGCAGC
Chapter 6	McLS cloning, CoTXSS helix mutagenesis and cloning	Sequence
HSC035	Helix1 mutagenesis PF	CCTGAATTTTGGGAATTTCCTTCCTTTTTGCCTTATCATCCAG
HSC036	Helix1 mutagenesis PR	AAGGAAGGAAATTCCCAAAATTCAGGAGGCAGAGGATTG
HSC005	Helix3 mutagenesis PF	AACCAGAATTAGAAGAATTGAACCCTTCAGAACTTTTTGC
HSC006	Helix3 mutagenesis PR	TCAATTCTTCTAATTCTGGTTTTGGAACTGGTGGCTC
HSC007	attB1-CoTXSSF	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCATGTGGAAATTAAAAATTGGTG
HSC008	attB2-CoTXSSR	GGGGACCACTTTGTACAAGAAAGCTGGGTTTCAGACTTGTTGCTTTGGC
HSC021	pDONR207_SeqF	gcagttccctactctcgc
HSC022	pH9GW_SeqF	taatacgactcactataggg
HSC041	pDONR207_SeqR	gtgtctcaaaatctctgatg

HSC042	pH9GW_SeqR	caaaaaacccctcaagac
HSC063	Helix2-SMH3-mut PF	TGAAAATTACCTTTTCAACGAGGTTGAACCTCTTTTGAATAAATGGCCCT TGAATAAG
HSC064	Helix2-SMH3-mut PR	CGTTGAAAAGGTAATTTTCAAGTTCGATTTGCTCTTCGGTATGAGGGTA GTATAGATCC
HSC067	Helix2-SME1-mut PF	TGGAGGATTATTTATATAACGAAGTTGAGCCGCTTGTGAATAAATGGCC CTTGAATAAG
HSC068	Helix2-SME1-mut PR	TCGTTATATAAATAATCCTCCAGTTCTATCAGCTCTTCGGTATGAGGGTA GTATAGATCC
HSC071	Helix2-SMH6-mut PF	TTTACGACTATCTCTACAACGAGGTCGAACCTTTACTTAACAAATGGCC CTTGAATAAG
HSC072	Helix2-SMH6-mut PR	TCGTTGTAGAGATAGTCGTAAAGATCAAGTAACTCCTGGCTATGAGGGT AGTATAGATCC
HSC075	Helix2-RDH4-mut PF	TATGGGATTCACTTCACCAAGAGTGTGAGCCTCAAGTGAAAAAATGGCC CTTGAATAAG
HSC076	Helix2-RDH4-mut PR	TCTTGGTGAAGTGAATCCCATAACTCATCCTGCTGTGATGAATGAGGGT AGTATAGATCC
HSC081	Helix2-SME2-mut PF	TGCAAGATCAGCTTTACAACGAGGTAGAACCGCTGGTAGAGAAATGGC CCTTGAATAAGC
HSC082	Helix2-SME2-mut PR	TTGTAAAGCTGATCTTGCAAGGCCAAGAGCTCATCAGTATGAGGGTAGT ATAGATCCAAC
HSC083	Helix2-RDH2-mut PF	TATGGGATTCCCTCCATCAACAGTGCGAACCTCAAGTCAAGAAATGGC CCTTGAATAAGC
HSC084	Helix2-RDH2-mut PR	TGATGGAGGGAATCCCATAACTGGTCCTGTTGTGAAGAATGAGGGTAG TATAGATCCAAC
HSC087	SapI-McLS-PF	ttGCTCTTCgtctctaATGAAGCACTTGTTGTCA
HSC088	SapI-McLS-PR	ttGCTCTTCgtctctAAGCTCAACGTCTATATCCAGTG
HSC091	SapI_CoTXSS_F	GGGCTCTTCgtctcGAATGTGGAAATTAAAAATTGGTGA
HSC092	SapI_CoTXSS_R	GGGCTCTTCgtctcCAAGCtcaGACTTGTTGCTTTGGC
HSC093	SapI_(GS)5-YFP_F1	ccggttctggaagtggatccTTAGTGAGCAAGGGC
HSC094	SapI_(GS)5-YFP_F2	GGGCTCTTCgtctcGGGTAgcggatccggttctggaagtg

HSC095	SapI_YFP_R	GGGCTCTTCgtctcCAAGCTCACTTGTACAGC
HSC099	SapI_CoTXSS_NoSTOP_R	GGGCTCTTCgtctcCTACCGACTTGTTGCTTTGGC
HSC100	SapI_McLS_NoSTOP_R	GGGCTCTTCgtctcCTACCACGTCTATATCCAGTGTCA
HSC103	attB1_McLS_PF	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCATGAAGCACTTGTTGTCACTCC
HSC104	attB2_McLS_PR	GGGGACCACTTTGTACAAGAAAGCTGGGTTTCAACGTCTATATCCAGTGTCATG
HSC105	SapI_mCherry_HDEL_F_AGGT	GGGCTCTTCgtctcGAGGTATGGTGAGCAAGGGC
HSC106	SapI_mCherry_HDEL_R_GCTT	GGGCTCTTCgtctcCAAGCTTACAGTTCATCATGCTTGTACAGCTCGTCCAT
HSC107	SapI_OsRAmy3A_F_AATG	GGGCTCTTCgtctcGAATGGGGAAGCAAATGGC
HSC108	SapI_OsRAmy3A_R_AGGT	GGGCTCTTCgtctcCACCTGACGCGACGTCG
HSC109	SapI_CoTXSS_NoSTOP_McLSCtermaware_R	GGGCTCTTCgtctcGGACTTGTTGCTTTGGC
HSC110	SapI_McLSCterm_YFP_CoTXSSaware_F	GGGCTCTTCgtctcGagtcGCTGGCCCGG

Supplementary Table 2 Primers used in this thesis

Plasmid name	Backbone	Insert	5' Overhang	3' Overhang
pEPHS0CM0038	pUAP4	pCoTXSS	GGAG	AATG
pEPHS0CM0039	pUAP4	pCoCYP2	GGAG	AATG
pEPHS0CM0040	pUAP1	pCoACT1	GGAG	AATG
pEPHS0CM0041	pUAP1	pCoACT2	GGAG	AATG
pEPHS0CM0046	pUAP4	CoMYB24	AATG	GCTT
pEPHS0CM0047	pUAP4	CoMYB111	AATG	GCTT
pEPHS0CM0054	pTwist	TdTXSS	AATG	TTCG
pEPHS0CM0067	pUAP4	McLS	AATG	GCTT
pEPHS0CM0070	pUAP4	CoTXSS_SME1	AATG	GCTT
pEPHS0CM0071	pUAP4	CoTXSS_SME2	AATG	GCTT
pEPHS0CM0072	pUAP4	CoTXSS_SMH3	AATG	GCTT
pEPHS0CM0073	pUAP4	CoTXSS_SMH6	AATG	GCTT
pEPHS0CM0074	pUAP4	CoTXSS_RDH2	AATG	GCTT
pEPHS0CM0075	pUAP4	CoTXSS_RDH4	AATG	GCTT
pEPHS0CM0077	pUAP4	mCherry_HDEL	AGGT	GCTT
pEPHS0CM0078	pUAP4	McLS (No stop codon)	AATG	GGTA
pEPHS0CM0100	pUAP4	OsRAmy3A Signal Peptide	AATG	AGGT
pEPHS0CM0102	pUAP4	(GS)5 + YFP	GGTA	GCTT
pEPHS0CM0112	pUAP4	McLS C-terminal + (GS)5 + YFP	AGTC	GCTT
pEPHS0CM0113	pUAP4	CoTXSS (No stop codon)	AAGT	AGTC
pEPHS0CM0114	pUAP4	CoTXSS_SME1 (No stop codon)	AAGT	AGTC
pEPHS0CM0115	pUAP4	CoTXSS_SME2 (No stop codon)	AAGT	AGTC
pEPHS0CM0116	pUAP4	CoTXSS_SMH3 (No stop codon)	AAGT	AGTC

pEPHS0CM0117	pUAP4	CoTXSS_SMH6 (No stop codon)	AAGT	AGTC
pEPHS0CM0118	pUAP4	CoTXSS_RDH2 (No stop codon)	AAGT	AGTC
pEPHS0CM0119	pUAP4	CoTXSS_RDH4 (No stop codon)	AAGT	AGTC

Supplementary Table 3 Level 0 plasmids used in this thesis

Plasmid name	Assembly method	Backbone	Promoter + 5' UTR	N-terminal tag	CDS	C-terminal tag	Terminator
pEPHS1CB0001	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS G380T		CaMV35s
pEPHS1CB0002	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS D385E		CaMV35s
pEPHS1CB0003	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS H492Q		CaMV35s
pEPHS1CB0004	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS P751A		CaMV35s
pEPHS1CB0005	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS G380D		CaMV35s
pEPHS1CB0006	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS D385A		CaMV35s
pEPHS1CB0007	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS H492A		CaMV35s
pEPHS1CB0008	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS P751D		CaMV35s
pEPHS1CB0021	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS T374G		CaMV35s
pEPHS1CB0010	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS E379D		CaMV35s
pEPHS1CB0011	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS Q486H		CaMV35s
pEPHS1CB0012	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS A745P		CaMV35s
pEPHS1CB0013	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS T374D		CaMV35s
pEPHS1CB0014	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS E379A		CaMV35s

pEPHS1CB0015	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS Q486A		CaMV35s
pEPHS1CB0016	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS A745D		CaMV35s
pEPHS1CB0017	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS D385E H492Q		CaMV35s
pEPHS1CB0020	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS G380T D385E H492Q P751A		CaMV35s
pEPHS1CB0022	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS E379D Q486H		CaMV35s
pEPHS1CB0027	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS T374G E379D Q486H A745P		CaMV35s
pEPHS1CB0028	MoClo	pICH47732	CaMV35s + TMV omega		TdTXSS		CaMV35s
pEPHS1CB0029	MoClo	pICH47732	CaMV35s + TMV omega		CoMAS I367M		CaMV35s
pEPHS1CB0030	MoClo	pICH47732	CaMV35s + TMV omega		CoMAS E371D		CaMV35s
pEPHS1CB0031	MoClo	pICH47732	CaMV35s + TMV omega		CoMAS I367M E371D		CaMV35s
pEPHS1CB0032	MoClo	pICH47732	CaMV35s + TMV omega		CoTXSS Y273F		CaMV35s
pEPHS1CB0033	MoClo	pICH47732	CaMV35s + TMV omega		TkTXSS Y267F		CaMV35s
pEPHS1KN0042	Loop	pCk1	pCoTXSS		LucN	3xFLAG	AtuNOS
pEPHS1KN0043	Loop	pCk1	pCoCYP2		LucN	3xFLAG	AtuNOS
pEPHS1KN0044	Loop	pCk1	pCoACT1		LucN	3xFLAG	AtuNOS
pEPHS1KN0045	Loop	pCk1	pCoACT2		LucN	3xFLAG	AtuNOS

pEPHS1KN0048	Loop	pCk2	CaMV35s + TMV omega		CoMYB24		CaMV35s
pEPHS1KN0049	Loop	pCk2	CaMV35s + TMV omega		CoMYB111		CaMV35s
pEPHS1KN0068	Loop	pCk2	CaMV35s + TMV omega		McLS		CaMV35s
pEPHS1KN0086	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SME1		CaMV35s
pEPHS1KN0087	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SME2		CaMV35s
pEPHS1KN0088	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SMH3		CaMV35s
pEPHS1KN0089	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SMH6		CaMV35s
pEPHS1KN0090	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_RDH2		CaMV35s
pEPHS1KN0091	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_RDH4		CaMV35s
pEPHS1KN0101	Loop	pCk2	CaMV35s + TMV omega	OsRAmy3 signal peptide	mCherry	HDEL	CaMV35s
pEPHS1KN0104	Loop	pCk2	CaMV35s + TMV omega		McLS	(GS)5 + YFP	CaMV35s
pEPHS1KN0120	Loop	pCk2	CaMV35s + TMV omega		CoTXSS	McLS C-terminal + (GS)5 + YFP	CaMV35s
pEPHS1KN0121	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SME1	McLS C-terminal + (GS)5 + YFP	CaMV35s
pEPHS1KN0122	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SME2	McLS C-terminal + (GS)5 + YFP	CaMV35s

pEPHS1KN0123	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SMH3	McLS C-terminal + (GS)5 + YFP	CaMV35s
pEPHS1KN0124	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_SMH6	McLS C-terminal + (GS)5 + YFP	CaMV35s
pEPHS1KN0125	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_RDH2	McLS C-terminal + (GS)5 + YFP	CaMV35s
pEPHS1KN0126	Loop	pCk2	CaMV35s + TMV omega		CoTXSS_RDH4	McLS C-terminal + (GS)5 + YFP	CaMV35s

Supplementary Table 4 Level 1 plasmids used in this thesis

Entry plasmid	Backbone	Insert
pEPHSdGM0050	pDONR207	CoTXSS
pEPHSdGM0051	pDONR207	CoTXSS_H1_mutated
pEPHSdGM0052	pDONR207	CoTXSS_H3_mutated
pEPHSdGM0053	pDONR207	CoTXSS_H1_H3_mutated
pEPHSdGM0055	pDONR207	CoTXSS_SME1
pEPHSdGM0056	pDONR207	CoTXSS_SME2
pEPHSdGM0057	pDONR207	CoTXSS_SMH3
pEPHSdGM0058	pDONR207	CoTXSS_SMH6
pEPHSdGM0059	pDONR207	CoTXSS_RDH2
pEPHSdGM0060	pDONR207	CoTXSS_RDH4
pEPHSdGM0105	pDONR207	McLS
Expression plasmid	Backbone	Insert
pEPHSeKN0104	pH9GW	CoTXSS
pEPHSeKN0061	pH9GW	CoTXSS_SME1
pEPHSeKN0062	pH9GW	CoTXSS_SME2
pEPHSeKN0063	pH9GW	CoTXSS_SMH3
pEPHSeKN0064	pH9GW	CoTXSS_SMH6
pEPHSeKN0065	pH9GW	CoTXSS_RDH2
pEPHSeKN0066	pH9GW	CoTXSS_RDH4
pEPHSeKN0106	pH9GW	McLS

Supplementary Table 5 Gateway entry and expression plasmids used in this study