Investigating the biosynthesis of major saponariosides in soapwort (*Saponaria officinalis*)

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

Seohyun Jo

December 2022

Abstract

Soapwort (*Saponaria officinalis*) is a flowering plant in the Caryophyllaceae family with a long history of use in human civilization as a traditional source of soap. Even today, soapwort extracts are still used in laundry detergents, cosmetics, herbal medicine, and as food additives. The well-known detergent properties of soapwort are due to the large amounts of bioactive saponins produced by this plant. Saponins present in soapwort extracts are triterpenoid glycosides which often have important pharmaceutical, nutraceutical and agronomical potentials. However, the properties of individual saponin components in soapwort are not well understood as these compounds are present in complex mixtures and thus difficult to isolate. Metabolic engineering may provide an alternative supply of pure soapwort saponins. At the commencement of this project, nothing was known about the biosynthesis of soapwort saponins. The overall aim of this project was to investigate the biosynthesis of major saponins (saponariosides A and B) found in soapwort. Metabolic analysis of soapwort organs revealed flowers as a potential major site of saponarioside biosynthesis. Based on this knowledge, RNA-Seq and genome sequence resources were generated for the discovery of saponarioside biosynthetic genes. Using these new sequence resources, a total of 13 saponarioside biosynthetic enzymes were identified, completing the biosynthetic pathway to saponarioside B. Only one step remains to be discovered to complete the pathway to saponarioside A. The newly characterized biosynthetic genes presented in this project open-up opportunities for metabolic engineering of soapwort saponins and analogues in heterologous systems, which may lead to large-scale production and biochemical studies of these biologically active saponins in the future.

Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my supervisor Prof. Anne Osbourn for her guidance, wisdom, and kindness. I am incredibly grateful and honoured to have finished my PhD with her, with a project I so dearly love and thoroughly enjoy. If I were to repeat my PhD again (let's hope not) I would, without hesitation, repeat it with Anne. I would also like to thank my secondary supervisor Prof. Andrew Truman for his help and advice.

I would also like to thank everyone in the Osbourn group for making the last four years such an enjoyable and educational experience. Thank you everyone for always answering my random questions and never hesitating to offer help. In particular, I would like to thank Dr. James Reed for being a walking triterpene encyclopaedia and always giving helpful and kind advice. I would also like to express my gratitude to Dr. Amr El-Demerdash for his help in everything chemistry related and isolating all the sticky sugars. I would also like to thank Dr. Hannah Hodgson and Dr. Charlotte Owen, not only for all the scientific help they have given but also for their truly wonderful friendship.

I am indebted to the JIC metabolomics team, especially Dr. Lionel Hill and Dr. Paul Brett, for always having our backs whenever the LC or GC or nitrogen failed us. I would also like to thank my fellow Rotation 2018 friends, Anna Backhaus, Jiawen Chen, Lauren Grubb and Thomas Gate, for being the very best group of people I could have shared this journey with. Also thank you to Andy Chen for driving me to JIC on New Year's Eve so I can get the submission form scanned.

I would also like to thank my family and friends. Especially Jane, who has continuously nagged me about writing my thesis and encouraged me by sending pictures of her really cute dog. I would also like to thank my brother Yooshin for his support; I am constantly in awe of his kindness and strength, and I cannot imagine my life without him. Lastly, I am forever grateful to my parents who have sacrificed everything to give my brother and I the best education we can receive. They have always supported everything I ever wanted to pursue, and their endless love and encouragements have given me confidence to challenge anything.

Abbreviations

$[M-H]^-$	Negative ion, minus hydrogen
1D	One dimensional
1KP	1000 Plants project
1n	Haploid
2D	Two dimensional
2n	Diploid
aa	Amino acid
AD	Anno Domini
AHRD	Automatic assignment of human readable description
ANOVA	One-way analysis of variance
AT	Acyltransferase
ATP	Adenosine triphosphate
BAHD	BEAT-AHCT-HCBT-DAT
bAS	β-Amyrin synthase
BGC	Biosynthetic gene cluster
BLAST	Basic local alignment search tool
BLASTP	Protein BLAST
bp	Base pair
BUSCO	Benchmarking universal single-copy orthologs
CAZymes	Carbohydrate-active enzymes
CCC	Chromosome conformation capture
CCS	Circular consensus sequences
cDNA	Copy DNA
CesA	Cellulose synthase
CSL	Cellulose synthase-like
CsyGT	Cellulose synthase superfamily-derived GT
CYP	Cytochrome P450 monooxygenase
dd-MS^2	Data dependent-tandem mass spectrometry
DMAPP	Dimethylallyl diphosphate
DNA	Deoxyribonucleic acid

DW	Dry weight
EI	Earlham Institute
EIC	Extracted ion chromatogram
ESI	Electrospray ionization
EST	Expressed sequence tag
FCC	Flash column chromatography
FPP	Farnesyl diphosphate
FPPS	Farnesyl diphosphate synthase
FRET	Förster resonance energy transfer
GC	Gas chromatography
gDNA	Genomic DNA
GFPP	Geranylfarnesyl diphosphate
GFPPS	Geranylfarnesyl diphosphate synthase
GGPP	Geranylgeranyl diphosphate
GGPPS	Geranylgeranyl diphosphate synthase
GH	Glycosyl hydrolase
GPP	Geranyl diphosphate
GPPS	Geranyl diphosphate synthase
GT	Glycosyltransferase
Hi-C	High-throughput chromosome conformation capture
HMW	High molecular weight
HPLC	High pressure liquid chromatography
HR	High resolution
IPP	Isopentenyl diphosphate
Iso-Seq	Isoform sequencing
JGI	Joint Genome Institute
JIC	John Innes Centre
LB	Lysogeny broth
LC	Liquid chromatography
m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser desorption ionization
MEP	Methylerythritol phosphate
MS	Mass spectrometry

MS/MS, MS^2	Tandem mass spectrometry
MVA	Mevalonic acid
MW	Molecular weight
NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	National Centre for Biotechnology Information
NGS	Next-generation sequencing
NMR	Nuclear magnetic resonance
OD	Optical density
ORF	Open reading frame
OS	2,3-Oxidosqualene
OSC	2,3-Oxidosqualene cyclase
PCA	Principal component analysis
PCC	Pearson correlation coefficients
PCR	Polymerase chain reaction
QA	Quillaic acid
QS	*Quillaja saponaria*
RIP	Ribosome inactivating protein
RNA	Ribonucleic acid
RNA-Seq	Ribonucleic acid sequencing
RP	Reverse phase
rpm	Rotation per minute
RSM	Rockland St. Mary
RT	Retention time
SCP	Serine carboxypeptidase
SCPL	Serine carboxypeptidase-like
SDR	Short-chain dehydrogenase/reductase
SMRT	Single molecule real-time
SOC	Super optimal broth with catabolite
SpA	Saponarioside A
SpB	Saponarioside B
spp	Species
SQE	Squalene epoxidase
TG	Transglycosidase

Th1	Type 1 T helper cell
Th2	Type 2 T helper cell
TIC	Total ion chromatogram
UGT	UDP-dependent glycosyltransferase
UPD	Uridine diphosphate
UV	Ultraviolet
VIGS	Virus-induced gene silencing

1

General introduction

1.1 Plant specialized metabolites

Plants produce an enormous array of compounds with diverse chemical structures. These compounds are often divided into ‘primary’ or ‘specialized’ metabolites. Primary metabolites have direct roles associated with growth, development, photosynthesis and respiration while specialized metabolites have been long overlooked as ‘secondary metabolites’ and were deemed as nonessential, accessory metabolites (Hartmann, 2007). However, specialized metabolites have important ecological functions, for example by providing protection against biotic and abiotic stresses, including pathogens, herbivores, and UV radiation (Osbourn and Lanzotti, 2009). They can also aid in attracting pollinators, allelopathy, and in intra- and inter-species communication (Osbourn and Lanzotti, 2009).

Plant specialized metabolites can be broadly grouped into three major categories: phenylpropanoids, alkaloids and terpenoids. However, as plant specialized metabolites show enormous structurally diversity, not all metabolites strictly fall into the above three categories, such as glucosinolates and other sulfur-containing metabolites. Specialized metabolites are thought to have derived from various areas of primary metabolism (Weng, 2014). Phenylpropanoids, including phenolic polymers such as lignin and tannins, cinnamic acid derivatives and flavonoids, are biosynthesised from precursors produced by the shikimate pathway (Springob and Kutchan, 2009). Alkaloids are nitrogen containing compounds biosynthesized primarily from amino acids, for example tyrosine (morphine), arginine (nicotine), and tryptophan (monoterpenoid indole alkaloids), but can also utilise nucleotides, which is the case for purine alkaloids such as caffeine (Ziegler and Facchini, 2008).

As mentioned earlier, plant specialized metabolites have important ecological roles. For example, anthocyanins (flavonoids) are responsible for red, pink, blue and purple pigments of many flowers, fruits and leaves, which attract pollinators and seed dispersers (Springob and Kutchan, 2009). Flavonoids also act as UV-B filters by absorbing 280-315 nm wavelength energy and are found in high concentrations in the epidermal layers of leaves and fruits (Harborne and Williams, 2000). Many alkaloids are bitter-tasting antifeedants or exhibit neurotoxicity and cell signalling disruption if ingested. Glycoalkaloids, such as solanine and chaconine, are commonly found in the nightshade (Solanaceae) family and cause gastrointestinal and neurological disorders (Matsuura and Fett-Neto, 2015). Some specialized metabolites are highly volatile and are emitted during specific conditions. For example, isoprene (terpenoid) is emitted into the atmosphere by trees, such as poplar and aspen, to offer protection from high temperatures, reactive oxygen species and ozone (Sharkey, Wiberley and Donohue, 2008).

Due to their wide spectrum of bioactivities, plant specialized metabolites have been exploited by humans for centuries as medicines, flavours, fragrances, and for other applications (Bourgaud *et al.*, 2001). For instance, volatile phenolic compounds such as cinnamaldehyde and eugenol (phenylpropanoids) from cinnamon (*Cinnamomum ceylanicum*) and clove (*Syzygium aromaticum*) trees, respectively, have been long employed as spices and herbal remedies (Springob and Kutchan, 2009). In modern medicine, paclitaxel, a diterpene first discovered in the bark of Pacific yew trees (*Taxus brevifolia*), is used in cancer chemotherapy (McGuire *et al.*, 1989). In agriculture, seed extracts of neem trees (*Azadirachta indica*), which contain high levels of the limonoid azadirachtin (terpenoid), have a long history of use in traditional methods of crop protection (Morgan, 2009).

Recent advances in next-generation sequencing (NGS), metabolite analysis and synthetic biology have greatly enhanced our understanding of the biosynthesis of plant specialized metabolism in the last decade. These new technologies have advanced our knowledge of specialized metabolism and accelerated biosynthetic pathway discovery in non-model plant species (Torrens-Spence, Fallon and Weng, 2016). As more enzymes involved in biosynthesis of plant specialized metabolites and their precursors have been discovered, heterologous production of these high-value molecules in

heterologous hosts is becoming more feasible. For example, the full opiate biosynthetic pathway has been engineered in yeast, providing an alternative source for this essential medicine (Galanie *et al.*, 2015).

Investigations of plant specialized metabolic pathways have also raised interesting evolutionary questions, such as a how chemo-diversity has risen in plants. As mentioned above, plant specialized metabolism is hypothesized to have diverged from primary metabolism through gene duplication followed by non-deleterious mutations (Weng, Philippe and Noel, 2012). This resulted in promiscuity in enzyme activity, leading to broadened substrate specificity and numerous products. Coupled investigation of enzyme biochemistry and phylogenetic analysis has also led to deeper understanding of evolution of specific biochemical pathways, such as the iridoid biosynthesis in the mint (Lamiaceae) family. Despite being widespread amongst species of the mint family, iridoid biosynthesis has been lost in the subfamily Nepetoideae expect in one genus, *Nepeta*, where nepetalactones are made. Simultaneous investigation into enzymology of ancestral biosynthetic enzymes and comparative genomics of iridoid producing and non-producing species in Nepetoideae provided evidence that gene duplication of promiscuous ancestral genes followed by recruitment into gene clusters led to the re-emergence of iridoid biosynthesis in *Nepeta* (Lichman *et al.*, 2020). However, the state of understanding of biosynthesis across the different types of specialized metabolites are widely variable. Many specialized metabolic pathways remain unidentified with little to no genetic information available.

1.2 Terpenes

Terpenes (also referred to as terpenoids) constitute the largest and most structurally diverse group of phytochemicals with more than 80,000 compounds identified so far (Zhou and Pichersky, 2020). Although the term ‘terpene’ refers to simple hydrocarbons made up of isoprene units while ‘terpenoids’ are modified terpenes with different functional groups, both terms are often used interchangeably and will be used as such throughout this thesis. All terpenes are composed of the basic 5-carbon (C-5) isoprene precursors which are biosynthesized in different compartments of the cell such as cytosol, plastid and mitochondria (Bouvier, Rahier and Camara, 2005). The condensation of multiple isoprene precursors give rise to the huge structural diversity of terpenes, ranging from monoterpenes (C-10),diterpenes (C20), sesterterpenes (C25)

triterpenes (C-30), tetraterpenoids/carotenoids (C-40) and polyterpenoids which are made up from more than eight isoprene units and are encountered in materials such as rubber. Plants use two distinct pathways for terpene biosynthesis: the cytosolic mevalonic acid (MVA) pathway and the plastidial methylerythritol phosphate (MEP) pathway (Fig. 1.2.1). The MVA pathway is ubiquitously found in most eukaryotes and archaea but only in some bacteria as most bacteria employ the alternative MEP pathway (Kuzuyama and Seto, 2012). Although the two pathways are unique, they both synthesize the common isoprene precursors isopentenyl diphosphate (IPP), which is also isomerized to dimethylallyl diphosphate (DMAPP) by isopentenyl diphosphate isomerase (IPPI).

Plants are unique in possessing both the MVA and MEP pathway, a feature likely inherited through the endosymbiosis of a cyanobacterium-like cell that served as ancestors to modern-day chloroplasts (Hemmerlin, Harwood and Bach, 2012). The retainment of the two pathways may be advantageous to sessile organisms like plants, by providing advanced regulation of fixed carbon and ATP sources accordingly based on availability (Vranová, Coman and Gruissem, 2013).

Metabolic crosstalk between the MVA and MEP pathways exists but is very limited and is strictly controlled at both transcript and protein level, with additional feedback regulatory mechanisms (Vranová, Coman and Gruissem, 2013). Monoterpenes, diterpenes and sesquiterpenes can be synthesized by precursors produced by both pathways under specific conditions, while triterpenes and tetraterpenes follow strict compartmentalization (Hemmerlin, Harwood and Bach, 2012). Broadly, the MVA pathway provides precursors for the cytosolic biosynthesis of sesquiterpenoids (C-15) and triterpenoids (C-30), while IPP and DMAPP produced by the MEP pathway are used for the biosynthesis of monoterpenoids (C-10), diterpenoids (C-20), sesterterpenoids (C-25) and tetraterpenoids (C-40) in the plastid (Tholl, 2015). IPP and DMAP serve as C-5 building blocks for subsequent terpene biosynthesis through three chemical reactions: 1. “head-to-tail” elongation of isoprene units, 2. “head-to-head” condensation of isoprene units, and 3. cyclization of linear precursors (Hemmerlin, Harwood and Bach, 2012). As such, linear prenyl pyrophosphates with various chain lengths arise from the condensation of DMAPP and IPP by the activity of prenyltransferases.

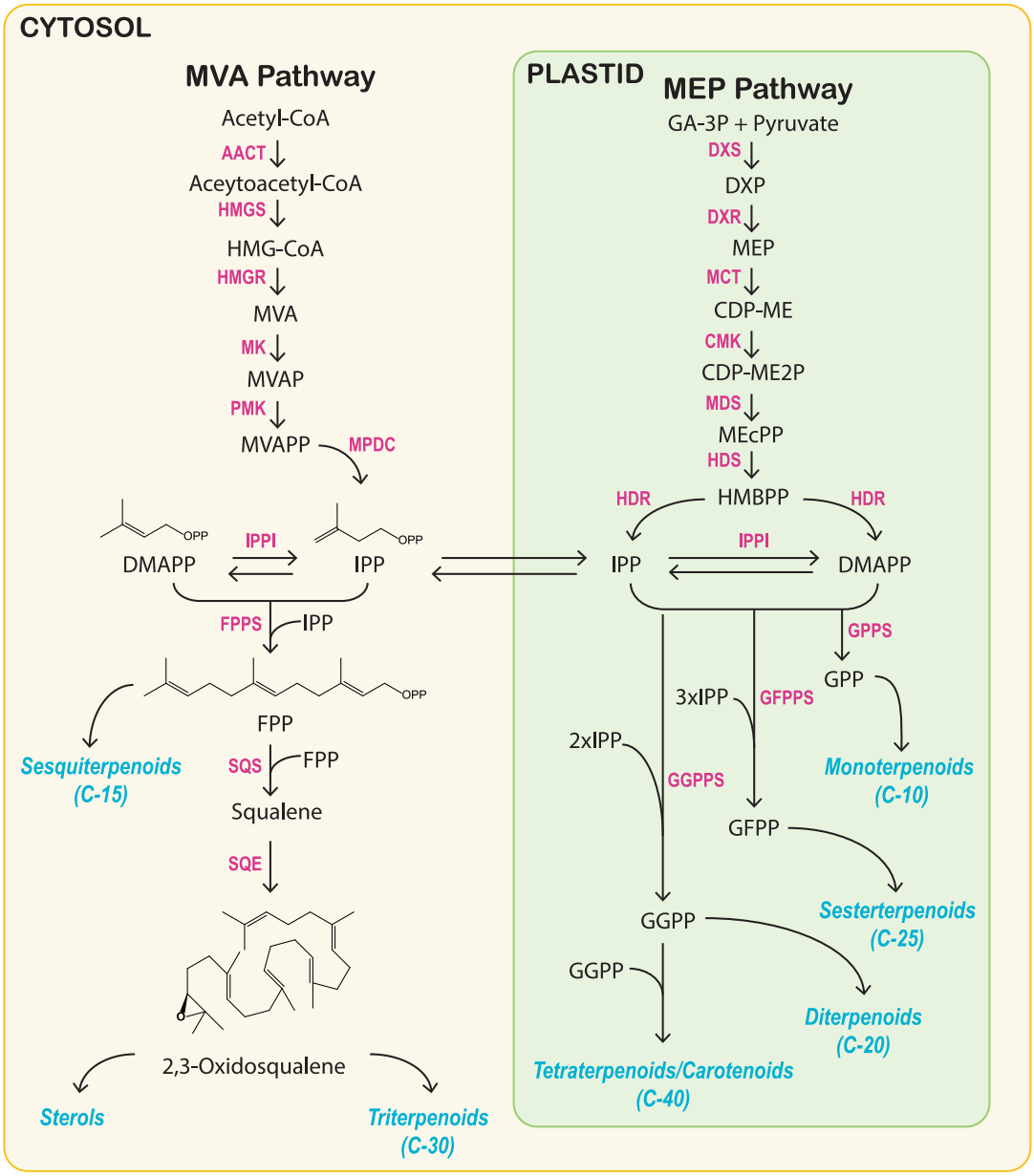


Figure 1.2.1. The plant mevalonate (MVA) and methylerythritol (MEP) pathways. The MVA pathway is in the cytosol (yellow) while MEP pathway occurs in the plastid (green). Trafficking of IPP and DMAP between the two compartments are shown by double arrows. Enzymes are in pink and terpene products are in blue. Abbreviations: MVA, mevalonic acid; MEP, 2-C-methyl-d-erythritol 4-phosphate; IPP, isopentyl diphosphate; DMAPP, dimethylallyl diphosphate; IPPI, isopentenyl diphosphate isomerase; AACT, acetoacetyl-CoA thiolase; HMG-CoA, 3-hydroxy-3-methylglutaryl-CoA; HMGS, HMG-CoA synthase; HMGR, HMG-CoA reductase; MVAP, melavonate-5-phosphate; MK, mevalonate kinase; MVAPP, mevalonate-diphosphate; PMK, phosphomevalonate kinase; MPDC, mevalonate diphosphate decarboxylase; GA-3P, glyceraldehyde-3-phosphate; DXP, 1-deoxy-d-xylulose 5-phosphate; DXS, DXP synthase; DXR, DXP reductoisomerase; MCT, MEP cytidylyltransferase; CDP-ME, 4-(cytidine 5′-diphospho)-2-C-methyl-D-erythritol; CDP-ME2P, 2-phospho-4-(cytidine 5′-diphospho)-2-C-methyl-D-erythritol; CMK, CDP-ME kinase; MEcPP, 2-C-methyl-D-erythritol-2,4-cyclodiphosphate; MDS, MEcPP synthase; HMBPP, 4-hydroxy-3-methylbut-2-enyl-diphosphate; HDS,

HMBPP synthase; HDR, 4-hydroxy-3-HMBPP reductase; FPP, farnesyl diphosphate; FPPS, FPP synthase; SQS, squalene synthase; SQE, squalene epoxidase; GPP, geranyl diphosphate; GPPS, GPP synthase; GFPP, geranylfarnesyl diphosphate; GFPPS, GFPP synthase; GGPP, geranylgeranyl diphosphate; GGPPS, GGPP synthase. Figure adapted from (Hemmerlin, Harwood and Bach, 2012) and (Vranová, Coman and Gruissem, 2013).

For example, monoterpenes (C-10) are derived from geranyl diphosphate (GPP) produced by GPP synthase (GPPS), sesquiterpenes (C-15) are derived from farnesyl diphosphate (FPP) backbone produced by FPP synthase (FPPS), diterpenes (C-20) are derived from geranylgeranyl diphosphate (GGPP) produced by GGPP synthase (GGPPS), and sesterterpenes (C-25) are derived from geranylfarnesyl diphosphate (GFPP) by GFPP synthase (GFPPS). The backbones of triterpenes (C-30) and tetraterpenes (C-40) are produced from the condensation of two FPPs and GGPPs, respectively (Fig. 1.2.1). The linear backbones are then cyclized to diverse array of skeletons by the activity of terpene synthases, and may undergo further oxidation and rearrangements, leading to the endless diversity of terpene structures (Hemmerlin, Harwood and Bach, 2012; Hemmerlin, 2013).

1.3 Triterpenes

1.3.1. Introduction

Triterpenes are structurally diverse C-30 molecules found widespread in all organisms. The term ‘triterpenes’ constitutes two classes of molecules: sterol and non-steroidal triterpenes. Sterols play essential roles in physiology of eukaryotic organisms, such as in regulating membrane fluidity and permeability, as well as in hormone signalling and transduction particularly via brassinosteroids (Hartmann, 1998). Sterols in plants also serve as precursors for synthesis of wide range of specialized metabolites such as steroidal alkaloids, saponins and cardenolides (Kreis and Müller-Uri, 2010). In additional to sterols, plants also produce non-steroidal triterpenes that are often associated with specialized metabolism. This thesis will focus on these non-steroidal triterpenes, which the term ‘triterpenes’ will refer to hereafter.

Triterpenes display a wide range of bioactivities associated with variety of functions in plants. Undecorated triterpenes, such as the pentacyclic β-amyrin, are lipophilic molecules that may play structural roles by maintaining the flexibility of waxy plant cuticles (Buschhaus and Jetter, 2012). They are also implicated in growth and

development as altered accumulation of these simple triterpenes have been associated with physiological changes in the plant. For example, in oat (*Avena strigosa*), increased accumulation of β-amyrin leads to the formation of roots with super-hairy phenotype (Kemen *et al.*, 2014). Another example is in the model legume species, *Lotus japonicus*, where lupeol, another pentacyclic triterpene, is involved in suppression of nodule formation (Delis *et al.*, 2011). Triterpenes are also well known for their roles in plant defence, exhibiting cytotoxicity (e.g. betulin and betulinic acid derivatives) and antifeedant (e.g. limonoids and quassinoids) activities to name a few (González-Coloma *et al.*, 2011). In particular, glycosylated triterpenes (known as saponins) are associated with antimicrobial, antifungal and molluscicidal activities and are characterized by their ability to form stable foams in water. (Osbourn, 1996a; Sparg, Light and Van Staden, 2004). For instance, oat roots produce avenacins with antifungal properties that have been implicated in the resistance of oats to soil-borne diseases such as take-all (Papadopoulou *et al.*, 1999).

Humans have also benefited from the bioactivities of triterpenes (Fig. 1.3.1). Ginsenosides found in the *Panax* spp. (ginseng), a popular ingredient in traditional Chinese medicine, show potential anticarcinogenic, anti-inflammatory and antidiabetic properties (Christensen, 2008). Triterpenes have also played important religious roles; for example, boswellic acids are the major constituent of frankincense, and are additionally anti-inflammatory compounds (Shah, Qazi and Taneja, 2009). Beyond their traditional uses, triterpenes have generated significant interests as agents of novel therapeutics. For example, celastrol, a pentacyclic triterpene extracted from *Tripterygium* spp., displays a wide variety of biological properties, including anti-obesity activities in mice (Liu *et al.*, 2015). Another area of interest is the use of saponins as vaccine adjuvants. QS-21, a complex triterpene saponin isolated from the Chilean soapbark tree (*Quillaja saponaria*), is an immunostimulatory adjuvant and is approved for human use as a vaccine adjuvant in Shingrix® and Mosquirix®, vaccines for shingles and malaria, respectively (Wang, 2021). QS-21 and other QS-saponins are known to stimulate several immune responses required for prolonged immunity after vaccination. These include the stimulation of mixed T-cell helpers, Th1 and Th2, which leads to cellular and humoral immunity, respectively, and the production of cytotoxic T-lymphocytes against exogenous antigens (Sun, Xie and Ye, 2009).

However, only limited knowledge is available regarding the molecular mechanisms of how QS-saponins elicit these immune responses (Fernández-Tejada *et al.*, 2014).

The potential use of triterpenes for human health has attracted considerable interests in the chemical synthesis and development of semi-synthetic derivatives of these naturally occurring triterpenes. However, the complexity of triterpene structures makes chemical synthesis of these molecules difficult. For example, the total synthesis of azadirachtin took 22 years to achieve and requires over 70 steps, yielding 0.00015% of the final product (Jauch, 2008). Thus, the primary source of these high-value compounds are usually the producing plants itself (Reed and Osbourn, 2018). Many of these compounds are lowly accumulating and can be present in complex mixtures, preventing extraction at a commercial scale. Furthermore, the cultivation of the source plant may be restricted to specific climates and can be time consuming. Another method of accessing potential therapeutic triterpenes is through metabolic engineering in a heterologous host; however, this method is also restricted by the limited knowledge of enzymes involved in triterpene biosynthetic pathways.

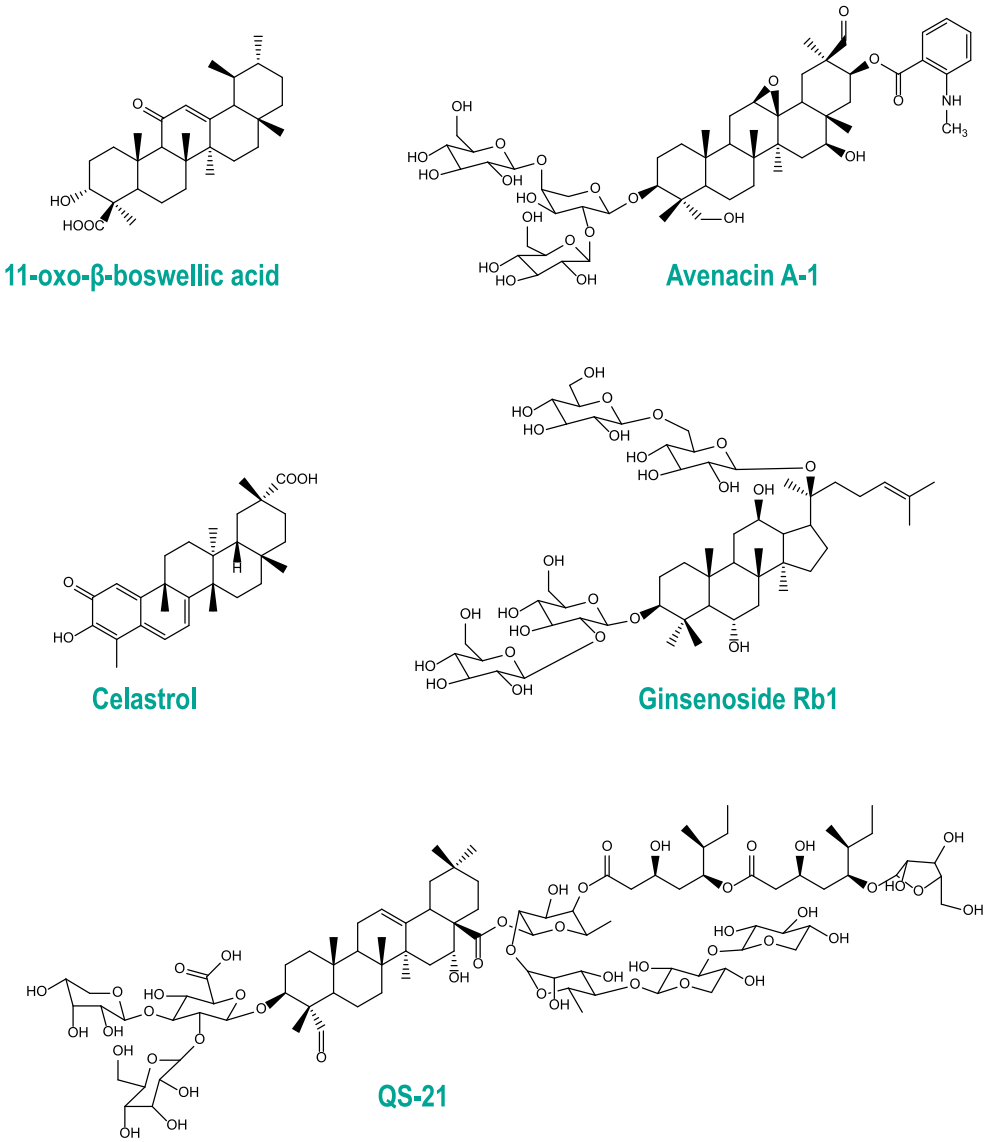
1.3.2. Biosynthesis

Generation of the triterpene scaffold

The biosynthesis of sterols and triterpenes begin with the cyclization of squalene or 2,3-oxidosqualene in the cytosol (Fig. 1.3.2). In bacteria, squalene can be directly cyclized by the activity of squalene-hopene cyclases (Siedenburg and Jendrossek, 2011). In eukaryotes such as plants, squalene is typically further oxidized to 2,3-oxidosqualene (OS) by squalene epoxidase (SQE), giving rise to the characteristic C-3 oxygen species of sterols and triterpenes (Augustin *et al.*, 2011). This substrate is then subsequently cyclized by 2,3-oxidosqualene cyclases (OSCs) by cascade of cationic attacks leading to various one to five ringed terpenes (Augustin *et al.*, 2011). The cyclization of the A, B, and C rings of 2,3-oxidosqualene through the *chair-chair-chair* conformation leads to the formation of dammerenyl cation which give rise to most triterpenes, while sterols originate from the protosteryl cation, formed by the cyclization of 2,3-oxidosqualene to the *chair-boat-chair* configuration (Fig. 1.3.2), (Phillips *et al.*, 2006). The different cyclization mechanisms of OSCs give rise to the vast diversity of triterpene scaffolds, with more than 100 different skeletons reported

so far (Xu, Fazio and Matsuda, 2004). The diversity of triterpenes is further enriched by tailoring enzymes that introduce functional groups to the scaffold. Oxygenation of the scaffold is one of the crucial modifications that improves the solubility and polarity of the molecule (Cramer, Sager and Ernst, 2019). Furthermore, enzymatic oxidations can generate bioactive molecules from the inactive triterpene scaffolds by functionalizing the usually chemically inert C-H bonds of the triterpene scaffold.

Figure 1.3.1. Bioactive triterpenes from nature showing wide-range of structural complexity.



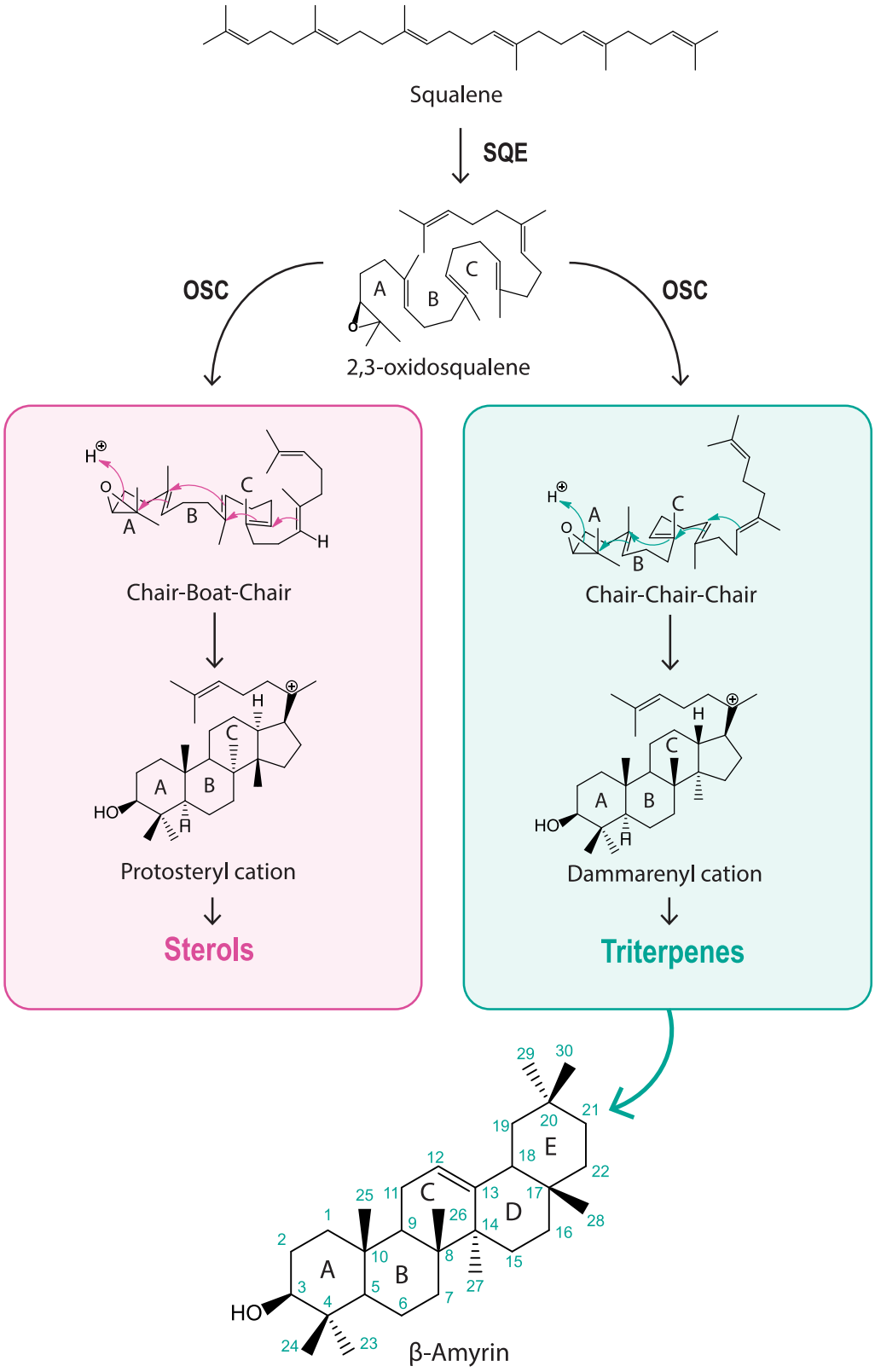


Figure 1.3.2. Cyclization of 2,3-oxidosqualene to form sterols and triterpenes. Prior to cyclization, 2,3-oxidosqualene adapts to different conformations in the active site of OSC. The dammarenyl cation give rise to triterpenes such as β-amyrin (rings are labelled in black, carbon numbers are labelled in teal). SQE, squalene epoxidase; OSC, 2,3-oxidosqualene cyclase.

Oxidation

The enzymatic oxidation of triterpenes is most commonly carried out by cytochrome P450 monooxygenases (CYPs). CYPs are one of the largest enzyme superfamilies in plant metabolism, to the extent that 1% of most plant genomes represent CYP encoding genes (Nelson and Werck-Reichhart, 2011). To date, CYP members in clans CYP51, CYP71, CYP72 and CYP85 have been associated with modifications of triterpene scaffolds (Ghosh, 2017). All CYPs have a common catalytic core of a heme iron group with a thiolate as the axial ligand, characterizing them as heme-thiolate enzymes (Bak *et al.*, 2011). As monooxygenases, CYPs split molecular oxygen (O_2) and insert one oxygen atom to the substrate while reducing the second to yield water (Munro *et al.*, 2013). In most cases, this reaction results in the formation of a hydroxylated product, as represented by the following scheme:

$$RH + O^2 + 2^{e-} + 2H^+ \rightarrow ROH + H_2O$$

Electrons required by this mechanism are commonly donated from reduction of nicotinamide adenine dinucleotide phosphate (NADPH) (McLean *et al.*, 2005). CYP450 proteins are usually located on the endoplasmic reticulum, but have also been reported to be localised in mitochondria and chloroplasts. The CYP superfamily performs wide range of highly regio- and stereospecific reactions. In addition to hydroxylation, they can catalyse a variety of reactions such as bond cleavage, bridge formation, dehydration, epoxidation as well as deamination (Munro *et al.*, 2013). Further oxidations can also lead to the formation of carboxylic acids, ketones and aldehydes (Cramer, Sager and Ernst, 2019). These functional groups act as handles for downstream tailoring enzymes such as sugar transferases and acyltransferases for further modifications.

Glycosylation

Sugar transferases are typically involved in the transfer of a sugar moiety to a hydroxyl or carboxylic group of the substrate. Conjugations with sugars affects the reactivity, stability and solubility of a molecule and may also act as a biological flag for compartmentalization of metabolites (Louveau and Osbourn, 2019). Glycosylation plays a crucial role in the ability of triterpene saponins and steroidal glycoalkaloids to permeabilize and integrate into the plasma membrane (Bowyer *et al.*, 1995). Aside

from these, glycosylation is also involved in inactivation and detoxification of harmful compounds, as well as in hormone regulation (Gachon, Langlois-Meurinne and Saindrenan, 2005). Enzymes involved in the building or degradation of glycosylated molecules are collectively referred to as carbohydrate-active enzymes (CAZymes) (Cantarel *et al.*, 2009). Most glycosylation in plant specialized metabolism is carried out by enzymes of the glycosyltransferase family 1 (GT1) enzymes, which use uridine diphosphate (UPD)-activated sugar donors, and thus are also referred to as UPD-dependent glycosyltransferases (UGTs) (Vogt and Jones, 2000). Plant GT1s commonly show high sugar donor specificity but may be more promiscuous regarding their acceptors (Osmani, Bak and Møller, 2009). They are classified into 17 monophyletic groups (A-Q), and the members of each group typically have similar functions and/or use structurally related acceptors (Louveau and Osbourn, 2019). Although there are many exceptions, the group D GT1s are often involved in triterpene glycosylation (Louveau and Osbourn, 2019). In addition to the GT1 family, recent reports have also implicated members of other types of enzyme families, such as glycosyl hydrolase 1 (GH1) transglycosidases (TGs) and cellulose synthase-like (CSL) enzymes in glycosylation of triterpenes. For example, AsTG1, a GH1 from oat, is responsible for the last glucosylation in the avenacin A-1 biosynthetic pathway (Orme *et al.*, 2019). Several members of the CSL superfamily from legume species and spinach have been reported in the 3-*O*-glucuronosylation of oleanane-type triterpenes (Chung *et al.*, 2020; Reed *et al.*, 2023; Jozwiak *et al.*, 2020).

Acylation

Another common modification of plant natural products is acylation, which involves the transfer of an acyl group from an activated energy-rich donor to an acceptor. Like glycosylation, acylation also influences the functions and properties, such as solubility and stability of a compound (Bontpart *et al.*, 2015). Acylation of plant specialized metabolites are catalysed by two enzyme families: BEAT-AHCT-HCBT-DAT (BAHD) and Serine CarboxyPeptidase-Like (SCPL) acyltransferases (ATs). Both AT families can acylate a variety of compounds belonging to the same families in the same plant species (Bontpart *et al.*, 2015). The main differences between the two families are the localization and the type of energy-rich donors they use.

BAHD-ATs are generally cytosolic enzymes with affinity towards acyl-CoA thioester donors (Bontpart *et al.*, 2015). The members of the BAHD family share a low protein sequence identity except for the two conserved motifs, the HXXXD motif and the DFGWG motif (St-Pierre and De Luca, 2000). To date, there are over 160 characterized BAHD ATs that are grouped into 8 phylogenetic clades based on similarities in sequence and acyl acceptors (Kruse *et al.*, 2022). Most of the characterized BAHD ATs are reported to be cytosolic enzymes, with few exceptions, such as CER2 which localizes to the endoplasmic reticulum (Haslam *et al.*, 2012). Moglia et al., (2014, 2016) have reported that the BAHD enzyme Hydroxycinnamoyl CoA Qinate Transferase (HQT) is located both in the cytosol and the vacuole. In the cytosol it undertakes a BAHD acyl transfuse activity, but in the vacuole it functions as a chorogenate-chlorogenate transferase to form dicaffeoyl quinate. This alternative activity in the vacuole can occur because of the relatively low pH and the high concentrations of chlorogenic acid the vacuole.

Unlike BAHD ATs, only a small number of SCPL ATs are characterized so far. SCPL ATs are reported to date are vacuolar in localization and use 1-*O*-β-D-glucose ester donors while BAHD ATs are cytosolic and use acyl-CoA thioester donors (Bontpart *et al.*, 2015). They are believed to have emerged from serine carboxypeptidases (SCPs) through divergent evolution, replacing the hydrolytic activity of SCPs with transferase activity (Stehle *et al.*, 2006). Both classes of ATs have been reported to acylate triterpenes. For example, SOAP10 is a BAHD AT involved in the acylation step of the yossoside biosynthetic pathway in spinach (Jozwiak *et al.*, 2020), and AsSCPL1 (SAD7) is a SCPL AT from oat that is involved in avenacin biosynthesis (Mugford *et al.*, 2009).

With more than 20,000 different compounds reported, triterpenes are one of the largest classes of plant specialized metabolites (Ghosh, 2020). Together with OSCs, the combined actions of various CYPs, GTs and ATs lead to the immense structural diversity of triterpenes from a single precursor, 2,3-oxidosqualene. Biosynthesis of complex triterpenes, such as saponins, involves the activity of several enzyme families as they require at least three reactions (cyclization, oxidation and glycosylation) and is often acylated. These enzyme families are known to localize to various compartments in the plant cell. As aforementioned, triterpenoid biosynthesis begins with the cyclization of 2,3-oxidosqualene produced by the MVA pathway in the

cytosol, and catalysed by OSCs (Augustin *et al.*, 2011). The oxygenations of triterpenoid aglycones are commonly carried out by CYPs, which are reported to be membrane bound proteins, usually anchored to the cytoplasmic side of the endoplasmic reticulum (Bak *et al.*, 2011). Glycosylation of triterpenoids has been reported to be carried out by several GT families, most notably by UGTs, which are generally known to be soluble, cytosolic proteins (Caputi *et al*., 2012). Other GTs involved in triterpenoid biosynthesis include GH1 TGs, where most of the characterized enzymes so far are reported to localize to the vacuole, and CSLs, which are thought to localize to the endoplasmic reticulum (Orme *et al.*, 2019; Jozwiak *et al.*, 2020; Chung *et al.*, 2020). Triterpenoids can be further decorated with different acyl groups by the activities of both BAHD and SCPL ATs, which are generally reported to be active in the cytosol and the vacuole, respectively (Kruse *et al.*, 2022). However, these statements are general observations based on the enzymes characterized so far, and continued discovery of new enzymes involved in triterpenoid biosynthesis may reveal exceptions to these cases.

Triterpenoid biosynthesis is further complicated by intricate regulation, involving transport and storage of these compounds in different organs, tissues and cells, depending on the life stage of the plant (da Silva Magedans *et al*., 2020). For example, in *Medicago truncatula*, triterpenoid saponins are hypothesized to be primarily biosynthesized in the palisade parenchyma cells within the leaves, and then transported via the phloem to different plant organs for storage or defensive purposes (Zannino, *et al*., 2024). Like many other plant specialized metabolites, the vacuole is considered as the likely storage compartment for many triterpenoids, as a strategy to store these potentially toxic, water-soluble compounds. (da Silva Magedans *et al*., 2020). Transporter families such as ATP-binding cassette (ABC) and multidrug and toxic compound extrusion (MATE) have been identified as candidates involved in transport of triterpenoid saponins into the vacuole (de Brito Francisco and Martinoia, 2018). Recently a MATE transporter, GmMATE100, was identified as transporting soyasaponins to the vacuole of soyabean (Ma et al., 2024). The full elucidation of the biosynthetic pathways of these complex metabolites presents significant challenges although, recent breakthroughs in NGS together with advances in genomics and bioinformatics greatly expedite the discovery of plant natural product pathways, making these bioactive compounds more accessible than ever (Owen *et al.*, 2017).

1.4 Soapwort (*Saponaria officinalis*)

Saponaria officinalis, commonly known as soapwort, is a perennial flowering plant in the Caryophyllaceae family native to Eurasia (Fig. 1.4.1). As a member of the 'pinks' family, soapwort produces delicate, white to pinkish flowers that bloom from July to early October (Eastman, 2014). Although soapwort prefers moist and bright conditions, they can grow in diverse conditions as they are drought-tolerant and highly adaptive to their environment (Henry, 1989). As such, soapwort is widely found along roadsides, railroad tracks, meadows and desolate areas otherwise considered as wastelands (Lubke and Cavers, 1969). The flowers are sequentially unisexual, generally the male parts develop and perish before the female parts mature. Although this mechanism avoids self-pollination for the most part, self-pollination can still occur when the sexual stages overlap by chance (Eastman, 2014). Soapwort can also propagate asexually via their reddish rhizomes. The plant is sometimes considered as an invasive weed because of its hardiness and their ability to spread rapidly. Soapwort is grown for both ornamental and functional uses. The name *saponaria* is derived from the Latin word *sapo* or *saponis,* which means soap, and the word *officinalis* is a Latin word for herbal medicine (Jones, 1996). As the name suggests, *S. officinalis* has been a traditional source of soap and medicine. In reference to this, other common names of soapwort include latherwort, soaproot, scourwort and bruisewort (Mitich, 1990). Soapwort is also known as fuller's herb and bouncing bet for the plant's usage in laundry, and wild sweet william or sweet betty due to the fragrance of the flowers (Mitich, 1990).

The ancient Greeks, Romans and Egyptians used soapwort extracts to clean and wash clothing and later, the first American colonists brought soapwort plants from Europe to North America for their household uses (Eastman, 2014). The gentle detergent properties of soapwort extracts made it a popular choice for washing fragile fabrics, such as silks, wools and delicate linen (Eastman, 2014). In fact, soapwort extract is believed to have been used as a gentle soap to treat the Shroud of Turin, an antique linen dating 1260-1390 AD depicting an image of a naked man believed to be Jesus. In addition to their detergent properties, soapwort extracts have been used in folk medicine to treat a wide variety of conditions such as syphilis, gout, rheumatism, coughs and bronchitis, bile disorders, jaundice and various skin ailments (Rees, 1819). Soapwort also play an important role in Middle Eastern cultures as the extracts are

used as emulsifiers to make desserts such as tahini halvah and Turkish delight (Korkmaz and Özçelik, 2011). Today, soapwort extracts are still used in cosmetic, nutraceutical and phytomedicinal products (Böttger and Melzig, 2011).

Figure 1.4.1. *Saponaria officinalis* (soapwort). Reproduced from (Curtis, 1777). Photo taken by Phil Robinson.

1.4.1 Saponins from soapwort

As introduced earlier, glycosylated triterpenes are collectively referred to as saponins. Saponins are composed of a lipophilic aglycone with linked hydrophilic sugar chains, resulting in their amphiphilic nature and their soap-like characteristics (Osbourn, 1996a). Monodesmodic saponins have a single sugar chain typically attached to the C-3 position of the aglycone, while bisdesmosidic saponins have two sugar chains typically at C-3 and C-28 positions (Osbourn, 1996a). Generally, saponins are membrane permeabilizing agents as the hydrophobic aglycon can insert into the cell membrane and complex with cholesterol in the lipid bilayer, leading to pore formation

and subsequent cell lysis (Moses, Papadopoulou and Osbourn, 2014). However, the membrane permeabilizing activity of saponins is greatly dependent on type of the aglycone backbone (Augustin *et al.*, 2011).

The well-known detergent properties of soapwort are due to the large amounts of bioactive saponins in the plant extracts. Indeed, the term 'saponin' was named after *Saponaria officinalis* due to soapwort's age-old use as soap (Jia *et al.*, 2002). Befittingly, soapwort is a rich source of oleanane-based (β-amyrin derived) triterpenoid saponins with either one or two sugar chains (Jia *et al.*, 2002). The first soapwort saponins to be reported, saponasides A-D, were isolated in the 1970s (Jia *et al.*, 2002). The chemical structures of saponasides A and D were partially resolved to be gypsogenin-based saponins with sugar units attached at the C-3 and C-28 positions of the aglycone (Chirva and Kintya, 1969; Chirva and Kintya, 1970). Impressively, saponaside D contains ten sugar units composed of two D-galactoses, L-arabinose, two D-xyloses, two L-rhamnoses, D-glucuronic acid, D-glucose and D-fucose, while saponaside A contains four sugar units including D-glucuronic acid and three D-glucoses (Chirva and Kintya, 1969; Chirva and Kintya, 1970). During this time, two more saponins, gypsogenin and gypsoside-based saponins named as saponaroside and glycoside B, respectively, were also identified (Bukharov and Shcherbak, 1969). The structure of saponaroside was elucidated as 3-*O*-β-D-glucopyranosyl gypsogenin, while only the sugar units, including D-glucose, D-galactose, and L-arabinose, of glycoside B were identified (Bukharov and Shcherbak, 1969) Over the years, more than 40 additional saponins have been isolated from extracts of *S. officinalis* and are described in Chapter 3. The major saponins found in soapwort extracts are reported to be saponariosides A and B (Jia, Koike and Nikaido, 1998). Both are composed of a quillaic acid aglycone with two similar sugar chains (Fig. 1.4.2). Saponarioside A (SpA) is chemically defined as 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)-β-D-4-*O*-acetylquinovopyranosyl-(1→4)]-*β*-D-fucopyranosyl ester}-quillaic acid, and saponarioside B (SpB) is defined as 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-[β-D-4-*O*-acetylquinovopyranosyl-(1→4)]-*β*-D-fucopyranosyl ester}-quillaic acid (Jia,

Koike and Nikaido, 1998). The only structural difference between the two saponariosides is an additional D-xylose attached to D-quinovose in SpA.

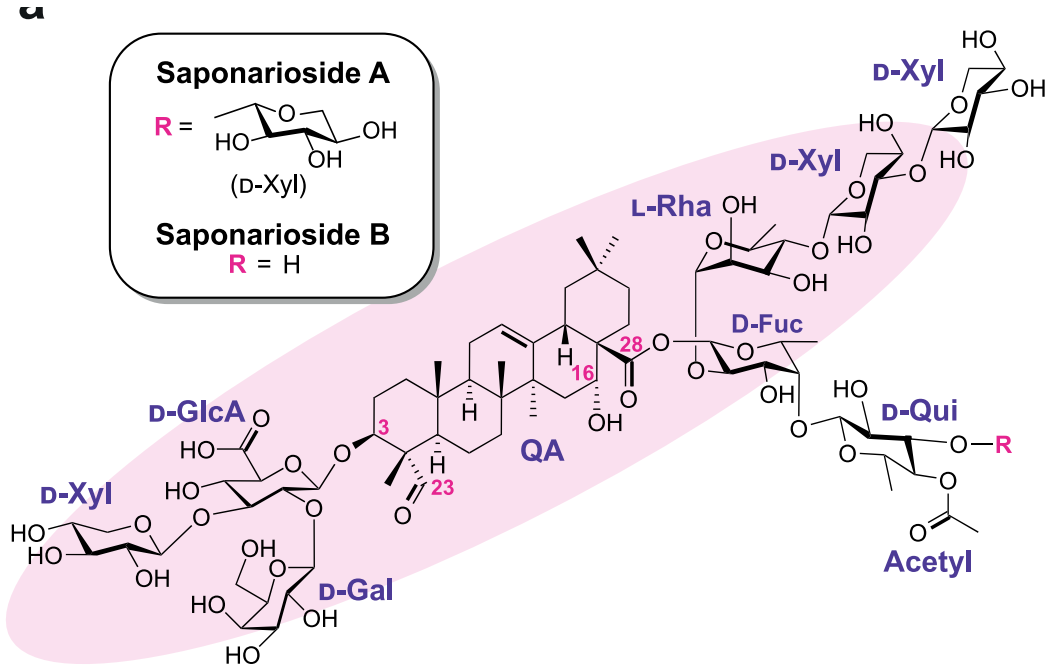


Figure 1.4.2. Major saponins reported from *S. officinalis*. Structures of saponariosides A (SpA) and B (SpB), both consist of a quillaic acid aglycone with branched trisaccharide at C-3 (composed of D-glucuronic acid, D-galactose and D-xylose) and a linear tetrasaccharide at C-28 (composed of D-fucose, L-rhamnose, D-xylose and D-xylose) with an acetylquinovose moiety attached to D-fucose. In SpA, an additional D-xylose is attached to D-quinovose. The common saponin scaffold found across various flowering plants is highlighted in pink.

Saponins with structural similarities to SpA and SpB are widespread in the Caryophyllaceae family. For example, saponins such as tunicosaponins from *Psammosilene tunicoides* (Wen *et al.*, 2020), pachystegiosides from *Acanthophyllum pachystegium* (Haddad *et al., 2004*), and various saponins from *Gypsophila* (Frechet *et al.* 1991; Chen, Luo and Kong, 2011) and *Silene* spp. (Lacaille-Dubois *et al.*, 1999; Fu *et al.*, 2005) share a common saponin scaffold (3-*O*-{β-D-galactopyranosyl-(1→2)-β-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranosyl ester}-quillaic acid) as SpA and SpB (Fig. 1.4.2). In fact, this common scaffold is found across the Angiosperms. Of note, the aforementioned QS-21 fraction of triterpenoid saponins from *Q. saponaria* shares a striking chemical resemblance to SpA and SpB, despite being from a distant plant family (Fig. 1.4.3). Although less potent compared to activities of QS-21, saponin mixtures extracted from *S. officinalis* are also able to form immunostimulating complexes (Bomford *et al.*, 1992). However, the details of which soapwort saponins

possess this adjuvant activity are unknown since only impure saponin mixtures were tested.

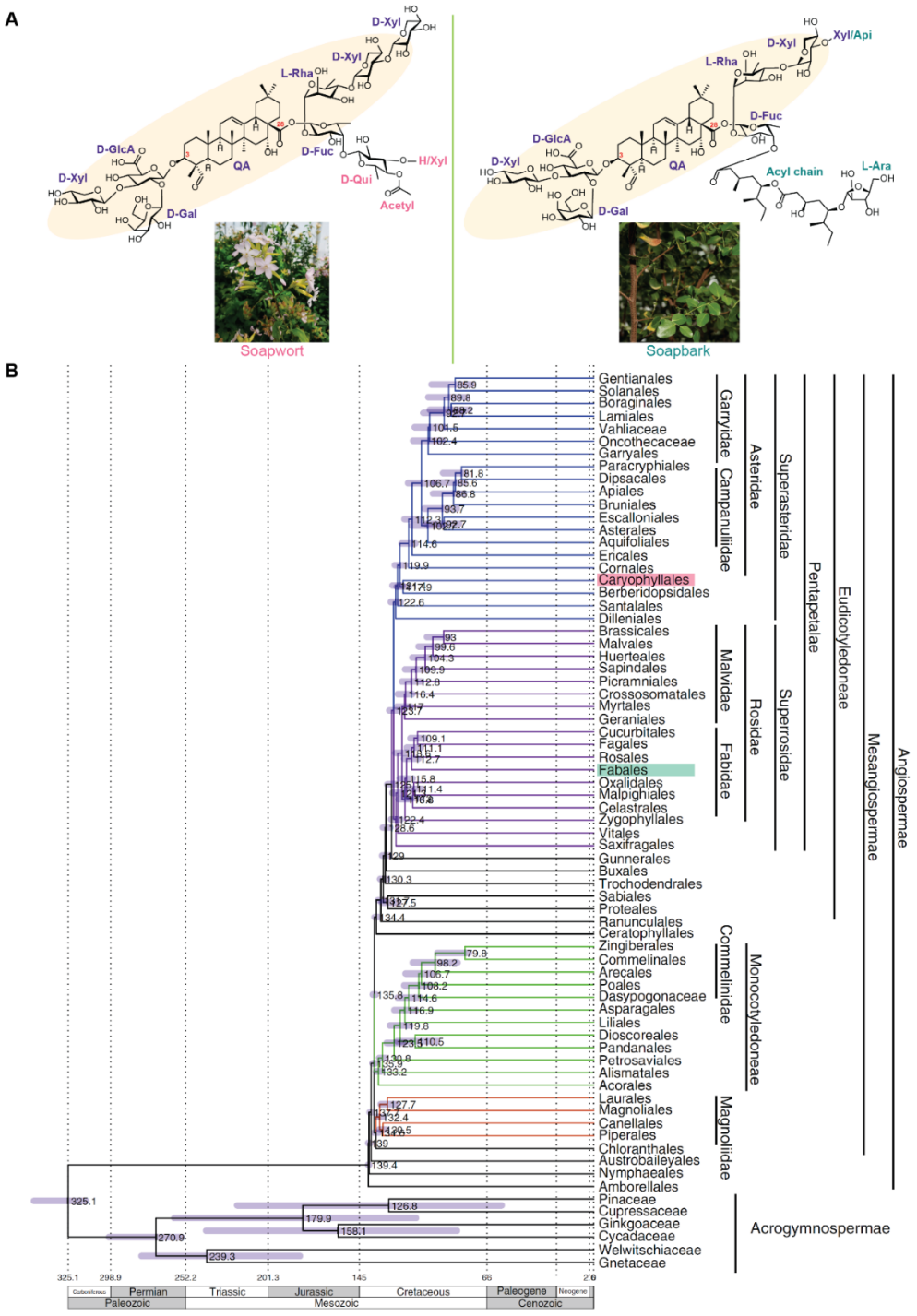


Figure 1.4.3. Soapwort and soapbark produce structurally similar saponins despite phylogenetical differences. (A) Comparison of saponariosides produced by soapwort and QS-21 produced by soapbark. Chemical moieties found in both compounds are labelled in purple, those specific to saponariosides and specific to QS-21 are labelled in pink and teal, respectively. SpA has terminal D-xylose attached to D-quinovose, which is absent in SpB. QS-21 is a 65:35 mixture of D-apiose and D-xylose variants at the terminal C-28 sugar chain. The common saponin scaffold found across

angiosperms are highlighted in pale yellow. QA, quillaic acid; D-GlcA, D-glucuronic acid; D-Gal, D-galactose; D-Xyl, D-xylose; D-Fuc, D-fucose; L-Rha, L-rhamnose; D-Qui, D-quinovose; L-Ara, L-arabinose; Api, D-apiose. **(B)** Angiosperm time-tree collapsed to show orders. This figure is reproduced from (Magallón *et al.*, 2015) with minor edits to highlight the order containing soapwort (Caryophyllales, pink) and soapbark (Fabales, teal) to illustrate the phylogenetic distance between soapwort and soapbark. The two orders are estimated to have diverged 123.7 million years ago.

The saponin components of soapwort extracts have also been investigated for their potential use in bioremediation (Smulek *et al.*, 2017), as food surfactants (Jurado-Gonzalez and Sörensen, 2019) and for their anti-fungicidal (Sadowska *et al.*, 2014). Furthermore, quillaic acid-based soapwort saponins (i.e., SO1861) have been reported to augment the cytotoxicity of saporin, a type I ribosome inactivating protein (RIP) found in soapwort (Gilabert-Oriol *et al.*, 2016). Evaluation of pure soapwort saponins, such as SpA and SpB, may provide insights into the pharmaceutical and nutraceutical properties of these compounds. However, isolation of individual saponin component is challenging as these saponins are present in complex mixtures (Jia *et al.*, 2002). This has limited the use of saponins (from soapwort and other species) and hampered attempts to fully investigate the pharmaceutical potential of these valuable metabolites.

1.5 PhD Overview

Despite the promising commercial potential of saponariosides, nothing is known about their biosynthesis. The overall aim of this project was to investigate the biosynthesis of the major saponins (saponariosides A and B) found in *S. officinalis*. Metabolite analysis of soapwort organs focused on SpA and SpB revealed flowers as a potential major site of saponarioside biosynthesis (Chapter 3). Building on this knowledge, RNA-Seq and genome sequence resources were generated for the discovery of saponarioside biosynthetic genes (Chapter 4). These newly generated sequence resources were used to identify a total of 13 genes involved in saponarioside biosynthesis. Co-expression of these genes in *N. benthamiana* by transient expression led to the biosynthesis of saponarioside B (Chapter 5). Collectively, this project offers novel insights into saponin biosynthesis in a non-model plant species in the Caryophyllaceae family. It opens-up for the first time, opportunities to produce saponariosides and analogues in an alternative host system using metabolic engineering approaches.

2

Materials and methods

2.1 Common reagents and plant maintenance

2.1.1 Standards

Standards of echinocystic acid, coprostanol and digitoxin were purchased from Sigma-Aldrich. Oleanolic acid and β-amyrin standards were purchased from Extrasynthese. Saponariosides A and B (purified from *S. officinalis* (Section 2.3.8)) and QA-Tri standards were kindly provided by Dr. Amr El-Demerdash.

2.1.2 Media and antibiotics

Media (both liquid and solid) used in this work were prepared by the JIC Laboratory Support team. Standard lysogeny broth (LB) and Super Optimal broth with Catabolite repression (SOC) were used.

Antibiotics were prepared as detailed in Table 2.1.1.

Table 2.1.1. Antibiotics used in this work. The same antibiotic concentrations were used for both liquid and solid agar LB media. All stocks were stored at -20 °C in 1 mL aliquots.

Antibiotic	Stock (mg/mL)	Final (µg/mL)	Dilution Factor	Solution
Gentamicin	50	50	1/1000	H_2O
Kanamycin	50	50	1/1000	H_2O
Streptomycin	100	100	1/1000	H_2O
Rifampicin	50	50	1/1000	Dimethylformamide
Chloramphenicol	30	30	1/1000	Ethanol
Ampicillin	100	100	1/1000	H_2O

2.1.3 *S. officinalis* plants

Two different accessions of soapwort (*Saponaria officinalis*) were used in this project, JIC and RSM. The JIC accession is a single-flowered assession originally purchased from Norfolk Herbs, Norfolk, UK in August 2018. The RSM accession is a double-flowered assession from Rockland St. Mary, Norfolk, NR14 7H5, UK collected by Prof. Anne Osbourn (also in August 2018). Both accessions were maintained in pots with cereal mix and were grown in an outdoor glasshouse at the John Innes Centre with natural light conditions and seasonal temperature. Decayed above ground organs such as leaves, and stems were removed in the winter. The plants were vegetatively propagated (separated by rhizomes) and individually re-potted.

2.2 General molecular biology methods

2.2.1 Gene cloning

The coding sequences of candidate soapwort genes were either PCR-amplified from a *S. officinalis* cDNA pool (described in Section 2.4.1) using gene specific primers (Appendix A.1) or the gene fragments were synthesized either by Twist Bioscience or IDT. For PCR amplification, iProofTM High-fidelity DNA polymerase (Bio-Rad) was used following the manufacturer's protocol. The thermocycling condition was as follows: initial denaturation at 98 °C for 30 s; 35 cycles of denaturation at 98 °C for 10 s, annealing at 60 °C for 20 s, extension at 72 °C for 45 s; final extension at 72 °C for 5 min. PCR amplified products were separated by agarose gel electrophoresis to confirm the expected molecular weight.

Colony PCR was performed using GoTaq Green Mastermix (Promega) following the manufacturer's protocol using attL primers (Appendix A.1). Cells from selected bacterial colonies were directly transferred to 10 µL of reaction mix using a sterile pipette.

PCR products were purified using QIAquick PCR Purification kit following manufacturer's protocol. Gateway technology (Invitrogen) was used to transfer the purified PCR products or synthesized gene fragments into the pDONR206 entry vector (Invitrogen) and eventually into the pEAQ-HT-DEST1 expression vector, kindly

provided by the Lomonossoff laboratory (Sainsbury, Thuenemann and Lomonossoff, 2009). To generate entry clones, BP recombination reaction was performed following the manufacturer's instructions, with equal ratios (~150 ng each) of pDONR207 vector and purified PCR product or synthesized gene fragments. The reaction was subsequently heat-shock transformed into chemically competent *Escherichia coli* cells (DH5α, ThermoFisher Scientific) as described in Section 2.2.2. To recover plasmids from positive bacterial colonies, liquid overnight cultures were made by transferring selected colonies to 10 mL LB media with appropriate antibiotics using a sterile pipette tip. Cultures were grown at 37 °C overnight and plasmids were recovered using QIAprep Spin Miniprep Kit (Qiagen) following manufacturer's protocol. Recovered plasmids were sequence verified using attL1 and attL2 primers (Appendix A.1). To generate expression clones, LR recombination reactions were performed following the manufacturer's protocol with equal ratios (~150 ng each) of the entry clones carrying the gene of interest and pEAQ-HT-DEST1. Positive clones were recovered as described above.

2.2.2 *E. coli* transformation

For transformation, 25 µL of chemically competent *E. coli* cells (DH5α; ThermoFisher Scientific) were incubated with 1 µL of purified plasmid (~300 ng) and incubated on ice for 30 min. Heat shock was performed at 42 °C for 45 s and incubated with 200 µL SOC media for 1 hour at 37 °C shaking at 200 rpm. After incubation, cells were plated onto LB agar plates with appropriate antibiotics and incubated at 37 °C overnight. Gentamycin was used to select for positive pDONR207 entry clones, and kanamycin was used for pEAQ-HT-DEST1 expression clones.

2.2.3 Agrotransformation and agroinfiltration

Agrobacteria preparation and maintenance

Chemically competent *Agrobacterium tumefaciens* cells (strain LBA4404) were purchased from Invitrogen. To generate a stock, cells were streaked onto LB agar plates containing rifampicin and streptomycin and incubated overnight at 28 °C. Cells in the exponential phase of growth were obtained by preparing overnight cultures in 50 mL LB media containing rifampicin and streptomycin with shaking at 200 rpm at 28 °C. Cultures were centrifuged at 4 °C 4,500 × g for 5 min and the supernatant was

discarded. Cell pellets were resuspended in 1 mL of ice-cold 20 mM $CaCl_2$. The centrifugation step was repeated, and the cells were again resuspended in 1 mL of ice-cold 20 mM $CaCl_2$. For transformation, 50 µL cells were aliquoted in 1 mL Eppendorf tubes, flash frozen in liquid N_2, and stored in -80 °C until further use.

Agrotransformation

Agrotransformation was performed as previously described (Reed *et al.*, 2017). Chemically competent cells aliquoted in 50 µL were thawed on ice and incubated with ~300 ng of pEAQ-HT-DEST1 vector carrying the gene of interest. After 30 min of incubation on ice, cold shock treatment was performed in liquid nitrogen for 30 s. Cells were allowed to thaw at room temperature and were incubated in 200 µL of SOC media at 28 °C for 2 hours before plating on LB agar plates with rifampicin, streptomycin and kanamycin. After 3 days of incubation in 28 °C, overnight liquid cultures were prepared by transferring cells from a single colony to 5 mL liquid LB media containing appropriate antibiotics. Cultures were grown at 28 °C shaking at 200 rpm. After, 500 µL of culture was mixed with 500 µL of 20% (v/v) aqueous glycerol and stored at -80 °C as glycerol stock.

Manual agroinfiltration

Small-scale manual agroinfiltration was performed as described in (Reed *et al.*, 2017). Transformed *A. tumefaciens* cells stored as glycerol stocks were streaked onto LB agar plates containing appropriate antibiotics and incubated at 28 °C for 3 days. Liquid cultures were prepared by transferring cells from a single colony to 10 µL of liquid LB media containing appropriate antibiotics using a sterile pipette tip. Cultures were incubated at 28 °C with shaking (200 rpm) for at least 24 hours. Cells were harvested by centrifugation at 4,500 × g for 10 min. The resulting supernatants were discarded, and cell pellets were resuspended in enough volume of MMA buffer (10 mM MES (2-[N-morpholino] ethanesulfonic acid) pH 5.6, 10 mM $MgCl_2$, 100 µM acetosyringone) to submerge the cell pellet. The bacterial suspensions were diluted 1:10 in MMA buffer and the optical density at 600 nm (OD_{600}) was measured using a spectrometer. The measured OD_{600} was then used to calculated appropriate volumes required for the infiltration mixtures so that each re-suspended strain reached a final OD_{600} of 0.2. Leaves of 5-week-old *Nicotiana benthamiana* plants maintained by the JIC horticultural services were used for infiltration. The adaxial side of each leaf was

lightly punctured with a needle to make a guide hole prior to infiltration with the *A. tumefaciens* mixture using a 1 mL syringe without a needle. The infiltrated leaves were harvested 5 days after and freeze-dried using LyoDry Midi freeze dryer (MechaTech Systems) for 2 days.

Vacuum-mediated agroinfiltration

Large-scale agroinfiltration was carried out as described in (Stephenson *et al.*, 2018), using a purpose-built vacuum infiltration system. A video showing the process is also available in (Stephenson *et al.*, 2018). The bath of the vacuum infiltration apparatus was filled with 10 L of agroinfiltration culture (prepared as described above). Five-week-old *N. benthamiana* plants were positioned onto the acrylic holder and inverted into the bath containing the bacterial suspension so that aerial parts of the plant were completely submerged. A vacuum of ~ 26 mmHg was applied to draw out the air in the interstitial leaf space. The vacuum chamber was isolated to allow the chamber to re-equilibrate with the external atmosphere to allow the infiltration suspension to be drawn into the submerged leaves. Plants were carefully taken out of the vacuum manifold and monitored in a greenhouse at 25 °C with 16 h light per day for 5 days. Infiltrated leaves were then harvested and freeze-dried using LyoDry Midi freeze dryer (MechaTech Systems) for 4 days.

2.3 Metabolite extraction and analysis

2.3.1 Harvested *S. officinalis* material

Six different organs (flowers, flower buds, young leaves, old leaves, stem and root; Fig. 2.3.1) were harvested from a total of four *S. officinalis* plants (accession JIC) in July 2019. In November 2019, plants were also harvested for both JIC and RSM accessions (seeds, leaf, stem and root), again using four plants from each accession. The harvested materials were flash-frozen in liquid nitrogen. The frozen samples were ground into fine powder in liquid nitrogen using a mortar and pestle and stored at -80 °C until further use.

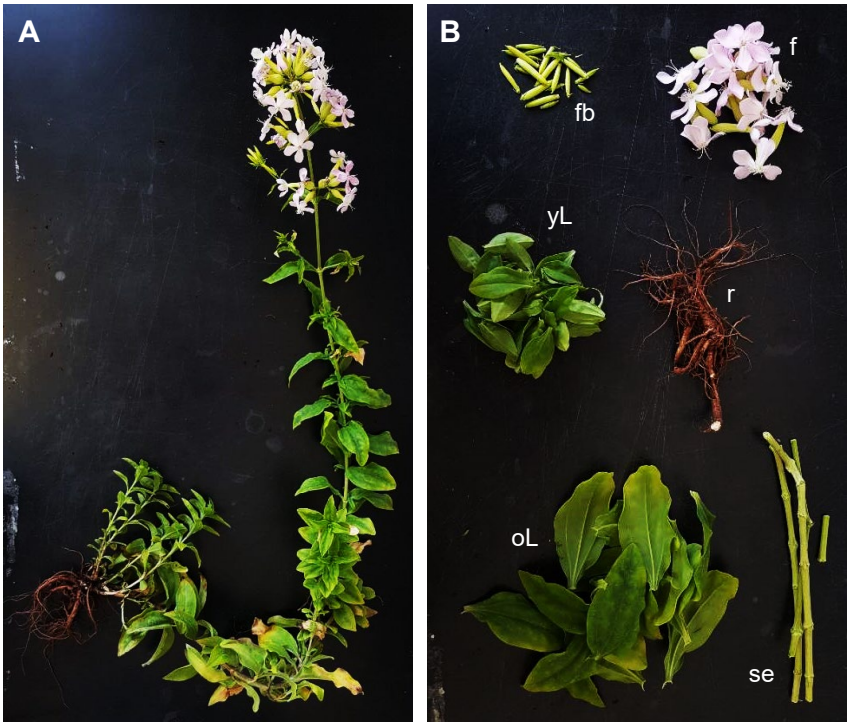


Figure 2.3.1. Representative image of a soapwort plant harvested in July 2019. (**A)** Whole plant. (**B)** The different organs harvested: flower bud (fb), flower (f), young leaf (yL), old leaf (oL), root (r), stem (se).

2.3.2 Extraction from *S. officinalis*

For metabolite analysis, 1 mL volumes of frozen ground plant materials were dried in a freeze-dryer (MechaTech Systems) for 4 days. Aliquots (10 mg) were then extracted using 1 mL of extraction buffer (80% (v/v) MeOH/H_2O, 10 µg/mL digitoxin (Sigma-Aldrich)) with shaking at 1,400 rpm for 2 h at room temperature. Following centrifugation at 12,000 × *g* for 5 min, supernatants were filtered using 0.2 µm Costar® Spin-X® microcentrifuge tube filters (Merck). Filtered samples were transferred to Teflon-sealed, screw-capped 1 mL glass vials (Agilent) with glass inserts for LC-MS analysis.

2.3.3 Extraction from *N. benthamiana*

Metabolite analysis was performed using 10 mg of dried *N. benthamiana* leaves (see Section 2.2.3). The weighed leaf samples were homogenized with two 3 mm tungsten beads using the TissueLyser (Qiagen) at 1000 rpm for 1 min. For GC-MS analysis, ground samples were extracted using 550 µL of ethyl acetate containing 50 µg/mL coprostanol (Sigma-Aldrich) as the internal standard by agitating intermittently using

a vortexer (StarLab) for 30 min at room temperature. After centrifugation at 12,000 x *g* for 1 min, the supernatants were recovered and transferred into new 2 mL Eppendorf tubes. The samples were de-fatted by the addition of 400 µL of hexane and were vortexed briefly. To separate the organic phase, samples were centrifuged at 12,000 x *g* for 1 min, and the top aqueous layer was recovered and filtered using 0.2 µm Costar® Spin-X® microcentrifuge tube filters (Merck). The filtered samples were dried using a Genevac EZ-2 evaporator (SP Scientific) and derivatized prior to GC-MS analysis using 50 µL of 1-(Trimethylsilyl)imidazole - Pyridine mixture (Sigma-Aldrich). The derivatized samples were transferred to Teflon-sealed, screw-capped 1 mL glass vials (Agilent) with glass inserts. For LC-MS analysis, sample preparation was carried out as described above (Section 2.3.2) for *S. officinalis* plant samples.

2.3.4 Saponification of *S. officinalis* plant extracts

For saponification experiments, 10 mg of dried *S. officinalis* samples (see Section 2.3.2) were saponified using saponification solution (ethanol/KOH/water, 9:1:1, (v/v or w/v)) with shaking at 1,400 rpm, 75 °C. The samples were vortexed for 10 s every 15 min and neutralized by the addition of formic acid (~40 µL, sample dependent) after 1 hour. The pH was checked using universal indicator strips to confirm a neutral pH. The samples were centrifuged at 12,000 × g for 5 min, the supernatants were filtered using 0.2 µm Costar® Spin-X® microcentrifuge tube filters (Merck) and transferred to Teflon-sealed, screw-capped 1 mL glass vials (Agilent) with glass inserts for LC-MS analysis. A standard curve was generated using 8 different concentrations of hederacoside C (1, 2.5, 7.5, 10, 15, 25, 40, 50 µg/mL; Sigma-Aldrich), to give a linear trendline with $R^2 = 1.00$.

2.3.5 GC-MS analysis

GC-MS analysis was performed with an Agilent 7890B GC system fitted with a Zebron AB5-HT Inferno Column (Phenomenex) using a 20-minute method program described in (Reed *et al.*, 2023). Briefly, 1 µL of each sample was injected into the inlet (250 °C) in pulse splitless mode (pulse pressure 30 psi). The oven temperature was held for 2 min at 170 °C, then increased to 300 °C at rate of 20 °C/min and held at 300 °C for 11.5 mins for a total run time of 20 minutes. The mass spectrometry was performed using Agilent 5977A Mass Selector Detector in scan mode from 60-800 *m/z*

after a solvent delay of 8 mins. MassHunter workstation (Agilent) was used to analyse the resulting data.

2.3.6 LC-MS analysis

LC-MS analysis was performed using a ThermoFisher Q Exactive™ HPLC system fitted with a Hybrid Quadrupole-Orbitrap™ mass spectrometer (ThermoFisher). Samples were analysed using a Kinetex XB-C18 100A (50 × 2.1 mm, 2.6 μm; Phenomenex) column at 40 °C with an injection volume of 5 μL. The mass spectrometer was equipped with electrospray in negative ionization mode (capillary temperature 250 °C, nebulizing gas 1.3 min/L, heat block temperature 300 °C, spray voltage -3.5 kV). Full MS spectra was collected with scan range of 200-3000 *m/z*, and dd-MS^2 data was collected for top three most abundant ions at a given period (Top^N = 3). All data processing was performed using Thermo FreeStyle ver. 1.6 (ThermoFisher).

Different elution gradients were used as appropriate (see below). For all gradients, acetonitrile was used as solvent A and 0.1% (v/v) aqueous formic acid was used as solvent B:

i) Gradient 'Terpenes' 30 min

To analyse the products of SoC23 following heterologous expression with other *S. officinalis* pathway genes in *N. benthamiana*, the 'Terpenes' gradient was used. The elution profile was as the following: 0 – 1.5 min, 15% B in A; 1.5 – 26.0 min, 60% B in A; 26.0 – 26.5 min, 100% B; 26.5 – 28.5 min, 100% B; 28.5 – 29.0 min, 15% B in A; 29.0 – 30.0 min, 15% B in A. The flow rate was set to 0.5 mL/min.

ii) Gradient 'Saponin' 15.6 min

To analyse metabolites extracted from soapwort, and the majority of products generated by heterologous expression of candidate soapwort genes (except for SoC23 and SoBAHD1) in *N. benthamiana*, the 'Saponin' gradient was used. The elution profile was the following: 0 – 1.0 min, 5% B in A; 1.0 – 10.0 min, 55% B in A; 10.0 – 12.0 min, 100% B; 12.0 – 13.0 min, 100% B; 13.0 – 13.3 min, 5% B in A; 13.3 – 15.6 min, 5% B in A. The flow rate was set to 0.5 mL/min.

iii) Gradient 'QS-21' 16.5 min

To analyse the products generated by SoBAHD1 following heterologous expression with other *S. officinalis* pathway genes in *N. benthamiana,* the 'QS-21' gradient was used (Reed *et al.*, 2023). The elution profile was the following: 0 – 0.75 min, 15% B in A; 0.75 – 13.0 min, 60% B in A; 13.0 – 13.25 min, 100% B; 13.25 – 14.25 min, 100% B; 14.25 – 14.5 min, 15% B in A; 14.5 – 16.5 min, 15% B in A. The flow rate was set to 0.6 mL/min.

2.3.7 Internal standard-based quantification

The peaks for the target compounds and the internal standard were extracted and the relative concentration of the target analyte was calculated based on the concentration of internal standard (digitoxin, 10 µg/mL) used. This value was then scaled based on the extracted dry mass to estimate the amount of compound in the starting material.

All statistical analyses were performed using SigmaPlot 11.0 for Windows (Systat Software Inc. 2008). Raw data was first transformed (^0.25) to fit the normal distribution. One-way analysis of variance (ANOVA) was performed followed by a Tukey Test to test for significant differences in the different soapwort organs. Paired t-tests were used to test for significant differences between the same plant organs.

2.3.8 Saponarioside extraction, purification, and structural elucidation

Purification from agroinfiltrated N. benthamiana

A total of 110 *N. benthamiana* plants were agroinfiltrated using the vacuum manifold as described in Section 2.2.3. The harvested leaf material was lyophilized to yield 90.5 g of dried material. This was subsequently extracted with 80% (v/v) MeOH/H_2O and saponins were partitioned from the aqueous methanolic extract using n-butanol. Flash column chromatography (FCC) using RP C-18 column (120 g) was performed using a linear gradient of water/acetonitrile acidified with 0.1 formic acid (2400 mL) [100/0→30/70]. A total of 12 subfractions (200 mL each) were collected and monitored by HR-LC-ESI-MS. Fractions containing a peak of interest were combined and subjected to further purification using preparative HPLC via a linear gradient of water/acetonitrile acidified with 0.1% formic acid [100/0→30/70].

Purification of saponariosides from S. officinalis

As soapwort is a perennial plant, the availability of green organs was dependent on the season. To mitigate this problem during the non-growing season, dried *S. officinalis* leaf material was purchased from Joannas Garden (Germany, https://joannasgarden.com). Extraction, purification and structural confirmation of saponariosides A and B were kindly performed by Dr. Amr El-Demerdash. For extraction, 17 g of leaf material were extracted under reflux using a combination of 80% (v/v) MeOH/H_2O at 110 °C. The aqueous methanolic extracted was collected and dried under reduced pressure, then re-suspended in the least amount of methanol to cover the dried product and complete to 1 L by distilled water. Then it was partitioned against hexane, ethyl acetate, and n-butanol. The butanolic layer was collected and completely dried under reduced pressure, then dissolved in the last amount of methanol and immediately saturated with cold acetone, where the crude saponins fraction was precipitated and collected by filtration. A portion of this saponin-enriched extract was subjected to purification by preparative HPLC (Luna C_{18} column; 250 x 21.2, i.d., 5 µm) using a linear gradient of water/acetonitrile ([90/10→10/90], 25 mL/min, over 32 min) acidified with 0.1% formic acid. Fractions containing the presumed target compounds were collected, dried, and further subjected to reparative semi-preparative HPLC (250 × 10 mm i.d.; 5 µm) using water/acetonitrile ([70/30→20/80, 4 mL/min, over 24 min]) acidified with 0.1% formic acid to afford 1.1 and 2.2 mg of saponariosides A and B respectively as white powder material. The identity of the isolated compounds was resolved based on a combination of extensive 1 and 2D NMR spectral data interpretations together with comparison with the literature (Jia, Koike and Nikaido, 1998) (Appendix B).

2.4 Nucleic acid extraction and sequencing

2.4.1 RNA extraction from *S. officinalis* and cDNA synthesis

The frozen ground plant materials generated from the four clonal JIC accession plants harvested in July 2019 (Section 2.3.1) were used for RNA extraction. Total RNA was extracted from individual organ samples (6 organs per 4 biological replicates) using the RNeasy Plant Mini kit (Qiagen) with a modified protocol as described in (Mackenzie *et al.*, 2005). The MacKenzie-modified protocol utilises a customised lysis buffer to optimize the removal of contaminating polyphenolics. The lysis buffer was composed of 4 M guanidine isothiocyanate (GITC), 0.2 M sodium acetate (NaAc)

pH 5.0, 25 mM EDTA, 2.5% (w/v) PVP-40 (polyvinylpyrollidone, average molecular weight 40,000) and 1% (v/v) β-mercaptoethanol (β-ME) was added immediately before use. Per 50 mg of frozen ground sample, 600 µL of lysis buffer and 1 sterile tungsten bead was added in 2 mL Eppendorf tubes. Samples were vortexed vigorously and 60 µL of 20% (v/v) Sarkosyl/H_2O solution was added to each. Samples were subsequently incubated for 10 min in a heat block at 70 °C with vigorous shaking. After centrifugation at 18,000 × *g* for 1 min, the resulting supernatants were transferred to QIAshredder columns (Qiagen), and a standard ethanol wash and RNA elution was performed following the manufacturer's protocol. On-column DNase digestion was performed using RQ1 RNase-Free DNase (Promega). The quality of the extracted RNA was measured using a NanoDrop (ThermoFisher) and gel electrophoresis to ensure the samples met the standard requirements for sequencing by the Earlham Institute (260/280 nm ≥ 2.00; 260/230 nm ≥ 1.80; concentration > 3 ng).

For subsequent cloning, cDNAs were generated from 0.8 µg of DNase-treated RNAs from each organ of JIC 2 soapwort plant using GoScriptTM Reverse Transcriptase (Promega) with oligodT primers following the manufacturer's instructions. The resulting cDNAs were then diluted 1:20 with distilled water and a cDNA pool was produced by combining equal volumes of diluted cDNA from each plant organ.

2.4.2 DNA extraction from *S. officinalis*

For draft genome sequencing, genomic DNA (gDNA) was extracted from the leaves of four soapwort plants (JIC accession) harvested in Section 2.3.1 using a DNeasy® Plant Kit (Qiagen) following the manufacturer's protocol.

For PacBio long-read sequencing, high molecular weight (HMW) gDNA was extracted from leaves of soapwort (accession JIC 2) using a modified CTAB protocol including addition of proteinase K and RNAse K (Giolai *et al.*, 2016). The CTAB buffer was prepared with 100 mM Tris-HCl pH 8, 2% (w/v) cetyltrimethyl ammonium bromide (CTAB), 1.4 M NaCl, 20 mM EDTA in distilled water. Per 1 g of frozen ground material, 10 mL CTAB buffer and 20 µL proteinase K (10 mg/mL) were added, followed incubation at 55 °C for 1 hour. The samples were cooled on ice for 5 min and 5 mL chloroform was added in a 2 mL Eppendorf tube. The samples were inverted gently about 10 times and were subsequently centrifuged at 2,000 × *g* for 30 min. The resulting upper phase was transferred to a fresh tube and a 1x volume of

phenol/chloroform/isoamyl alcohol (25:24:1; SigmaAldrich) was added. After gently inverting several times, samples were centrifuged 2,000 × *g* for 30 min. The upper phase was then collected and 10% volume of 3 M NaOAc pH 5.2 was added, followed by addition of 2.5x volume of ice-cold ethanol. After inverting gently, samples were incubated on ice for 30 min. Following centrifugation at 2,000 × *g* for 30 min at 4 °C supernatants were discarded, and the remaining pellet was washed with ice-cold 70% ethanol. The centrifugation and ethanol wash steps were repeated twice more. After discarding the supernatants from the final wash, the gDNA pellets were dried overnight by gently inverting the sample tubes onto a paper towel. Dried pellets were resuspended in 300 µL water and 3 µL RNase. After incubating overnight at 4 °C, quality control was performed by NanoDrop and gel electrophoresis.

2.4.3 Sequencing, assembly, and annotation

Transcriptome

All transcriptome sequencing, assembly and annotation was performed by EI. A total of 24 RNA samples (extracted as described in Section 2.4.1) were sent to EI for transcriptome sequencing. The RNA-Seq library was prepared using NEBNext Ultra II Directional RNA-Seq library preparation kit and was subsequently sequenced on two lanes of NovaSeq 6000 SP flow cell (150 pair-end reads). Transcriptome assembly was performed using Trinity *de novo* assembler (Grabherr *et al.*, 2011), and ORF prediction and functional annotation was assigned using TransDecoder (http://transdecoder.github.io) and AHRD (automatic assignment of human readable description; https://github.com/groupschoof/AHRD), respectively. Transcript reads were quantified using Salmon (Patro, Duggal and Kingsford, 2015).

Genome

All genome sequencing, pseudochromosome assembly and annotation was performed by the Joint Genome Institute (JGI; Shengqiang Shu, Chris Plott, Jerry Jenkins, Melissa Willaims, Lori-Beth Boston, Jane Grimwood, Jeremy Schmutz) as a collaborative effort with Professor James Leebens-Mack (University of Georgia). Along with the extracted gDNAs, frozen leaf materials from lines 2 and 3 JIC soapwort plants were sent to JGI for Hi-C library preparation and sequencing.

Draft genomes were sequenced on Illumina Novaseq from gDNAs extracted in Section 2.4.2. Histograms of 24-mer frequencies were analysed by Genomescope (Vurture *et al.*, 2017) to assess heterozygosity. HMW gDNA of line 2 was chosen to be sequenced for the main genome. The main genome was sequenced and assembled with 41.61x (coverage against the haploid genome size) PacBio HiFi reads (mean length = 17,825 bp) using HiFiAsm (Cheng *et al.*, 2021) and polished with RACON with 59x Illumina 2 x 150 paired end reads. The resulting contigs were oriented, ordered and joined into chromosomes using the JUICER pipeline with 65.5x HiC reads, which indicated no mis-joins in the initial assembly. A total of 44 joins were informed from JUICER and applied to the initial assembly to form the final assembly consisting of 14 chromosomes, which contained 99.46% of the assembled sequences. Chromosomes were numbered largest to smallest, with the p-arm oriented to the 5′ end.

Genome annotation was aided by using Illumina RNA-seq reads using PERTRAN (JGI) (Wu *et al.*, 2016). PacBio Iso-Seq circular consensus sequences (CCS) was performed on the cDNA produced from the RNA pool (composed of RNAs from flower, flower bud, young leaf, old leaf, stem, root) of JIC 2 soapwort plant material and was used to obtain putative full-length transcripts. Gene models were predicted by homology-based predictors and AUGUSTUS (Stanke and Morgenstern, 2005). The transcripts were further selected using C-score, a protein BLASTP score ratio to the mutual best hit BLASTP score, as well as protein and expressed sequence tag (EST) coverage. The filtered gene models were subjected to Pfam analysis and models with weak gene models and more than 30% of transposable element domains were removed. Gene models with low homology, short single exons without protein domains, and low expression were also manually filtered.

2.5 Identification of gene candidates

2.5.1 Processing RNA-seq data and co-expression analysis

Transcript quantification was performed by EI using Salmon (ver. 3.3.3) on the *de novo* transcriptome generated in Section 2.4.3. This read count data was used as the basis for further processing and co-expression analysis, performed in R. The Salmon read counts were read into R using tximport (ver. 1.18.0) (Soneson, Love and

Robinson, 2016). Read counts of zero were removed and the remaining read counts were normalised using DEseq2 (ver. 1.30.1) (Love, Huber and Anders, 2014) by 'median of ratios' method which estimates the library size factor. The normalised read counts were used for transcriptome-based co-expression analysis. The characterised *S. officinalis* β-amyrin synthase (*SobAS1*), implicated as initiating saponarioside biosynthesis (Chapter 5), was used as a bait gene to extract other candidate genes with similar expression across different soapwort organs. Co-expression analysis was performed by generating Pearson correlation coefficient (PCC) for each gene to *SobAS1*. For heat-map generation, DEseq2 was used to perform $\log_2$ transformation on the normalised read counts with a pseudo count of one. Subsequently, heatmaps were generated using Heatmap3 (ver. 1.1.9) (Zhao *et al.*, 2014) by hierarchical clustering method.

2.5.2 BLAST searches against *S. officinalis* sequence resources

While the *S. officinalis* genome was not available until the near end of this thesis work, the EI transcriptome was ready to be used earlier on. Thus, all candidate gene mining was performed using the translated EI transcriptome. Once the genome was completed, all candidate genes identified from the EI transcriptome was used as queries to perform reciprocal BLASTP against the new genome for sequence verification. This also checked for the presence of any additional candidates that may be present in the genome but not in the transcriptome.

The translated *S. officinalis* transcriptome was mined for candidate genes using BLAST+ (ver. 2.7.1) (Camacho *et al.*, 2009). Based on the predicted biosynthetic pathway (Chapter 5), a variety of classes of enzymes were hypothesised to be involved in saponaroside biosynthesis, including the scaffolding OSCs and tailoring CYPs, CSLs, UGTs and ATs (Chapter 5). Therefore for each enzyme class, a selection of functionally characterised literature amino acid (aa) sequences (detailed in Chapter 5) were used as BLASTP queries, with the results filtered based on amino acid sequence length (OSCs ≥ 700 aa; CYPs, CSLs ≥ 400 aa; UGTs, ATs ≥ 300aa). The resulting candidate lists were further filtered based on their annotations, expression profiles across different soapwort organs, as well as co-expression to *SobAS1* where relevant (detailed in and Chapter 5).

2.5.3 Phylogenetic tree construction and sequence analysis

For phylogenetic analysis, protein sequences of identified candidates were aligned to query sequences in MAFFT (Katoh *et al.*, 2002) with a maximum of 1,000 iterations (local pair). Phylogenetic trees were generated from alignments using RaXML (Stamatakis, 2014) using the PROTGAMMAAUTO model and 100 bootstraps.

Signal peptide predictions of SoGH1 was performed using SignalP 5.0 (Almagro Armenteros *et al.*, 2019). The amino acid sequences were submitted to SignalP 5.0 with default parameters. The reported score is the D-score (discrimination score), discriminating signal peptides from non-signal peptides (Almagro Armenteros *et al.*, 2019). A low score represents likely non-secretory proteins.

2.5.4 plantiSMASH analysis of *S. officinalis*

plantiSMASH 1.0 (Kautsar *et al.*, 2017) was run on *S. officinalis* genome (generated in Section 2.4.3). Default parameters were used, which define a cluster as locus where at least three different enzyme subclasses (within at least two different enzyme classes) are co-located, with less than 50% amino acid identity.

3

Metabolite profiling of *S. officinalis*

3.1 Introduction

The accumulation of plant specialized metabolites often reflects where and when they are biosynthesized in the plant, and therefore can show organ and/or time specific patterns in the plant (Hartmann, 1996). For example, the biosynthesis of flavonoids responsible for flower pigmentation in petunia (*Petunia hybrida*) is coordinated with flower development, leading to differential accumulation of flavonoids in different floral parts (Van Tunen *et al.*, 1988). In another example, the biosynthesis of monoterpenes in spearmint (*Mentha spicata*) is restricted to the glandular trichomes that are formed early in leaf development, resulting in higher concentration of monoterpenes in young leaves compared to mature leaves (Gershenzon, Maffei and Croteau, 1989). Furthermore, in many saponin producing plant species, the accumulation pattern of saponins correspond (to varying degrees) with the expression profiles of the biosynthetic genes known to produce them (Zhao *et al.*, 2010).

Soapwort extracts have been exploited for many decades as natural detergent and medicine and is still used in pharmaceutical and nutraceutical sectors (Böttger and Melzig, 2011). In both past and present, the usage of soapwort extracts is dependent on the high content of saponins in the plant extracts; however, comprehensive study of metabolite content of soapwort is limited. Soapwort extract is a rich source of saponins with various aglycone cores such as quillaic acid, gypsogenin, gypsogenic acid and more (Fig. 3.1.1). Over the years, more than 40 different saponins have been isolated from soapwort extracts, which can be divided based on their aglycone core into quillaic acid (Fig. 3.1.2A), gypsogenin (Fig. 3.1.2B), gypsogenic acid (Fig. 3.1.2C), 16α-hydroxygypsogenic acid based saponins (Fig. 3.1.2D), and saponins with

miscellaneous aglycones (Fig. 3.1.2E) (Jia, Koike and Nikaido, 1998; Jia, Koike and Nikaido, 1999; Sadowska *et al.*, 2014; Moniuszko-Szajwaj *et al.*, 2016; Lu *et al.*, 2015; Takahashi *et al.*, 2022).

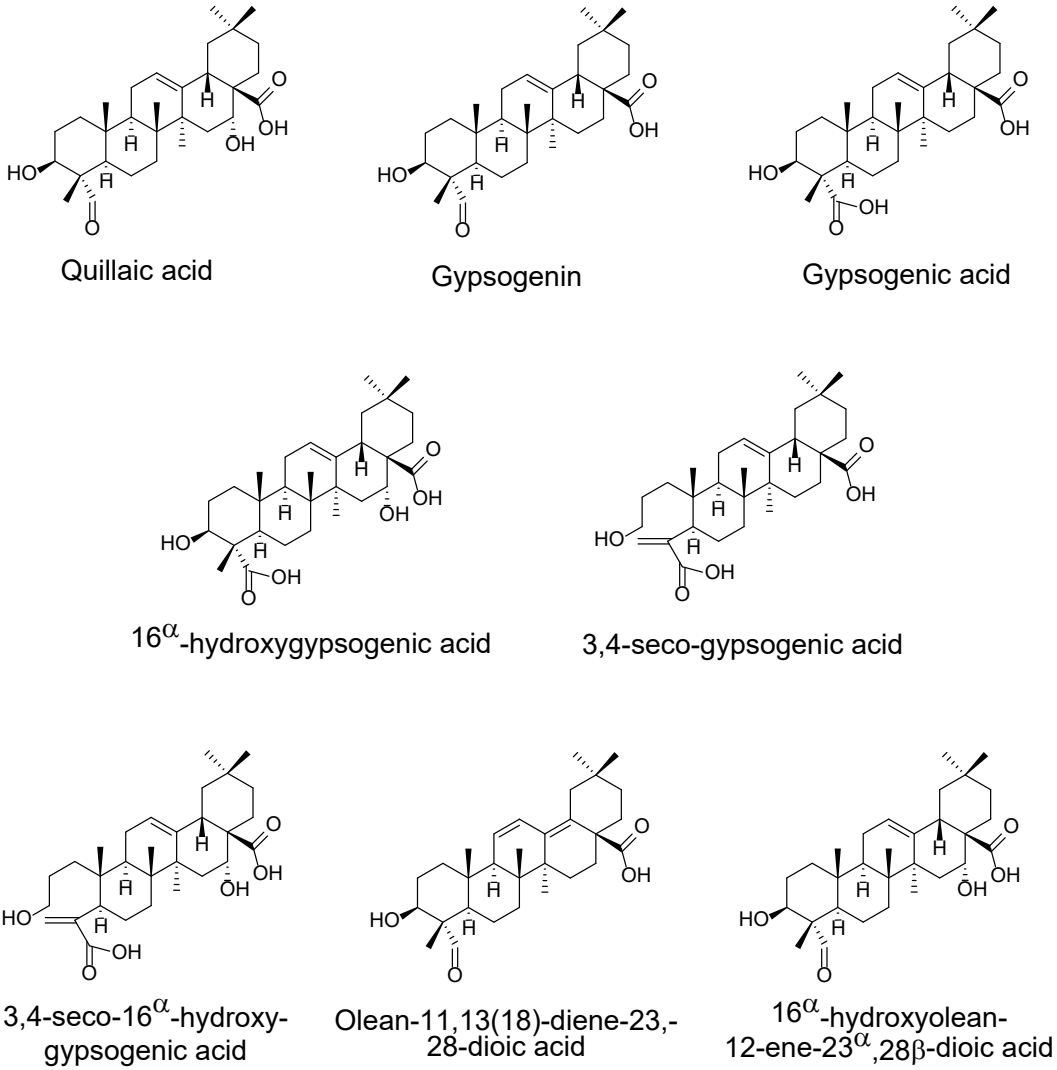


Figure 3.1.1. Aglycones of saponins isolated from soapwort.

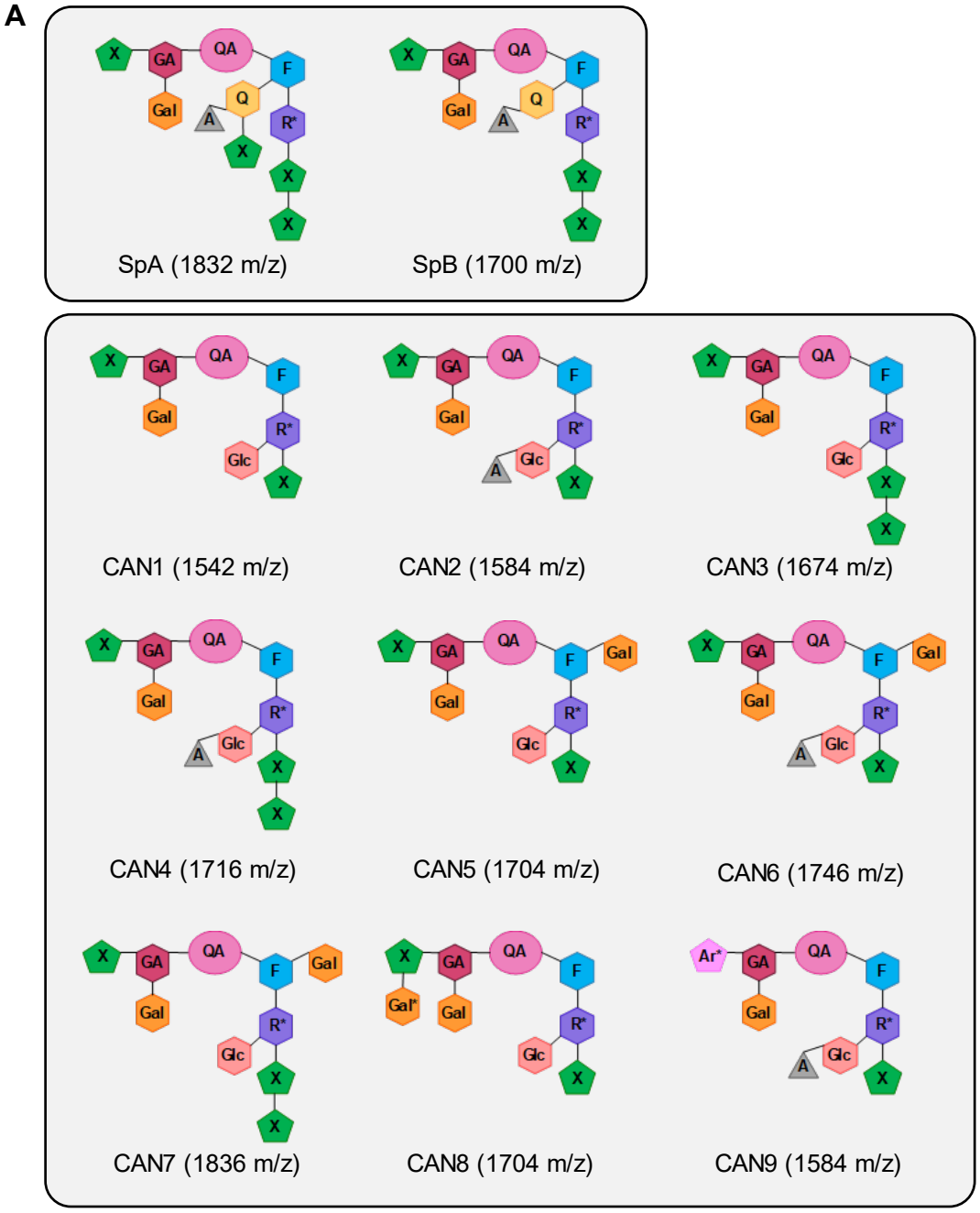
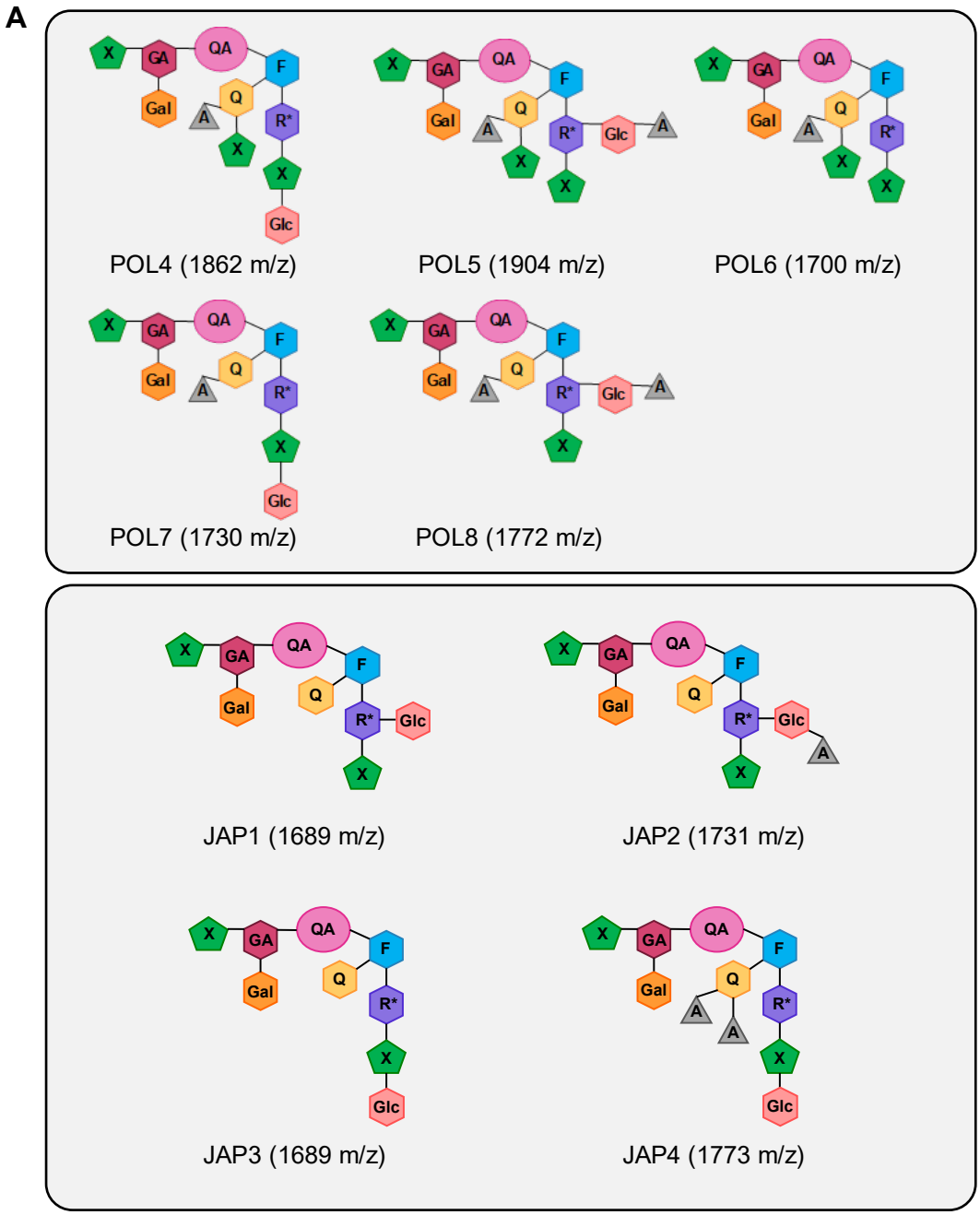
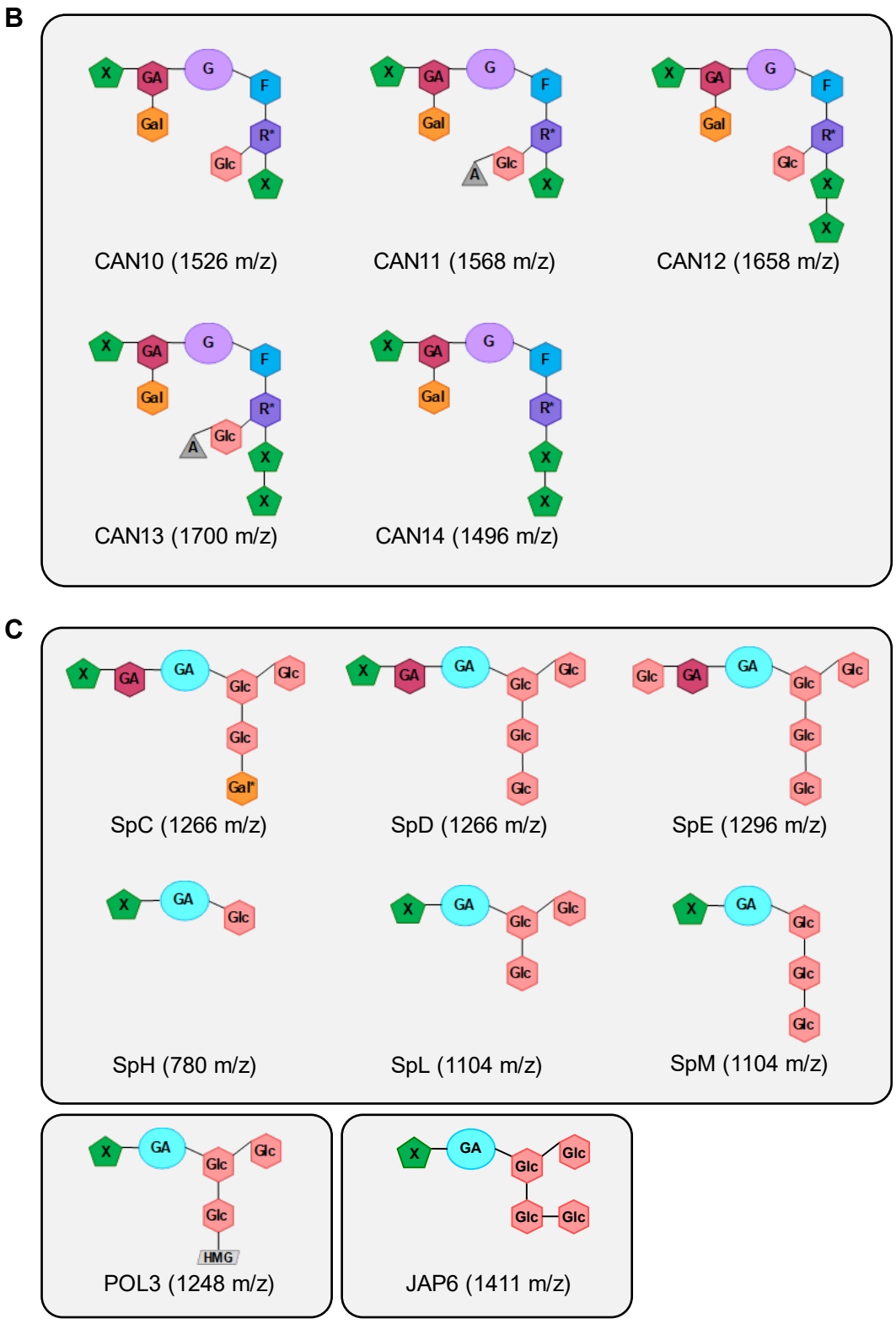


Figure 3.1.2. Schematic structures of saponins reported from soapwort. Saponins are grouped based on their aglycones: **(A)** quillaic acid, **(B)** gypsogenin, **(C)** gypsogenic acid, **(D)** 16α-hydroxygypsogenic, **(E)** miscellaneous aglycones. Corresponding mass-to-charge ratios (m/z) are shown with compound names. Details can be found in Appendix A.2. Sugars in α-L-configuration are noted with asterisks (*), without are in β-D-configuration. QA, quillaic acid; G, gypsogenin; GA (sky-blue), gypsogenic acid; OHGA, 16α-hydroxygypsogenic acid; ODDA, olean-11,13(18)-diene-23,28-dioic acid; secoOHGA, 3,4-seco-16α-hydroxy-gypsogenic acid; OHODA, 16α-hydroxyolean-12-ene-23,28 β-dioic acid; secoGA, 3,4-seco-gypsogenic acid; F, fucose; R, rhamnose; X, xylose; Q, quinovose; A, acetyl moiety; GA (red), glucuronic acid; Gal, galactose; Ar, arabinose; Glc, glucose. HMG, hydroxymethylglutaryl.

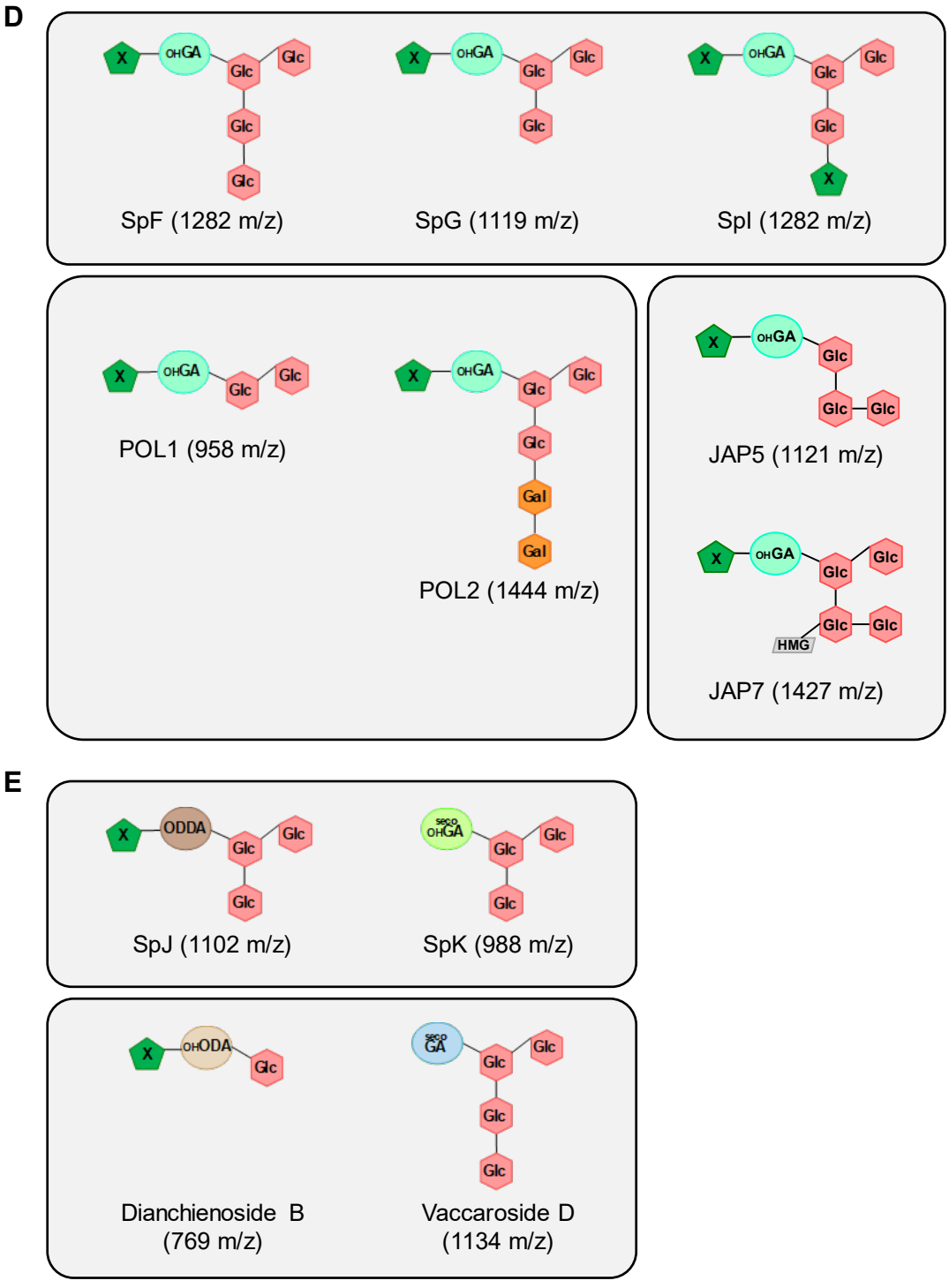
(Fig. 3.1.2. continued)



(Fig. 3.1.2. continued)



(Fig. 3.1.2. continued)



Although the major saponins found in soapwort extracts are reported as saponariosides A and B (Jia, Koike and Nikaido, 1998), this statement may be outdated as numerous soapwort saponins were identified later. In 1998, the Nikaido group obtained 8.3 mg/kg and 3.3 mg/kg of saponariosides A and B, respectively, from fresh whole soapwort plants (Jia, Koike and Nikaido, 1998). A year later, the Nikaido group reported a new saponin, saponarioside I, to be abundantly present (87 mg/kg) in the extracts while it was not detectable in their previous work (Koike, Jia and Nikaido, 1999). The only differences between the earlier and later study were the usage of fresh whole plants versus air-dried whole plants, and the location of the plant harvest (University of Tokyo versus Toho University) The findings from the Nikaido group suggests that specific saponin constituents of *S. officinalis* extracts may be highly variable depending on the origin and preparation of the plant material. The variability of soapwort saponin content may have important consequences to the end usage of the soapwort extracts.

The majority of studies of soapwort saponins have been conducted using either roots, leaves or whole plants and do not provide direct comparisons of saponin levels in different organs. Furthermore, saponins are often globally quantified using methods involving butanol extraction rather than targeting specific compounds (Budan *et al.*, 2014). Budan and co-workers have identified this problem and have attempted to provide a clearer resolution on saponin distribution in soapwort by dividing the plant into aerial parts and root (Budan *et al.*, 2014). The plant extracts were analysed using high-pressure liquid chromatography coupled with mass spectrometry (HPLC-MS) and saponin identification was based on the aglycone core, lacking details in specific chemical structure. Saponins were then quantified based on their peak areas using an external standard curve of hederacoside C, a bidesmodic triterpenoid saponin with a hederagenin aglycone (Góral and Wojciechowski, 2020). Using this quantification method, they have reported the saponin content of soapwort aerial parts to be 224.0 mg/g and 693.8 mg/g in the dried root material (Budan *et al.*, 2014). However, in this study, the organs composing the aerial parts were not disclosed, and the origin of aerial parts (field harvested) and root material (commercial powder) were from different sources.

Furthermore, in most studies regarding soapwort, information about the soapwort variety used is not provided and it is unknown if these studies have used single- or double-flowered plants. Double-flowers are often found in plants of the Caryophyllaceae order as result of stamen modification to petaloid staminodes (Marks, 1967). The double-flowered soapwort plants have flowers that appear to have several extra petals compared to the single-flowered form, which is five-petalled (Fig. 3.1.3). As such, the double-flowered soapwort plants are sometimes referred to as the *flore pleno* (double flower) variety (Cherevach and Shchekaleva, 2020). Although both double- and single- flowered forms of soapwort are found in the wild, the former has likely escaped from household gardens as double-flowers are unlikely to be selected in nature as they ultimately reduce the seed number (Corbet *et al.*, 2001). Additionally, double-flowers of soapwort produce minimal to no nectar while single-flowers produce high amounts (Corbet *et al.*, 2001), further suggesting that the double-flowered soapwort variety is a garden escapee. There is very little reported on the comparison of saponin content in the single- and double- flowered soapwort. A study performed by Cherevach and Shchekaleva in 2020 reported that the saponin content of roots of the double-flowered soapwort (30%) was higher than of the single-flowered variety (23%) (Cherevach and Shchekaleva, 2020). However, this study lacked technical or biological replicates as well as any statistical analysis to confirm the significance of the observed difference.

Figure 3.1.3. Flowers of soapwort plants. The white to pale pink flowers can exist as **(A)** single- or **(B)** double-formed.

Absolute quantification of saponins in soapwort is challenging due to the numerous types of saponins present in soapwort and lack of any commercially available standards. Nonetheless, of the identified soapwort saponins, quillaic acid-based saponins are the most numerous (Fig. 3.1.2). Within these saponins, all but one compound (CAN9) share the common prosapogenin, 3-*O*-{β-D-xylopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-quillaic acid (abbreviated as QA-Tri; Fig. 3.1.2A). As such, these saponins may share common early pathway steps. Instead of profiling each saponin species, saponification can be performed on the plant extracts to release the prosapogenin QA-Tri, which then can be analysed and quantified in the soapwort extracts (Fig. 3.1.4). The differing accumulation pattern of QA-Tri in addition to saponariosides A and B will provide foundational knowledge for designing RNA sequencing (RNA-Seq) experiments, which will be crucial for elucidating the biosynthetic pathway of saponariosides.

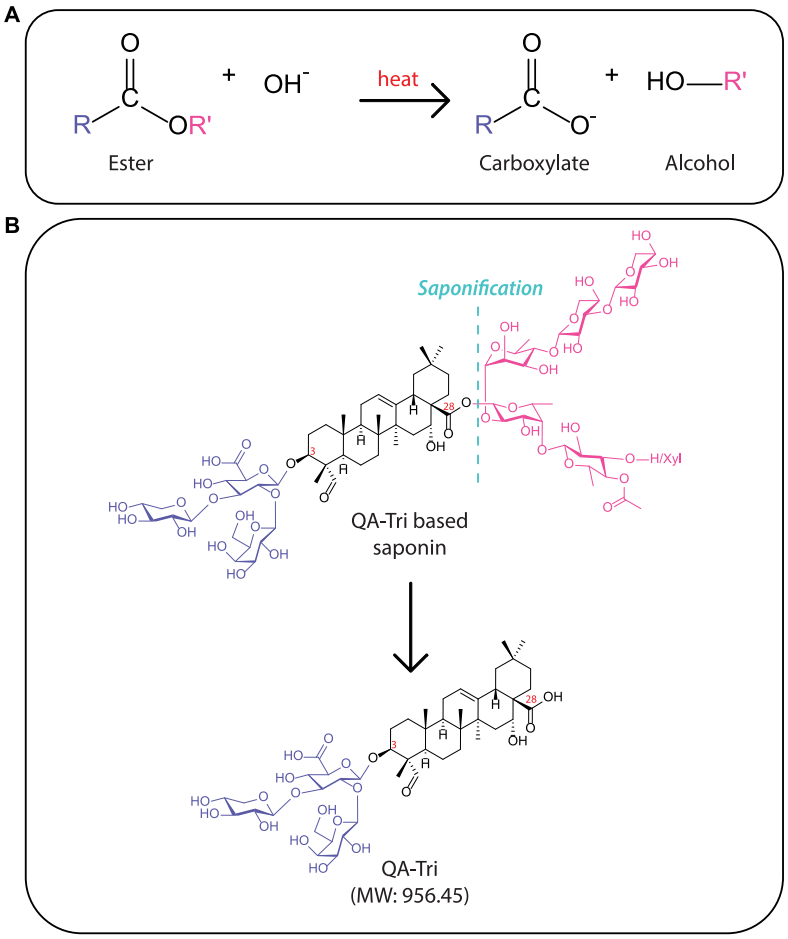


Figure 3.1.4. Schematic of saponification reaction. **(A)** Saponification is a chemical process where an ester bond reacts with a strong base resulting in a carboxylate and alcohol. **(B)** Saponification performed on bidesmodic quillaic acid-based saponin can release the prosapogenin, QA-Tri. MW, molecular weight.

3.2.1 Aims

In this Chapter, focused metabolite analysis of saponariosides A and B was performed on six different organs of *S. officinalis* to establish which organs of the plant are likely to be most biosynthetically active regarding saponarioside biosynthesis. Additionally, saponarioside A and B content was compared between the summer and winter months as well as in two varieties of soapwort. Although soapwort plants have been analysed for their saponin content by several groups, these studies lacked information about organ-specific saponin levels and levels in different development stages of the plant. As such, the aims of this chapter were to analyse:

1. Content of saponariosides A and B in different soapwort organs
2. Saponarioside levels during different developmental stages of soapwort
3. Saponarioside content in different soapwort varieties
4. Prosapogenin (QA-Tri) levels in different soapwort organs

3.2 Results and discussion

3.2.1 Spatiotemporal distribution of saponariosides A and B

To investigate the biosynthesis of SpA and SpB and their spatial accumulation patterns, metabolite analysis was conducted on six different organs (flower, flower bud, young leaf, old leaf, stem, root) of single-flowered soapwort plants (JIC accession) harvested in July 2019 (Fig. 3.2.1). The different plant organs were harvested separately, freeze dried and extracted for metabolite analysis. The resulting extracts were analysed using HPLC-MS in negative ionization mode. It is worthwhile noting that any minor differences in *m/z* observed in the LC-MS analyses presented in this thesis were due to the signals generated by isotopologues of the same chemical entity. For example, carbon-13 makes up 1% of all carbon atoms, thus 1:100 of carbons will contain carbon-13 rather than carbon 12 (i.e. 1:100 of the carbon-based molecules will have a mass of +1 units).

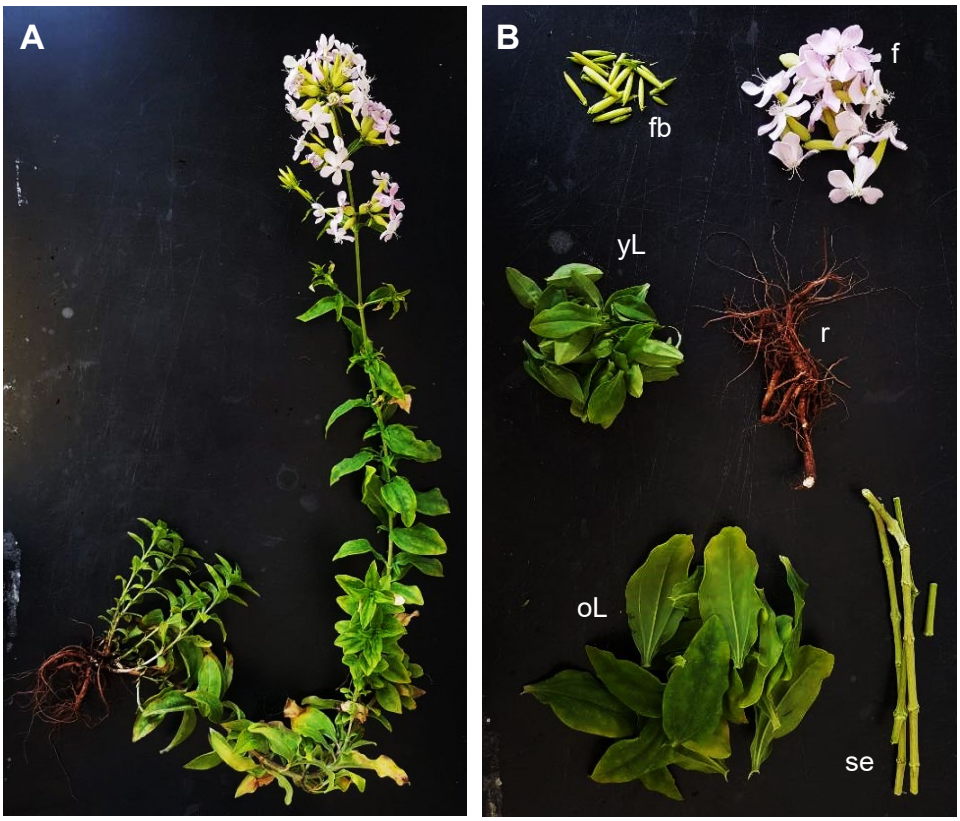


Figure 3.2.1. Representative image of a soapwort plant harvested in July 2019. **(A)** Whole plant. **(B)** The different organs harvested: flower bud (fb), flower (f), young leaf (yL), old leaf (oL), root (r), stem (se).

Throughout this project, none of the saponins from soapwort were commercially available to be used as standards. To mitigate this, authentic standards for

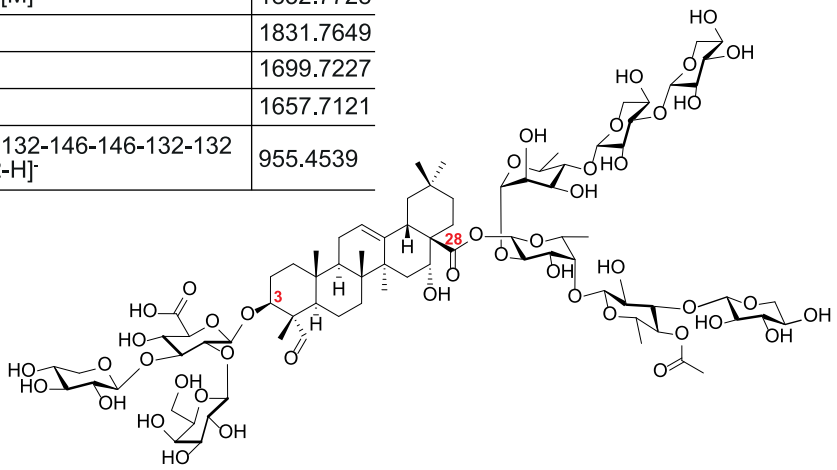
saponariosides A and B were kindly generated by Dr. Amr El-Demerdash. As this is a destructive process and only limited numbers of JIC plants were available in the beginning of the project, we opted to use commercially available soapwort material. Dried soapwort leaves purchased from a commercial supplier were extracted with 80% methanol, and butanol was used to separate the saponins from the aqueous methanolic extract. The crude saponin fraction was purified using preparative HPLC and the chemical structures of saponariosides A and B were resolved by extensive 1D and 2D NMR (Appendix B). These authentic standards were then used to identify saponariosides A and B from the extracts of various soapwort organs (Figs. 3.2.2 and 3.2.3).

In extracts of all soapwort organs analysed, two metabolite peaks (**a** and **b**) were detected in the extracted ion chromatograms (EIC) at *m/z* 1831.76 and 1699.72, respectively (Figs. 3.2.2B and 3.2.3B). The *m/z* of peaks **a** and **b** correspond to the reported *m/z* of SpA and SpB in negative ionization mode, respectively (Jia, Koike and Nikaido, 1998). Additionally, the MS/MS fragmentation of **a** and **b** revealed the major fragment ion as *m/z* 955.45, corresponding to the *m/z* of the prosapogenin QA-Tri, which was previously reported as the major fragment ion of SpA and SpB (Jia, Koike and Nikaido, 1998). Furthermore, the fragmentation pattern and the retention time (RT) of peaks **a** and **b** matched with the authentic SpA and SpB standards, respectively. Based on these results, **a** was identified as SpA, and **b** as SpB. However, in the EIC at *m/z* 1699.72, an additional prominent peak (**c**) was observed. Although **c** showed the same fragmentation pattern as the authentic SpB standard, they differed in the RT. The identity of **c** may be a positional isomer of SpB; for example, the terminal D-xylose moiety of the C-28 sugar chain may be attached to the D-quinovose moiety instead, or possibly the attachment of the acetyl moiety on D-quinovose to the terminal D-xylose moiety instead.

A

Saponarioside A ($C_{82}H_{128}O_{45}$)

Exact mass [M]	1832.7728
$[M-H]^-$	1831.7649
$[M-132-H]^-$	1699.7227
$[M-Ac-H]^-$	1657.7121
$[M-176-162-132-146-146-132-132-146-Ac-132-H]^-$	955.4539



B

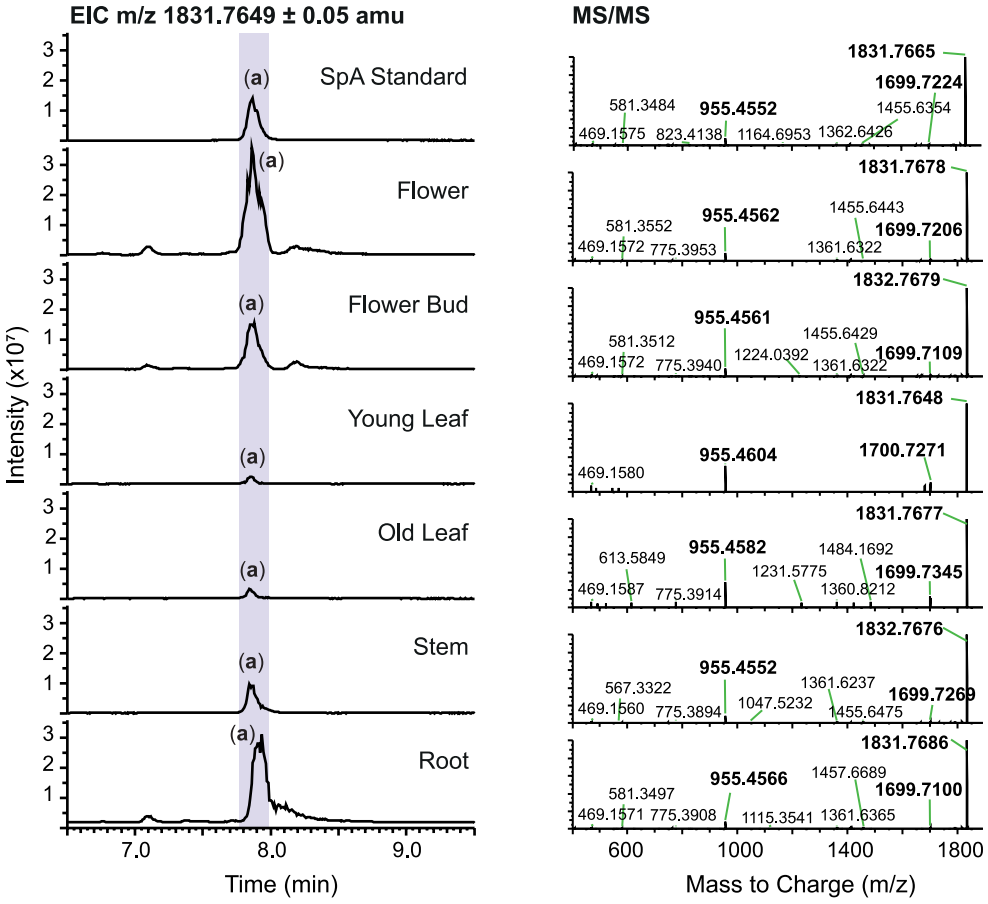


Figure 3.2.2. Detection of SpA in extracts of different soapwort organs. The plant extracts were analysed using HPLC-MS in negative ionization mode. **(A)** Structure of saponarioside A with a table showing relevant calculated adducts and fragments. **(B)** EIC displayed for *m/z* 1831.7649 (calculated $[M-H]^-$ of SpA) and MS/MS spectra of the highlighted peak in corresponding plant samples are shown.

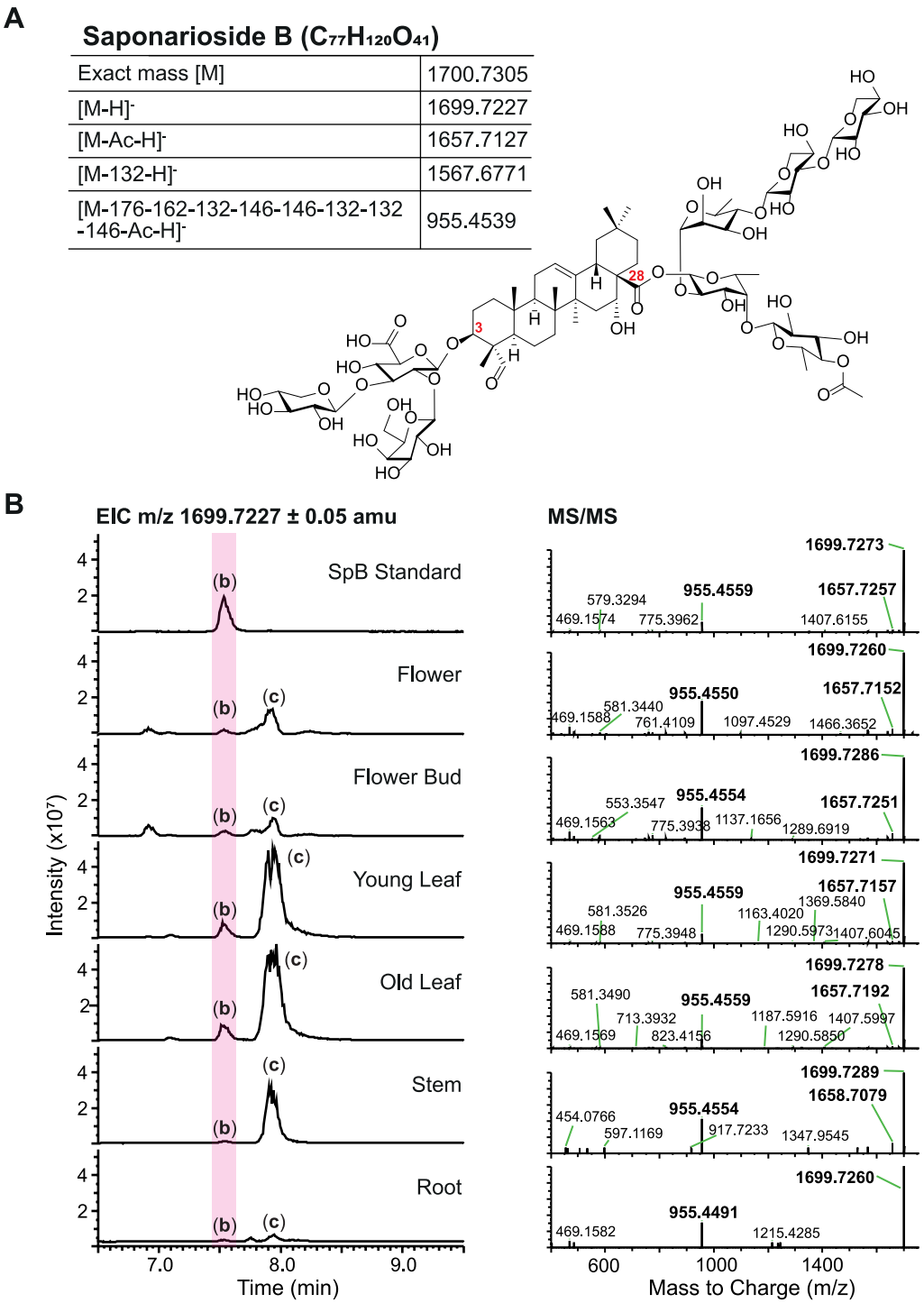


Figure 3.2.3. Detection of SpB in extracts of different soapwort organ. The plant extracts were analysed using LC-MS in negative ionization mode. **A.** Structure of saponarioside A with a table showing relevant calculated adducts and fragments. **B.** EIC displayed for *m/z* 1699.7227 (calculated $[M-H]^-$ of SpB) and MS/MS spectra of the highlighted peak in corresponding plant samples are shown. Identity of peak (**c**) may be a positional isomer of SpB.

Following the identification of SpA and SpB in soapwort extracts, the peak areas of **a** and **b** were normalized using the internal standard digitoxin, and the relative abundance (as mg of saponin per g of dried plant material extracted) was calculated (Fig. 3.2.4).

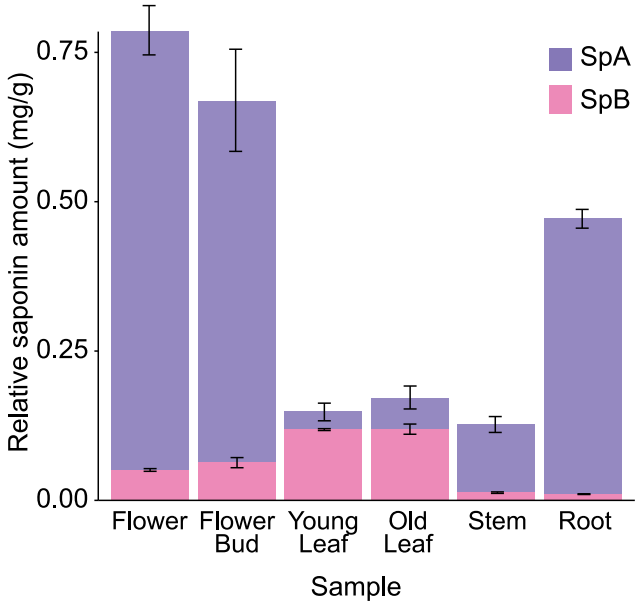


Figure 3.2.4. Relative abundance of SpA and SpB. Compounds were identified using authentic standards. Relative abundance was calculated using the internal standard digitoxin, based on dry weight. Each bar represents the mean of four biological replicates, and error bars indicate standard error. SpA shown in purple, SpB in pink.

The relative abundances of SpA and SpB differed between the different soapwort organs. SpA was most abundant in the flower and flower bud, while only trace amounts were detected in the young and old leaves. In contrast, SpB was more dominant in the young and old leaves, and only trace amounts were detected in the stem and root. Overall, SpA accumulated in greater levels compared to SpB in the flower, flower bud, stem, and root, while SpB was present in higher amounts in the leaves (Fig. 3.2.4). This contrast is interesting as saponariosides A and B are very similar in chemical structure (Figs. 3.2.2A and 3.2.3A). The only difference between the two compounds is the additional D-xylose sugar on the C-28 sugar chain of SpA. The unequal distribution of SpA and SpB may be due to several reasons. The biosynthesis of SpA and SpB may be occurring in the site of accumulation; for example, the biosynthesis of SpA may be more active in the flower organs compared to the leaves and vice versa for SpB. On the contrary, SpB may be a by-product of SpA hydrolysis. Another reason may be due to tightly governed transportation of SpA and SpB from site of

biosynthesis to modulate the accumulation profile in the different organs. The difference in the spatial distribution of saponariosides A and B may be due to the differences in their biological roles in the plant, which are as yet unknown. For example, SpA may provide protection against herbivory and soil microbes, while SpB may provide protection against harmful UV rays. Future investigation of the biological activities of saponariosides A and B may reveal valuable fundamental knowledge about the likely biological roles of these saponins. Overall, SpA was more abundant in the whole soapwort plant compared to SpB, and the highest accumulation of SpA was in the flower and the flower buds. If biosynthesis of saponariosides is occurring in the site of accumulation, this may suggest that the flower organs are the major site of saponarioside biosynthesis. Although soapwort roots are often reported as the organ with high saponin content, this may be due to sampling bias. Underground organs, especially the roots, are a preferential site of saponin accumulation in most saponin producing plants (Moses, Papadopoulou and Osbourn, 2014). As such, most saponin analyses in soapwort have been performed either on the root or the whole plant without comprehensive metabolite analysis, as mentioned in Section 3.1. Furthermore, aerial organs, such as flower and flower buds, are season-dependent and thus not easily accessible, hindering the sampling of such organs. However, when foamability of extracts from different soapwort organs was tested, extracts of fresh leaves and flowers produced the highest quality foams compared to fresh roots or rhizome extracts (Goral, Jurek and Wojciechowski, 2018). The authors of this experiment suggested that if the foamability of the extract was directly proportional to the saponin content, the leaves and flowers may be rich reservoirs of soapwort saponins (Goral, Jurek and Wojciechowski, 2018).

Next, as saponin content may also vary depending on the developmental stages of the plant, the saponin levels in extracts of soapwort plants harvested in the summer (July 2019) and winter (November 2019) were compared (Fig. 3.2.5). Plant organs harvested in the winter showed slightly higher accumulation of both SpA and SpB compared to plant organs harvested in the summer; however, the differences were mainly statistically non-significant. Furthermore, the overall accumulation pattern of SpA and SpB remained unchanged; SpA was most abundant in the root while SpB was most abundant in the leaf (flowers were not compared as they were not available in November). However, two time points are likely not sufficient to accurately observe

the accumulation pattern of saponariosides throughout the plant's developmental stages. Future experiments with increased numbers of time points distributed equally throughout the year may provide better insight into the timeframe of saponarioside biosynthesis in soapwort plants.

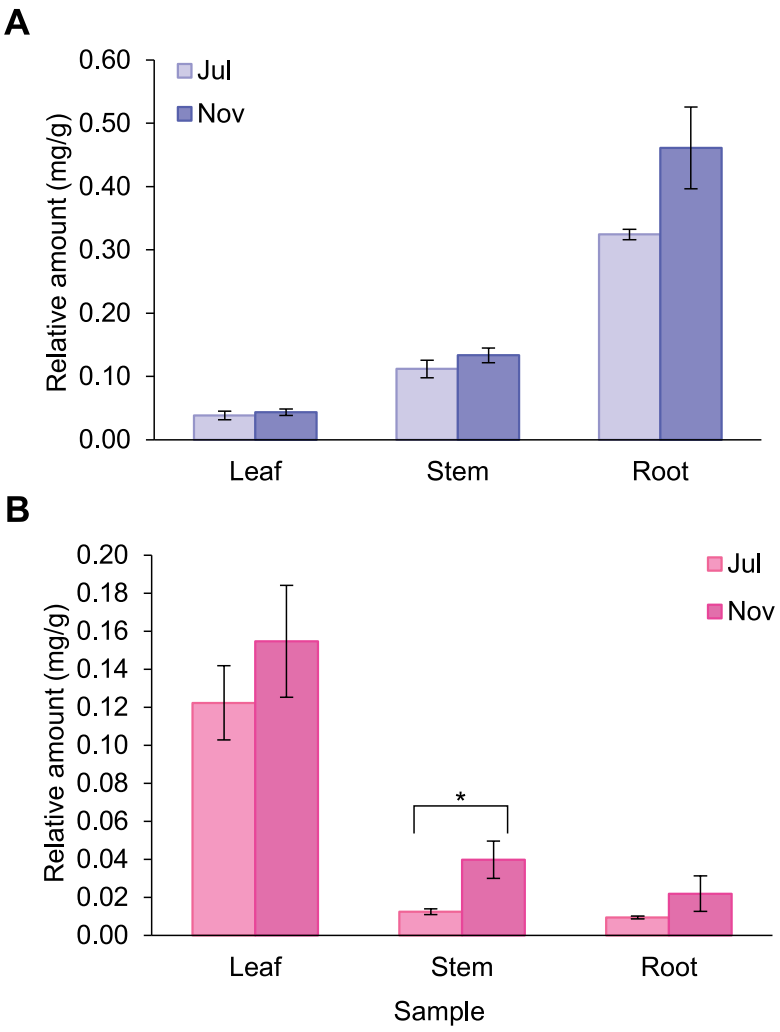


Figure 3.2.5. Comparison of relative amounts of SpA (A) and SpB (B) in different soapwort organs in summer and winter. Plants were harvested in July and November to represent summer and winter months, respectively. Extracted plant material were analysed using HPLC-MS. Saponariosides A and B were identified using authentic standards. Relative abundance was calculated using the internal standard digitoxin, based on dry weight. Each bar represents four biological replicates, the error bars indicate standard error. Asterisks (*) indicate significant differences between the different months determined by student's T-test (P-value ≤ 0.05).

3.2.2 Profiling of SpA and SpB in different *S. officinalis* varieties

The main soapwort accession used throughout this project was the single-flowered JIC accession originally purchased from Norfolk Herbs, Norfolk. However, a double-flowered soapwort, given the RSM accession, was collected at a location in Rockland St. Mary, Norfolk by Prof. Anne Osbourn. To investigate whether different soapwort accessions have differing accumulation patterns regarding SpA and SpB, metabolite analysis focused on these two saponins was performed to compare between the JIC and RSM accessions. However, as sample harvest was conducted in November (2019) for this experiment, instead of flowers, seeds from the two accessions were harvested and analysed, along with leaf, stem and root. The plant materials were extracted for metabolites and the resulting extracts were analysed using HPLC-MS in negative ionization mode. Saponariosides A and B were identified using authentic standards, and the relative amounts were calculated as explained above. The two different accessions showed similar levels of SpA and SpB in the extracts of soapwort organs compared, and the overall spatial pattern remained unchanged (Fig. 3.2.6). As more JIC soapwort plants were available, this accession was used for further experiments as the two accessions had little variability in SpA and SpB accumulation.

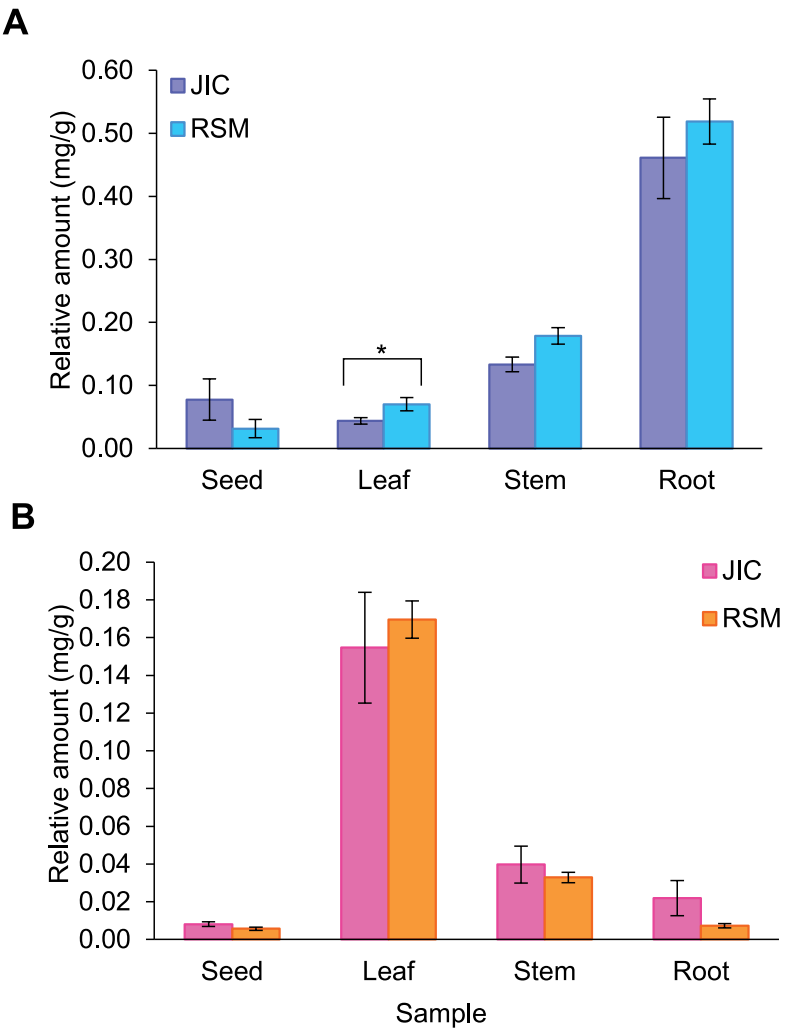


Figure 3.2.6. Relative amount of SpA (A) and SpB (B) in soapwort plants from different Norfolk areas. JIC accession plants were purchased from Norfolk Herbs based in Dereham, Norfolk. RSM accession plants were found naturally growing in Rockland St. Mary, Norfolk. Plant materials were extracted and analysed using HPLC-MS. Saponariosides A and B were identified using authentic standards. Relative abundance was calculated using the internal standard digitoxin, based on dry weight. Each bar represents four biological replicates, the error bars indicate standard error. Asterisks (*) indicate significant differences between the different location, determined by student's T-test (P-value ≤ 0.05). DW, dry weight.

3.2.3 Profiling of the common saponin scaffold QA-Tri

QA-Tri is a common prosapogenin for many quillaic acid-derived soapwort saponins (Fig. 3.1.2A). As such, the accumulation pattern of QA-Tri may provide insight into early saponarioside biosynthesis. Saponification was performed on six different organs of JIC accession plants harvested in July 2019 (Fig. 3.2.1). The resulting extracts were neutralized and analysed on HPLC-MS in negative ionization mode.

A dominant peak (**d**) in the EIC of *m/z* 955.46 ($[M-H]^-$ of QA-Tri) was observed only in the saponified plant extracts, suggesting that this peak was the result from saponification. The *m/z*, RT, and MS/MS fragmentation pattern of peak (**d**) observed in the plant extracts corresponded to the authentic QA-Tri standard, and thus was identified as QA-Tri (Fig. 3.2.7). Following identification, the peak areas of QA-Tri were quantified, and the relative amounts were calculated using an external standard curve of hederacoside C, as the internal standard digitoxin was affected by the saponification process. Soapwort organs with relatively high amounts of QA-Tri after saponification were revealed to be the flower organs, followed by leaf, root and stem (Fig. 3.2.7A). This suggests that quillaic acid-based saponins are present in high amounts in the flower parts compared to other soapwort organs, which was in agreement with the accumulation pattern of SpA and SpB (Fig. 3.2.4).

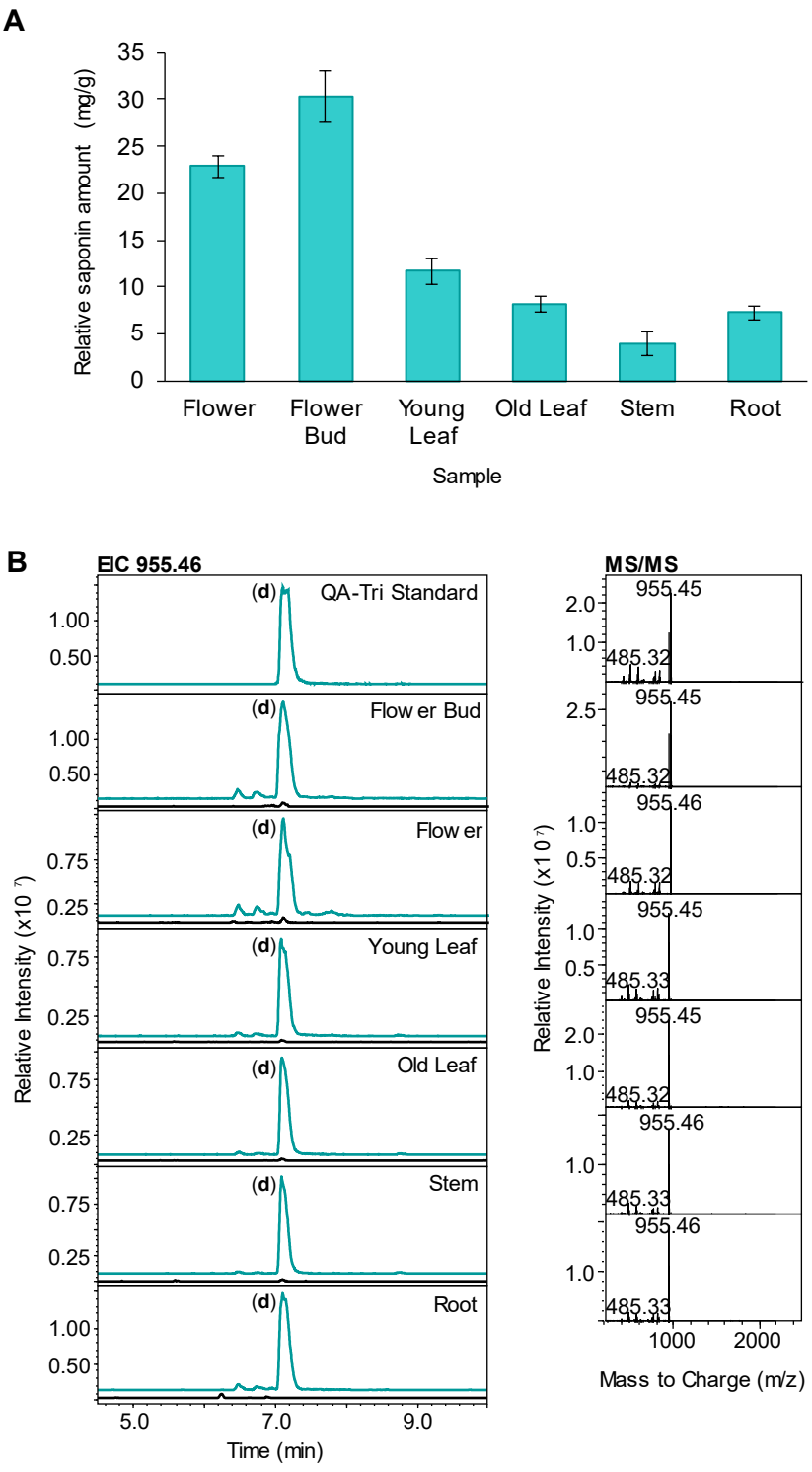


Figure 3.2.7. Prosapogenin QA-Tri in various soapwort organs treated by saponification. (A) Relative abundance of QA-Tri in different soapwort organs. External standard curve of hederacoside C was used to calculate the relative abundance of QA-Tri. Each bar represents four biological replicates, the error bars indicate standard error. **(B)** Extracted ion chromatograms (EIC) and MS/MS fragmentation of *m/z* 955.46 (calculated $[M-H]^-$ of QA-Tri) displaying peak (**d**). Black traces indicate samples before saponification, teal traces correspond to samples after saponification. Peak (**d**) is identified as QA-Tri by comparison to authentic QA-Tri standard.

3.3 Conclusion

Throughout the different metabolite analyses, SpA was found to be consistently more abundant in the flower organs while SpB was more abundant in the leaves. The difference between the accumulation profiles of SpA and SpB may be due to several reasons, such as uneven biosynthesis of SpA and SpB in those organs, the hydrolysis of SpA into SpB, or the differences in transportation of SpA and SpB to various soapwort organs from site of biosynthesis. This difference in the spatial distribution of SpA and SpB may also hint at a difference in their biological roles in the plant. Future work investigating the biological activities of SpA and SpB may reveal their *in planta* roles, which are not yet known. Overall, the flower organs contained the highest combined relative amounts of SpA and SpB compared to the other organs analysed. If the site of accumulation and biosynthesis are the same for SpA and SpB, results of this chapter may suggest that flower organs are the active sites of saponarioside biosynthesis. The knowledge gained from these metabolite analysis experiments can be used to guide transcriptomic analysis. The differential accumulation of SpA and SpB in different soapwort organs will pave the way for RNA-sequencing (RNA-Seq) experiments, which will aid in identification of candidate genes involved in saponarioside biosynthesis based on differential gene expression in high and low saponin content organs.

4

Generation of sequence resources for *S. officinalis*

4.1 Introduction

Genomic and transcriptomic resources are crucial tools in biosynthetic gene discovery. On the commencement of this project, the only publicly available soapwort sequence data was a transcriptome from the 1000 Plants (1KP) project (www.onekp.com) (Wickett *et al.*, 2014). The 1KP project provides transcriptomic data from a wide variety of plant species; however, it often lacks details about the material used and the methodology due the broad nature of the project. As a result, the *S. officinalis* 1KP transcriptome was produced from pooled plant organs and was, thus, unsuitable for further RNA sequencing (RNA-Seq) analysis such as gene expression comparison or co-expression analyses. In addition to the limiting transcriptomic data, there was no genome sequence available for *S. officinalis*. To date, there are a total of 40 chromosome or scaffold-level genome sequences for plants in the Caryophyllales order publicly available in the NCBI database. Of these, only seven are members of the Caryophyllaceae family (Table 4.1.1).

Although there was no genome sequence available for soapwort, there have been some attempts to determine the karyotype and genome size of this species. Using flow cytometry, two separate studies confirmed *S. officinalis* to be a diploid species (2n = 28) with DNA content of 2C = 4.65 pg (Di Bucchianico *et al.*, 2008) or 2C = 4.51 pg (Pustahija *et al.*, 2013). The C-value refers to the amount of DNA content within a haploid nucleus, and picograms of DNA can be converted to base pairs by applying the formula (Dolezel, 2003):

$$Genome\ size\ (bp) = (0.978 \times 10^9) \times DNA\ content\ (pg)$$

Table 4.1.1. Sequenced plant genomes in the Caryophyllales order. Data was retrieved from NCBI genome list (https://www.ncbi.nlm.nih.gov/datasets/genome/).

Organism Name	Family	Size (Mbp)	Assembly	BioProject
Amaranthus cruentus	Amaranthaceae	370.91	Chromosome	PRJNA713964
Amaranthus hypochondriacus	Amaranthaceae	417.46	Chromosome	PRJNA214803
Amaranthus palmeri	Amaranthaceae	541.10	Scaffold	PRJNA380417
Amaranthus tricolor	Amaranthaceae	520.08	Chromosome	PRJNA891371
Bassia scoparia	Amaranthaceae	711.36	Scaffold	PRJNA526487
Beta vulgaris	Amaranthaceae	568.75	Chromosome	PRJNA780534
Chenopodium formosanum	Amaranthaceae	1629.75	Chromosome	PRJNA840947
Chenopodium pallidicaule	Amaranthaceae	337.01	Scaffold	PRJNA326220
Chenopodium quinoa	Amaranthaceae	1333.40	Scaffold	PRJNA306026
Chenopodium suecicum	Amaranthaceae	536.95	Scaffold	PRJNA326219
Dysphania ambrosioides	Amaranthaceae	468.77	Chromosome	PRJNA821252
Spinacia oleracea	Amaranthaceae	894.26	Chromosome	PRJNA598728
Suaeda aralocaspica	Amaranthaceae	451.68	Scaffold	PRJNA428881
Carnegiea gigantea	Cactineae	980.40	Scaffold	PRJNA318822
Cereus fernambucensis	Cactineae	912.18	Scaffold	PRJNA587492
Lophocereus schottii	Cactineae	797.93	Scaffold	PRJNA318822
Pachycereus pringlei	Cactineae	629.66	Scaffold	PRJNA318822
Pereskia humboldtii	Cactineae	414.05	Scaffold	PRJNA318822
Selenicereus undatus	Cactineae	1326.73	Scaffold	PRJNA664414
Stenocereus thurberi	Cactineae	853.35	Scaffold	PRJNA318822
Talinum fruticosum	Cactineae	620.37	Scaffold	PRJNA659383
Corrigiola litoralis	Caryophyllaceae	304.40	Scaffold	PRJEB25561
Dianthus caryophyllus	Caryophyllaceae	636.30	Chromosome	PRJNA796118
Heliosperma pusillum	Caryophyllaceae	1208.43	Scaffold	PRJNA739571
Silene latifolia	Caryophyllaceae	665.28	Scaffold	PRJNA289891
Silene noctiflora	Caryophyllaceae	2598.31	Scaffold	PRJNA550146
Silene uniflora	Caryophyllaceae	769.23	Scaffold	PRJNA699303
Spergula arvensis	Caryophyllaceae	146.12	Scaffold	PRJEB25562
Drosera capensis	Droseraceae	263.79	Scaffold	PRJNA291419
Kewa caespitosa	Kewaceae	664.63	Scaffold	PRJEB24909
Limeum aethiopicum	Limeaceae	1153.03	Scaffold	PRJEB24908
Macarthuria australis	Macarthuriaceae	482.89	Scaffold	PRJEB24906
Microtea debilis	Microteaceae	475.43	Scaffold	PRJEB24907
Pharnaceum exiguum	Molluginaceae	287.97	Scaffold	PRJEB24886
Nepenthes mirabilis	Nepenthaceae	691.41	Scaffold	PRJNA869487
Limonium bicolor	Plumbaginaceae	2925.44	Chromosome	PRJNA753199
Fagopyrum esculentum	Polygonaceae	1211.00	Scaffold	PRJNA487881
Fagopyrum tataricum	Polygonaceae	505.88	Chromosome	PRJNA381676
Polygonum aviculare	Polygonaceae	351.57	Chromosome	PRJEB51791
Simmondsia chinensis	Simmondsiaceae	831.54	Scaffold	PRJNA694450

Thus, the genome size of *S. officinalis* can be estimated to be about 2.2 Gb. This genome size can be viewed as large when compared with some species such as *Quillaja saponaria* (1C = 0.42), a plant in the Fabales order producing saponins of similar chemical structure as saponariosides, and other plants within the same Caryophyllales order like *Beta vulgaris* (1C = 0.92) (Table 4.1.2). However, the estimated average C-value of 125 members of the Caryophyllaceae family is 1.46 ± 0.97 (https://cvalues.science.kew.org), placing the estimated genome size of soapwort slightly larger than average overall.

Table 4.1.2. C-values of selected members of the Caryophyllaceae family. Examples from other plant families are included for comparison. Data was retrieved from the Kew Plant DNA C-values database (https://cvalues.science.kew.org).

Organism name	Order	Family	1C (pg)
Arabidopsis thaliana 'Columbia'	Brassicales	Brassicaceae	0.16
Beta vulgaris	Caryophyllales	Amaranthaceae	0.92
Spinacia oleracea 'Greenmarket'	Caryophyllales	Amaranthaceae	1.00
Dianthus caryophyllus 'Master'	Caryophyllales	Caryophyllaceae	0.70
Dianthus caryophyllus 'Reina'	Caryophyllales	Caryophyllaceae	1.30
Dianthus hispanicus	Caryophyllales	Caryophyllaceae	0.86
Gypsophila repens	Caryophyllales	Caryophyllaceae	0.70
Gypsophila spergulifolia	Caryophyllales	Caryophyllaceae	0.50
Saponaria bellidifolia	Caryophyllales	Caryophyllaceae	1.85
Saponaria officinalis	Caryophyllales	Caryophyllaceae	2.26
Silene dioica	Caryophyllales	Caryophyllaceae	2.70
Silene vulgaris	Caryophyllales	Caryophyllaceae	1.13
Viscaria vulgaris	Caryophyllales	Caryophyllaceae	2.15
Quillaja saponaria	Fabales	Quillajaceae	0.42
Avena strigosa	Poales	Poaceae	4.54
Triticum aestivum 'Chinese Spring'	Poales	Poaceae	17.30
Zea mays 'W64A'	Poales	Poaceae	2.70

Sequencing the genome of *S. officinalis* would provide insights into genome size and physical organization. It would further enable systematic mining for genes encoding members of enzyme families implicated in specialized metabolism which, in combination with transcriptome data, would allow identification and prioritisation of candidate genes for saponarioside biosynthesis. Genes for plant metabolic pathways can sometimes be co-localized in the genome as biosynthetic gene clusters (BGCs) (Field and Osbourn, 2008). There are numerous reports of BGCs from plants, including for triterpenes such as avenacin from oat (*Avena strigosa*), ellarinacin from wheat (*Triticum aestivum*), and yossosides from spinach (*Spinacia oleracea*) (Qi *et*

al., 2004; Polturak *et al.*, 2022; Jozwiak *et al.*, 2020). Furthermore, genes involved in the biosynthesis of QS-21, a saponin with similar chemical structures to saponariosides A and B, partially form BGCs in the soapbark (*Quillaja saponaria*) genome (Reed *et al.*, 2023). With a high-quality genome, tools such as plantiSMASH can be used to mine for candidate BGCs in the genome, which has the potential to rapidly speed up the process of discovering biosynthetic pathways (Kautsar *et al.*, 2017).

4.1.1 Aims

At the start of this project, there were no genome assembly nor high-quality RNA-Seq data available for soapwort. Thus, the aims of this chapter were to generate:

1. Comprehensive RNA-Seq data from multiple *S. officinalis* organs
2. A pseudochromosome-level genome assembly for *S. officinalis*

4.2 Results and discussion

4.2.1 Generation of RNA-Seq resources for *S. officinalis*

The soapwort transcriptome data available from the 1KP project database was derived from pooled plant organs, thus limiting the usefulness of the data. In Chapter 3, six different soapwort organs (flower, flower bud, young leaf, old leaf, stem, root) were determined to contain different levels of saponins. Based on this result, RNA was extracted from these six organs from four JIC accession plants and sent to the Earlham Institute (EI) for sequencing, assembly, and annotation. The four sequenced plants were then treated as four separate lines (JIC 1-4) of the JIC accession. A total of 24 RNA samples were sequenced on two lanes of NovaSeq 6000 (Illumina) to produce 532 million base pairs (Table 4.2.1). The sequence reads were assembled using Trinity *de novo* assembler (Grabherr *et al.*, 2011), and individual assemblies from each organ were merged into a single reference transcriptome. The reference transcriptome was annotated by identifying open reading frames (ORFs) using TransDecoder (http://transdecoder.github.io), and the functional annotations were assigned with human readable descriptions (AHRD; https://github.com/groupschoof/AHRD), yielding a total of 445,547 genes (Table 4.2.1). For downstream gene analysis, transcript quantification using Salmon was also provided by EI.

Table 4.2.1. Summary statistics of EI *de novo* transcriptome assembly.

Overview	
Total number of genes	445547
Total number of transcripts	661522
GC (%)	40.06
All transcripts (bp)	
N50	1505
Median length	404
Average length	804.95
Total assembled bases	532489271

Using the RNA-Seq read counts, quality control was performed on the EI-generated transcriptome using hierarchical clustering to view the segregation of the replicates by sample type. The dendogram of the hierarchical clustering analysis revealed Old Leaf 3 as an outlier as it was not grouped together with other leaf samples (Fig. 4.2.1). Furthermore, although Root 2 sample was grouped together with other root samples in the dendogram, principal component analysis (PCA) revealed that it was responsible for almost 20% of the variance in component 2 (Fig. 4.2.1). Based on these observations, Old Leaf 3 and Root 2 were clear outliers compared to the rest of the replicates and were removed from further analysis (Fig. 4.2.2).

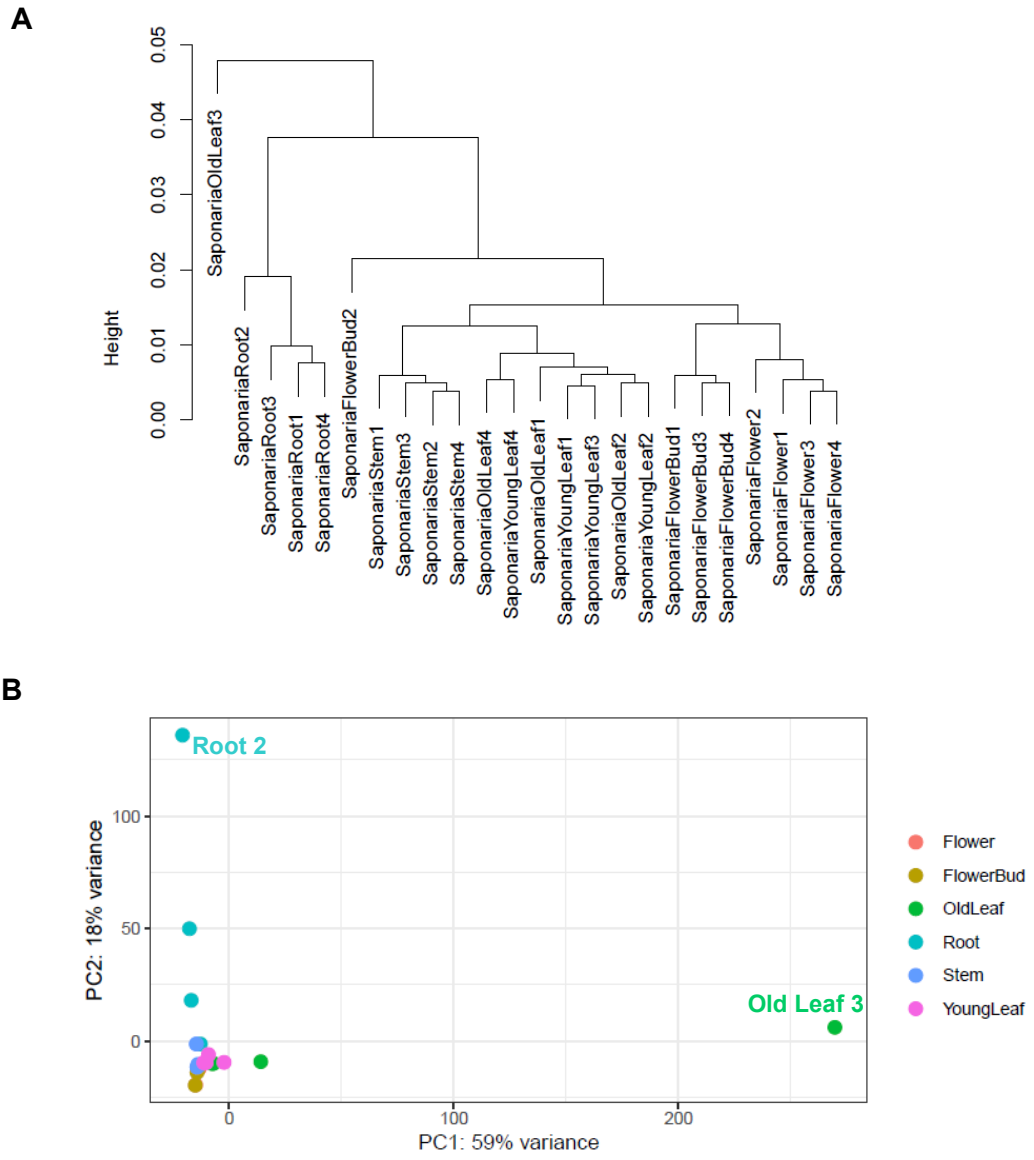


Figure 4.2.1. Hierarchical clustering of select RNA-seq samples. (A) Dendogram of hierarchical clustering and **(B)** principal component analysis (PCA) of rlog-transformed read counts generated by DEseq2. The outlier samples are labelled.

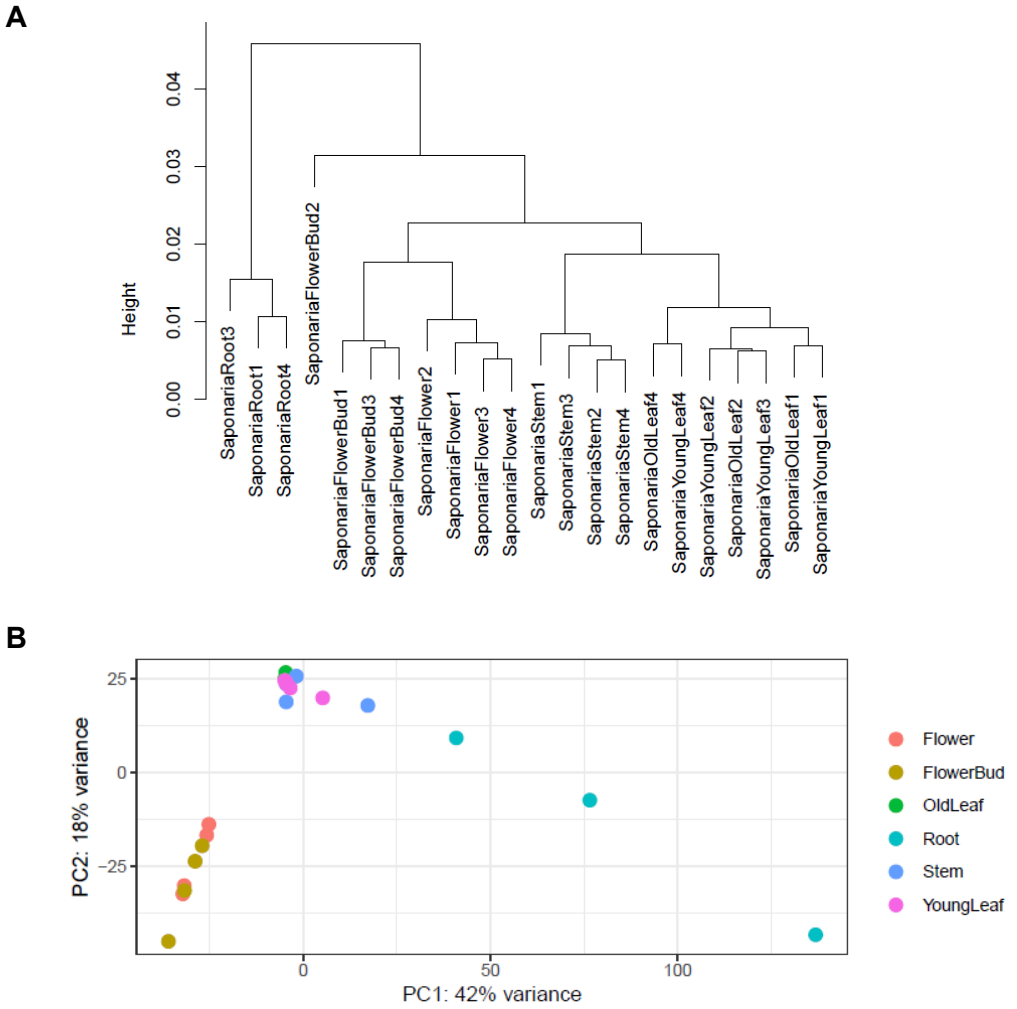


Figure 4.2.2. Hierarchical clustering of RNA-seq samples after quality control. **(A)** Dendogram of hierarchical clustering and **(B)** principal component analysis (PCA) of rlog-transformed read counts generated by DEseq2.

4.2.2 Generation and annotation of *S. officinalis* genome assembly

High levels of heterozygosity may provide challenges in genome assembly and downstream analyses (Pryszcz and Gabaldón, 2016). To determine the heterozygosity of JIC accession soapwort, genomic DNAs (gDNAs) of four JIC accession plants (JIC 1-4) were extracted and sent to the Joint Genome Institute (JGI) where short read sequencing was performed using Novaseq 6000 (Illumina). To access heterozygosity, histograms of 24-mer frequencies were analysed using GenomeScope (Vurture *et al.*, 2017) (Fig. 4.2.3). The relative heights of the first two peaks in the k-mer profile plot are directly proportional to the heterozygosity of the species (Vurture *et al.*, 2017). For a diploid species, low heterozygosity will result in a smaller first peak and a higher second peak, which is observed for all soapwort plants sequenced (Fig. 4.2.3).

Heterozygosity was also reported as 0.191%, 0.188%, 0.196%, 0.193% for JIC 1-4 by GenomeScope, respectively.

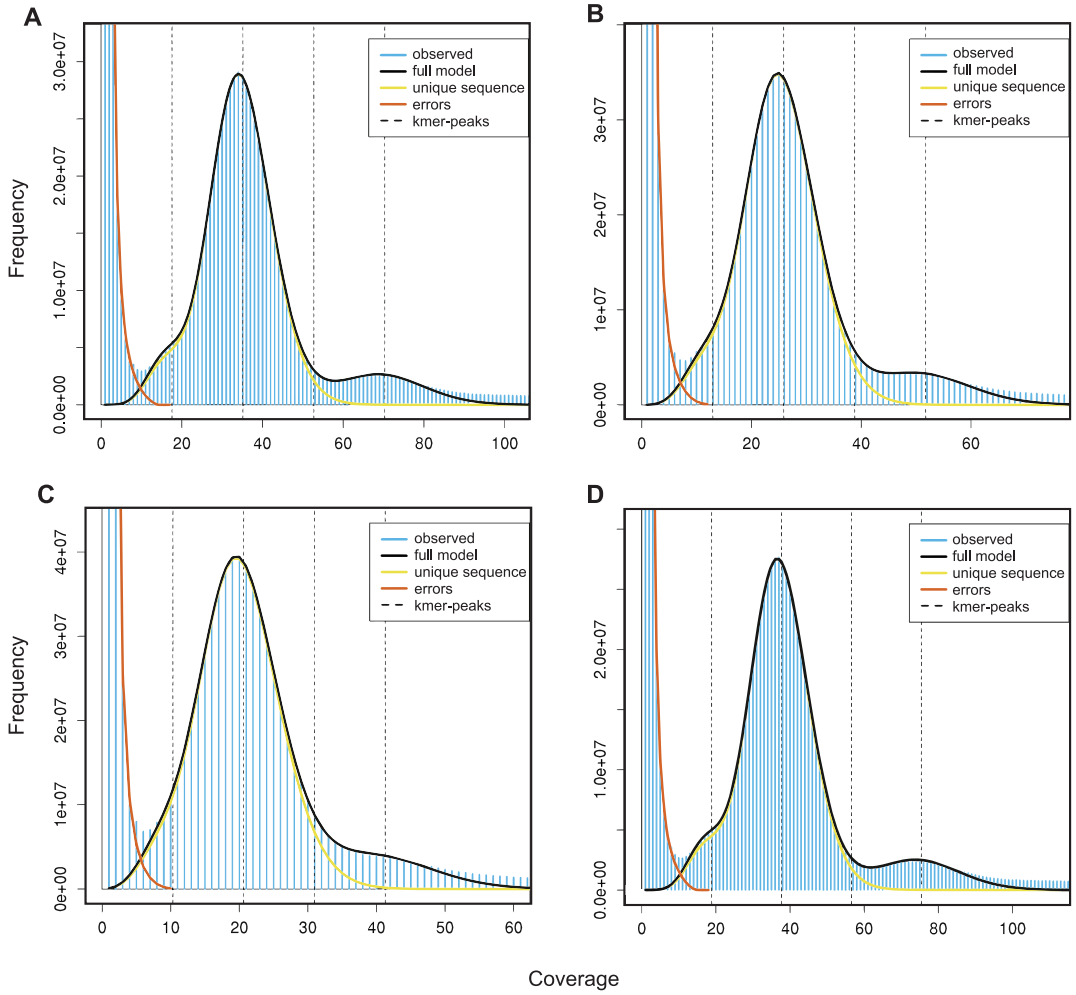


Figure 4.2.3. GenomeScope 24-mer profile plot of JIC accession soapwort plants. Four plants **(A)** 1, **(B)** 2, **(C)** 3, **(D)** 4, were sequenced to assess the heterozygosity. The extreme peak at the left is due to sequencing errors. The dotted lines represent the estimated centres of the k-mer peaks. The first two peaks show the bi-model distribution expected from a heterozygous diploid genome. The small shoulder to the left of the main peak represents the heterozygous portions of the genome, while the main peak represents the homozygous portions. The two dotted lines to the right of the main peak corresponds to duplicated heterozygous and homozygous regions.

As all four soapwort plants showed low levels of heterozygosity, high molecular weight (HMW) gDNA was extracted from leaves of JIC 2 soapwort plant as excess material was available for this individual. HMW gDNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) method (Giolai *et al.*, 2016; Hodgson *et al.*, 2019) and was sent to JGI for sequencing and assembly. PacBio Single Molecule Real-Time (SMRT) Circular Consensus Sequencing (CCS) was performed with average read length of 17.8 kb. Frozen leaf material of JIC 2 was also sent to JGI

for High-throughput Chromosome Conformation Capture (Hi-C) library construction and sequencing. The Hi-C technique exploits the relationship between genomic and physical distances, as sequences that are in proximity of the same chromosome are closer in physical space (Belton *et al.*, 2012). A Hi-C library was prepared by cross-linking chromatin, trapping sequence interactions across the chromosome and genome, followed by fragmentation, biotinylation and ligation. The library was subsequently sequenced, providing insight into chromosome architecture and allow for the chromosome scaffolding. The genome assembly was performed by JGI using HiFiasm (Cheng *et al.*, 2021) and mis-joins were identified using Hi-C data, generating a total of 129 scaffolds, 168 contigs, and genome size of 2.1 Gb (Table 4.2.2).

Table 4.2.2. Summary statistics of *S. officinalis* genome assembly.

Pseudochromosome (1n)	14
Total length	2089.5 Mb
Number of scaffolds	129
Scaffold L/N50	7/148.8 Mb
Number of contigs	168
Contig L/N50	9/91.6 Mb
BUSCO Score	95.2 %

Contigs were then ordered and oriented into chromosomes, resulting in the final assembly of 14 pseudochromosomes containing 99.46% of the assembled sequences (Fig. 4.2.4; Table 4.2.3). Both the genome size and chromosome number of *S. officinalis* reported here (2.0895 Gb, 1n = 14) are supported by the previously reported chromosome number of 1n = 14 and the estimated genome size of 2.2 Gb (Di Bucchianico *et al.*, 2008; Pustahija *et al.*, 2013). To aid in genome annotation, JGI performed Illumina RNA sequencing on pooled RNA of JIC 2 soapwort, and PacBio Iso-Seq CCS to sequence the full-length cDNA. These new sequence data, together with the EI transcriptome assembly, were used to annotate the genome. Gene models were predicted using homology-based predictors and were filtered for transcripts with C-score (BLASTP score) larger or equal to 0.5, and the selected gene models were subjected to Pfam analysis to identify protein families. Genome completeness was assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool, which determines the presence or absence of highly conserved universally single-copy genes (Seppey, Manni and Zdobnov, 2019). BUSCO scores for genomes of model

organisms are typically at least 95% complete, while genomes of non-model organisms range from 50% to 95% complete (Seppey, Manni and Zdobnov, 2019). The BUSCO score of the *S. officinalis* genome presented here was reported by JGI to be 95.2% complete. As soapwort is a non-model plant species, this BUSCO score provides confidence that the assembly and annotation of this genome is of high quality.

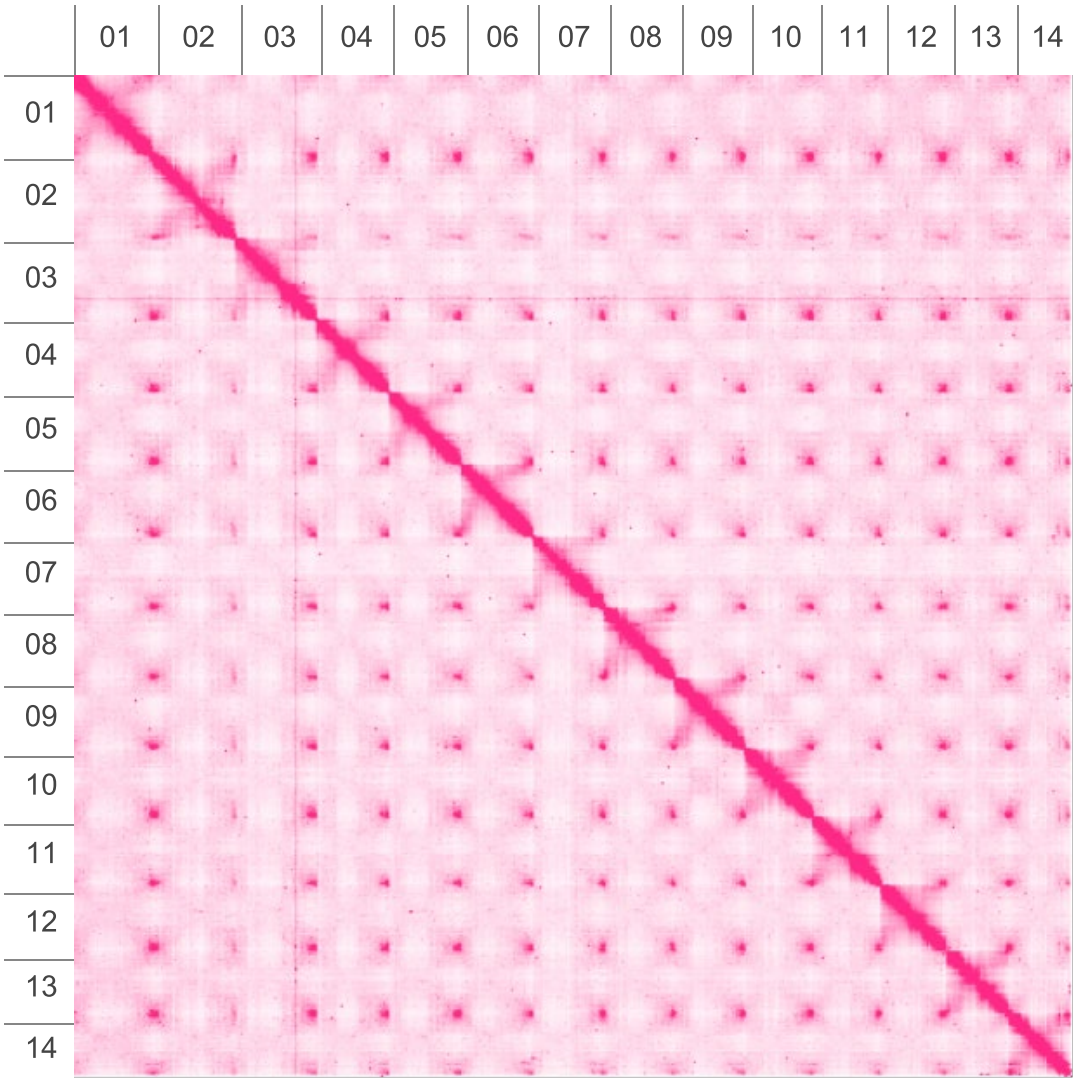


Figure 4.2.4. Hi-C post-scaffolding heatmap of *S. officinalis* genome. The interactions among *S. officinalis* chromosomes with a resolution of 4 Mb are shown. The intensity of the interactions corresponds to the colour intensity from light (weak) to dark (strong). The 14 squares represent 14 chromosome pairs of *S. officinalis*. Analysis and generation of the heatmap was performed by JGI.

Table 4.2.3. Summary statistics of the assembled pseudochromosome-level genome of *S. officinalis*.

Scaffold	Contigs	Scaffold Size (bp)	GC (%)
Chr 01	3	171,998,431	39.41
Chr 02	8	168,346,886	39.37
Chr 03	6	167,862,214	39.00
Chr 04	3	151,746,589	39.07
Chr 05	2	150,658,900	38.97
Chr 06	3	148,914,749	39.13
Chr 07	4	148,844,233	39.40
Chr 08	1	146,996,705	39.28
Chr 09	6	146,090,521	39.10
Chr 10	1	142,121,670	39.17
Chr 11	2	141,745,259	39.30
Chr 12	6	136,675,473	39.22
Chr 13	5	129,907,096	39.30
Chr 14	3	126,650,553	38.99
Remaining	115	11,376,882	

4.2.3 plantiSMASH analysis of *S. officinalis* genome

As plant genome sequencing is becoming more prevalent, more BGCs of plant specialized pathways are being discovered. plantiSMASH is an online bioinformatics tool that the Osbourn lab was involved in developing (with Dr. Marnix Medema, University of Wageningen) that enables identification of predicted BGCs within plant genomes (Kautsar *et al.*, 2017). When 47 plant genomes with chromosome assemblies were analysed using plantiSMASH, an average of 42 putative BGC gene clusters per plant were identified (Kautsar *et al.*, 2017). Of the analysed genomes, 41 genomes were from diploid plant species, and had an average of 40 putative BGC gene clusters (Kautsar *et al.*, 2017). To access putative BGCs in the newly generated *S. officinalis* genome, plantiSMASH was performed and the results were viewed using the plantiSMASH web browser (http://plantismash.secondarymetabolites.org/). A total of 15 putative BGCs were predicted by plantiSMASH, suggesting that BGCs may be relatively infrequent in the *S. officinalis* genome (Table 4.2.4).

Of the identified soapwort BGCs, those annotated as ‘saccharides’ were the dominant type and interestingly, no clusters annotated as ‘terpene’ were identified. The annotation as ‘terpene’ type BGC is based on the presence of a terpene synthase, one of the ‘signature genes’ that create the backbone of different classes of specialized

metabolites, in this case, the terpene class (Osbourn, 2010). Terpene biosynthetic genes are often found organized in gene clusters (Boutanaev *et al.*, 2015). The absence of a terpene BGC and the low number of total BGCs in the soapwort genome may suggest that the genes involved in saponarioside biosynthesis are not arranged in metabolic gene clusters, thus an alternative gene discovery approach was required.

Table. 4.2.4. Summary of plantiSMASH output of *S. officinalis* genome. Details of each cluster are available in Appendix A.3.

Cluster	Location	Type	From	To	Size (kb)	Core domains (Pfam)
1	Chr01	Saccharide	12020470	12374627	354.16	UDPGT_2, p450
2	Chr01	Saccharide	165518546	165845984	327.44	Glycos_transf_1, Methyltransf_7, Transferase
3	Chr03	Saccharide	159914605	160063858	149.25	DAHP_synth_2, UDPGT_2
4	Chr03	Saccharide	162616388	162689396	73.01	Glycos_transf_1, Transferase, adh_short
5	Chr04	Saccharide	90570156	90809975	239.82	2OG-FeII_Oxy, DIOX_N, Methyltransf_2, UDPGT_2
6	Chr05	Alkaloid	138302893	138388816	85.92	Aminotran_1_2, Methyltransf_11, Str_synth
7	Chr08	Lignan-Saccharide	113416118	114063019	646.9	2OG-FeII_Oxy, DIOX_N, Dirigent, UDPGT_2
8	Chr08	Putative	129107454	129738088	630.63	2OG-FeII_Oxy, DIOX_N, Transferase
9	Chr09	Saccharide	139658470	139817018	158.55	Aminotran_1_2, UDPGT_2
10	Chr10	Saccharide	119289280	119349263	59.98	Transferase, UDPGT_2
11	Chr11	Saccharide	128440666	128716923	276.26	DAHP_synth_2, UDPGT_2
12	Chr12	Saccharide	134521469	134613620	92.15	Glycos_transf_2, Methyltransf_11, p450
13	Chr12	Saccharide	136050093	136204946	154.85	Glycos_transf_1, p450
14	Chr13	Lignan	13774086	13982602	208.52	Dirigent, adh_short, adh_short_C2, p450
15	Chr13	Polyketide	114523798	114713386	189.59	Chal_sti_synt_C, Peptidase_S10

4.3 Conclusion

This chapter reports the first comprehensive multi-organ *S. officinalis* RNA-Seq database, as well as the first pseudochromosome-level genome assembly for this plant species. The RNA-Seq dataset generated here contains transcript-reads from multiple organs of soapwort with varying degrees of saponin content. Using this new transcriptome, differential gene expression analysis may be performed to identify subsequent saponarioside biosynthetic genes. Sequencing of the soapwort genome resolved the genome size of *S. officinalis* as 2.1 Gb, which is supported by the hypothesized genome size of 2.2 Gb based on previously reported C-values. The pseudochromosome-level genome assembly of *S. officinalis* adds to the limited genome databased of the Caryophyllaceae family. plantiSMASH analysis of the soapwort genome predicted only 15 putative BGCs and no triterpene clusters, which suggests that saponarioside biosynthetic genes are not co-localized in the genome. Alternative approaches for pathway gene discovery, such as phylogenetic and co-expression analyses, will be explored in the next chapter.

5

Elucidation of saponarioside biosynthetic pathway genes from *S. officinalis*

5.1 Introduction

5.1.1 Predicted steps of saponarioside biosynthesis

Although very little is known regarding saponarioside biosynthesis, the likely biosynthetic route can be speculated based on previous knowledge of saponin biosynthesis. Saponariosides A and B are both composed of a quillaic acid backbone, decorated with sugar chains at the C-3 and C-28 positions. Biosynthesis of the quillaic acid aglycone is likely to be the first stage of saponarioside biosynthesis, as glycosylation and other modifications of saponin scaffolds are generally believed to occur subsequently (Haralampidis, Trojanowska and Osbourn, 2002). Furthermore, as saponariosides share similar chemical structure to QS-2 (Fig. 1.4.1), I hypothesized that they may share similar pathway steps and intermediates (Reed *et al.*, 2023; Martin *et al.*, 2024). Thus, the biosynthesis of saponariosides A and B can be conceptually divided into two stages: 1. the biosynthesis of the quillaic acid aglycone, and 2. the decoration of quillaic acid (Fig. 5.1.1).

All triterpenes are biosynthesized from a common precursor, 2,3-oxidosqualene. The cyclization of 2,3-oxidosqualene by the activity of oxidosqualene cyclases (OSCs) leads to diverse triterpene skeletons, such as β-amyrin (Thimmappa *et al*., 2014). More than 90 plant OSCs have been biochemically characterized so far and although multi-functional OSCs have been observed, the majority of the characterized OSCs are mono-functional, showing specificity for a single triterpene product (Ghosh, 2016). β-Amyrin is a typical pentacyclic triterpene found widely amongst angiosperms, and many β-amyrin synthases (bASs) have been characterized, including from the

Caryophyllaceae members *Saponaria vaccaria* (Meesapyodsuk *et al.*, 2007) and *Spinacia oleracea* (Jozwiak *et al.*, 2020). Oxidations at the C-28, C-16α and C-23 positions of β-amyrin then lead to the production of quillaic acid (Fig. 5.1.1).

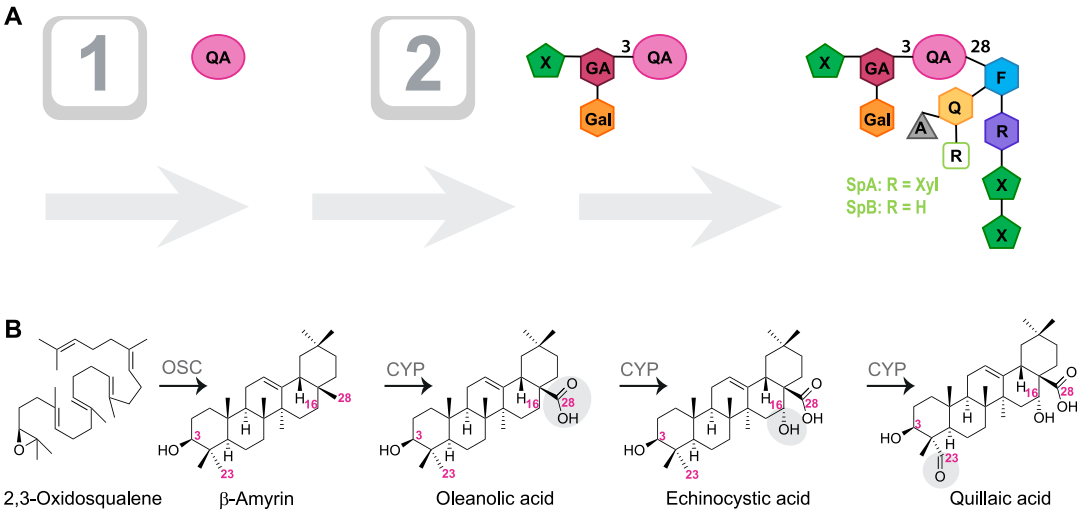


Figure 5.1.1. Schematic of predicted biosynthetic pathway for SpA and SpB. (A) Predicted two stages of saponarioside biosynthesis. QA, quillaic acid; F, D-fucose; R, L-rhamnose; X, D-xylose; Q, D-quinovose; A, acetyl moiety; GA, D-glucuronic acid; Gal, D-galactose. Numbers 3 and 28 correspond to carbon positions of the QA scaffold. **(B)** Biosynthesis of the aglycone quillaic acid from 2,3-oxidosqualene. Quillaic acid is a β-amyrin derived triterpene that is oxidized at the C-28, C-16α and C-23 positions. OSC, oxidosqualene cyclase; CYP, cytochrome P450 monooxygenase. Numbers correspond to carbon positions.

The quillaic acid core is decorated with two sugar chains, at the C-3 and C-28 positions. The C-3 sugar chain consists of β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucopyranosiduronic acid, and the C-28 sugar chain consists of β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)-β-D-4-*O*-acetylquinovopyranosyl-(1→4)]-β-D-fucopyranoside for saponarioside A, and β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-4-*O*-acetylquinovopyranosyl-(1→4)]-β-D-fucopyranoside for saponarioside B (Fig. 5.1.1). The presence of a C-3 sugar chain is a common feature shared by many triterpenoid saponins (Haralampidis, Trojanowska and Osbourn, 2002). Additionally, the majority of monodesmodic (single sugar chain) saponins have sugar chains attached to the C-3 position of the aglycone (Yu and Sun, 2009). Thus, in saponarioside biosynthesis, the attachment of the C-3 sugar chain may occur first, followed by the addition of C-28 sugar chain.

As introduced in Chapter 1, triterpene scaffolds are commonly modified by enzyme families such as cytochrome P450s, (CYPs), UDP-glycosyltransferases (UGTs) and acyltransferases (ATs) (Thimmappa *et al.*, 2014). However, other enzyme families, such as cellulose synthase superfamily-derived glycosyltransferase (CsyGT) and glycoside hydrolase family 1 transglycosidase (GH1 TG), have also been reported to be involved in the biosynthesis of complex triterpene saponins (Jozwiak *et al.*, 2020; Chung *et al.*, 2020; Reed *et al.*, 2023; Orme *et al.*, 2019). Based on the structure of saponariosides A and B, the biosynthesis of these compounds is likely to involve the combined activities of an OSC (a β-amyrin synthase) and CYPs to produce quillaic acid, and subsequential decorations by sugar transferases and an acyltransferase.

5.1.2 Strategies for biosynthetic gene discovery

Once potential enzyme classes have been assigned to the proposed hypothetical pathway, different methods can be used to facilitate the discovery of the genes encoding these biosynthetic enzymes. A standard method of gene discovery is by mining transcriptomic and/or genomic resources using homology search tools such as Basic Local Alignment Search Tool (BLAST) (Owen *et al.*, 2017). BLAST searches can be performed against a sequence database using a query enzyme known to catalyse a similar reaction, leading to a list of potential candidate genes encoding a particular enzyme family (Torrens-Spence, Fallon and Weng, 2016). This search output can be inspected manually and curated based on the annotations of each candidate gene. Further candidate genes may be discovered by identifying genes physically located in the vicinity of previously identified (i.e. based on biosynthetic gene clustering) candidates. However, based on the results from Chapter 4, saponarioside biosynthetic genes are unlikely to be organized in biosynthetic gene clusters in *S. officianalis*.

Another gene discovery approach involves co-expression analysis, which can be used to rapidly identify candidate biosynthetic genes from large-scale transcriptomic datasets. Genes involved in the same specialized metabolic pathways often display similar gene expression patterns across different plant organs, developmental stages, or environmental conditions (Tohge and Fernie, 2012). The differential expression pattern of a known gene in a specific biosynthetic pathway, typically a gene involved in early stages of the pathway, enables the gene to be used as a bait to build a network of co-expressed genes. This approach is known as directed co-expression analysis

(Aoki, Ogata and Shibata, 2007). The degree of co-expression can be statistically determined by performing correlation analysis, such as assigning Pearson correlation coefficients (PCCs) which can describe the strength and direction of the relationship between two samples. (Rao and Dixon, 2019). Co-expression analysis has been used to identify key genes and regulators for many plant metabolic pathways, such as saponin biosynthesis in barrel medic (*Medicago truncatula*) (Naoumkina *et al.*, 2010) and carotenoid metabolism in foxtail millet (*Setaria italica*) (Li *et al.*, 2022). Complete biosynthetic pathways of plant natural products have also been successfully elucidated using co-expression analysis, for example, mogroside V in monk fruit (*Siraitia grosvernorri*) (Itkin *et al.*, 2016). Furthermore, metabolomics can be coupled with RNA-seq experiments to identify genes that are differentially expressed between samples containing varying amounts of the end-product (Torrens-Spence, Fallon and Weng, 2016).

Following the identification of candidate genes, the activity of the encoded enzymes can be characterized by functional analysis typically using recombinant expression in alternative microbial or plant systems (Eljounaidi and Lichman, 2020). However, this is not sufficient to determine the *in planta* role of the candidate pathway gene, for which reverse genetic approaches such as gene silencing in the original plant are needed.

5.1.3 Aims

As saponarioside biosynthetic genes have not yet been identified from *S. officinalis*, the main aim of this chapter were to identify candidate biosynthetic saponarioside genes and perform functional characterization of the encoded enzymes using *Agrobacterium tumefaciens*-mediated transient expression in *Nicotiana benthamiana*. The specific aims were as follows:

1. Identify candidate genes using phylogenetic and co-expression analysis
2. Characterize selected candidate genes by transient expression in *N. benthamiana*

5.2 Results and discussion

5.2.1 Identifying the gene encoding the scaffold-generating enzyme

The newly generated sequence resources (Chapter 4) were mined to identify candidate genes using gene family homology. As the genome was not available until near the end of this thesis work, all candidate gene mining was performed using the EI transcriptome (generated as described in Section 2.4.1 and 2.4.3). Once the genome was completed, all candidate sequences from the transcriptome were searched against the high-quality genome by reciprocal BLASTP for sequence verification, as well as to check if any additional genome specific candidates or alternative variants had been missed in the transcriptome analysis.

The first committed step of saponarioside biosynthetic pathway was likely to be the production β-amyrin catalysed by an oxidosqualene cyclase (OSC), β-amyrin synthase (bAS) (Fig. 5.1.1B). The translated EI soapwort transcriptome was searched for candidate bAS sequences by performing BLASTP search using previously characterized OSCs reviewed in (Thimmappa *et al.*, 2014) from other plant species as search queries, as well as the recently characterized bAS (QsbAS) from *Q. saponaria* (Reed *et al.*, 2023) (Table C.1.1). The resulting list was refined by removing any candidates less than 500 amino acids (aa) in length, as lengths of the query OSCs ranged from 700-800 aa. This list was then manually curated by their annotation (AHRD and InterPro assignments), leading to only three candidate OSCs. These remaining candidates were used to produce a phylogenetic tree with published OSCs from other plant species (Table C.1.1; Fig. 5.2.1). Of the three candidate OSCs, only one (*TRINITY_DN1084_c0_g4*) grouped together with OSCs reported to produce β-amyrin, while the other two (*TRINITY_DN5932_c0_g1* and *TRINITY_DN27404_c0_g1*) grouped with OSCs known to produce cycloartenol and lupeol, respectively (Fig. 5.2.1). The soapwort bAS candidate also showed high amino acid similarity (93.8%) with a bAS sequence, previously characterized from the closely related species, *Saponaria vaccaria* (Meesapyodsuk *et al.*, 2007).

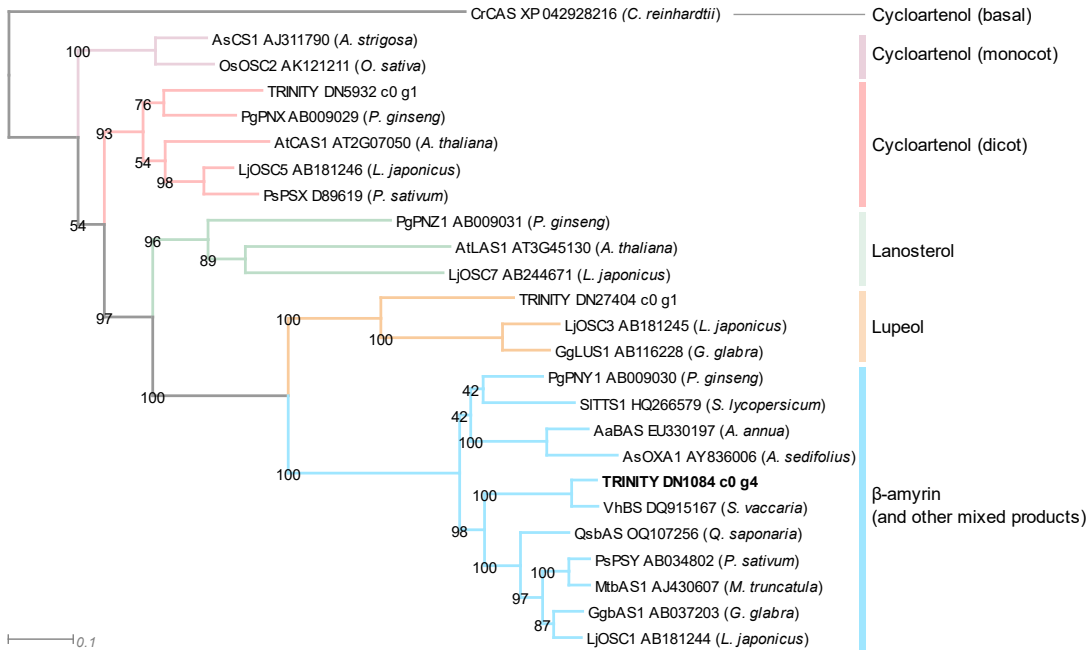


Figure 5.2.1. Phylogenetic analysis of candidate *S. officinalis* OSCs. The maximum-likelihood tree was constructed using an amino acid alignment of putative OSCs from *S. officinalis* (primary candidate in bold) and previously characterized OSCs from other plant species (listed in Table C.1.1). Bootstrap values are shown beside each node. The scale bar indicates the number of amino acid substitutions per amino acid site. Common enzyme products produced by each clade are labelled on the right. A cycloartenol synthase from *Chlamydomonas reinhardtii* (CrCAS) was used as a basal outgroup. *P. ginseng, Panax ginseng; L. japonicus, Lotus japonicus, A. thaliana, Arabidopsis thaliana; O. sativa, Oryza sativa; A. strigosa, Avena strigosa; P. sativum, Pisum sativum; G. glabra, Glycyrrhiza glabra; A. sedifolius, Aster sedifolius; A. annua, Artemisia annua; S. lycopersicum, Solanum lycopersicum; S. vaccaria; Saponaria vaccaria; Q. saponaria, Quillaja saponaria; M. truncatula; Medicago truncatula*.

The candidate soapwort bAS was renamed as SobAS1 and its functional activity was tested by *A. tumefaciens* mediated transient expression in *N. benthamiana*. In all transient expression experiments, the truncated *HMG-CoA reductase* from *A. strigosa* (*AstHMGR*) was co-expressed with the candidate gene to increase the metabolite flux towards the MVA pathway (Reed and Osbourn, 2018). The infiltrated leaves were harvested after 5 days post-infiltration, and leaf extracts were derivatized with trimethylsilyl (TMS) prior to GC-MS analysis. Transient expression of *SobAS1* in *N. benthamiana* led to the formation of a peak (**1**) with mass-to-charge (*m/z*) of 498.4, which corresponds to the mass of derivatized β-amyrin and was consistent with the derivatized commercial β-amyrin standard (**1**) in both retention time (RT) and mass spectra (MS) (Fig. 5.2.2). This new peak (**1**) was not present in extracts from leaves expressing *AstHMGR* only, which served as a negative control (Fig. 5.2.2A). Thus,

based on these results, SobAS1 was identified as an OSC capable of cyclizing oxidosqualene into β-amyrin.

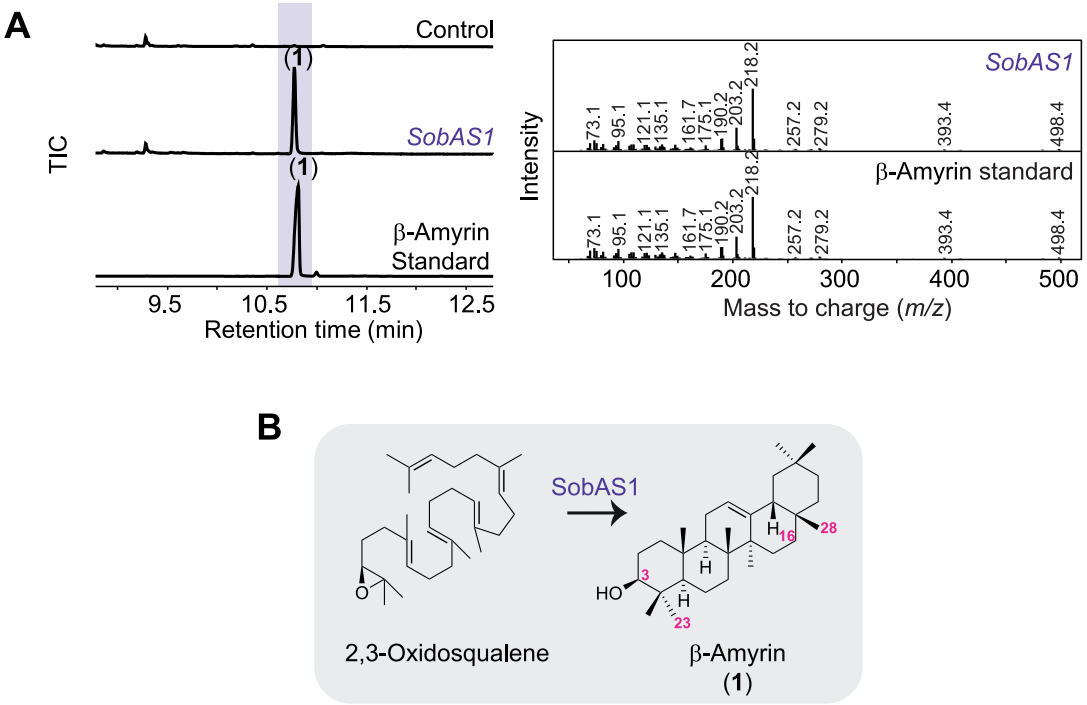


Figure 5.2.2. Transient expression of *SobAS1* in *N. benthamiana* leaves. (A) GC-MS total ion chromatograms (TICs) of leaf extracts co-expressing *AstHMGR* and *SobAS1*, along with a control (leaf only expressing *AstHMGR*) and a commercial standard of β-amyrin (**1**) are shown. The full TIC range is available in Fig. C.2.1. Mass spectra for leaf extracts expressing *SobAS1* and commercial β-amyrin standard are also given. **(B)** Activity of SobAS1 in converting 2,3-oxidosqualene to β-amyrin (**1**).

5.2.2 Identification of genes encoding β-amyrin modifying enzymes

Following the biosynthesis of β-amyrin, the next predicted step in saponarioside biosynthetic pathway was likely to be the oxidation of β-amyrin to quillaic acid by the activity of cytochrome P450s (CYPs) (Fig. 5.1.1). To create a list of candidate CYPs implicated in these modifications, BLASTP was performed against the translated EI transcriptome using 81 previously characterized triterpene oxygenating CYPs from the TriForC database (http://bioinformatics.psb.ugent.be/triforc/) (Miettinen *et al.*, 2017) as queries. Although SobAS1 could be readily discovered by phylogenetic analysis due to the small number of predicted OSC candidates in the soapwort transcriptome and genome assemblies, over 250 candidate CYPs were identified in soapwort. A different approach was therefore required to filter this large number of candidate genes and prioritize them for functional analysis. One such alternative was

to exploit the potential sequence similarity with characterized genes from the QS-21 biosynthetic pathway of *Q. saponaria* (Reed *et al.*, 2023). As QS-21 and saponariosides A and B are structurally similar, both having a quillaic acid core, the genes involved in quillaic acid biosynthesis might be similar in sequence. Indeed, SobAS1 and QsbAS1 share a high protein sequence identity of 79.7%. To search for related genes of *Q. saponaria* CYPs in soapwort, a characterized CYP from *Q. saponaria* able to oxidize the C-28 position of β-amyrin (QsC28), was used as a BLASTP search query against the translated soapwort transcriptome. The two soapwort C-28 CYP candidates that showed highest sequence identity with QsC28 were TRINITY_DN651_c0_g3 and TRINITY_DN13626_c1_g2, which were renamed as SoC28-1 and SoC28-2, respectively. SoC28-1 and SoC28-2 respectively shared 74.8% and 54.2% protein sequence identity with QsC28. Additionally, the amino acid sequences of SoC28-1 and SoC28-2 were searched against the translated soapwort genome for any potential genome specific alternative variants, which were not found.

The functional activities of *SoC28-1* and *SoC28-2* were tested by performing *A. tumefaciens*-mediated transient expression in *N. benthamiana*. Leaves were harvested and extracted, followed by derivatization with TMS prior to GC-MS analysis. The co-expression of *AstHMGR* and *SobAS1* with *SoC28-1* in *N. benthamiana* led to the formation of a new peak (**2**) with *m/z* 585.5, corresponding to the mass of derivatized oleanolic acid (Fig. 5.2.3). This peak also had the same RT and mass spectra as oleanolic acid standard, and thus peak (**2**) was identified as oleanolic acid. Interestingly, co-infiltration of *AstHMGR* and *SobAS1* with *SoC28-2* produced a different peak (**3**) with *m/z* 570.4, corresponding to the expected mass of derivatized echinocystic acid. Peak (**3**) produced by heterologous gene expression had the same RT, *m/z*, and mass spectra as the echinocystic acid standard, and was identified as such (Fig. 5.2.3). Based on these results, SoC28-1 may be a CYP with C-28 oxidation activity, leading to the formation of oleanolic acid from β-amyrin, while SoC28-2 may be a CYP with both C-28 and C-16α oxidation activity, leading to the production of echinocystic acid from oleanolic acid. Based on its dual activity, SoC28-2 was renamed SoC28C16. Interestingly, although SoC28 was a much more efficient C-28 oxidase in converting β-amyrin to oleanolic acid compared to SoC28C16, transient co-expression of both SoC28 and SoC28C16 together in *N. benthamiana* did not lead

to the increased accumulation of echinocystic acid (Fig. C.2.2.). Thus, to reduce the number of co-expressed genes, further experiments were performed without SoC28, as the additional C-28 oxidase activity of SoC28 did not translate into increased end-product (echinocystic acid in this case). This might be due to a potential feed-back mechanism in *N. benthamiana* regulating the level of echinocystic acid accumulation.

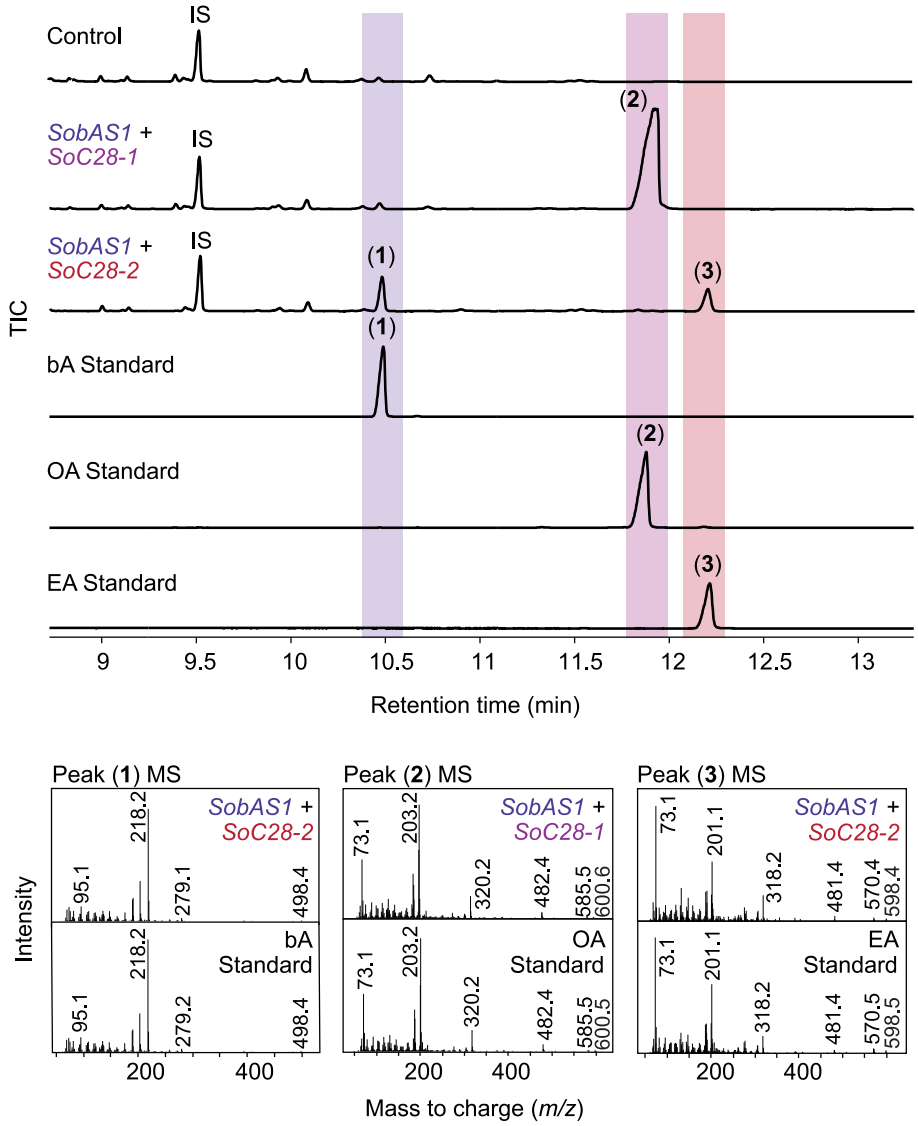


Figure 5.2.3. Transient expression of *SoC28-1* and *SoC28-2* in *N. benthamiana*. GC-MS total ion chromatograms (TIC) of leaf extracts co-expressing *SobAS1* with either *SoC28-1* or *SoC28-2* are shown, along with a control (leaf expressing only *AstHMGR*) and the following commercial standards: bA (**1**, β-amyrin), OA (**2**, oleanolic acid) and EA (**3**, echinocystic acid). Mass spectra (MS) of bA (**1**), OA (**2**) and EA (**3**) for leaf extracts expressing *SobAS1* with either *SoC28-1* or *SoC28-2* and for relevant commercial standards are also shown. SoC28-1 is synonymous with SoC28 and SoC28-2 is synonymous with SoC28C16 based on its observed dual activity as both C-28 and C-16α oxidase. The full TIC range is available in Fig. C.2.2.

To assess if using QS-21 biosynthetic genes from *Q. saponaria* as search queries was a valid method for finding downstream saponarioside biosynthetic genes, the amino acid sequence identity of *Q. saponaria* and *S. officinalis* genes encoding enzymes with shared functions were compared (Table 5.2.1). Although SoC28C16 was identified using sequence similarity with QsC28, its sequence similarity with QsC16, a CYP from *Q. saponaria* with C-16α oxidation activity, was also compared to determine if it shared higher sequence identity with QsC16 rather than QsC28. While SobAS1 and SoC28 showed high sequence identity with their *Q. saponaria* counterparts, SoC28C16 shared only about 50% sequence identity with both QsC28 and QsC16 (Table 5.2.1). Thus, based on this observation and the phylogenetic distance between *S. officinalis* and *Q. saponaria*, using QS-21 biosynthetic genes as search queries might be an unsuitable strategy to find downstream saponarioside biosynthetic genes. Another method to search and filter for candidate biosynthetic genes is to perform co-expression analysis with an already known gene involved early in the biosynthetic pathway. SobAS1 catalyses the first committed step of saponarioside biosynthetic pathway – the cyclization of 2,3-oxidosqualene into β-amyrin – and so is an ideal bait for co-expression analysis to discover downstream saponarioside biosynthetic genes.

Table 5.2.1. Shared amino acid sequence identity of *S. officinalis* and *Q. saponaria* enzymes with shared functional activities. *Q. saponaria* enzymes are described in (Reed *et al.*, 2023).

S. officinalis	***Q. saponaria***	**AA Identity**
SobAS1	QsbAS	79.7%
SoC28	QsC28	74.8%
SoC28C16	QsC28	48.6%
SoC28C16	QsC16	54.2%

5.2.3 Selection of candidate genes using co-expression and differential expression analysis

In addition to gene family homology searches, candidate saponarioside biosynthetic genes were further identified by using co-expression analysis. The RNA-seq dataset generated in Chapter 4 was used for co-expression analysis to identify genes with similar expression profiles to *SobAS1*, the gene encoding the enzyme implicated in the first committed step of saponarioside biosynthesis. Gene expression patterns were correlated using the Pearson correlation coefficient (PCC), and only those with positive PCC values were selected for further analysis (Appendix D). Candidate genes were filtered by arbitrary PCC cut-off values based on the number of candidate genes found in the gene family.

Cytochrome P450s

The combined activities of *SobAS1*, *SoC28*, and *SoC28C16* transiently expressed in *N. benthamiana* lead to the production of echinocystic acid, which needs to be oxidized at the C-23 position to form quillaic acid (Fig. 5.1.1). To identity candidate CYP450s with potential C-23 oxidation activity, a total of 254 CYP sequences (identified in Section 5.2.2) were evaluated for co-expression with *SobAS1*. Of those, expression of 109 candidates positively correlated with *SobAS1* expression pattern across the different soapwort organs. To further reduce the number of candidates, only those that were highly co-expressed with *SobAS1* with PCC values greater than 0.9 were chosen for further analysis, resulting in a total of 15 candidates (Table 5.2.2). The soapwort biosynthetic genes identified above, *SoC28* (PCC = 0.969) and *SoC28C16* (PCC = 0.946), were highly co-expressed with *SobAS1*, suggesting that co-expression analysis was indeed a reliable approach to discover downstream candidate saponarioside biosynthetic genes (Table 5.2.2).

The expression patterns of the newly identified CYP candidates, together with *SobAS1*, *SoC28* and *SoC28C16,* were examined across the different soapwort organs. *SobAS1*, *SoC28* and *SoC28C16* showed good expression in all organs, with the highest expression in the flower, and the lowest expression in the root or stem (Table 5.2.2). As such, the new CYP candidates were refined based on their gene expression across different soapwort organs, and those with overall low expression were discarded. For example, although *TRINITY_DN53369_c0_g1* (PCC = 0.946) showed strong co-

expression with *SobAS1*, it was deemed as a poor candidate as the overall gene expression levels were low in all soapwort organs (Table 5.2.2). Based on the gene expression patterns, seven candidates (*TRINITY_DN645_c1_g2, TRINITY_DN5729_c1_g1, TRINITY_DN2993_c0_g1, TRINITY_DN58802_c0_g3, TRINITY_DN5664_c0_g3, TRINITY_DN8790_c0_g3, TRINITY_DN5664_c0_g1*) were renamed *CYP1-7*, respectively, and selected for functional characterisation.

Table 5.2.2. List of candidate CYPs co-expressed with *SobAS1*. Only candidates with PCC values to *SobAS1* greater than 0.9 are shown. The absolute transcript read count for each candidate across different soapwort organs are shown with colours matching to the relative levels of transcript abundance (high = magenta; middle = white; low = green). Simplified gene names of candidates tested for functional activity are also given. PCC, Pearson correlation coefficient.

			Transcript Read Count					
Name	**TRINITY ID**	**PCC**	**Flower**	**Flower Bud**	**Young Leaf**	**Old Leaf**	**Stem**	**Root**
SobAS1	TRINITY_DN1084_c0_g4	1.000	14915	11822	3604	1417	1404	1251
CYP1	TRINITY_DN645_c1_g2	0.989	16128	13470	6232	4713	5411	5629
SoC28	TRINITY_DN651_c0_g3	0.969	27880	22133	7975	3840	1721	1581
CYP2	TRINITY_DN5729_c1_g1	0.967	151	57	16	9	10	9
CYP3	TRINITY_DN2993_c0_g1	0.954	2626	2852	285	161	130	92
SoC28C16	TRINITY_DN13626_c1_g2	0.946	27861	11592	5406	1555	391	288
	TRINITY_DN53369_c0_g1	0.946	4	4	1	0	1	0
CYP4	TRINITY_DN58802_c0_g3	0.931	1058	409	2	0	1	1
CYP5	TRINITY_DN5664_c0_g3	0.921	5800	3443	583	113	156	8
	TRINITY_DN283414_c0_g1	0.916	5915	335	4	0	11	0
CYP6	TRINITY_DN8790_c0_g3	0.909	1379	1577	241	249	100	104
	TRINITY_DN47434_c0_g2	0.908	1	1	0	0	0	0
	TRINITY_DN111518_c0_g1	0.907	1	1	0	0	0	0
	TRINITY_DN349059_c0_g1	0.907	1	2	0	0	0	0
	TRINITY_DN234703_c0_g1	0.904	1	2	0	0	0	0
CYP7	TRINITY_DN5664_c0_g1	0.903	1019	556	209	241	195	207
	TRINITY_DN171936_c0_g1	0.903	1	2	0	0	0	0
	TRINITY_DN26666_c0_g1	0.902	4	1	0	0	0	0

Cellulose-synthase like glycosyltransferases

The following step after the biosynthesis of the quillaic acid aglycone is likely to be the attachment of the C-3 sugar chain (Fig. 5.1.1). The sugar that is directly attached to the C-3 position of quillaic acid is D-glucuronic acid. Uridine diphosphate-dependent (UDP) sugar transferases belonging to glycosyltransferase family 1 (GT1) are typically responsible for glycosylation of plant natural products (Louveau and Osbourn, 2019). However, several enzymes from a subgroup of cellulose synthase-

like (CSL) family, CSyGT (cellulose-synthase superfamily-derived glycosyltransferase), have been recently reported to be involved in the 3-*O*-glucuronidation of triterpene aglycones. These include SOAP5 from *Spinach oleracea* (Jozwiak *et al.*, 2020), GmCSyGT1 from *Glycine max* and its homologues in *Glycyrrhiza uralensis* (GuCSyGT) and *Lotus japonicus* (LjCSyGT) (Chung *et al.*, 2020), and QsCSL identified from *Quillaja saponaria* (Reed *et al.*, 2023). Thus, the translated soapwort transcriptome was searched for candidate CSLs by performing BLASTP search using 30 previously identified CSL sequences retrieved from published work (Carroll and Specht, 2011), in addition to the characterized CSLs mentioned above, as queries (Table C.1.2). A total of 232 candidates with amino acid length greater than 400 aa was identified, as lengths of CSL sequences ranged from 600-1000 aa (Carroll and Specht, 2011). Of those, the expression patterns of 66 candidates showed positive correlations with the *SobAS1* expression pattern. The list of candidate CSLs were further refined by removing those with PCC values less than 0.85, resulting in seven candidates (Table 5.2.3). CSyGTs are categorised as a subfamily of cellulose synthase-like (CSL) family M (Chung *et al.*, 2020). Phylogenetic analysis was performed by generating a phylogenetic tree from these candidate soapwort CSLs together with other CSLs and cellulose synthases (CesAs) (Table C.1.2). Only a single candidate (*TRINITY_DN23622_c0_g2*) grouped within the CSyGT subgroup and was renamed as *SoCSL1* (Fig. 5.2.4).

Table 5.2.3. List of CSL candidates co-expressed with *SobAS1*. Candidates with PCC values greater than 0.85 are shown. The absolute transcript read count for each candidate across different plant organs are shown with colours matching to the relative levels of transcript abundance (high = magenta; middle = white; low = green). SoCSL1 was selected for functional analysis. PCC, Pearson correlation coefficient.

			Transcript Read Count					
Name	**TRINITY ID**	**PCC**	**Flower**	**Flower Bud**	**Young Leaf**	**Old Leaf**	**Stem**	**Root**
	TRINITY_DN345366_c0_g1	0.969	499	184	38	10	32	20
SoCSL1	TRINITY_DN23622_c0_g2	0.915	13360	9473	7290	3666	2497	1704
	TRINITY_DN46549_c0_g1	0.900	82	19	1	1	1	1
	TRINITY_DN11658_c0_g2	0.894	485	578	0	0	2	0
	TRINITY_DN57970_c0_g1	0.879	204	21	2	2	2	1
	TRINITY_DN86505_c0_g1	0.855	80	17	1	1	0	2
	TRINITY_DN19883_c0_g5	0.852	826	847	48	81	123	33

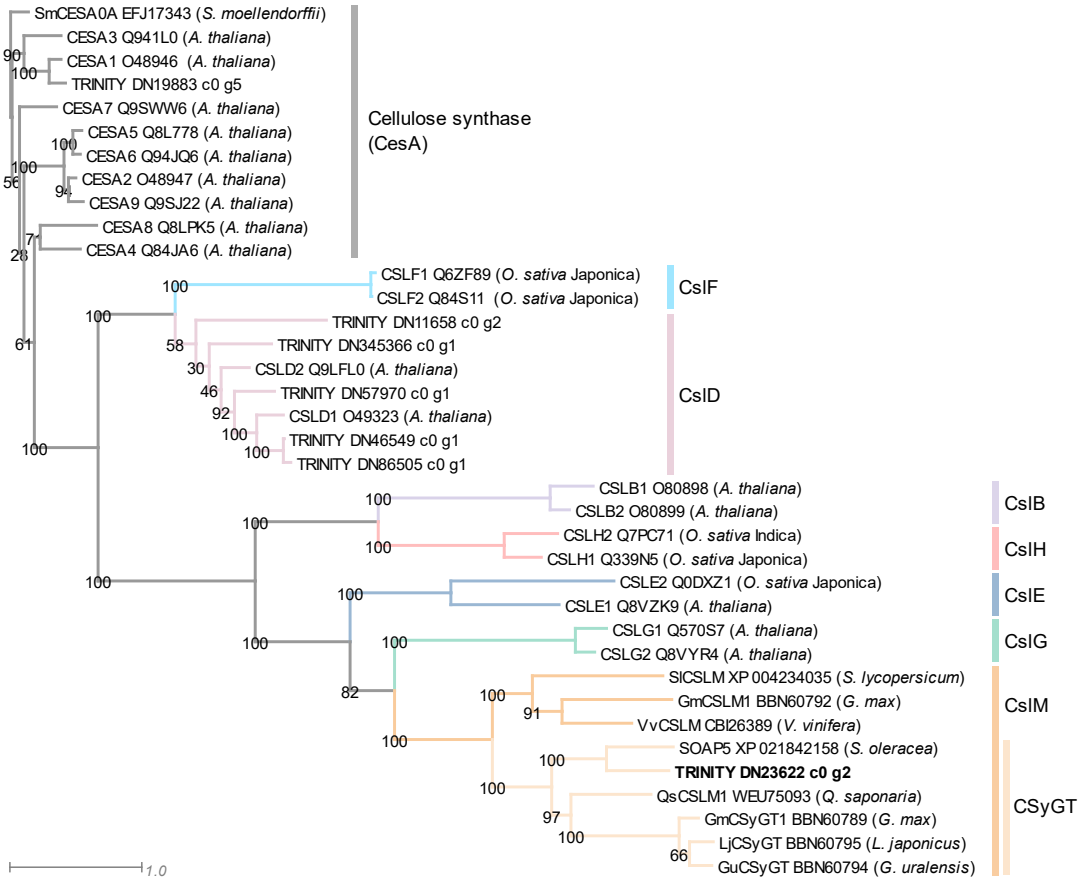


Figure 5.2.4. Phylogenetic analysis of candidate *S. officinalis* CSLs. The maximum-likelihood tree was constructed using an amino acid alignment of putative CSLs from *S. officinalis* (prioritized candidate in bold) and previously identified CesA and CSLs from other plant species (Table C.1.2). Bootstrap values are shown beside each node. The scale bar indicates the number of amino acid substitutions per amino acid site. A cellulose synthase from *Selaginella moellendorffii* (SmCESA0A) was used as a basal outgroup. CesA and different CSL subfamilies are colour coded and labelled. *O. sativa* Japonica*, Oryza sativa* Japonica*; A. thaliana, Arabidopsis thaliana; S. lycopersicum, Solanum lycopersicum; V. vinifera, Vitis vinifera; G. max; Glycine max; S. oleracea, Spinach oleracea; Q. saponaria, Quillaja saponaria; G. uralensis, Glycyrrhiza uralensis; L. japonicus, Lotus japonicus.*

UDP-glycosyltransferases

To create a list of candidate soapwort UGTs, a BLASTP search was performed against the soapwort transcriptome using 86 characterized UGTs retrieved from (Louveau and Osbourn, 2019). The list was refined by removing candidates less than 300 aa in length, as lengths of the literature sequences ranged from 400-500 aa, resulting in a total of 938 UGT candidates. Of those, 279 candidates had expression patterns positively correlating with *SobAS1*. To refine the list of candidates UGTs, candidates with PCC values (to *SobAS1*) greater than 0.9 were selected, resulting in 17 candidates (Table 5.2.4). Additionally, the gene expression patterns of the chosen 17 UGT candidates were investigated, and four candidates were discarded as they showed low expression profile across all soapwort organs (Table 5.2.4). The remaining 13 candidates were renamed *UGT1-13*.

Table 5.2.4. List of candidate UGTs co-expressed with *SobAS1*. Only candidates with PCC values to *SobAS1* greater than 0.9 are shown. The absolute transcript read count for each candidate across different plant organs are shown with colours matching to relative levels of transcript abundance (high = magenta; middle = white; low = green). Simplified gene names of candidates tested for functional activity are also given. PCC, Pearson correlation coefficient.

			Transcript Read Count					
Name	**TRINITY ID**	**PCC**	**Flower**	**Flower Bud**	**Young Leaf**	**Old Leaf**	**Stem**	**Root**
UGT1	TRINITY_DN1618_c1_g2	0.986	1819	2244	806	482	487	551
UGT2	TRINITY_DN28657_c0_g1	0.981	11188	2701	1083	180	209	196
UGT3	TRINITY_DN5570_c0_g3	0.979	759	892	290	124	90	74
UGT4	TRINITY_DN5701_c1_g1	0.975	9677	9989	3483	2222	1605	1452
UGT5	TRINITY_DN54808_c0_g7	0.972	8186	4986	1433	582	466	946
UGT6	TRINITY_DN5570_c0_g1	0.961	1043	1199	332	168	76	69
UGT7	TRINITY_DN51550_c0_g1	0.960	8596	7779	1208	551	446	173
UGT8	TRINITY_DN347728_c0_g1	0.956	84	51	2	0	1	0
	TRINITY_DN2822_c1_g3	0.954	12	8	1	0	0	0
UGT9	TRINITY_DN41181_c0_g1	0.954	430	204	227	86	50	53
UGT10	TRINITY_DN342_c0_g1	0.953	26199	23337	11627	2583	2530	1515
UGT11	TRINITY_DN5422_c7_g1	0.950	606	74	4	0	0	1
UGT12	TRINITY_DN14107_c4_g1	0.938	20278	10234	4593	2612	996	1891
	TRINITY_DN31287_c0_g2	0.914	27	6	1	0	0	0
UGT13	TRINITY_DN586_c1_g1	0.909	29670	13539	9937	5714	4254	2154
	TRINITY_DN47337_c0_g1	0.907	1	1	0	0	0	0
	TRINITY_DN4499_c3_g1	0.903	54	108	3	0	3	5

Acyltransferases

Both BAHD and SCPL families of ATs were considered as potential gene families involved in the acylation step of saponarioside biosynthesis. The translated soapwort transcriptome was searched for candidate ATs by performing separate BLASTP searches using 16 characterized BAHD ATs retrieved from (Bontpart *et al.*, 2015) (Table C.1.3) and 25 characterized SCPL ATs retrieved from (Bontpart *et al.*, 2018) (Table C.1.4) from other plant species as queries. The resulting candidate lists were refined by removing any sequences less than 300 aa in length as lengths of literature ATs ranged from 400-500 aa. This refined the candidate lists to 331 BAHD AT and 99 SCPL AT candidates. Co-expression analysis revealed that 121 BAHD AT and 26 SCPL AT gene candidates showed positive correlation with the expression pattern of *SobAS1*. To reduce the number of BAHD candidates, candidates with PCC values less than 0.8 were discarded, resulting in 21 candidates (Table 5.2.5). The overall expression profiles of these candidates were evaluated, and 11 candidates with overall low transcript levels were discarded (Table 5.2.5). The remaining 10 candidates were renamed *BAHD1-10*.

To refine the list of candidate SCPL genes, phylogenetic analysis was performed to identify Clade I SCPLs, as members of this clade are reported to be involved in plant specialized metabolism (Fraser, Rider and Chapple, 2005). Of the 26 SCPL candidates, only 7 grouped with known SCPL Clade I members (Fig. 5.2.5). Based on the overall expression patterns of each candidate, a total of 5 candidates were chosen to be examined for their biochemical functions. The 5 SCPL candidates were renamed *SCPL 1-5* (Table 5.2.6).

Table 5.2.5. List of candidate BAHD ATs co-expressed with *SobAS1*. Only candidates with PCC values to *SobAS1* greater than 0.8 are shown. The absolute transcript read count for each candidate across different plant organs are shown with colours matching to relative levels of transcript abundance (high = magenta; middle = white; low = green). Simplified gene names of candidates tested for functional activity are also given. PCC, Pearson correlation coefficient.

			Transcript Read Count					
Name	**TRINITY ID**	**PCC**	**Flower**	**Flower Bud**	**Young Leaf**	**Old Leaf**	**Stem**	**Root**
BAHD1	TRINITY_DN1473_c3_g1	0.980	5121	3579	733	409	248	414
	TRINITY_DN69958_c0_g1	0.946	1	1	0	0	0	0
BAHD2	TRINITY_DN3011_c0_g3	0.941	1950	763	628	69	97	32
BAHD3	TRINITY_DN3341_c0_g1	0.921	1663	1855	1091	273	138	312
BAHD4	TRINITY_DN221488_c0_g1	0.916	94	81	45	17	46	18
BAHD5	TRINITY_DN10898_c0_g1	0.914	894	1412	201	197	109	91
	TRINITY_DN40880_c0_g3	0.908	13	11	1	0	0	1
BAHD6	TRINITY_DN5184_c0_g1	0.889	4654	1814	349	318	54	46
BAHD7	TRINITY_DN5384_c0_g3	0.886	2404	439	7	9	14	8
	TRINITY_DN29178_c1_g1	0.882	126	47	2	0	0	6
	TRINITY_DN86763_c0_g1	0.880	2	2	0	0	0	0
	TRINITY_DN41684_c0_g4	0.880	143	216	71	52	8	8
BAHD8	TRINITY_DN3341_c0_g4	0.863	597	230	11	1	3	36
	TRINITY_DN320_c1_g2	0.861	461	576	0	0	7	0
	TRINITY_DN9977_c0_g1	0.841	18	8	0	0	0	2
BAHD9	TRINITY_DN1707_c1_g2	0.835	3186	2719	137	110	405	3
BAHD10	TRINITY_DN316011_c0_g1	0.834	878	473	1943	10	57	1
	TRINITY_DN370449_c0_g1	0.825	1	1	0	0	0	0
	TRINITY_DN70669_c0_g2	0.822	1	1	0	0	0	0
	TRINITY_DN18871_c1_g1	0.820	139	37	46	18	14	3
	TRINITY_DN31985_c0_g1	0.812	12	7	0	0	2	0

Table 5.2.6. List of candidate SCPL ATs co-expressed with *SobAS1*. Only candidates with positive PCC values to *SobAS1* are shown. The absolute transcript read count for each candidate across different plant organs are shown with colours matching to relative levels of transcript abundance (high = magenta; middle = white; low = green). Simplified gene names of candidates tested for functional activity are also given. PCC, Pearson correlation coefficient.

			Transcript Read Count					
Name	**TRINITY ID**	**PCC**	**Flower**	**Flower Bud**	**Young Leaf**	**Old Leaf**	**Stem**	**Root**
	TRINITY_DN252840_c0_g1	0.930	400	877	73	43	11	14
SCPL1	TRINITY_DN2803_c0_g1	0.927	487	1242	23	10	32	4
	TRINITY_DN22101_c1_g1	0.838	1795	917	635	562	500	771
	TRINITY_DN23990_c0_g1	0.831	11	59	4	3	1	0
	TRINITY_DN21494_c1_g1	0.741	22	3	0	0	2	1
	TRINITY_DN198768_c0_g1	0.737	1	9	0	0	0	0
	TRINITY_DN35732_c0_g1	0.737	2	41	0	0	0	0
	TRINITY_DN2803_c1_g1	0.724	259	9	1	0	9	1
	TRINITY_DN17173_c0_g1	0.666	1480	796	748	800	548	253
SCPL2	TRINITY_DN4822_c0_g2	0.637	6174	3099	3031	3577	1589	1854
	TRINITY_DN970_c0_g1	0.631	1910	683	1487	853	83	115
	TRINITY_DN2104_c1_g1	0.615	107	36	18	10	90	0
SCPL3	TRINITY_DN12262_c0_g1	0.603	501	291	520	422	17	6
	TRINITY_DN90412_c0_g1	0.568	22	46	26	23	24	37
	TRINITY_DN5879_c0_g1	0.495	4672	5232	2306	2540	3564	5209
SCPL4	TRINITY_DN5672_c0_g1	0.482	1317	1104	926	1067	1232	806
	TRINITY_DN7155_c0_g1	0.460	55	116	38	41	73	68
SCPL5	TRINITY_DN105438_c0_g1	0.445	1625	1173	1499	1273	615	1307
	TRINITY_DN2104_c0_g2	0.410	1155	2425	1938	4082	531	31
	TRINITY_DN45061_c0_g1	0.393	1101	1920	604	1370	1326	376
	TRINITY_DN1129_c0_g1	0.383	2233	2987	3453	2594	1437	1679
	TRINITY_DN970_c0_g2	0.157	589	187	272	192	233	639
	TRINITY_DN761_c5_g1	0.120	2130	2976	1075	1402	2502	3264
	TRINITY_DN227947_c0_g1	0.107	0	0	0	0	0	0
	TRINITY_DN7756_c0_g1	0.100	672	1605	1271	1336	978	586
	TRINITY_DN25360_c1_g1	0.035	1809	2705	4879	5110	4843	486

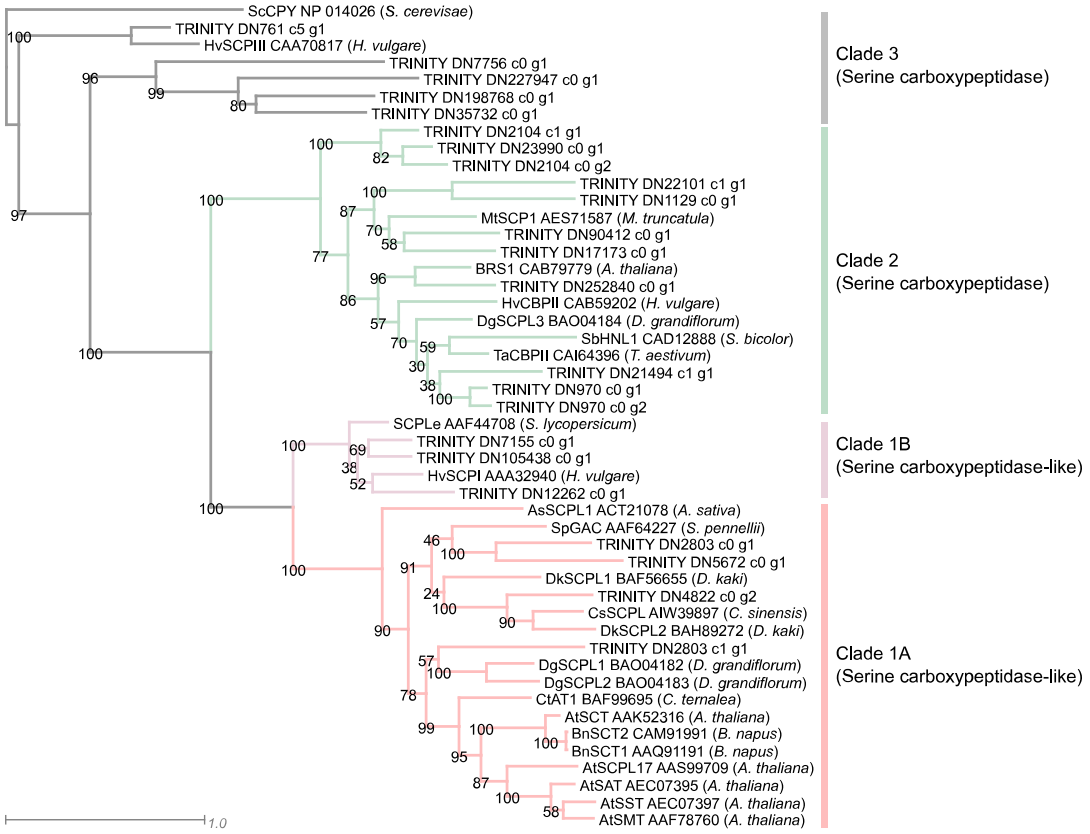


Figure 5.2.5. Phylogenetic analysis of candidate *S. officinalis* SCPL ATs. The maximum-likelihood tree was constructed using an amino acid alignment of putative SCPL ATs from *S. officinalis* and previously identified SCPL ATs from other plant species (listed in Table C.1.4). Bootstrap values are shown beside each node. The scale bar indicates the number of amino acid substitutions per amino acid site. A SCPL AT from *Saccharomyces cerevisae* (ScCPY) was used as a basal outgroup. Different SCPL AT clades are colour coded and labelled. *H. vulgare, Hordeum vulgare; S. lycopersicum, Solanum lycopersicum; S. pennellii, Solanum pennellii; A. thaliana, Arabidopsis thaliana; B. napus, Brassica napus; A. sativa, Avena sativa; M. truncatula, Medicago truncatula; C. sinensis, Camellia sinensis; D. kaki, Diospyros kaki; C. ternalea, Clitoria ternalea; D. grandiflorum, Delphinium grandiflorum; H. vulgare, Hordeum vulgare; S. bicolor, Sorghum bicolor; T. aestivum, Triticum aestivum.*

Once the high-quality genome was completed, the translated amino acid sequences of all candidates were searched against the genome by reciprocal BLASTP. In addition to sequence verification, this ensured that no genome specific candidates with similar sequence identity or alternative variants, such as other isoforms, splicing variants or variants arising from differences in the assembly, were overlooked. The sequences of all candidates resulted in a 100% match against the genome sequences without any new additional candidates, except *CYP1*, *BAHD9*, *BAHD12*, *BAHD13*, *SCPL1* and *SCPL3*. These sequences were present as partial sequences in the transcriptome as they were missing the C-terminus compared to their respective genome variants. The soapwort genome contained full open reading frames (ORFs) from start to end codon of *CYP1*, *BAHD9*, *BAHD12*, *BAHD13*, *SCPL1* and *SCPL3*, and thus the genome variants were used for further functional analysis.

5.2.4 Functional characterization of candidate biosynthetic genes using transient expression in *N. benthamiana*

The candidate saponarioside biosynthetic genes identified in Section 5.2.3 were transiently expressed in *N. benthamiana* to investigate their enzymatic activity. The ORFs of candidate genes were PCR-amplified from a cDNA pool of six soapwort organs with upstream 5′ attb sites to allow for Gateway® cloning. However, *CYP1*, *SoGH1*, and *BAHD10* were synthesized as gene fragments (Twist Biosciences and IDT). The PCR-amplified or synthesized gene fragments were cloned into pDONR207 and transferred into the plant expression vector pEAQ-HT-DEST1 (Sainsbury and Lomonossoff, 2014). The expression constructs were individually transformed into *A. tumefaciens* (LBA4404) for transient expression in *N. benthamiana*. In all experiments, an *A. tumefaciens* strain carrying an expression construct for a truncated *HMGR* (*tHMGR*) from *Avena strigosa* was co-infiltrated to enhance triterpene production (Reed *et al.*, 2017). Leaves were harvested 5 days after infiltration and extracted for metabolites. Leaf extracts were analysed by HPLC-MS in negative ionization mode ($[M-H]^-$). In most cases, commercial or authentic standards of the expected products were not available. As saponariosides and QS-21 from *Q. saponaria* are similar in chemical structure, both pathways share biosynthetic intermediates up to the last common intermediate, 3-*O*-{β-D-xylopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranosyl

ester}-quillaic acid. Thus, to mitigate the lack of commercially available saponarioside biosynthetic intermediates, QS-21 biosynthetic genes identified from *Q. saponaria* (Reed *et al.*, 2023) were used as positive comparisons. The saponarioside pathway genes were generally identified in a stepwise manner where the gene encoding the enzyme likely involved in the previous step of the hypothesized pathway (Fig. 5.1.1) was identified first, and was then used to produce the precursor compound required for the identification of gene encoding the enzyme responsible for the next step of the pathway. However, some pathway genes were identified at the same time as genes expected to be involved in previous steps or even at later steps. In these cases, the counterpart *Q. saponaria* genes were used to fill in for yet to be identified *S. officinalis* genes to produce the precursor compound. After identifying the complete gene set involved in saponarioside B biosynthesis, the pathway was reconstructed in *N. benthamiana* using the identified *S. officinalis* genes only. For clarity, the below sections will describe the discovery of saponarioside biosynthetic pathway in a consecutive order.

Biosynthesis of quillaic acid backbone

With the identification of *SobAS1*, *SoC28* and *SoC28C16*, an additional CYP with C-23 oxidation activity is potentially required to complete the biosynthesis of quillaic acid (Fig. 5.1.1). As mentioned in Section 5.2.2, further experiments were performed without *SoC28* as *SoC28C16* alone can carry out both C-28 and C-16 oxidation activity, and the co-expression of both genes together did not result in any noticeable increase of echinocystic acid in *N. benthamiana*. Thus, all 7 CYP candidates were tested in combination with *SobAS1* and *SoC28C16* for their potential ability to oxidize the C-23 position of echinocystic acid to produce quillaic acid. Only one candidate, *CYP1*, showed C-23 oxidation activity and thus was renamed as *SoC23*. It is worthwhile noting that in the transcriptome assembly, this candidate was missing 3 amino acids at the C-terminus compared to the genome variant. However, both variants of CYP1/SoC23 displayed the same activity. Other CYP candidates did not display any noteworthy, unexpected activities. Co-expression of *SobAS1*, *SoC28C16* and *SoC23* led to the production of a new peak (**1**) with *m/z* 485.3, corresponding to the expected $[M-H]^-$ of quillaic acid (**1,** QA) (Fig. 5.2.6). The retention time (RT) and mass spectra (MS) of peak (**1**) matched with the quillaic acid standard (Fig. 5.2.6). Peak (**1**) was not detected in the negative control, where only *SobAS1* and *SoC28C16*

were expressed. Based on these results, *SoC23* is likely to be a CYP with C-23 oxidation activity, and together with *SobAS1* and *SoC28C16*, completes the pathway to the biosynthesis of quillaic acid, the aglycone of saponariosides A and B.

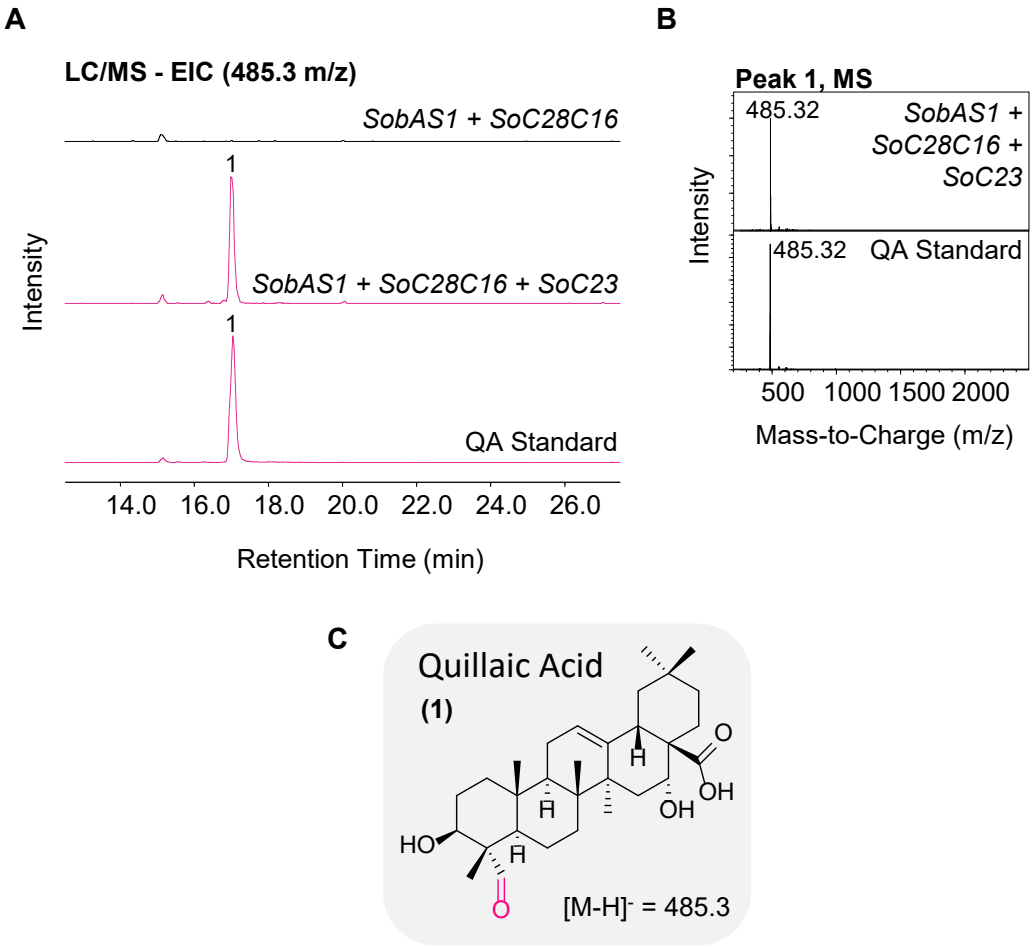


Figure 5.2.6. Transient expression of *SoC23* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* strains containing expression constructs for *SobAS1*, *SoC28C16* and *SoC23*. Leaves were harvested five days after infiltration and extracts were analysed by HPLC-MS. **(A)** Extracted ion chromatograms (EIC) and **(B)** mass spectra (MS) are shown. An extract from *N. benthamiana* leaves co-expressing *SobAS1* and *SoC28C16* was used as a negative control. The additional activity of *SoC23* produced a peak corresponding to quillaic acid (*m/z* 485.3). **(C)** Structure of quillaic acid (QA, **1**), the expected product of SoC23 when acting in combination with SobAS1 and SoC28C16. Modification performed by SoC23 is highlighted in colour. Peak (**1**), identified as quillaic acid (QA). The full TIC range is available in Fig. C.3.1.

Building of the C-3 sugar chain

The C-3 trisaccharide chain of saponariosides A and B consists of D-glucuronic acid, D-galactose, and D-xylose. Several members of the cellulose-synthase superfamily-derived glycosyltransferase (CSyGT) have been reported to be involved in the 3-*O*-glucuronidation of triterpene aglycones (Jozwiak *et al.*, 2020; Chung *et al.*, 2020; Reed *et al.*, 2023). Thus, the candidate *SoCSL1* identified above was co-expressed with genes required to produce quillaic acid (*SobAS1* + *SoC28C16* + *SoC23*). The negative control consisted of the genes for QA biosynthesis only. As the expected product, 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid (**2**, QA-Mono), is not commercially available, a CSL characterized from *Q. saponaria* (QsCSL) was co-expressed with soapwort genes required to produce QA as a positive comparison. Co-expression of *SoCSL1* with *SobAS1*, *SoC28C16,* and *SoC23* led to the production of a new peak (**2**) with *m/z* 661.3, the expected $[M-H]^-$ of QA-Mono (Fig. 5.2.7). Peak (**2**) was not detected in the negative control, and the RT and MS of peak (**2**) matched with peak (**2**) produced by expression of *QsCSL* (Fig. 5.2.7). The MS/MS fragmentation pattern of peak (**2**) revealed the main fragment ion to be *m/z* 485.33, which corresponds to the expected $[M-H]^-$ of quillaic acid (Fig. 5.2.7D). Based on these results, peak (**2**) was identified as QA-Mono, and SoCSL1 was assigned as a CSL able to glucuronidate quillaic acid at the C-3 position.

Next, the 13 UGT candidates identified in Section 5.2.3 were tested for their potential ability to elongate the C-3 sugar chain. Of these, two candidates, *UGT7* and *UGT13* (hereafter named *SoC3Gal* and *SoC3Xyl* respectively), were able to glycosylate the glucuronic moiety of QA-Mono. When *SoC3Gal* was co-expressed with the soapwort genes required to produce QA-Mono (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1*), production of a new peak (**3**) was observed (Fig. 5.2.8). The expected product, 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid}-quillaic acid (**3**, QA-Di) is not commercially available, thus a QA-3-*O*-glucuronoside-*β*-1,2-galactosyltransferase identified from *Q. saponaria* (*QsC3Gal*) was co-expressed with soapwort genes involved in QA-Mono biosynthesis to produce QA-Di. The new peak (**3**) produced by the additional expression of *SoC3Gal*, displayed *m/z* of 823.4 which corresponded to the $[M-H]^-$ of QA-Di (**3**, Fig. 5.2.8). Furthermore, the RT and MS of peak (**3**) produced by the addition of *SoC3Gal* matched with peak (**3**) produced by the additional expression of *QsC3Gal* to QA-Mono biosynthetic genes (Fig. 5.2.8). The

MS/MS fragmentation pattern revealed the major fragment ion of peak (**3**) to be *m/z* 485.32, corresponding to $[M-H]^-$ of QA, which suggests the fragmentation of the C-3 sugar chain from QA-Di (Fig. 5.2.8). Peak (**3**) was not observed in the negative control only co-expressing genes up to QA-Mono biosynthesis. Based on these results, peak (**3**) was identified as QA-Di, and SoC3Gal is likely to be a QA-3-*O*-glucuronoside-*β*-1,2-galactosyltransferase from *S. officinalis*.

Subsequently, *SoC3Xyl* was co-expressed with QA-Di biosynthetic genes (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1* + *SoC3Gal*). As a positive control, a xylosyltransferase (*QsC3Xyl*) from *Q. saponaria* that adds D-xylose to D-galactose of QA-Di, was co-expressed instead of *SoC3Xyl* as the expected product, 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (**4**, QA-Tri) is not commercially available. The additional expression of *SoC3Xyl* produced a new peak (**4**) with *m/z* 955.4, which corresponds to the expected $[M-H]^-$ of QA-Tri (Fig. 5.2.9). This new peak (**4**) was not detected in the negative control, where only the genes required to produce QA-Di were expressed. Not only did peak (**4**) have the same RT and MS as peak (**4**) present in the positive control, MS/MS fragmentation also revealed the major ions to be *m/z* 823.42 $[M\text{-pentose-H}]^-$ and *m/z* 485.33 $[M\text{-pentose-hexose-H}]^-$ (Fig. 5.2.9D). Based on these results, peak (**4**) was identified as QA-Tri and SoC3Xyl is likely a QA-3-*O*-glucuronoside-*β*-1,2-galactose-*β*-1,3-D-xylosyltransferase. Together with SoCSL1, the identification of SoC3Gal and SoC3Xyl completed the biosynthetic route to the C-3 sugar chain present in saponariosides A and B.

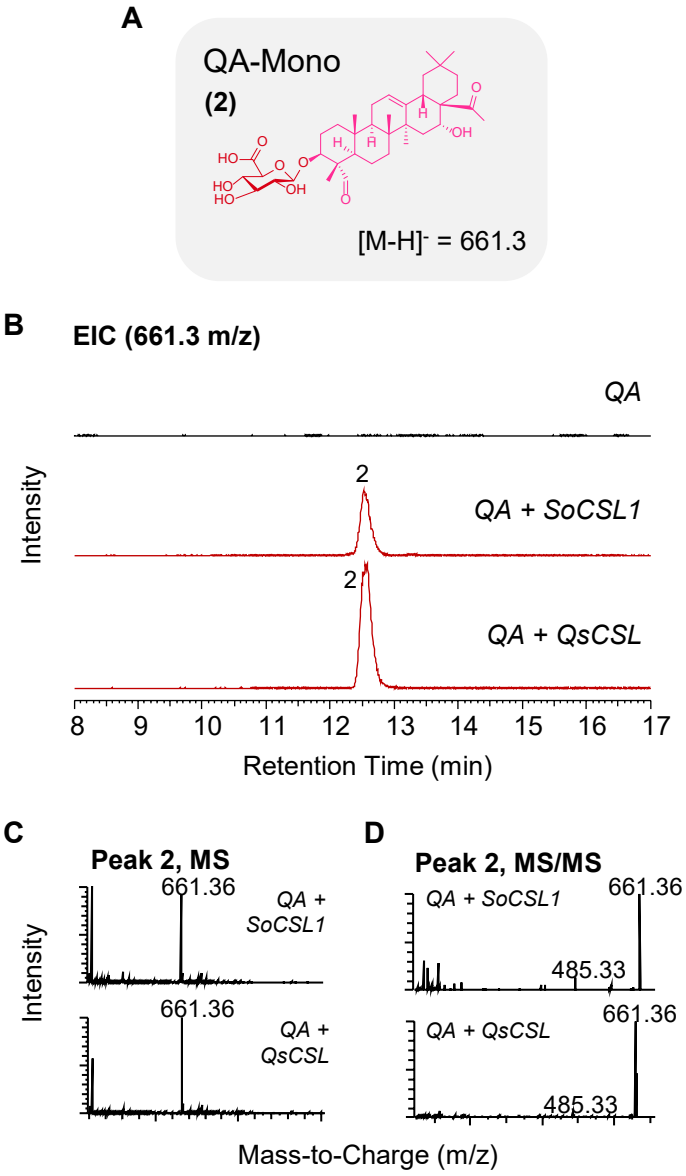


Figure 5.2.7. Activity of *SoCSL1* transiently expressed in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* strains containing expression constructs for *SobAS1, SoC28C16, SoC23* and *SoCSL1*. Metabolites were extracted from leaves harvested five days after infiltration and analysed by HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid (**2**, QA-Mono), the expected product of SoCSL1 when acting in combination with QA biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23). **(B)** Extracted ion chromatogram (EIC) at *m/z* 661.3. **(C)** Mass spectra (MS) of peak (**2**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peak (**2**). Extract from *N. benthamiana* leaves only co-expressing QA biosynthetic genes was used as negative control. As a positive comparison, QsCSL identified from *Q. saponaria* was co-expressed with QA biosynthetic genes from *S. officinalis* to produce QA-Mono. Peak (**2**), identified QA-Mono.

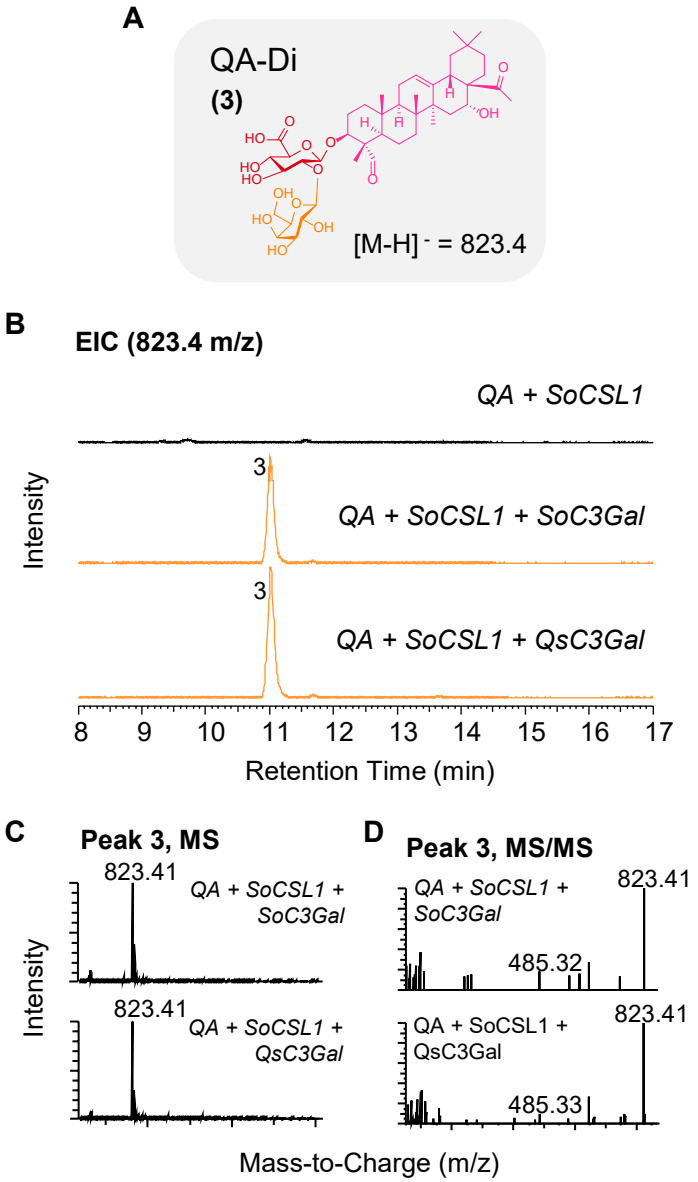


Figure 5.2.8. Activity of *SoC3Gal* transiently expressed in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* strains containing expression constructs for *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1* and *SoC3Gal*. Metabolites were extracted from leaves harvested five days after infiltration, and analysed on HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid}-quillaic acid (QA-Di), the expected product of SoC3Gal when acting in combination with QA-Mono biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1). **(B)** Extracted ion chromatogram (EIC) at *m/z* 823.4. **(C)** Mass spectra (MS) of peak (**3**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peak (**3**). Extract from *N. benthamiana* leaves co-expressing QA-Mono biosynthetic genes (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1*) was used as a negative control. As a positive comparison, *QsC3Gal* identified from *Q. saponaria* was co-expressed with QA-Mono biosynthetic genes from *S. officinalis* to produce QA-Di. Peak (**3**), identified as QA-Di. The full TIC range is available in Fig.C.3.2.

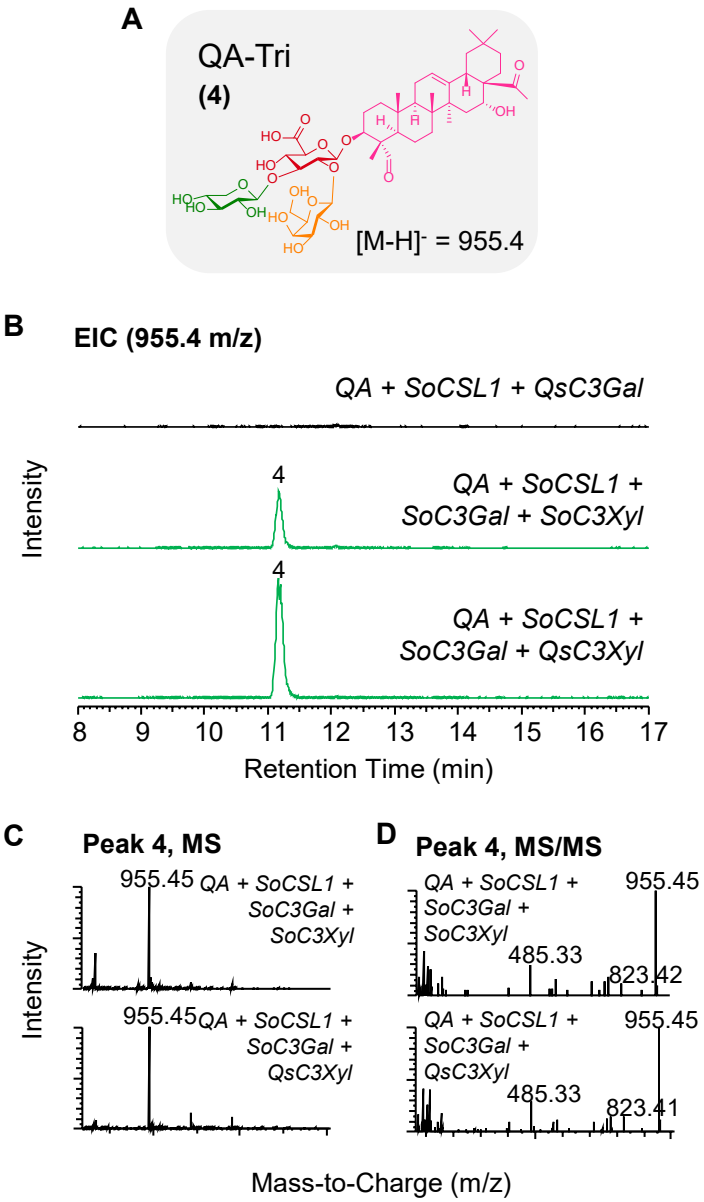


Figure 5.2.9. Transient expression of *SoC3Xyl* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* strains containing expression constructs for *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1, SoC3Gal* and *SoC3Xyl*. Metabolites were extracted from leaves harvested four days after infiltration, and analysed on HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (QA-Tri), the expected product of SoC3Xyl when acting in combination with QA-Di biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1 + SoC3Gal). **(B)** Extracted ion chromatogram (EIC) at *m/z* 955.4. **(C)** Mass spectra (MS) of peak (**4**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peak (**4**). Extract from *N. benthamiana* leaves co-expressing QA-Di biosynthetic genes only was used as a negative control. As a positive comparison, *QsC3Gal* identified from *Q. saponaria* was co-expressed with QA-Di biosynthetic genes from *S. officinalis* to produce QA-Tri. Peak (**4**), identified as QA-Tri. The full TIC range is available in Fig.C.3.3.

Building of C-28 sugar chain

With the completion of the C-3 sugar chain, the next focus was to identify genes involved in the biosynthesis of the main C-28 tetrasaccharide chain consisting of D-fucose, L-rhamnose, and two D-xyloses. As such, the suite of 13 candidate UGTs were screened again for their potential ability to biosynthesize the C-28 sugar chain. Of the 13 candidates, four - *UGT10*, *UGT4*, *UGT13* and *UGT2* (hereafter referred to as *SoC28Fu*, *SoC28Rha*, *SoC28Xyl1* and *SoC28Xyl2*, respectively) displayed enzymatic activities involved in the attachment and elongation of the C-28 sugar chain and are discussed further below.

The steps to fucosylation in saponin biosynthesis were characterised only recently by the Osbourn group (Reed *et al.*, 2023). Two QS-saponin biosynthetic enzymes, QsC28Fu and QsFucSyn from *Q. saponaria* are involved in the addition of D-fucose to QA-Tri (Reed *et al.*, 2023). This study revealed that instead of UDP-D-fucose, QsC28Fu transfers UPD-4-keto-6-deoxy-glucose to the saponin substrate, while QsFucSyn is a keto-reductase that converts UDP-4-keto-6-deoxy-glucose to UDP-D-fucose (Reed *et al.*, 2023). During the efforts to characterize QsFucSyn, a soapwort homologue of QsFucSyn was identified by performing BLASTP search against the translated soapwort transcriptome using QsFucSyn as query. The closest SDR candidate from soapwort, TRINITY_DN10791_c0_g2 (hereafter referred to as SoSDR1), shared 57.2% amino acid sequence identity with QsFucSyn. Furthermore, when *SoSDR1* was transiently co-expressed in *N. benthamiana* with QS-7 biosynthetic genes identified from *Q. saponaria*, *SoSDR1* was also able to increase the fucosylated product similarly to QsFucSyn. In addition to these findings (Reed et al., 2023), *SoSDR1* also showed strong co-expression with *SobAS1* (PCC = 0.941), suggesting its role in saponin biosynthesis in *S. officinalis,* and thus fucosylation in saponarioside biosynthesis may occur in similar steps to QS-saponin biosynthesis in *Q. saponaria*.

As D-fucose is suspected to be present in limiting amounts in *N. benthamiana* (Jozwiak *et al.*, 2020; Reed *et al.*, 2023), *SoSDR1* was co-expressed when testing for activities of candidate genes hereafter. When *SoC28Fu* was co-expressed with the QA-Tri producing genes (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1* + *SoC3Gal* + *SoC3Xyl*) and *SoSDR1*, a new product peak (**5**) with *m/z* 1101.5 was observed (Fig. 5.2.10). This peak (**5**) was not detected when only *SoSDR1* was co-expressed with QA-Tri

producing genes, and the *m/z* of peak (**5**) corresponded with the expected $[M-H]^-$ of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-fucopyranosyl ester}-quillaic acid (**5**, QA-TriF). Furthermore, the RT and MS of peak (**5**) produced by additional expression of *SoC28Fu* corresponded with the peak (**5**) produced by co-expression of *QsC28Fu* with *S. officinalis* QA-Tri biosynthetic genes. The MS/MS fragmentation pattern of peak (**5**) revealed the major daughter ions to be *m/z* 955.4, corresponding to $[M-H]^-$ of QA-Tri, and *m/z* 485.3, corresponding to $[M-H]^-$ of QA (Fig. 5.2.10D). Based on these results, peak (**5**) was identified as QA-TriF, and SoC28Fu as a sugar transferase involved in the addition of D-fucose to QA-Tri. Based on the findings from QS-saponin biosynthetic pathway (Reed *et al*., 2023), SoC28Fu is likely to use UDP-4-keto-6-deoxy-glucose as a substrate to transfer 4-keto-6-deoxy-glucose to QA-Tri, which is then likely to be reduced by SoSDR1 acting as a 4-ketoreductase. However, *in vitro* enzyme assays with purified SoSDR1 and SoC28Fu should be performed in the future to confirm the substrates of these two enzymes.

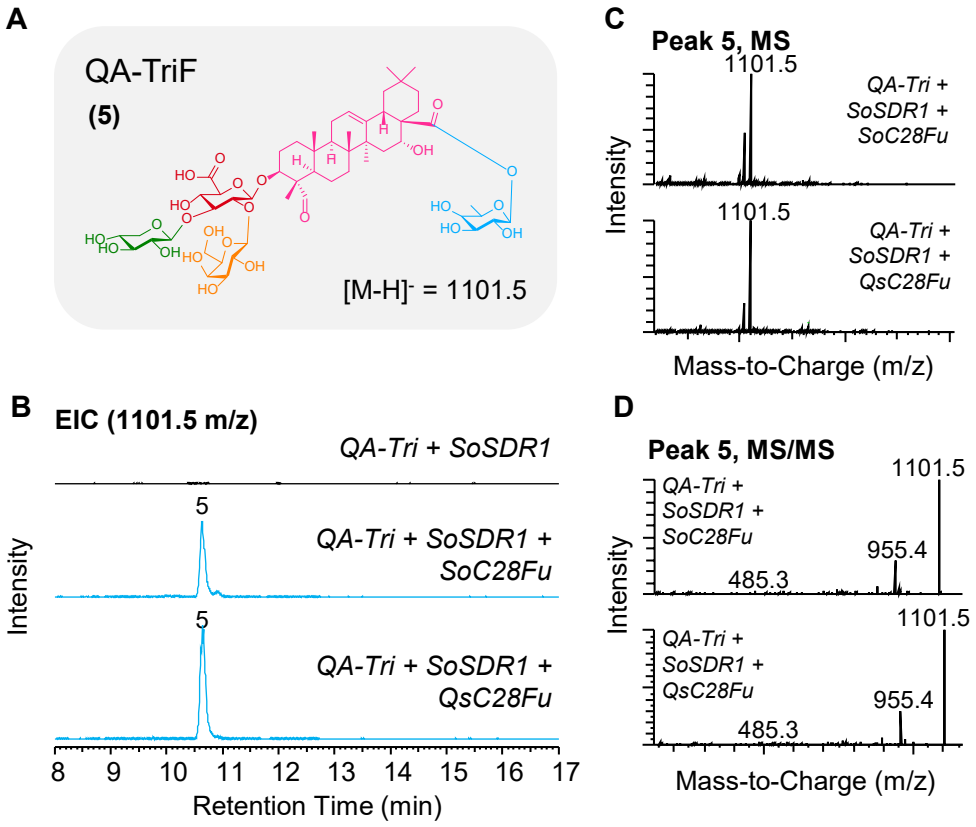


Figure 5.2.10. Transient expression of *SoC28Fu* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* containing expression constructs strains for *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1, SoC3Gal*, *SoC3Xyl*, *SoSDR1* and *SoC28Fu*. Leaves were harvested five days after infiltration and leaf extracts were analysed using HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-fucopyranosyl ester}-quillaic acid (**5**, QA-TriF), the expected product of SoC28Fu when acting in combination with QA-Tri biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1 + SoC3Gal + SoC3Xyl) and SoSDR1 to enhance the availability of fucose in *N. benthamiana*. **(B)** Extracted ion chromatogram (EIC) at *m/z* 1101.5. **(C)** Mass spectra (MS) of peak (**5**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peak (**5**). Extract from *N. benthamiana* leaves co-expressing *S. officinalis* genes required to produce QA-Tri and *SoSDR1* was used as a negative control. As a positive comparison, *QsC28Fu* identified from *Q. saponaria* was co-expressed with QA-Tri biosynthetic genes identified from *S. officinalis* together with *SoSDR1* to produce QA-TriF. Peak (**5**), identified as QA-TriF. The full TIC range is available in Fig. C.3.4.

Following the discovery of SoC28Fu, the co-expression of *SoC28Rha* with the combination of genes required to produce QA-TriF (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1* + *SoC3Gal* + *SoC3Xyl* + *SoSDR1* + *SoC28Fu*) resulted in the production of a new peak (**6**, Fig. 5.2.11). Peak (**6**) was not observed when only QA-TriF producing genes were co-expressed without *SoC28Rha*, and displayed *m/z* of 1247.5, corresponding to the expected $[M-H]^-$ of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid (**6**, QA-TriFR). As QA-TriFR is not commercially available, *QsC28Rha*, a rhamnosyltransferase from *Q. saponaria*, was co-expressed with QA-TriF biosynthetic genes from soapwort to produce QA-TriFR. Peak (**6**) produced by the co-expression of *SoC28Rha* with QA-TriF biosynthetic genes corresponded with peak (**6**) produced by the activity of QsC28Rha in both RT and MS (Fig. 5.2.11). Tandem MS of peak (**6**) revealed the major fragment ions to be *m/z* 955.4, corresponding to $[M-H]^-$ of QA-Tri, and *m/z* 485.3, corresponding to $[M-H]^-$ of quillaic acid, suggesting the fragmentation of the C-28 sugar chain, followed by the C-3 sugar chain (Fig. 5.2.11D). Based on these observations, peak (**6**) was identified as QA-TriFR, and SoC28Rha is likely a rhamnosyltransferase with the ability to catalyse the addition of L-rhamnose to QA-TriF.

The last two sugar moieties of the main C-28 sugar chain are both D-xyloses, thus the UGT candidates were screened for potential xylosyltransferase activities on QA-TriFR (**6**). Co-expression of *SoC28Xyl1* with combination of genes to produce QA-TriFR (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1* + *SoC3Gal* + *SoC3Xyl* + *SoSDR1* + *SoC28Fu* + *SoC28Rha*) resulted in the formation of peak (**7**) with *m/z* 1379.6, corresponding to the expected $[M-H]^-$ of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid (**7**, QA-TriFRX, Fig. 5.2.12). As a positive comparison *QsC28Xyl1*, a xylosyltransferase from *Q. saponaria* with the ability to transfer D-xylose to QA-TriFR, was co-expressed with QA-TriFR biosynthetic genes identified from *S. officinalis*. The RT and MS for peak (**7**)**,** both produced by the additional activity of *SoC28Xyl1* or *QsC28Xyl1*, corresponded to each other, and peak (**7**) was only detected in samples either expressing *SoC28Rha* or *QsC28Rha* with genes required to produce

the acceptor substrate, QA-TriFR (Fig. 5.2.12B). Furthermore, tandem MS of peak (**7**) revealed the major fragment ions as *m/z* 955.4 and *m/z* 485.3, which suggested the loss of the C-28 sugar chain to yield QA-Tri, followed by the loss of the C-3 sugar chain, yielding quillaic acid (Fig. 5.2.12D). These results suggested the identity of peak (**7**) as QA-TriFRX and SoC28Xyl1 is likely a xylosyltransferase with the ability to transfer D-xylose to the C-28 L-rhamnose of QA-TriFR.

Additional rounds of candidate gene screening revealed formation of a new product peak (**8**) when *SoC28Xyl2* was co-expressed with *SoC28Xyl1* and QA-TriFR producing genes (Fig. 5.2.13). Peak (**8**) was not detected in the absence of *SoC28Xyl2* expression, and displayed a *m/z* of 1511.6, corresponding to the expected $[M-H]^-$ of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid (**8**, QA-TriFRXX, Fig. 5.2.13). As a positive comparison, *QsC28Xyl2*, a xylosyltransferase from *Q. saponaria* with the ability to transfer a D-xylose to QA-TriFRX, was co-expressed with QA-TriFRX producing genes identified from soapwort (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1* + *SoC3Gal* + *SoC3Xyl* + *SoSDR1* + *SoC28Fu* + *SoC28Rha* + *SoC28Xyl1*). The RT and MS of peak (**8**) produced by the activity of *SoC28Xyl2* corresponded with the product peak (**8**) of *QsC28Xyl2* acting in combination with *S. officinalis* QA-TriFRX biosynthetic genes (Fig. 5.2.13). Furthermore, MS/MS analysis revealed the major fragment ions of peak (**8**) to be *m/z* 1379.6, *m/z* 955.4 and *m/z* 485.3, which correspond to the $[M-H]^-$ of QA-TriFRX, QA-Tri and QA, respectively (Fig. 5.2.13D). This fragmentation pattern suggested the loss of a terminal D-xylose in the C-28 sugar chain, followed by the loss of the remaining C-28 sugar chain, and finally, the loss of the C-3 sugar chain, resulting in quillaic acid. Collectively these results suggested that peak (**8**) is QA-TriFRXX, and that SoC28Xyl2 is therefore a xylosyltransferase involved in the addition of the last D-xylose of the C-28 chain.

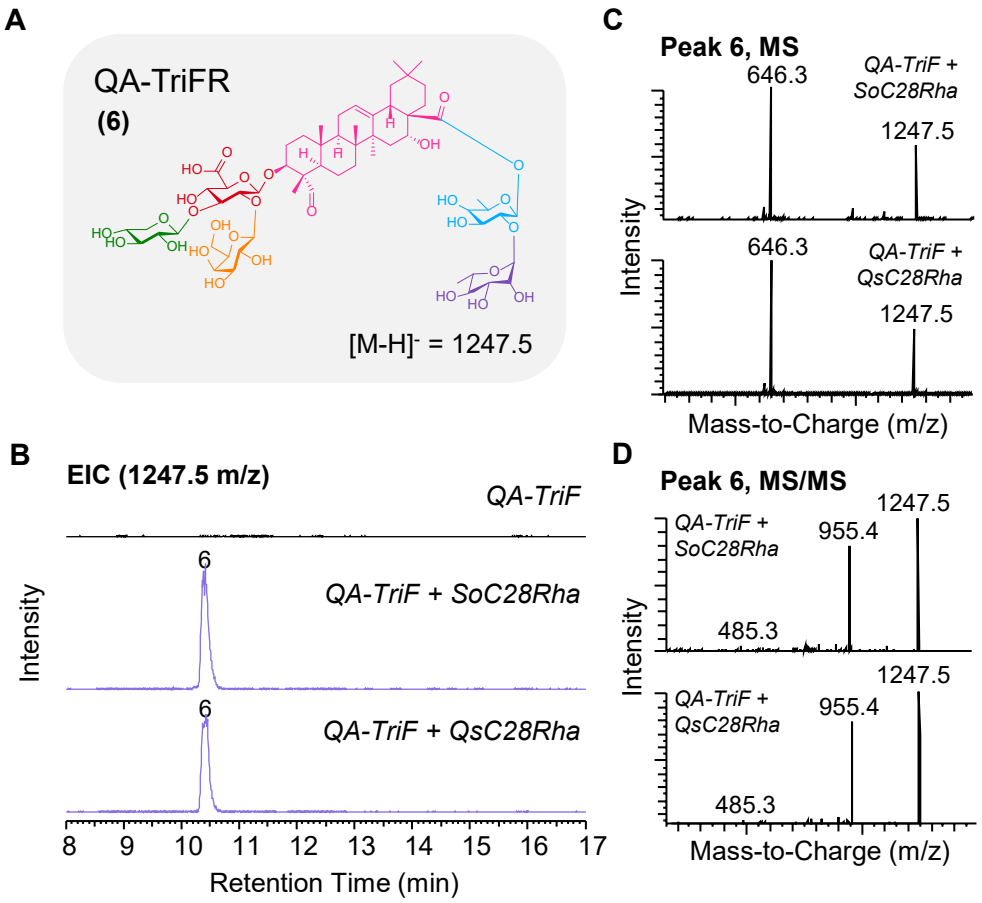


Figure 5.2.11. Transient expression of *SoC28Rha* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* containing expression constructs strains for *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1, SoC3Gal*, *SoC3Xyl*, *SoSDR1. SoC28Fu* and *SoC28Rha*. Leaves were harvested five days after infiltration and leaf extracts were analysed using HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid (**6**, QA-TriFR), the expected product of SoC28Rha when acting in combination with QA-TriF biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1 + SoC3Gal + SoC3Xyl + SoSDR1 + SoC28Fu). **(B)** Extracted ion chromatogram (EIC) at *m/z* 1247.5. **(C)** Mass spectra (MS) of peak (**6**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peak (**6**). Extract from *N. benthamiana* leaves co-expressing *S. officinalis* genes required to produce QA-TriF only was used as a negative control. As a positive comparison, *QsC28Rha* identified from *Q. saponaria* was co-expressed with QA-TriF biosynthetic genes identified from *S. officinalis* to produce QA-TriFR. Peak (**6**), identified as QA-TriFR. The full TIC range is available in Fig. C.3.5.

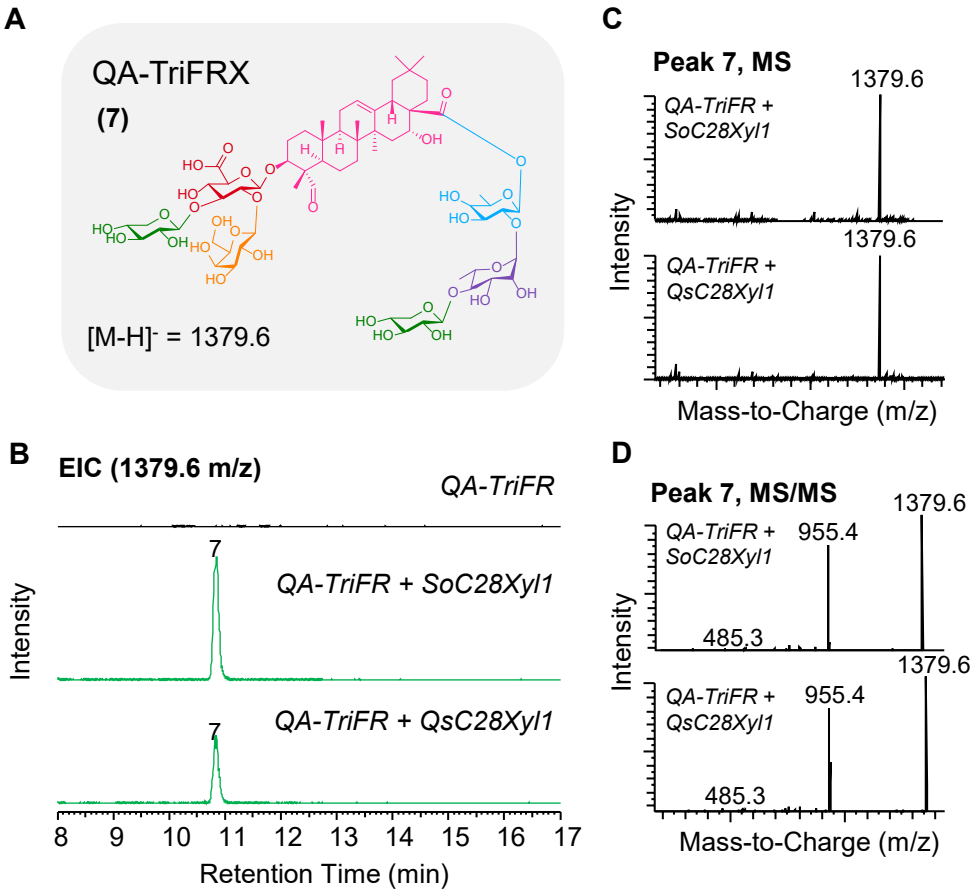


Figure 5.2.12. Transient expression of *SoC28Xyl1* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* containing expression constructs strains for *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1, SoC3Gal*, *SoC3Xyl*, *SoSDR1. SoC28Fu, SoC28Rha* and *SoC28Xyl1*. Leaves were harvested five days after infiltration and leaf extracts were analysed using HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid (**7**, QA-TriFRX), the expected product of SoC28Xyl1 when acting in combination with QA-TriFR biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1 + SoC3Gal + SoC3Xyl + SoSDR1 + SoC28Fu + SoC28Rha). **(B)** Extracted ion chromatogram (EIC) at *m/z* 1379.6. **(C)** Mass spectra (MS) of peak (**7**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peak (**7**). Extract from *N. benthamiana* leaves co-expressing *S. officinalis* genes required to produce QA-TriFR only was used as a negative control. As a positive comparison, *QsC28Xyl1* identified from *Q. saponaria* was co-expressed with QA-TriFR biosynthetic genes identified from *S. officinalis* to produce QA-TriFRX. Peak (**7**), identified as QA-TriFRX. The full TIC range is available in Fig. C.3.6.

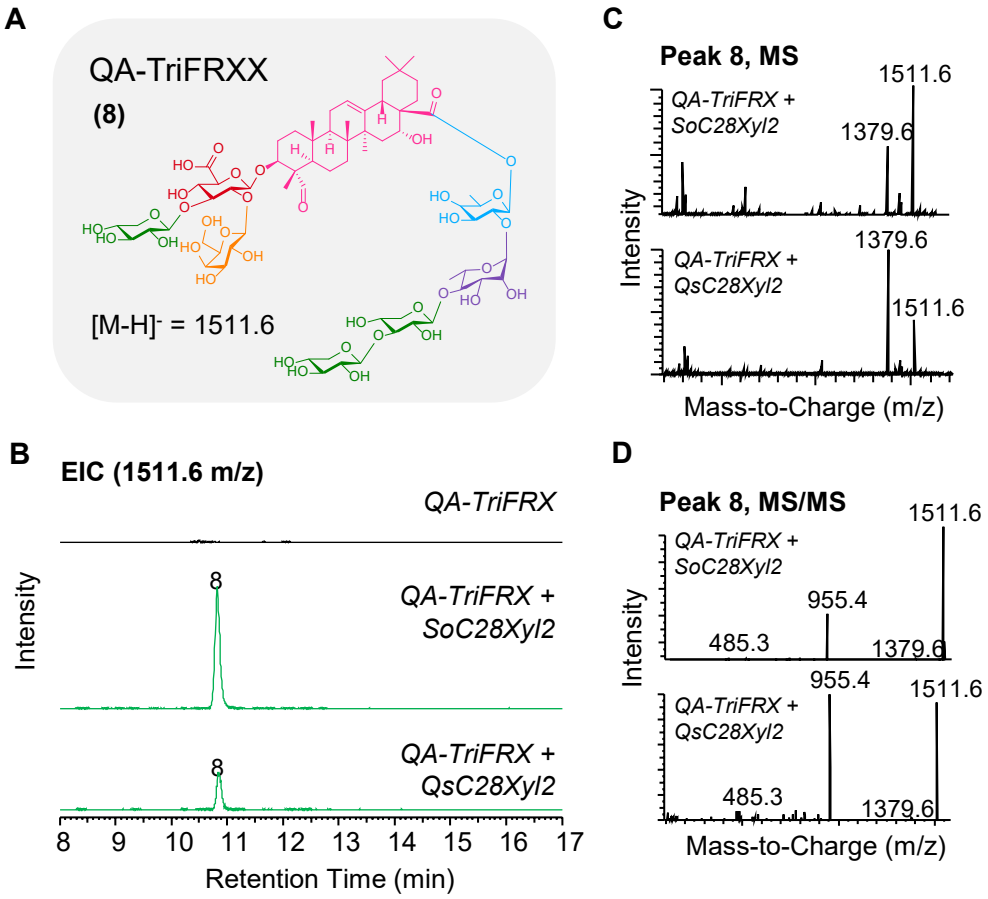


Figure 5.2.13. Transient expression of *SoC28Xyl2* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* containing expression constructs strains for *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1, SoC3Gal*, *SoC3Xyl*, *SoSDR1. SoC28Fu, SoC28Rha, SoC28Xyl1* and *SoC28Xyl2*. Leaves were harvested five days after infiltration and leaf extracts were analysed using HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid (**8**, QA-TriFRXX), the expected product of SoC28Xyl2 when acting in combination with QA-TriFRX biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1 + SoC3Gal + SoC3Xyl + SoSDR1 + SoC28Fu + SoC28Rha + SoC28Xyl1). **(B)** Extracted ion chromatogram (EIC) at *m/z* 1511.6. **(C)** Mass spectra (MS) of peak (**8**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peak (**8**). Extract from *N. benthamiana* leaves co-expressing *S. officinalis* genes required to produce QA-TriFRX only was used as a negative control. As a positive comparison, *QsC28Xyl2* identified from *Q. saponaria* was co-expressed with QA-TriFRX biosynthetic genes identified from *S. officinalis* to produce QA-TriFRXX. Peak (**8**), identified as QA-TriFRXX. The full TIC range is available in Fig. C.3.7.

The discovery of *SoC28Fu*, *SoC28Rha*, *SoC28Xyl1* and *SoC28Xyl2* leads to the complete set of genes required to produce the main linear part of the C-28 sugar chain in saponariosides A and B. Furthermore, together with *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1*, *SoC3Gal* and *SoC3Xyl,* represents a set of genes with the ability to produce QA-TriFRXX, the last shared intermediate between QS-saponins and the saponarioside biosynthetic pathways in *Q. saponaria* and *S. officinalis*, respectively. Although *Q. saponaria* genes were used to confirm the identity of enzyme products produced by soapwort candidates, the gene discovery was performed *de novo* and did not rely on the *Q. saponaria* genes, but rather relied on the close co-expression of the saponarioside biosynthetic genes to *SobAS1*. In fact, despite their shared biochemical activities, *S. officinalis* and *Q. saponaria* enzymes involved in glycosylation show low amino acid identities (Table 5.2.7).

Table 5.2.7. Shared amino acid sequence identity of glycosylation related enzymes in *S. officinalis* and *Q. saponaria* with shared functional activities. *Q. saponaria* enzymes are described in (Reed *et al.* 2023).

S. officinalis	***Q. saponaria***	**AA Identity (%)**
SoCSL1	QsCSL	56.0
UGT73DL1	UGT73CU3	46.3
UGT73CC6	UGT73CX1	47.8
UGT74CD1	UG74BX1	43.0
SoSDR1	QsFucSyn	57.2
UGT79T1	UGT91AR1	29.2
UGT79L3	UGT91AQ1	31.1
UGT73M2	UGT73CY3	41.2

Addition of D-quinovose by a non-canonical glycosyl hydrolase

To complete the biosynthetic pathway to saponarioside B, the steps responsible for the transfer of 4-*O*-acetylquinovose to QA-TriFRXX (**8**) still need to be elucidated. Although D-quinovose is a common sugar found in specialised metabolites produced by marine animals such as starfishes and sea cucumbers (Stonik and Elyakov, 1988), it is considered a rare sugar as a component of plant metabolites (Augustin *et al.*, 2011). Consequently, little is known about the types and mechanisms of GTs involved in the biosynthesis of saponins containing D-quinovose (Vogt and Jones, 2000). UGTs involved in the biosynthesis of plant specialized metabolites typically belong to family 1 of the GT enzyme superfamily, one of the largest groups of plant enzymes involved

in specialized metabolism (Louveau and Osbourn, 2019). However, none of the soapwort UGT candidates showed quinovosyltransferase activity towards QA-TriFRXX. Thus, the search for this candidate sugar transferase expanded outside the enzyme families (OSC, CYP, CSL, UGT and AT) originally hypothesized to be involved in saponarioside biosynthesis. During re-evaluation of all soapwort genes co-expressed with *SobAS1*, a candidate gene (*TRINITY_DN530_c2_g1*) predicted to encode a member of a different class of carbohydrate-active enzymes, glycoside hydrolase family 1 (GH1), with strong co-expression with *SobAS1* (PCC = 0.971) was identified. Although GH1 enzymes are typically β-glycosidases, several have been reported to have transglycosidase (TG) activity and function in biosynthesis of glycoconjugates (Cairns *et al.*, 2015). Thus, the soapwort candidate annotated as a member of the GH1 family was renamed as SoGH1 and its function was investigated for potential quinovosyltransferase activity. When *SoGH1* was co-expressed together with soapwort genes anticipated to produce QA-TriFRXX (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1* + *SoC3Gal* + *SoC3Xyl* + *SoSDR1* + *SoC28Fu* + *SoC28Rha* + *SoC28Xyl1* + *SoC28Xyl2*), two new peaks (**9**) and (**10**) were observed (Fig. 5.2.14). Both peaks (**9**) and (**10**) displayed *m/z* of 1657.7, which corresponds to the expected $[M-H]^-$ of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-[*β*-D-quinovopyranosyl-(1→4)]-*β*-D-fucopyranosyl ester}-quillaic acid (**10**, QA-TriF(Q)RXX). Although the RT differed between the two product peaks, tandem MS produced a same fragmentation pattern. The main fragment ions were *m/z* 1525.7 (expected $[M-H]^-$ of QA-TriFRXX) and *m/z* 955.4 (expected $[M-H]^-$ of Tri-QA), which suggested a loss of a deoxyhexose, followed by the loss of the entire C-28 sugar chain, resulting in QA-Tri (Fig. 5.2.14D). Based on MS analysis alone, the identity of these two product peaks were indistinguishable. To elucidate the identities of peaks (**9**) and (**10**), large-scale agroinfiltration of 110 *N. benthamiana* plants was carried out. Leaves were harvested 5 days after infiltration and were lyophilized. The resulting 90.5 g of dried leaf material was extracted with 80% methanol, and saponins were partitioned from the aqueous methanolic extract using butanol. Flash column chromatography (FCC) was performed and fractions containing peak (**10**) were purified using preparative HPLC. The structure of peak (**10**) was elucidated by Dr. Amr El-Demerdash by ^{1}H NMR and

was determined to be QA-TriF(Q)RXX (Appendix B.3). Although the structure of peak (**9**) was not elucidated due to limited purified material, the MS/MS fragmentation suggested the addition of a deoxyhexose moiety to the C-28 sugar chain (Fig. 5.2.14D). The identity of peak (**9**) may be a result of the addition of a different deoxyhexose unit (D-fucose of L-rhamnose), or the addition of D-quinovose to a different part of the C-28 sugar chain. Future experiments should involve the purification of peak (**9**) to resolve its exact chemical structure. Nonetheless, the structural identification of peak (**10**) confirmed that SoGH1 can catalyse the addition of D-quinovose to QA-TriFRXX.

GH1 transglycosidases (TGs) are emerging as a new class of sugar transferases with roles in plant specialized metabolism. These enzymes use acyl sugars rather than nucleotide sugars as the sugar donors (Cairns *et al.*, 2015). So far, all characterized GH1 TGs are involved in the transfer of glucose (Matsuba *et al.*, 2010; Miyahara *et al.*, 2012; Miyahara *et al.*, 2013; Nishizaki *et al.*, 2013; Miyahara *et al.*, 2014; Luang *et al.*, 2013; Orme *et al.*, 2019), exception of one galactosyltransferase (Moellering, Muthan and Benning, 2010). Furthermore, GH1 enzymes typically have N-terminal signal peptides (Xu *et al.*, 2004) and all reported GH1 natural product sugar transferases contain signal peptides predicted to target the vacuole (Orme *et al.*, 2019). To investigate the potential localization of SoGH1, signal peptide analysis was performed using SignalP 5.0 (Almagro Armenteros *et al.*, 2019). Transit peptides are known to be highly variable in sequence and length; however, most N-terminal signal peptides are cleaved during or after translocation (Jarvis, 2008; Almagro Armenteros *et al.* 2019). SignalP is an algorithm-based tool that predicts N-terminal signal peptides based on the detection of cleavage sites by signal peptidases (Almagro Armenteros *et al.*, 2019). The full amino acid sequence of SoGH1 was submitted to SignalP using default parameters, which predicted a very low likelihood (score: 0.003) of presence of a signal peptide sequence (Fig. 5.2.15). Although this result suggested that SoGH1 may be a cytosolic protein, SignalP can only predict N-terminal signal peptides that direct the protein across the ER membrane and is unsuitable for the detection of transit peptides that act as import signals to mitochondria, plastids and vacuoles (Almagro Armenteros *et al.*, 2019). Thus, fluorescent marker-based experiments would provide more insight into the localization of SoGH1.

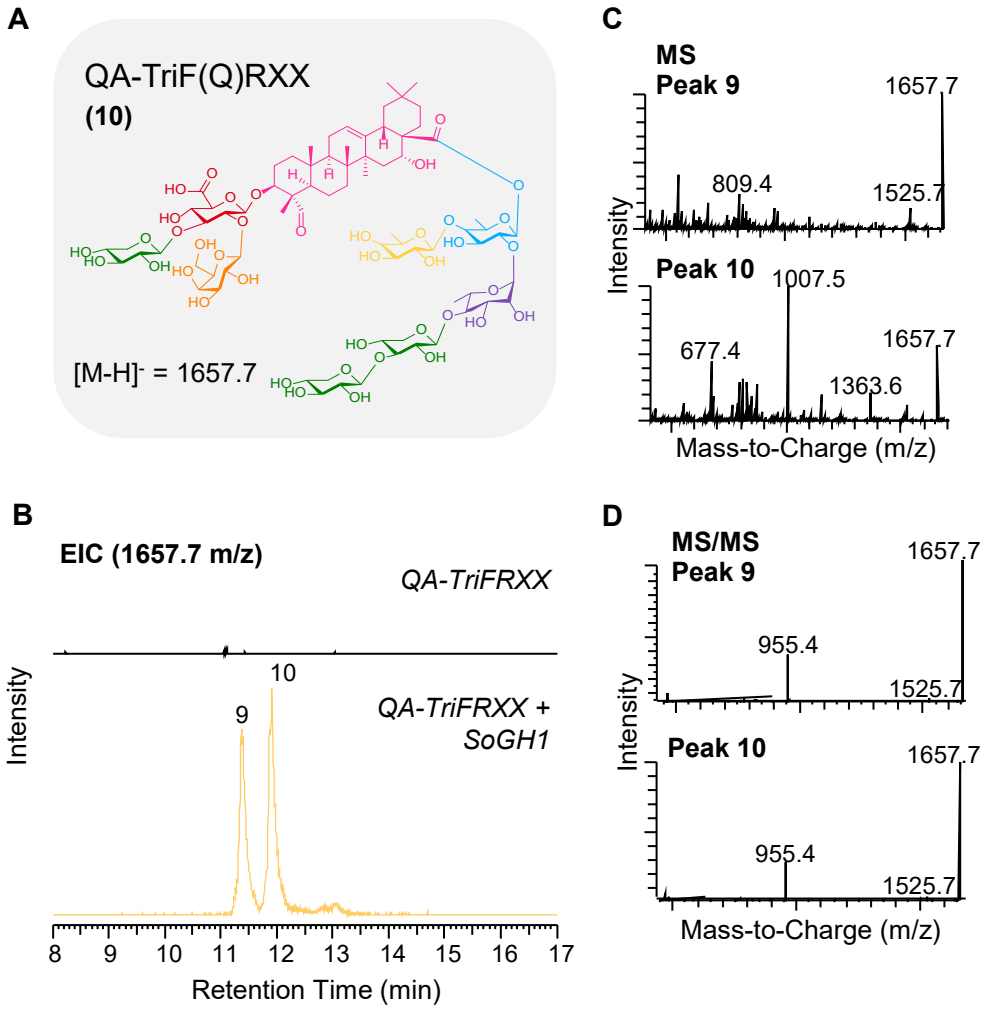


Figure 5.2.14. Transient expression of *SoGH1* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* containing expression constructs strains for *SobAS1*, *SoC28C16*, *SoC23*, *SoCSL1, SoC3Gal*, *SoC3Xyl*, *SoSDR1. SoC28Fu, SoC28Rha, SoC28Xyl1, SoC28Xyl2,* and *SoGH1*. Leaves were harvested five days after infiltration and leaf extracts were analysed using HPLC-MS. **(A)** Structure of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)- [*β*-D-quinovopyranosyl-(1→4)]-*β*-D-fucopyranosyl ester}-quillaic acid (**10**, QA-TriF(Q)RXX), the expected product of SoGH1 when acting in combination with QA-TriFRXX biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1 + SoC3Gal + SoC3Xyl + SoSDR1 + SoC28Fu + SoC28Rha + SoC28Xyl1 + SoC28Xyl2). **(B)** Extracted ion chromatogram (EIC) at *m/z* 1657.7. **(C)** Mass spectra (MS) of peaks (**9**, **10**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peaks (**9**, **10**). Extract from *N. benthamiana* leaves co-expressing *S. officinalis* genes required to produce QA-TriFRXX only was used as a negative control. Peak (**10**), identified as QA-TriF(Q)RXX by ^{1}H NMR. Peak (**9**) may be a result of D-quinovose attached at different position on the C-28 sugar chain, or the attachment of a different deoxyhexose sugar.

Protein type	Signal Peptide (Sec/SPI)	Other
Likelihood	0.003	0.997

Download: PNG / EPS / Tabular

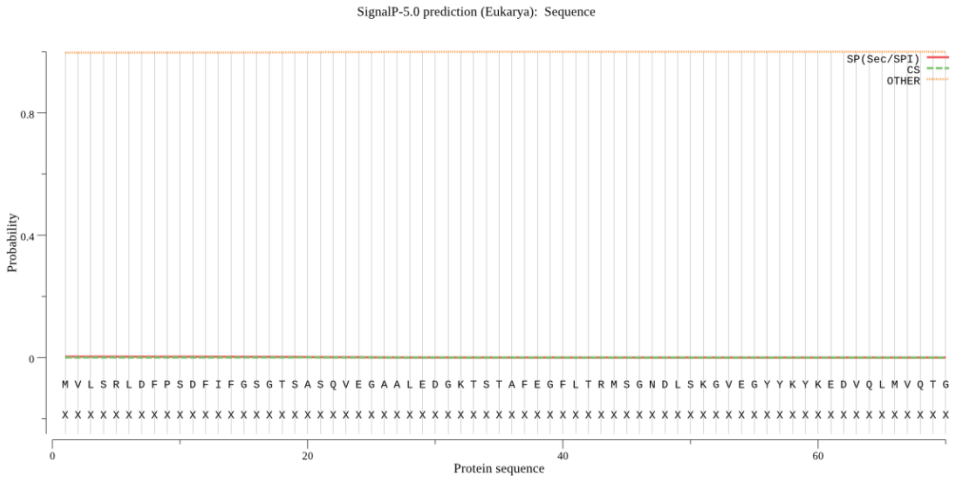


Figure 5.2.15. SoGH1 analyzed by SignalP 5.0. The first 70 amino acids of SoGH1 are not predicted to contain a signal peptide, indicated by a low SignalP score (0.003). The score shown is the discrimination score (D-score) which is a weighted average of the mean S-score (signal peptide score) and the maximum Y-score (combined cleavage site score). Low SignalP score represents the low likelihood of a signal peptide.

Complete biosynthetic pathway to SpB

The final remaining step left to complete the biosynthetic pathway to saponarioside B is the acetylation of the D-quinovose moiety of QA-TriF(Q)RXX. Although it is possible that 4-*O*-acetylquinovose is directly attached to QA-TriFRXX rather than the consecutive addition of D-quinovose followed by acetylation, non-acetylated quinovose-containing saponins have been reported from *S. officinalis* before (Takahashi *et al.* 2023) (illustrated in Chapter 3). Thus, I hypothesize that the consecutive addition of D-quinovose and acetyl-group may be more likely.

The candidate ATs (both BAHD and SCPL) identified in Section 5.2.3 were screened for their potential acetylation activity towards QA-TriF(Q)RXX (Figs. C.3.8 and C.3.9). Of the 10 BAHD AT and 5 SCPL AT candidates tested, only one candidate (*BAHD6*, hereafter referred to as *SoBAHD1*), was able to acetylate the D-quinovose moiety of QA-TriF(Q)RXX. When *SoBAHD1* was co-expressed with QA-TriF(Q)FRXX producing genes (*SobAS1* + *SoC28C16* + *SoC23* + *SoCSL1* + *SoC3Gal* + *SoC3Xyl* + *SoSDR1* + *SoC28Fu* + *SoC28Rha* + *SoC28Xyl1* + *SoC28Xyl2* + *SoGH1*), two new product peaks, (**11**) and (**12**), were observed (Fig. 5.2.16). Both peaks displayed *m/z* values of 1699.7, corresponding to the expected $[M-H]^-$ of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-[*β*-D-4-*O*-acetylquinovopyranosyl-(1→4)]-*β*-D-fucopyranosyl ester}-quillaic acid (**12**, QA-TriF(Q-Ac)FRXX/saponarioside B). Furthermore, tandem MS analysis revealed the same fragmentation pattern for both peaks (**11**) and (**12**), which also corresponded to the MS/MS of the authentic SpB standard. The major fragment ions were *m/z* 1657.7 (expected $[M-H]^-$ of QA-TriF(Q)RXX) and *m/z* 955.5 (expected $[M-H]^-$ of QA-Tri), suggesting the fragmentation of an acetyl moiety, followed by the loss of the entire C-28 sugar chain (Fig. 5.2.16D). However, only the RT of peak (**11**) produced by transient expression of *SoBADH1* in *N. benthamiana* corresponded to the RT of SpB standard (Fig. 5.2.16E). Based on these results, peak (**11**) was identified as SpB, and SoBAHD1 is likely an acetyltransferase that transfers an acetyl-group to D-quinovose of QA-TriF(Q)RXX, resulting in the formation of saponarioside B. Although the identity of

peak (**12**) is unknown, based on the MS/MS fragmentation pattern, peak (**12**) may be QA-TriF(Q)RXX acetylated at a different position of the C-28 sugar chain.

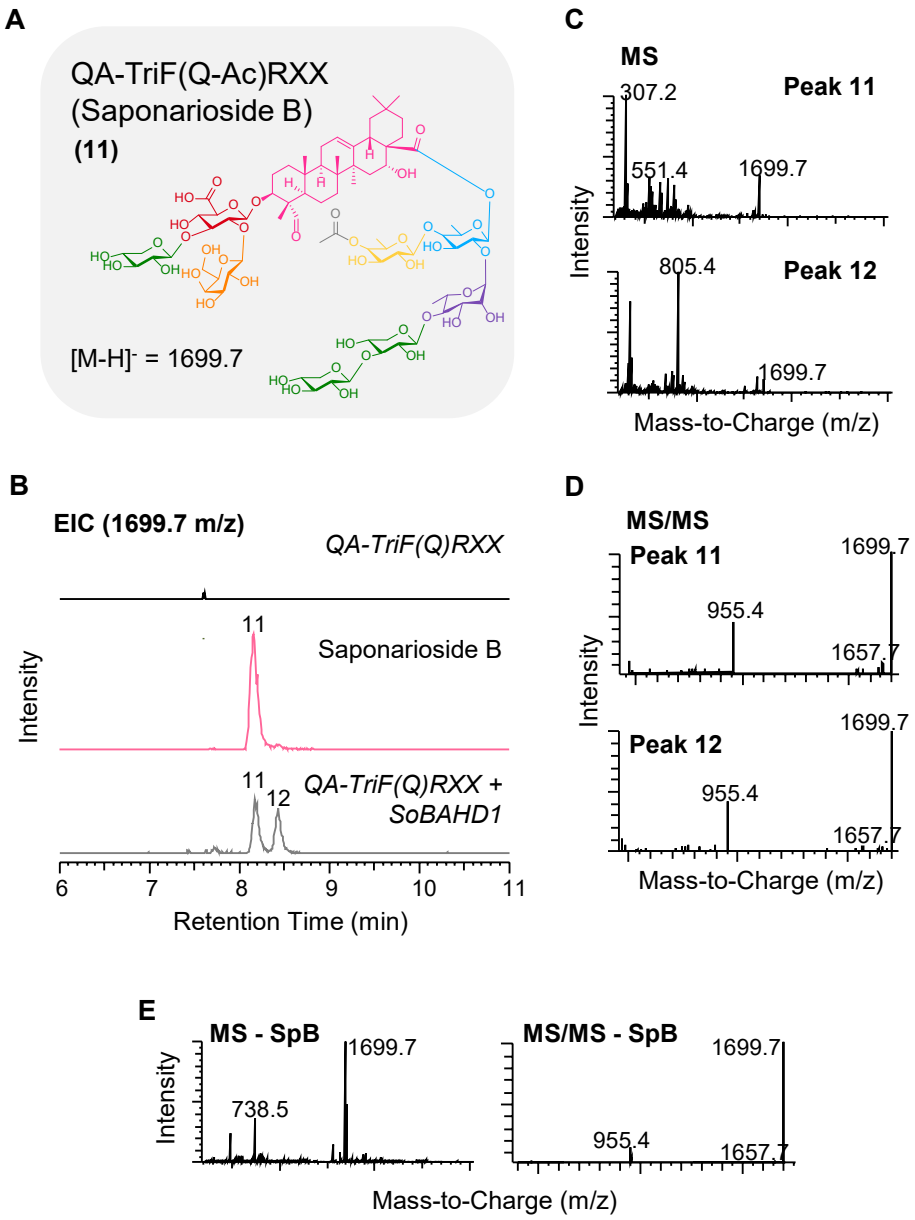


Figure 5.2.16. Transient expression of *SoBAHD1* in *N. benthamiana*. Leaves were infiltrated with *A. tumefaciens* strains containing various expression constructs. **(A)** Structure of saponarioside B (**11**, QA-TriF(Q-Ac)RXX), the expected product of SoBAHD1 when acting in combination with QA-TriF(Q)RXX biosynthetic enzymes (SobAS1 + SoC28C16 + SoC23 + SoCSL1 + SoC3Gal + SoC3Xyl + SoSDR1 + SoC28Fu + SoC28Rha + SoC28Xyl1 + SoC28Xyl2 + SoGH1). **(B)** Extracted ion chromatogram (EIC) at *m/z* 1699.7. **(C)** Mass spectra (MS) of peaks (**11**, **12**). **(D)** Tandem mass spectra (MS/MS) showing fragmentation pattern of peaks (**11**, **12**). Extract from *N. benthamiana* leaves co-expressing *S. officinalis* genes required to produce QA-TriF(Q)RXX only was used as a negative control. Peak (**11**), identified as SpB by comparison with authentic SpB standard. Peak (**12**) may be a QA-TriF(Q)RXX acetylated at a different position than SpB. The full TIC range is available in Fig. C.3.8.

With the identification of SoBAHD1, a complete set of enzymes capable of saponarioside B biosynthesis had been identified. The expression of the genes encoding these pathway enzymes all show high correlation with the expression pattern of *SobAS1*, as indicated by their high PCC values (Tables 5.2.2-5.2.6). However, when the positions of these genes in the *S. officinalis* genome was investigated, none of them displayed close physical proximity (Fig. 5.2.17). Thus, the saponarioside biosynthetic genes identified are not organized in biosynthetic gene clusters (BGCs), which is in correspondence with the plantiSMASH output presented in Chapter 4. This contrasts with the situation for the QS-saponin biosynthetic pathway genes, which are partially clustered in the *Q. saponaria* genome (Reed *et al.*, 2023).

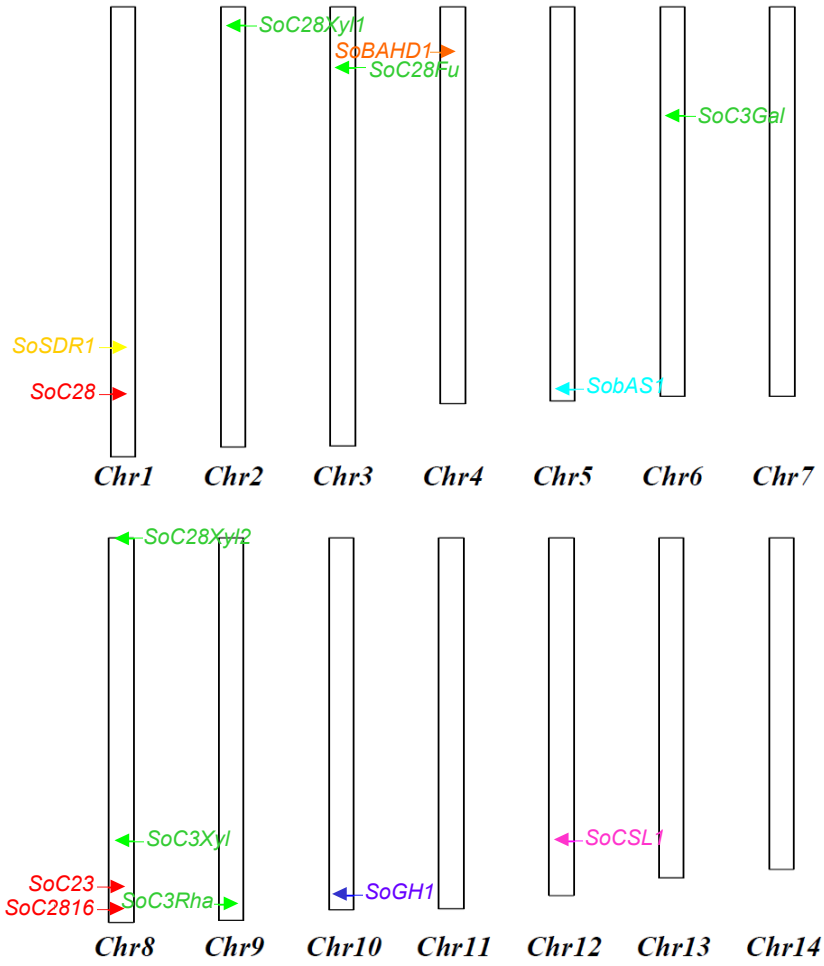


Figure 5.2.17. *S. officinalis* chromosome map showing the physical location of the characterized saponarioside biosynthetic genes. OSC, sky blue; CYP450s, red; UGTs, light green; CSL, pink; SDR, yellow; GH, purple; BAHD AT, orange.

5.3 Conclusion

Here, a set of enzymes with the ability to produce SpB have been identified. The enzymes characterized in this chapter include a OSC (SobAS1), 3 CYPs (SoC28, SoC28C16, SoC23), a CSL (SoCSL1), 6 UGTs (SoC3Gal, SoC3Xyl, SoC28Fu, SoC28Rha, SoC28Xyl1, SoC28Xyl2), a GH1 TG (SoGH1) and a BAHD AT (SoBAHD1). Although these genes are not organized as BGCs in the soapwort genome, they all show high co-expression (PCC > 0.88) with *SobAS1*, the gene encoding the first likely committed step in saponarioside biosynthesis, highlighting the power of co-expression analysis for pathway discovery. One further GT (as yet uncharacterized) is expected to complete the pathway to saponarioside A. The identification of these genes encoding enzymes capable of saponarioside biosynthesis can open-up opportunities for metabolic engineering of soapwort saponins in alternative systems, which may allow for large-scale production and biochemical studies of these understudied saponins.

6

General discussion

Soapwort (*Saponaria officinalis*) is an attractive flowering plant in the 'pinks' family (Caryophyllaceae) with a long history of use in human civilization. The cleansing soap-like properties of soapwort extracts have been exploited as a detergent, medicine, and food additive over thousands of years (Mitich, 1990). Soapwort extracts are still used in laundry detergent, cosmetics and in herbal medicine, and are an essential component of Turkish halvah. The bioactivities of soapwort extracts depend on the high amounts of triterpene saponins present in the extracts; however, the properties of the individual saponin components are poorly understood (Goral, Jurek and Wojciechowski, 2018). Purification of individual saponins from soapwort extract is a labour-intensive process as many of these compounds are similar in structure and are present in complex mixtures (Jia *et al.*, 2002). Soapwort saponins are often highly decorated with sugar chains, hindering chemical synthesis of these complex compounds at a commercial scale (Lambert, Faizal and Geelen, 2011). Thus, understanding the biosynthesis soapwort saponins may enable further research into the bioactivities of these potentially high value saponins. The aim of this project was to investigate the biosynthesis of the major saponins reported from soapwort, saponariosides A and B, and to elucidate the genes and enzymes involved in the biosynthetic pathway.

6.1 Biological roles of saponariosides

Detailed profiling of SpA and SpB in six different soapwort organs (flower, flower bud, young leaf, old leaf, stem, and root) revealed that SpA was most abundant in the flowers, while SpB accumulated mainly in the leaves (Chapter 3, Figure 3.2.4). Interestingly, SpA and SpB differ only by an additional D-xylose moiety in the C-28

sugar chain of SpA. Although the biological roles of saponariosides are not yet known, saponins are generally accepted as plant defence molecules through their ability to cause membrane perturbation (Augustin *et al.*, 2011). The combination of the hydrophobic aglycone and hydrophilic sugars allow saponins to incorporate into the biological membrane by forming complexes with membrane sterols, leading to membrane disruption and pore formation (Osbourn, 1996a). However, the degree of membrane permeabilizing ability of saponins is affected by numerous factors, such as the type of the aglycone and the characteristics (linkage, length and composition) of the saccharide chains (Augustin *et al.*, 2011). Furthermore, as glycosylation increases the solubility and chemical stability of a compound (Louveau and Osbourn, 2019), the additional D-xylose unit in SpA may serve a tactical purpose to store the bioactive saponariosides in their least active form. In addition to the levels of SpA and SpB, the total content of quillaic acid-based saponins were analysed by quantifying QA-Tri, the common quillaic acid prosapogenin, following saponification. Similar to the combined profiles of SpA and SpB, the highest level of quillaic acid-based saponins was also found in the flowers. Soapwort flowers may accumulate high levels of saponariosides to protect the fragile reproductive organs. Although the roots of soapwort are well known to contain high amounts of saponins, flowers have often been neglected in many soapwort metabolite studies, most likely due to their highly seasonal availability. The metabolite analysis in this study was performed using samples harvested in two different time points, in July and November to represent summer and winter months, respectively. However, these two time points are not sufficient to draw any conclusions regarding the time point of saponin biosynthesis in soapwort, especially if biosynthesis occurred before flowering. For example, the biosynthesis of montbretin A, a complex flavonol glycoside produced by *Crocosmia x crocosmiiflora*, occurs strictly during the few summer months when corm development begins (Irmisch *et al.*, 2018). Thus, regular sampling of soapwort plants for metabolite analysis may provide a better understanding of saponarioside biosynthesis during plant development and provide further insight into the biological roles of saponariosides.

Beyond the conventionally known bioactivities of saponins, quillaic acid-based soapwort saponins have been observed to augment the cytotoxicity of saporin, a type I ribosome inactivating protein (RIP) found in soapwort (Gilabert-Oriol *et al.*, 2016).

Like other type I RIPs, saporin by itself displays low cytotoxicity as type I RIPs lack the natural cell-binding B-domain (Gilabert-Oriol *et al.*, 2016). Interestingly, soapwort saponins drastically enhance the cytotoxicity of saporin by initiating endosomal escape of the internalized saporins into the cytosol where they exhibit their toxicity, rather than through membrane perturbation (Weng *et al.*, 2008). This synergistic defence mechanism is hypothesized to have evolved to deter herbivore feeding. Saporin is known to accumulate in the seeds, leaves and roots of soapwort (Ferreras *et al.*, 1993). If animals ingest plant material that is high in saporin but low in saponins such as leaves, the toxicity of saporin may be negligible since on its own, saporin is impermeable to the cell membrane (Gilabert-Oriol *et al.*, 2016; Bolshakov *et al.*, 2020). However, if animals were to feed on the entire plant, including high saponin containing parts such as flowers and roots, the resulting toxicity of saporin might be severe (Gilabert-Oriol *et al.*, 2016).

Thus, although SpA and SpB are closely related, the difference in the additional D-xylose unit in SpA, and the differential accumulation pattern of SpA and SpB may suggest different roles of these two saponins. Most anti-bacterial or anti-fungal assays require the usage of purified compounds, which involves labour-intensive and time-consuming processes. Using the biosynthetic pathway knowledge presented in this work, extracts of *N. benthamiana* leaves transiently expressing combinations of the newly discovered soapwort genes may be used instead of purified compounds. Furthermore, anti-insect effects of saponariosides may be assessed by a no-choice feeding assay. In these experiments, tobacco hornworm (*Manduca sexta*) larvae are grown and fed with *N. benthamiana* leaves accumulating the metabolite of interest, produced by the activity of the transiently expressed enzymes. Such insect-feeding assays have been previously used to assess the antifeedant effects of monodesmodic saponins (Liu *et al.*, 2019). However, as the biosynthetic knowledge presented in this thesis reaches only to SpB, to compare the bioactivities between SpA and SpB, an additional biosynthetic step still needs to be identified for the production of SpA.

6.2 Enzymes involved in saponarioside biosynthesis

Following the metabolite profiling, soapwort organs with varying levels of SpA and SpB were used in RNA-Seq analysis to search for candidate saponarioside biosynthetic genes. In addition to these transcriptomic resources, a high-quality *S.*

officinalis genome assembly was also generated to aid in gene discovery and to examine the physical locations of the biosynthetic genes. Sequencing of the soapwort genome confirmed the genome size of soapwort to be 2.2 Gb, consistent with previously reported C-values for soapwort (Di Bucchianico *et al.*, 2008; Pustahija *et al.*, 2013). Prior to these developments, the only publicly available soapwort sequence resource was a transcriptome from the 1000 Plants (1KP) project, which contains a single dataset generated from a pool of different organs. The newly generated soapwort sequence resources were used to mine for candidate pathway genes using different approaches, including searching for candidates based on expected enzyme families (SobAS1) and high protein homology to *Q. saponaria* enzymes (SoC28 and SoC28C16). However, the most successful approach relied on the degree of co-expression of the candidate genes to *SobAS1*, the gene involved in the first committed step of saponarioside biosynthesis, thus highlighting the power of co-expression analysis in pathway elucidation.

Overall, a total of 13 pathway genes have been elucidated in this work which allowed the synthesis of SpB in *N. benthamiana* (Fig. 6.2.1). As hypothesized, the biosynthesis of quillaic acid from 2,3-oxidosqualene required the activity of an OSC (SobAS1) and three CYPs (SoC28, SoC28C16 and SoC23). Although SoC28 and SoC28C16 had partial overlapping activities as a C-28 oxidase, the co-expression of both CYPs in *N. benthamiana* did not translate into the increased accumulation of echinocystic acid. In fact, SoC28C16 functioned as a sufficient C-28 oxidase by itself. This may be a limitation of the transient expression system in *N. benthamiana*, which may have internal feed-back mechanisms to limit the accumulation of echinocystic acid. Another possibility is that SoC28 may be a main player in the biosynthesis of oleanolic acid-based compounds instead of quillaic acid-based saponins. Silencing of *SoC28* via virus-induced gene silencing (VIGS) or agrobacterium-mediated hairy root transformation may provide further insight into the role of *SoC28* (further discussed in Section 6.3). Although there are no reports of VIGS performed in soapwort, generation of hairy roots from soapwort has been reported before (Hedayati *et al.*, 2022), as well as in a closely related species, *Saponaria vaccaria* (Schmidt *et al.*, 2007).

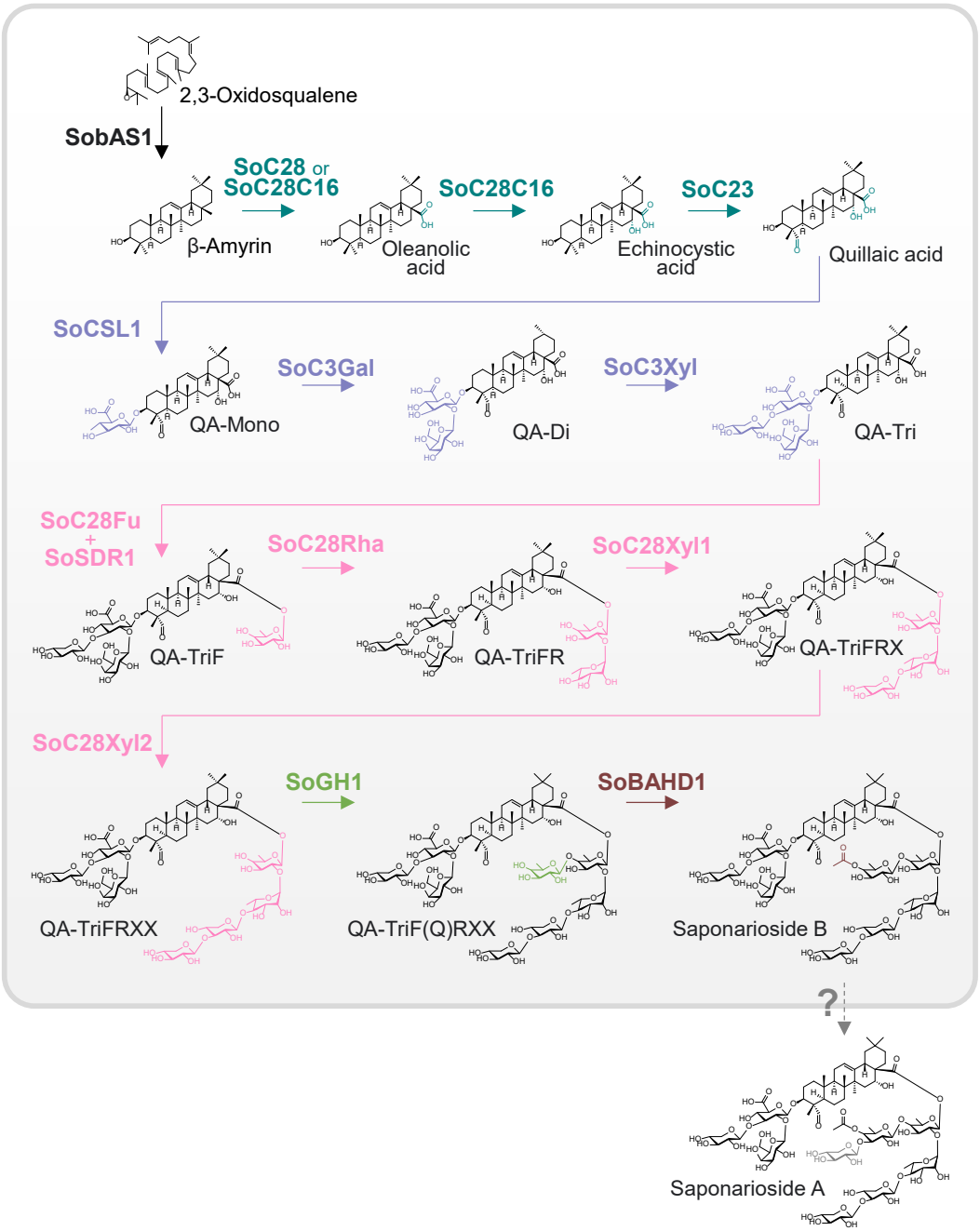


Figure 6.2.1. Predicted biosynthetic pathway for SpA and SpB. Saponarioside biosynthetic identified so far are labelled. The arrows represent accumulation of metabolite products after each addition of associated enzyme rather than specifying a biosynthetic order *in planta*. Enzymes involved in the oxidation of β-amyrin are labelled in teal; those involved in the building of the C-3 sugar chain are labelled in purple; those involved in the biosynthesis of the linear C-28 sugar chain are labelled in pink; SoGH1 is labelled in light green; SoBAHD1 in brown. SoSDR1 was previously identified by Dr. James Reed. Only one step to saponarioside A remains unidentified. Full sequences are available in Appendix E.

6.2.1 Cellulose synthase-derived enzyme from soapwort

Following the generation of the quillaic acid aglycone, the addition of the C-3 sugar chain was catalysed by SoCSL1, SoC3Gal and SoC3Xyl. Although SoC3Gal and SoC3Xyl are UGTs which are classically known to be involved in glycosylation of triterpenes, SoCSL1 is an interesting GT as it is a member of the CSyGT subfamily within the cellulose synthase like (CSL) family. Other members of CSyGT have recently been reported to be involved in the 3-*O*-glucuronidation of triterpene aglycones such as medicagenic acid in *S. oleracea* and quillaic acid in *Q. saponaria* (Jozwiak *et al.*, 2020; Reed *et al.*, 2023). Additionally, plants such as soybean (*Glycine max*), alfalfa (*Medicago sativa*), lotus (*Lotus japonicus*), liquorice (*Glycyrrhiza uralensis*), beetroot (*Beta vulgaris*) and quinoa (*Chenopodium quinoa*) also produce saponins with glucuronic acid attached at the C-3 position of the saponin aglycone. CSyGT candidates from each of these species have been identified and have been observed to be implicated in saponin biosynthesis (Jozwiak *et al.*, 2020; Chung *et al.*, 2020). Similarly to *SoCSL1*, CSyGTs with expression profiles (*SOAP5*, *QsCSL*, *GmCSyGT1, GuCSyGT, LjCSyGT*) were all highly co-expressed with genes involved in saponin biosynthesis in their respective plant species (Jozwiak *et al.*, 2020; Chung *et al.*, 2020; Reed *et al.*, 2023). While classical cellulose synthase (CesA) members are reported to localize to the plasma membrane, fluorescence tagging experiments have revealed SOAP5, GmCSyGT1, GuCSyGT and LjCSyGT to localize to the endoplasmic reticulum (Jozwiak *et al.*, 2020; Chung *et al.*, 2020). Based on their localization, CSyGTs are hypothesized to play a structural role in the formation of a potential saponin metabolon to facilitate the transport of ER localized CYP products to the cytosol for further UGT-mediated glycosylation (Jozwiak *et al.*, 2020; Chung *et al.*, 2020). To further investigate the role of SoCSL1 in *S. officinalis* saponin biosynthesis, future experiments involving fluorescence tagging of SoCSL1 may verify the cellular localization of SoCSL1. Furthermore, Förster resonance energy transfer (FRET) based experiments may provide insight into the potential interactions between SoCSL1 and other enzymes involved in saponarioside biosynthesis.

6.2.2 D-Fucosylation in plant specialized metabolism

Additional four UGTs (*SoC28Fu*, *SoC28Rha*, *SoC28Xyl1*, and *SoC28Xyl2*) involved in the elongation of the C-28 linear sugar chain were again discovered through co-

expression analysis with *SobAS1*. The first sugar moiety of the C-28 sugar chain is D-fucose. Until recently, the mechanism of fucosylation in plants has been elusive. The first report of a plant enzyme involved in fucosylation is a UGT (SOAP6) identified from *S. oleracea* (Jozwiak *et al.*, 2020). SOAP6 is found to be involved in the addition of a D-fucose to the C-28 carboxylic acid of medicagenic acid 3-*O*-glucuronide in the yossoside biosynthetic pathway. However, when *SOAP6* is transiently expressed in *N. benthamiana*, it showed broad substrate specificity towards other UDP-sugars such as pentose, hexose and ketodeoxyhexose. The authors hypothesized that this was due to the low abundance of the likely primary substrate of SOAP6, UDP-α-D-fucose, in *N. benthamiana*. Shortly after, another UGT from *Q. saponaria*, QsC28Fu, was identified to be involved in the C-28 fucosylation of QA-Tri (**4**) (Reed *et al.*, 2023). The transient expression of *QsC28Fu* in *N. benthamiana* led to a minuscule amount of the fucosylated product and showed substrate affinity for ketodeoxyhexose. However, the additional activity of a short-chain dehydrogenase/reductase (SDR), *QsFucSyn*, increased the accumulation of this fucosylated product. Extensive *in vitro* experiments revealed that UDP-α-D-fucose is not likely relevant in the production of the D-fucose moiety found in QS-saponins. Rather, UDP-4-keto-6-deoxy-glucose (an intermediate in UDP-L-rhamnose biosynthesis) acts as the sugar donor for the transfer of 4-keto-6-deoxy-glucose by QsC28Fu to the triterpene backbone before being reduced *in situ* to D-fucose by QsFucSyn, which functions as a 4-ketoreductase (Reed *et al.*, 2023). During this study, a similar keto-reductase from soapwort, SoSDR1 was identified by searching for a soapwort SDR with highest amino acid identity with QsFucSyn. The discovery of SoSDR1 and SoC28Fu from soapwort illustrates that the above fucosylation mechanism is present in species beyond *Q. saponaria*, perhaps even more broadly across the Angiosperms. However, before investigating the source of D-fucose, the exact mechanism of SoSDR1 and SoC28Fu must be verified. Although *N. benthamiana* assays performed previously on SoSDR1 by Dr. James Reed (unpublished data) and results presented here suggests that SoC28Fu transfers UDP-4-keto-6-deoxy-glucose to the saponin substrate which is then reduced by SoSDR1 to D-fucose, enzyme assays using purified SoC28Fu and SoSDR1 would provide much better understanding of their activities. Subsequently, a search for homologues of SoC28Fu and SoSDR1 in plant species which are known to produce specialized metabolites containing D-fucose, such as *Spinach oleracea* (Caryophyllaceae)

(Jozwiak *et al.*, 2020), *Amaranthus caudatus* (Amaranthaceae) (Mroczek, 2015), and *Digitalis* spp. (Plantaginaceae) (Kreis and Müller-Uri, 2010), followed by enzyme characterization may provide insight into the distribution of this fucosylation mechanism across the Plant Kingdom.

6.2.3 Involvement of an unexpected enzyme family

After the attachment of D-fucose, the C-28 sugar chain was further elongated by the enzymatic activities of three classical UGTs, SoC28Rha, SoC28Xyl1, and SoC28Xyl2, which was as predicted in Section 5.1.1.. However, the next step, attachment of D-quinovose to the D-fucose moiety, was performed by an enzyme (SoGH1) belonging to an unexpected enzyme family. SoGH1 was discovered during re-examination (aided by Dr. Charlotte Owen) of all soapwort genes co-expressed with *SobAS1* and was annotated as an enzyme belonging to the glycoside hydrolase family 1 (GH1). GHs (also referred to as glycosidases) are widely distributed enzymes that are found in almost all domains of life. Generally, they catalyse the breakage of glycosidic bonds and serve in a broad range of biological functions, such as in degradation of complex sugars (cellulase), in anti-bacterial defence mechanism (lysozyme) and in pathogenesis (neuraminidase) (Shrivastava, 2020). GH enzymes hydrolyse glycosidic linkages through general acid catalysis (requiring a proton donor and a base/nucleophile), giving rise to either net inversion or retention of the anomeric configuration (Davies and Henrissat, 1995). Family 1 GH members are retaining enzymes, where hydrolysis is achieved by a double-displacement mechanism in two steps: 1. the glycosylation step and 2. the de-glycosylation step (Fig. 6.2.2) (Withers *et al.*, 1986). The first step involves the nucleophilic attack to the anomeric centre to form a glycosyl-enzyme intermediate. In the second step, this intermediate is hydrolysed through another nucleophilic attack, typically carried out by water (Vuong and Wilson, 2010). However, if the second step is intercepted by another acceptor (such as a sugar), the reaction leads to the formation of a new glycosidic linkage (Crout and Vic, 1998).

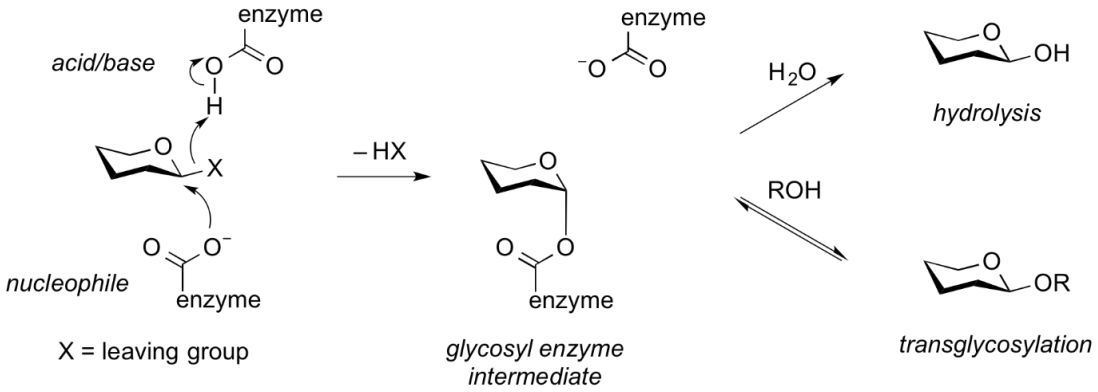


Figure 6.2.2. Generalized mechanism of a transglycosylase. Enzymatic cleavage of a substrate through a classical Koshland retaining mechanism results in formation of a glycosyl enzyme intermediate. This can partition to react with either water to cause hydrolysis (glycoside hydrolase activity) or to an alternative acceptor, often a sugar, to cause transglycosylation (transglycosylase activity). Figure and figure legend reproduced from (Williams, 2013).

Although most GHs favour hydrolase activity, transglycosidases (TGs) preferentially catalyse transglycosylation (Xu *et al.*, 2004). Unlike UGTs, these enzymes use acyl sugars rather than nucleotide sugars as sugar donors and are known to be vacuolar in their subcellular localization (Cairns *et al.*, 2015). Only a limited number of GH1 TGs have been reported so far, mainly involved in 7-*O*-glucosylation of anthocyanins (Sasaki *et al.*, 2014). However, the substrate specificity of GH1 TG is not limited to anthocyanins, as GH1 TG (Os9bglu31), involved in the transfer of acyl-glucose to phenylpropanoids, flavonoids and phytohormones, has been identified in *O. sativa* (Luang *et al.*, 2013). A GH1 TG from *A. strigosa* (AsGH1) involved in the glucosylation of avenacin A-1 has also been reported (Orme *et al.*, 2019). All GH1 TGs characterized so far are involved in the transfer of glucose (Matsuba *et al.*, 2010; Miyahara *et al.*, 2012; Miyahara *et al.*, 2013; Nishizaki *et al.*, 2013; Miyahara *et al.*, 2014; Luang *et al.*, 2013; Orme *et al.*, 2019), except for one galactosyltransferase (Moellering, Muthan and Benning, 2010). These characterized enzymes are either predicted or confirmed to be vacuolar in localization or localized to the chloroplast membrane, in case of the latter enzyme (Matsuba *et al.*, 2010; Miyahara *et al.*, 2012; Miyahara *et al.*, 2013; Nishizaki *et al.*, 2013; Miyahara *et al.*, 2014; Luang *et al.*, 2013; Orme *et al.*, 2019; Moellering, Muthan and Benning, 2010).

SoGH1 is a novel GH1 as it seems to be associated with the transfer of quinovose and is not predicted to contain the typical N-terminal signal peptide generally found in other members of the GH1 TG family. In fact, SoGH1 may be the first plant enzyme reported to be associated with quinovosyltransferase activity. Although D-quinovose

is commonly found in specialized metabolism of echinoderms such as starfish and sea cucumbers (Stonik and Elyakov, 1988), it is an uncommon sugar in plants. However, several triterpene saponins isolated from the Caryophyllaceae family are known to contain D-quinovose (Putieva *et al.*, 1977). Examples include acanthophyllosides B and C isolated from *Acanthophyllum gypsophiloides* (Putieva *et al.*, 1977), silenorubicunosides isolated from *Silene rubicunda* (Fu *et al.*, 2005), and several triterpenes from *Gypsophila* spp. (Elbandy, Miyamoto and Lacaille-Dubois, 2007; Chen, Luo and Kong, 2011; Pertuit *et al.*, 2014). Triterpene saponins with D-quinovose outside the Caryophyllaceae family have also been reported, such as aquilarinensides isolated from *Aquilaria sinensis* (Sun *et al.*, 2014) and avicin D isolated from *Acacia victoriae* (Jayatilake *et al.*, 2003). However, genes encoding enzymes associated with D-quinovose in plants remain elusive and the origin of D-quinovose is currently unknown. Sugar structural diversity is usually generated at the sugar nucleotide level. For example, TDP-D-quinovose is produced by the reduction of TDP-4-keto-6-deoxy-D-glucose in *Streptomyces venezuelae* (Han *et al.*, 2011). Furthermore, previous sugar nucleotide profiling of *N. benthamiana* has reported UDP-rhamnose as the only detectable UDP-deoxyhexose in this plant (Reed *et al.*, 2023; Pabst *et al.*, 2010). Given that D-fucose and D-quinovose are C-4 epimers, biosynthesis of D-quinovose may be similar to the mechanism reported for D-fucose (described in Section 6.2.3), requiring UDP-4-keto-6-deoxy-D-glucose as a sugar donor before being reduced *in situ* by an as-yet unidentified SDR. This reduction could occur once 4-keto-6-deoxy-D-glucose is transferred to the relevant acyl acceptor to form acyl-D-quinovose which is then utilized by SoGH1. Alternatively, this reduction may occur as the terminal step, with acyl-4-keto-6-deoxy-D-glucose serving as the donor for SoGH1, with reduction of 4-keto-6-deoxy-D-glucose to D-quinovose following attachment to a saponin acceptor. Searching for candidate SDRs in soapwort that are highly co-expressed with *SoGH1*, instead of *SobAS1*, may lead to new candidates potentially involved in the quinovosylation step. Once identified, protein purification of this new candidate and SoGH1, followed by enzyme characterization through series of activity assays, may provide further insight into the mechanism of SoGH1 such as substrate and donor specificities. Furthermore, resolving the subcellular localization of SoGH1 may provide more insight into the location of quinovosylation. SignalP analysis of SoGH1 protein sequence did not predict the presence of any N-terminal signal peptide; furthermore, the next step in the saponarioside pathway, the acetylation of D-

quinovose, is carried out by a BAHD AT (SoBAHD1), where most of the characterized members so far are reported to be localized in the cytosol (Moghe *et al.*, 2023). Furthermore, beyond the work presented in this thesis, fluorescent tagging revealed the cytoplasmic localization of SoBAHD1 (Jo *et al.*, 2024). Although these results suggest the cytoplasmic localization of SoGH1, other cellular targets such as the chloroplast, mitochondrion or vacuole cannot be ruled out, especially due to the limitation of the SignalP analysis tool. Thus, instead of relying on *in silico* data, fluorescence tagging of SoGH1 followed by confocal microscopic analysis would provide more robust data to address this question.

6.3 Future research

In addition to those mentioned in the above sections, the results presented in this work have revealed several exciting areas of future research and are discussed below.

Although the metabolite analysis performed in this thesis has been thorough, the focus was restricted to the accumulation of SpA and SpB. As presented in Chapter 3, more than 40 different saponins with varying aglycone cores have been identified from soapwort. As these saponins are highly decorated and are similar in structure, identifying and isolating these complex saponins without a standard is a difficult task. To exacerbate this, soapwort saponins are currently not commercially available as standards. However, the discovery of saponarioside biosynthetic enzymes presented in this work may open new avenues to produce authentic standards. The Osbourn group maintains a toolkit of triterpene biosynthetic enzymes which includes a wide variety of well-characterized enzymes across many organisms that function in the biosynthesis of simple to complex triterpenes (Reed *et al.*, 2017). The newly discovered set of soapwort genes could be expressed transiently in *N. benthamiana* in combination with genes in the triterpene toolkit to generate soapwort saponin standards. After generating a wide range of soapwort saponins, targeted and untargeted metabolite analysis of soapwort can be carried out by using mass spectrometry-based analysis tools such as Mzmine, which allows for the processing and visualization of mass spectrometry data through molecular profiling (Schmid *et al.*, 2023). A broader metabolite analysis such as this could provide an in-depth insight into the overall saponin biosynthesis in soapwort.

The expression levels of the newly discovered saponarioside biosynthetic genes all showed highest expression in the flower and flower buds, which are also the highest accumulating sites of quillaic acid-based saponins. These results may suggest that the flower and flower buds are likely the major sites of saponarioside biosynthesis. However, the cellular location of saponarioside biosynthesis remains unknown. As saponins are membrane permeabilizing compounds, to avoid autotoxicity, plants often sequester these metabolites in organelles like the vacuole, which is the case for the anti-microbial agent avenacin A-1 in oat (Osbourn, 1996b). Moreover, the site of saponin biosynthesis and accumulation may be different. Fluorescence-based localization analysis of saponarioside biosynthetic enzymes may provide further insight into the cellular location of saponarioside biosynthesis. Additionally, matrix-assisted laser desorption ionization (MALDI) imaging of different soapwort organs could provide visualization of the spatial distribution of SpA and SpB.

Furthermore, the expression profiles of saponarioside biosynthetic genes identified so far show high co-expression with *SobAS1*, suggesting that these genes are involved in the same metabolic pathway. Although the enzymes encoded by these newly discovered genes can carry out various steps in saponarioside biosynthesis, their true *in planta* roles may be different. As mentioned previously, reverse genetic approaches such as gene silencing via VIGS and hairy root transformation of *S. officinalis* may provide further insight into the roles of these genes in their native plant, rather than in an alternative host, in this case *N. benthamiana*. If soapwort genes discovered here are indeed involved in saponarioside biosynthesis, silencing the expression of these genes in soapwort should result in lower production and accumulation of saponariosides, thus providing 'in planta' evidence that these genes encode the biosynthetic enzymes involved in SpA and SpB biosynthesis in *S. officianalis*.

Interestingly, none of the saponarioside biosynthetic genes are found in close physical proximity in the soapwort genome. In contrast, QS-saponin biosynthetic genes in *Q. saponaria* are partially clustered (Reed *et al.*, 2023). Despite their overlapping biochemical functions, the enzymes involved in the shared biosynthetic pathway of saponariosides and QS-saponins overall have low amino acid sequence identity, except for those involved in the very early stage of the pathway. *S. officinalis* and *Q. saponaria* are phylogenetically distant species, yet both produce structurally related triterpene saponins. Although it is tempting to speculate that the two pathways may

have arisen by convergent evolution, caution must be exercised in making this assumption because of the challenges of interpreting the ancestral origins of the cognate pathway genes, given the taxonomic distance between the two species. Regardless, convergent evolution in plant specialized metabolism is not uncommon. For example, the biosynthesis of stilbenes, a group of anti-fungal compounds, occurs sporadically in unrelated genera such as *Pinus*, *Arachis*, and *Vitis* (Pichersky and Lewinsohn, 2011). Additionally, the biosynthesis of pyrrolizidine alkaloids, a group of anti-herbivory compounds, occurs in several distant plant families such as the Asteraceae, Boraginaceae, Apocynaceae, Fabaceae and Orchidaceae (Ober and Kaltenegger, 2009). However, compared to saponariosides and QS-saponins, these compounds are less decorated and thus much simpler in their chemical structure. Furthermore, as introduced in Chapter 1, quillaic acid-based saponins are also found in other families of the Caryophyllales order and are not limited to *S. officinalis*.

Many triterpenoid saponins have been isolated from members within the Caryophyllales order, for example, *Dianthus caryophyllus* and *Gypsophila paniculata* from the family Caryophyllaceae (Böttger and Melzig, 2011), and *Beta vulgaris* and *Spinacia oleracea* from the family Amaranthaceae (Mroczek, 2015). Interestingly, buckwheat (*Fagopyrum esculentum*), a member of the family Polygonaceae, is suspected to produce only non-glycosylated triterpenoids with no isolated saponins reported from this species yet (Jing *et al.*, 2015; Raguindin *et al.*, 2021). With massive developments in sequencing technologies, more and more plant genomes are becoming available, including many in the Caryophyllales order, including the species listed above (a full list is given in Chapter 4). As such, bioinformatic tools such as OrthoFinder (Emms and Kelly, 2019) and GENESPACE (Lovell *et al.*, 2022) have been developed to aid in orthogroup identification and MCscan (Python version), a package from JCVI utility libraries (github.com/tanghaibao/jcvi) can be used to investigate synteny between different genomes. The available Caryophyllales genomes can be used in synteny analysis with the newly generated soapwort genome to examine possible syntenic regions across different Caryophyllales members. Furthermore, investigating potential orthologous saponarioside biosynthetic genes in other Caryophyllales species may provide insight into the evolution of the biosynthesis of saponarioside-like compounds.

6.4 Concluding remarks

Saponins are often implicated in plant-pathogen defence as they display toxicity against microbes, fungi, insects, and other pests (Osbourn, 1996a; Sparg, Light and Van Staden, 2004). Additionally, these compounds exhibit a wide range of pharmacological activities, including anti-inflammatory, anti-cancerogenic, and adjuvant activities (Augustin *et al.*, 2011). Thus, the potential applications of soapwort saponins remain vastly untapped. Metabolic engineering of soapwort saponins may lead to sustainable large-scale production for in-depth biological studies of their biological properties. However, such a strategy depends on knowledge of soapwort saponin biosynthesis, which prior to this work was unknown.

This project has made major progress into understanding saponarioside biosynthesis in *Saponaria officinalis*. Detailed targeted metabolite analysis in six different organs of soapwort has revealed that the accumulation patterns of SpA and SpB are different: highest accumulation of SpA was observed in the flower, and SpB preferentially accumulated in the leaves. Using this newfound knowledge, a comprehensive multi-organ RNA-Seq dataset for *S. officinalis* was generated and used for co-expression analysis. Additionally, this project also reports the first pseudochromosome-level genome assembly for *S. officinalis*, adding to the genomic database of the Caryophyllaceae family. The newly generated sequence resources were utilized to elucidate a total of 13 genes involved in the biosynthesis of SpA and SpB, completing the suite of genes required to biosynthesize SpB. Only one step remains to be discovered to complete the biosynthetic pathway to SpA. The biosynthetic knowledge presented in this project paves the way for metabolic engineering of soapwort saponins in heterologous systems, which may lead to large-scale production and biochemical studies of these biologically active saponins in the future. Additionally, this project offers foundational knowledge of saponin biosynthesis within the order Caryophyllales and may aid in discovery of genes encoding biosynthetic enzymes for related saponins.

Bibliography

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G. and Nielsen, H. (2019) 'SignalP 5.0 improves signal peptide predictions using deep neural networks', *Nature Biotechnology,* 37(4), pp. 420-423.

Aoki, K., Ogata, Y. and Shibata, D. (2007) 'Approaches for extracting practical information from gene co-expression networks in plant biology', *Plant and Cell Physiology,* 48(3), pp. 381-390.

Augustin, J. M., Kuzina, V., Andersen, S. B. and Bak, S. (2011) 'Molecular activities, biosynthesis and evolution of triterpenoid saponins', *Phytochemistry,* 72(6), pp. 435-57.

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S. and Werck-Reichhart, D. (2011) 'Cytochromes P450', *The Arabidopsis Book/American Society of Plant Biologists,* 9.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y. and Dekker, J. (2012) 'Hi–C: a comprehensive technique to capture the conformation of genomes', *Methods,* 58(3), pp. 268-276.

Bolshakov, A. P., Stepanichev, M. Y., Dobryakova, Y. V., Spivak, Y. S. and Markevich, V. A. (2020) 'Saporin from *Saponaria officinalis* as a Tool for Experimental Research, Modeling, and Therapy in Neuroscience', *Toxins,* 12(9), pp. 546.

Bomford, R., Stapleton, M., Winsor, S., Beesley, J., Jessup, E., Price, K. and Fenwick, G. (1992) 'Adjuvanticity and ISCOM formation by structurally diverse saponins', *Vaccine,* 10(9), pp. 572-577.

Bontpart, T., Cheynier, V., Ageorges, A. and Terrier, N. (2015) 'BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds', *New Phytologist,* 208(3), pp. 695-707.

Bontpart, T., Ferrero, M., Khater, F., Marlin, T., Vialet, S., Vallverdù-Queralt, A., Pinasseau, L., Ageorges, A., Cheynier, V. and Terrier, N. (2018) 'Focus on putative serine carboxypeptidase-like acyltransferases in grapevine', *Plant Physiology and Biochemistry,* 130, pp. 356-366.

Böttger, S. and Melzig, M. F. (2011) 'Triterpenoid saponins of the Caryophyllaceae and Illecebraceae family', *Phytochemistry Letters,* 4(2), pp. 59-68.

Bourgaud, F., Gravot, A., Milesi, S. and Gontier, E. (2001) 'Production of plant secondary metabolites: a historical perspective', *Plant Science,* 161(5), pp. 839-851.

Boutanaev, A. M., Moses, T., Zi, J., Nelson, D. R., Mugford, S. T., Peters, R. J. and Osbourn, A. (2015) 'Investigation of terpene diversification across multiple sequenced plant genomes', *Proceedings of the National Academy of Sciences,* 112(1), pp. E81-E88.

Bouvier, F., Rahier, A. and Camara, B. (2005) 'Biogenesis, molecular regulation and function of plant isoprenoids', *Progress in Lipid Research,* 44(6), pp. 357-429.

Bowyer, P., Clarke, B., Lunness, P., Daniels, M. and Osbourn, A. (1995) 'Host range of a plant pathogenic fungus determined by a saponin detoxifying enzyme', *Science,* 267(5196), pp. 371-374.

Budan, A., Bellenot, D., Freuze, I., Gillmann, L., Chicoteau, P., Richomme, P. and Guilet, D. (2014) 'Potential of extracts from *Saponaria officinalis* and *Calendula officinalis* to modulate *in vitro* rumen fermentation with respect to their content in saponins', *Bioscience, Biotechnology, and Biochemistry,* 78(2), pp. 288-295.

Bukharov, V. G. and Shcherbak, S. P. (1969) 'Triterpene glycosides from *Saponaria officinalis*', *Chemistry of Natural Compounds,* 5(5), pp. 324-326.

Buschhaus, C. and Jetter, R. (2012) 'Composition and physiological function of the wax layers coating *Arabidopsis* leaves: β-amyrin negatively affects the intracuticular water barrier', *Plant Physiology,* 160(2), pp. 1120-1129.

Cairns, J. R. K., Mahong, B., Baiya, S. and Jeon, J.-S. (2015) 'β-Glucosidases: multitasking, moonlighting or simply misunderstood?', *Plant Science,* 241, pp. 246-259.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics,* 10, pp. 1-9.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009) 'The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics', *Nucleic Acids Research,* 37(suppl_1), pp. D233-D238.

Carroll, A. and Specht, C. D. (2011) 'Understanding plant cellulose synthases through a comprehensive investigation of the cellulose synthase family Sequences', *Frontiers in Plant Science,* 2, pp. 5.

Caputi, L., Malnoy, M., Goremykin, V., Nikiforova, S., and Martens, S. (2012) 'A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land', *The Plant Journal*, *69*(6), pp. 1030-1042.

Chen, Q., Luo, J.-G. and Kong, L.-Y. (2011) 'New triterpenoid saponins from the roots of *Gypsophila perfoliata* Linn', *Carbohydrate Research,* 346(14), pp. 2206-2212.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. and Li, H. (2021) 'Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm', *Nature Methods,* 18(2), pp. 170-175.

Cherevach, E. I. and Shchekaleva, R. K. (2020) 'Justification of *Saponaria officinalis* (*S. officinali*s) cultivation in the soil and climatic conditions of the Primorsky region (Russia) and analysis of saponin-containing root extracts', *Journal of Central European Agriculture,* 21(2), pp. 420-430.

Chirva, V. Y. and Kintya, P. K. (1969) 'Structure of saponaside A', *Chemistry of Natural Compounds,* 5(3), pp. 162-162.

Chirva, V. Y. and Kintya, P. K. (1970) 'The structure of saponaside D', *Chemistry of Natural Compounds,* 6(2), pp. 209-212.

Christensen, L. P. (2008) 'Ginsenosides: chemistry, biosynthesis, analysis, and potential health effects', *Advances in Food and Nutrition Research,* 55, pp. 1-99.

Chung, S. Y., Seki, H., Fujisawa, Y., Shimoda, Y., Hiraga, S., Nomura, Y., Saito, K., Ishimoto, M. and Muranaka, T. (2020) 'A cellulose synthase-derived enzyme catalyses 3-O-glucuronosylation in saponin biosynthesis', *Nature Communications,* 11(1), pp. 1-11.

Corbet, S. A., Bee, J., Dasmahapatra, K., Gale, S., Gorringe, E., La Ferla, B., Moorhouse, T., Trevail, A., Van Bergen, Y. and Vorontsova, M. (2001) 'Native or exotic? Double or single? Evaluating plants for pollinator-friendly gardens', *Annals of Botany,* 87(2), pp. 219-232.

Cramer, J., Sager, C. P. and Ernst, B. (2019) 'Hydroxyl groups in synthetic and natural-product-derived therapeutics: a perspective on a common functional group', *Journal of Medicinal Chemistry,* 62(20), pp. 8915-8930.

Crout, D. H. and Vic, G. (1998) 'Glycosidases and glycosyl transferases in glycoside and oligosaccharide synthesis', *Current Opinion in Chemical Biology,* 2(1), pp. 98-111.

Curtis, W. (1777) '*Saponaria officinalis.* Soapwort.', *Flora londinensis*. London: William Curtis, pp. 12.

da Silva Magedans, Y. V., Phillips, M. A. and Fett-Neto, A. G. (2021) 'Production of plant bioactive triterpenoid saponins: from metabolites to genes and back', *Phytochemistry Reviews*, *20*(2), pp. 461-482.

Davies, G. and Henrissat, B. (1995) 'Structures and mechanisms of glycosyl hydrolases', *Structure,* 3(9), pp. 853-859.

de Brito Francisco, R. and Martinoia, E. (2018), 'The vacuolar transportome of plant specialized metabolites', *Plant and Cell Physiology*, *59*(7), pp. 1326-1336.

Delis, C., Krokida, A., Georgiou, S., Peña-Rodríguez, L. M., Kavroulakis, N., Ioannou, E., Roussis, V., Osbourn, A. E. and Papadopoulou, K. K. (2011) 'Role of lupeol synthase in *Lotus japonicus* nodule formation', *New Phytologist,* 189(1), pp. 335-346.

Di Bucchianico, S., Venora, G., Lucretti, S., Limongi, T., Palladino, L. and Poma, A. (2008) '*Saponaria officinalis* karyology and karyotype by means of image analyzer and atomic force microscopy', *Microscopy Research and Technique,* 71(10), pp. 730-736.

Dolezel, J. (2003) 'Nuclear DNA content and genome size of trout and human', *Cytometry Part A,* 51, pp. 127-128.

Eastman, J. (2014) *Wildflowers of the Eastern United States: An Introduction to Common Species of Woods, Wetlands and Fields.* Stackpole Books.

Elbandy, M., Miyamoto, T. and Lacaille-Dubois, M. A. (2007) 'New triterpenoidal saponins from *Gypsophila repens*', *Helvetica chimica acta,* 90(2), pp. 260-270.

Eljounaidi, K. and Lichman, B. R. (2020) 'Nature's chemists: the discovery and engineering of phytochemical biosynthesis', *Frontiers in Chemistry,* 8, pp. 596479.

Emms, D. M. and Kelly, S. (2019) 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biology,* 20(1), pp. 238.

Fernández-Tejada, A., Chea, E. K., George, C., Pillarsetty, N., Gardner, J. R., Livingston, P. O., Ragupathi, G., Lewis, J. S., Tan, D. S. and Gin, D. Y. (2014) 'Development of a minimal saponin vaccine adjuvant based on QS-21', *Nature Chemistry,* 6(7), pp. 635-643.

Ferreras, J., Barbieri, L., Girbés, T., Battelli, M. G., Rojo, M. A., Arias, F. J., Rocher, M. A., Soriano, F., Mendéz, E. and Stirpe, F. (1993) 'Distribution and properties of major ribosome-inactivating proteins (28 S rRNA N-glycosidases) of the plant S*aponaria officinalis* L.(Caryophyllaceae)', *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression,* 1216(1), pp. 31-42.

Field, B. and Osbourn, A. E. (2008) 'Metabolic diversification—independent assembly of operon-like gene clusters in different plants', *Science,* 320(5875), pp. 543-547.

Fraser, C. M., Rider, L. W. and Chapple, C. (2005) 'An expression and bioinformatics analysis of the *Arabidopsis* serine carboxypeptidase-like gene family', *Plant Physiology,* 138(2), pp. 1136-1148.

Frechet, D., Christ, B., du Sorbier, B. M., Fischer, H. and Vuilhorgne, M. (1991) 'Four triterpenoid saponins from dried roots of *Gypsophila* species', *Phytochemistry,* 30(3), pp. 927-931.

Fu, H., Koike, K., Li, W., Nikaido, T., Lin, W. and Guo, D. (2005) 'Silenorubicosides A− D, Triterpenoid Saponins from *Silene rubicunda*', *Journal of Natural Products,* 68(5), pp. 754-758.

Gachon, C. M., Langlois-Meurinne, M. and Saindrenan, P. (2005) 'Plant secondary metabolism glycosyltransferases: the emerging functional analysis', *Trends in Plant Science,* 10(11), pp. 542-549.

Galanie, S., Thodey, K., Trenchard, I. J., Filsinger Interrante, M. and Smolke, C. D. (2015) 'Complete biosynthesis of opioids in yeast', *Science,* 349(6252), pp. 1095-1100.

Gershenzon, J., Maffei, M. and Croteau, R. (1989) 'Biochemical and histochemical localization of monoterpene biosynthesis in the glandular trichomes of spearmint (*Mentha spicata*)', *Plant Physiology,* 89(4), pp. 1351-1357.

Ghosh, S. (2016) 'Biosynthesis of structurally diverse triterpenes in plants: the role of oxidosqualene cyclases'. *Proceedings of the Indian National Science Academy*, pp. 1189-1210.

Ghosh, S. (2017) 'Triterpene structural diversification by plant cytochrome P450 enzymes', *Frontiers in Plant Science,* 8(1886), pp. 295540.

Ghosh, S. (2020) 'Triterpenoids: Structural diversity, biosynthetic pathway, and bioactivity', *Studies in Natural Products Chemistry,* 67, pp. 411-461.

Gilabert-Oriol, R., Thakur, M., Haussmann, K., Niesler, N., Bhargava, C., Gorick, C., Fuchs, H. and Weng, A. (2016) 'Saponins from *Saponaria officinalis* L. augment the efficacy of a rituximab-immunotoxin', *Planta Medica,* 82(18), pp. 1525-1531.

Giolai, M., Paajanen, P., Verweij, W., Percival-Alwyn, L., Baker, D., Witek, K., Jupe, F., Bryan, G., Hein, I. and Jones, J. D. (2016) 'Targeted capture and sequencing of gene-sized DNA molecules', *Biotechniques,* 61(6), pp. 315-322.

González-Coloma, A., López-Balboa, C., Santana, O., Reina, M. and Fraga, B. M. (2011) 'Triterpene-based plant defenses', *Phytochemistry Reviews,* 10, pp. 245-260.

Goral, I., Jurek, I. and Wojciechowski, K. (2018) 'How does the surface activity of soapwort (*Saponaria officinalis* L.) extracts depend on the plant organ?', *Journal of Surfactants and Detergents,* 21(6), pp. 797-807.

Góral, I. and Wojciechowski, K. (2020) 'Surface activity and foaming properties of saponin-rich plants extracts', *Advances in Colloid and Interface Science*, pp. 102145.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) 'Full-length

transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology,* 29(7), pp. 644-52.

Haddad, M., Miyamoto, T., Ramezani, M. and Lacaille-Dubois, M. A. (2004) 'New triterpene saponins from *Acanthophyllum pachystegium*', *Helvetica chimica acta,* 87(1), pp. 73-81.

Han, A.-R., Park, S.-R., Park, J.-W., Lee, E.-Y., Kim, D.-M., Kim, B.-G. and Yoon, Y.-J. (2011) 'Biosynthesis of glycosylated derivatives of tylosin in *Streptomyces venezuelae*', *Journal of Microbiology and Biotechnology,* 21(6), pp. 613-616.

Haralampidis, K., Trojanowska, M. and Osbourn, A. E. (2002) 'Biosynthesis of triterpenoid saponins in plants', *History and Trends in Bioprocessing and Biotransformation*: Springer, pp. 31-49.

Harborne, J. B. and Williams, C. A. (2000) 'Advances in flavonoid research since 1992', *Phytochemistry,* 55(6), pp. 481-504.

Hartmann, M.-A. (1998) 'Plant sterols and the membrane environment', *Trends in Plant Science,* 3(5), pp. 170-175.

Hartmann, T. 'Diversity and variability of plant secondary metabolism: a mechanistic view'. *Proceedings of the 9th International Symposium on Insect-Plant Relationships*: Springer, 177-188.

Hartmann, T. (2007) 'From waste products to ecochemicals: fifty years research of plant secondary metabolism', *Phytochemistry,* 68(22-24), pp. 2831-2846.

Haslam, T. M., Mañas-Fernández, A., Zhao, L. and Kunst, L. (2012) 'Arabidopsis ECERIFERUM2 is a component of the fatty acid elongation machinery required for fatty acid extension to exceptional lengths', *Plant Physiology*, *160*(3), pp. 1164-1174.

Hedayati, A., Naseri, F., Nourozi, E., Hosseini, B., Honari, H. and Hemmaty, S. (2022) 'Response of *Saponaria officinalis* L. hairy roots to the application of TiO2 nanoparticles in terms of production of valuable polyphenolic compounds and SO6 protein', *Plant Physiology and Biochemistry,* 178, pp. 80-92.

Hemmerlin, A. (2013) 'Post-translational events and modifications regulating plant enzymes involved in isoprenoid precursor biosynthesis', *Plant Science,* 203, pp. 41-54.

Hemmerlin, A., Harwood, J. L. and Bach, T. J. (2012) 'A raison d'être for two distinct pathways in the early steps of plant isoprenoid biosynthesis?', *Progress in Lipid Research,* 51(2), pp. 95-148.

Henry, M. (1989) '*Saponaria officinalis* L.: *in vitro* culture and the production of triterpenoidal saponins', *Medicinal and Aromatic Plants II*: Springer, pp. 431-442.

Hodgson, H., De La Peña, R., Stephenson, M. J., Thimmappa, R., Vincent, J. L., Sattely, E. S. and Osbourn, A. (2019) 'Identification of key enzymes

responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production', *Proceedings of the National Academy of Sciences,* 116(34), pp. 17096-17104.

Irmisch, S., Jo, S., Roach, C. R., Jancsik, S., Saint Yuen, M. M., Madilao, L. L., O'Neil-Johnson, M., Williams, R., Withers, S. G. and Bohlmann, J. (2018) 'Discovery of UDP-glycosyltransferases and BAHD-acyltransferases involved in the biosynthesis of the antidiabetic plant metabolite montbretin A', *The Plant Cell,* 30(8), pp. 1864-1886.

Itkin, M., Davidovich-Rikanati, R., Cohen, S., Portnoy, V., Doron-Faigenboim, A., Oren, E., Freilich, S., Tzuri, G., Baranes, N. and Shen, S. (2016) 'The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from *Siraitia grosvenorii*', *Proceedings of the National Academy of Sciences,* 113(47), pp. E7619-E7628.

Jarvis, P. (2008) 'Targeting of nucleus-encoded proteins to chloroplasts in plants', *New Phytologist,* 179(2), pp. 257-285.

Jauch, J. (2008) 'Total synthesis of azadirachtin—finally completed after 22 years', *Angewandte Chemie International Edition,* 47(1), pp. 34-37.

Jayatilake, G. S., Freeberg, D. R., Liu, Z., Richheimer, S. L., Blake, M. E., Bailey, D. T., Haridas, V. and Gutterman, J. U. (2003) 'Isolation and structures of avicins D and G: *in vitro* tumor-inhibitory saponins serived from *Acacia v ictoriae*', *Journal of Natural Products,* 66(6), pp. 779-783.

Jia, Z., Koike, K. and Nikaido, T. (1999) 'Saponarioside C, the first α-D-galactose containing triterpenoid saponin, and five related compounds from *Saponaria officinalis*', *Journal of Natural Products,* 62(3), pp. 449-453.

Jia, Z., Koike, K., Sahu, N. P. and Nikaido, T. (2002) 'Triterpenoid saponins from Caryophyllaceae family', *Studies in Natural Products Chemistry*: Elsevier, pp. 3-61.

Jia, Z. H., Koike, K. and Nikaido, T. (1998) 'Major triterpenoid saponins from *Saponaria officinalis*', *Journal of Natural Products,* 61(11), pp. 1368-1373.

Jing, R., Li, H. Q., Hu, C. L., Jiang, Y. P., Qin, L. P. and Zheng, C. J. (2016), 'Phytochemical and pharmacological profiles of three Fagopyrum buckwheats', *International journal of molecular sciences*, *17*(4), pp. 589.

Jones, F. A. (1996) 'Herbs–useful plants. Their role in history and today', *European Journal of Gastroenterology & Hepatology,* 8(12), pp. 1227-1231.

Jozwiak, A., Sonawane, P. D., Panda, S., Garagounis, C., Papadopoulou, K. K., Abebie, B., Massalha, H., Almekias-Siegl, E., Scherf, T. and Aharoni, A. (2020) 'Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery', *Nature Chemical Biology,* 16(7), pp. 740-748.

Jurado-Gonzalez, P. and Sörensen, P. M. (2019) 'Characterization of saponin foam from *Saponaria officinalis* for food applications', *Food Hydrocolloids*, pp. 105541.

Katoh, K., Misawa, K., Kuma, K. i. and Miyata, T. (2002) 'MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research,* 30(14), pp. 3059-3066.

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. and Medema, M. H. (2017) 'plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters', *Nucleic Acids Research,* 45(W1), pp. W55-W63.

Kemen, A. C., Honkanen, S., Melton, R. E., Findlay, K. C., Mugford, S. T., Hayashi, K., Haralampidis, K., Rosser, S. J. and Osbourn, A. (2014) 'Investigation of triterpene synthesis and regulation in oats reveals a role for β-amyrin in determining root epidermal cell patterning', *Proceedings of the National Academy of Sciences,* 111(23), pp. 8679-8684.

Koike, K., Jia, Z. and Nikaido, T. (1998) 'Triterpenoid saponins from *Vaccaria segetalis*', *Phytochemistry,* 47(7), pp. 1343-1349.

Koike, K., Jia, Z. H. and Nikaido, T. (1999) 'New triterpenoid saponins and sapogenins from *Saponaria officinalis*', *Journal of Natural Products,* 62(12), pp. 1655-1659.

Korkmaz, M. and Özçelik, H. (2011) 'Economic importance of *Gypsophila* L., *Ankyropetalum Fenzl* and *Saponaria* L.(Caryophyllaceae) taxa of Turkey', *African Journal of Biotechnology,* 10(47), pp. 9533-9541.

Kreis, W. and Müller-Uri, F. (2010) 'Biochemistry of sterols, cardiac glycosides, brassinosteroids, phytoecdysteroids and steroid saponins', *Annual Plant Reviews Volume 40: Biochemistry of Plant Secondary Metabolism*, pp. 304-363.

Kruse, L. H., Weigle, A. T., Irfan, M., Martínez-Gómez, J., Chobirko, J. D., Schaffer, J. E., Bennett, A. A., Specht, C. D., Jez, J. M. and Shukla, D. (2022) 'Orthology-based analysis helps map evolutionary diversification and predict substrate class use of BAHD acyltransferases', *The Plant Journal,* 111(5), pp. 1453-1468.

Kuzuyama, T. and Seto, H. (2012) 'Two distinct pathways for essential metabolic precursors for isoprenoid biosynthesis', *Proceedings of the Japan Academy, Series B,* 88(3), pp. 41-52.

Lacaille-Dubois, M.-A., Hanquet, B., Cui, Z.-H., Lou, Z.-C. and Wagner, H. (1999) 'A new biologically active acylated triterpene saponin from *Silene fortunei*', *Journal of Natural Products,* 62(1), pp. 133-136.

Lambert, E., Faizal, A. and Geelen, D. (2011) 'Modulation of triterpene saponin production: *in vitro* cultures, elicitation, and metabolic engineering', *Applied Biochemistry and Biotechnology,* 164(2), pp. 220-237.

Li, H.-y., Koike, K., Ohmoto, T. and Ikeda, K. (1993) 'Dianchinenosides A and B, two new saponins from *Dianthus chinensis*', *Journal of Natural Products,* 56(7), pp. 1065-1070.

Li, H., Han, S., Huo, Y., Ma, G., Sun, Z., Li, H., Hou, S. and Han, Y. (2022) 'Comparative metabolomic and transcriptomic analysis reveals a coexpression network of the carotenoid metabolism pathway in the panicle of *Setaria italica*', *BMC Plant Biology,* 22(1), pp. 105.

Lichman, B. R., Godden, G. T., Hamilton, J. P., Palmer, L., Kamileen, M. O., Zhao, D., Vaillancourt, B., Wood, J. C., Sun, M. and Kinser, T. J. (2020) 'The evolutionary origins of the cat attractant nepetalactone in catnip', *Science Advances,* 6(20), pp. eaba0721.

Liu, J., Lee, J., Hernandez, M. A. S., Mazitschek, R. and Ozcan, U. (2015) 'Treatment of obesity with celastrol', *Cell,* 161(5), pp. 999-1011.

Liu, Q., Khakimov, B., Cárdenas, P. D., Cozzi, F., Olsen, C. E., Jensen, K. R., Hauser, T. P. and Bak, S. (2019) 'The cytochrome P450 CYP72A552 is key to production of hederagenin-based saponins that mediate plant defense against herbivores', *New Phytologist,* 222(3), pp. 1599-1609.

Louveau, T. and Osbourn, A. (2019) 'The sweet side of plant-specialized metabolism', *Cold Spring Harbor Perspectives in Biology,* 11(12), pp. a034744.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology,* 15(12), pp. 550.

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M. and Schmutz, J. (2022) 'GENESPACE tracks regions of interest and gene copy number variation across multiple genomes', *Elife,* 11, pp. e78526.

Lu, Y., Van, D., Deibert, L., Bishop, G. and Balsevich, J. (2015) 'Antiproliferative quillaic acid and gypsogenin saponins from *Saponaria officinalis* L. roots', *Phytochemistry,* 113, pp. 108-120.

Luang, S., Cho, J.-I., Mahong, B., Opassiri, R., Akiyama, T., Phasai, K., Komvongsa, J., Sasaki, N., Hua, Y.-l. and Matsuba, Y. (2013) 'Rice Os9BGlu31 is a transglucosidase with the capacity to equilibrate phenylpropanoid, flavonoid, and phytohormone glycoconjugates', *Journal of Biological Chemistry,* 288(14), pp. 10111-10123.

Lubke, M. A. and Cavers, P. (1969) 'The germination ecology of S*aponaria officinalis* from riverside gravel banks', *Canadian Journal of Botany,* 47(4), pp. 529-535.

Ma, L., Yuan, J., Qin, H., Zhang, M., Zhang, F., Yu, F., Tian, Z. and Wang, G. (2024) *GmMATE100* is involved in the import of Soyasaponins A and B into vacuoles in soybean plants (*Glycine max* L.) *Journal of Agricultural and Food Chemistry* 72 (17), 9994-10004..

Mackenzie, P. I., Bock, K. W., Burchell, B., Guillemette, C., Ikushiro, S., Iyanagi, T., Miners, J. O., Owens, I. S. and Nebert, D. W. (2005) 'Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily', *Pharmacogenet Genomics,* 15(10), pp. 677-85.

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. and Hernández-Hernández, T. (2015) 'A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity', *New Phytologist,* 207(2), pp. 437-453.

Marks, G. 'Teratological Androecia of *Saponaria officinalis*'. *Proceedings of the Indiana Academy of Science*, 370-372.

Matsuba, Y., Sasaki, N., Tera, M., Okamura, M., Abe, Y., Okamoto, E., Nakamura, H., Funabashi, H., Takatsu, M. and Saito, M. (2010) 'A novel glucosylation reaction on anthocyanins catalyzed by acyl-glucose–dependent glucosyltransferase in the petals of carnation and delphinium', *The Plant Cell,* 22(10), pp. 3374-3389.

Matsuura, H. N. and Fett-Neto, A. G. (2015) 'Plant alkaloids: main features, toxicity, and mechanisms of action', *Plant Toxins,* 2(7), pp. 1-15.

McGuire, W. P., Rowinsky, E. K., Rosenshein, N. B., Grumbine, F. C., Ettinger, D. S., Armstrong, D. K. and Donehower, R. C. (1989) 'Taxol: a unique antineoplastic agent with significant activity in advanced ovarian epithelial neoplasms', *Annals of Internal Medicine,* 111(4), pp. 273-279.

McLean, K., Sabri, M., Marshall, K., Lawson, R., Lewis, D., Clift, D., Balding, P., Dunford, A., Warman, A. and McVey, J. (2005) 'Biodiversity of cytochrome P450 redox systems', *Biochemical Society Transactions,* 33(4), pp. 796-801.

Meesapyodsuk, D., Balsevich, J., Reed, D. W. and Covello, P. S. (2007) 'Saponin biosynthesis in *Saponaria vaccaria*. cDNAs encoding β-amyrin synthase and a triterpene carboxylic acid glucosyltransferase', *Plant Physiology,* 143(2), pp. 959-969.

Miettinen, K., Pollier, J., Buyst, D., Arendt, P., Csuk, R., Sommerwerk, S., Moses, T., Mertens, J., Sonawane, P. D. and Pauwels, L. (2017) 'The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis', *Nature Communications,* 8(1), pp. 1-13.

Mitich, L. W. (1990) 'Bouncingbet–the soap weed', *Weed Technology,* 4(1), pp. 221-223.

Miyahara, T., Sakiyama, R., Ozeki, Y. and Sasaki, N. (2013) 'Acyl-glucose-dependent glucosyltransferase catalyzes the final step of anthocyanin formation in *Arabidopsis*', *Journal of Plant Physiology,* 170(6), pp. 619-624.

Miyahara, T., Takahashi, M., Ozeki, Y. and Sasaki, N. (2012) 'Isolation of an acyl-glucose-dependent anthocyanin 7-O-glucosyltransferase from the monocot *Agapanthus africanus*', *Journal of Plant Physiology,* 169(13), pp. 1321-1326.

Miyahara, T., Tani, T., Takahashi, M., Nishizaki, Y., Ozeki, Y. and Sasaki, N. (2014) 'Isolation of anthocyanin 7-*O*-glucosyltransferase from Canterbury bells (*Campanula medium*)', *Plant Biotechnology*, pp. 14.0908a.

Moellering, E. R., Muthan, B. and Benning, C. (2010) 'Freezing tolerance in plants requires lipid remodeling at the outer chloroplast membrane', *Science,* 330(6001), pp. 226-228.

Moghe, G., Kruse, L. H., Petersen, M., Scossa, F., Fernie, A. R., Gaquerel, E. and d'Auria, J. C. (2023) 'BAHD company: the ever-expanding roles of the BAHD acyltransferase gene family in plants', *Annual Review of Plant Biology,* 74, pp. 165-194.

Moglia, A., Lanteri, S., Comino, C., Hill, L., Knevitt, D., Cagliero, C., Rubiolo, P., Bornemann, S. and Martin, C. (2014). Dual catalytic activity of hydroxycinnamoyl-coenzyme A quinate transferase from tomato allows it to moonlight in the synthesis of both mono-and dicaffeoylquinic acids. *Plant Physiology* 166, (4) pp. 1777-1787.

Moglia, A., Acquadro, A., Eljounaidi, K., Milani, A.M., Cagliero, C., Rubiolo, P., Genre, A., Cankar, K., Beekwilder, J. and Comino, C., (2016). Genome-wide identification of BAHD acyltransferases and in vivo characterization of HQT-like enzymes involved in caffeoylquinic acid synthesis in globe artichoke. *Frontiers in Plant Science*, *7*, p.1424.

Moniuszko-Szajwaj, B., Masullo, M., Kowalczyk, M., Pecio, L., Szumacher-Strabel, M., Cieslak, A., Piacente, S., Oleszek, W. and Stochmal, A. (2016) 'Highly polar triterpenoid saponins from the roots of *Saponaria officinalis* L', *Helvetica Chimica Acta,* 99(5), pp. 347-354.

Morgan, E. D. (2009) 'Azadirachtin, a scientific gold mine', *Bioorganic and Medicinal Chemistry,* 17(12), pp. 4096-4105.

Moses, T., Papadopoulou, K. K. and Osbourn, A. (2014) 'Metabolic and functional diversity of saponins, biosynthetic intermediates and semi-synthetic derivatives', *Critical Reviews in Biochemistry and Molecular Biology,* 49(6), pp. 439-462.

Mroczek, A. (2015) 'Phytochemistry and bioactivity of triterpene saponins from *Amaranthaceae* family', *Phytochemistry Reviews,* 14(4), pp. 577-605.

Mugford, S. T., Qi, X., Bakht, S., Hill, L., Wegel, E., Hughes, R. K., Papadopoulou, K., Melton, R., Philo, M., Sainsbury, F., Lomonossoff, G. P., Roy, A. D., Goss, R. J. M. and Osbourn, A. (2009) 'A serine carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats ', *The Plant Cell,* 21(8), pp. 2473-2484.

Munro, A. W., Girvan, H. M., Mason, A. E., Dunford, A. J. and McLean, K. J. (2013) 'What makes a P450 tick?', *Trends in Biochemical Sciences,* 38(3), pp. 140-150.

Naoumkina, M. A., Modolo, L. V., Huhman, D. V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L. W. and Dixon, R. A. (2010) 'Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in *Medicago truncatula*', *The Plant Cell,* 22(3), pp. 850-866.

Nelson, D. and Werck-Reichhart, D. (2011) 'A P450-centric view of plant evolution', *The Plant Journal,* 66(1), pp. 194-211.

Nishizaki, Y., Yasunaga, M., Okamoto, E., Okamoto, M., Hirose, Y., Yamaguchi, M., Ozeki, Y. and Sasaki, N. (2013) 'p-Hydroxybenzoyl-glucose is a zwitter donor for the biosynthesis of 7-polyacylated anthocyanin in *Delphinium*', *The Plant Cell,* 25(10), pp. 4150-4165.

Ober, D. and Kaltenegger, E. (2009) 'Pyrrolizidine alkaloid biosynthesis, evolution of a pathway in plant secondary metabolism', *Phytochemistry,* 70(15-16), pp. 1687-1695.

Orme, A., Louveau, T., Stephenson, M. J., Appelhagen, I., Melton, R., Cheema, J., Li, Y., Zhao, Q., Zhang, L., Fan, D., Tian, Q., Vickerstaff, R. J., Landon, T., Han, B. and Osbourn A. (2019) 'A noncanonical vacuolar sugar transferase required for biosynthesis of antimicrobial defense compounds in oat', *Proceedings of the National Academy of Sciences,* 116(52), pp. 27105-27114.

Osbourn, A. (1996a) 'Saponins and plant defence—a soap story', *Trends in Plant Science,* 1(1), pp. 4-9.

Osbourn, A. (1996b) 'Preformed antimicrobial compounds and plant defense against fungal attack', *The Plant Cell,* 8(10), pp. 1821.

Osbourn, A. (2010) 'Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation', *Trends in Genetics,* 26(10), pp. 449-457.

Osbourn, A. and Lanzotti, V. (2009) *Plant-derived natural products.* Springer.

Osmani, S. A., Bak, S. and Møller, B. L. (2009) 'Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling', *Phytochemistry,* 70(3), pp. 325-347.

Owen, C., Patron, N. J., Huang, A. and Osbourn, A. (2017) 'Harnessing plant metabolic diversity', *Current Opinion in Chemical Biology,* 40, pp. 24-30.

Pabst, M., Grass, J., Fischl, R., Leonard, R., Jin, C., Hinterkorner, G., Borth, N. and Altmann, F. (2010) 'Nucleotide and nucleotide sugar analysis by liquid chromatography-electrospray ionization-mass spectrometry on surface-conditioned porous graphitic carbon', *Analytical Chemistry,* 82(23), pp. 9782-9788.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J. and Osbourn, A. (1999) 'Compromised disease resistance in saponin-deficient plants', *Proceedings of the National Academy of Sciences,* 96(22), pp. 12923-12928.

Patro, R., Duggal, G. and Kingsford, C. (2015) 'Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment', *Biorxiv*, pp. 021592.

Pertuit, D., Avunduk, S., Mitaine-Offer, A.-C., Miyamoto, T., Tanaka, C., Paululat, T., Delemasure, S., Dutartre, P. and Lacaille-Dubois, M.-A. (2014) 'Triterpenoid saponins from the roots of two *Gypsophila* species', *Phytochemistry,* 102, pp. 182-188.

Phillips, D. R., Rasbery, J. M., Bartel, B. and Matsuda, S. P. (2006) 'Biosynthetic diversity in plant triterpene cyclization', *Current Opinion in Plant Biology,* 9(3), pp. 305-314.

Pichersky, E. and Lewinsohn, E. (2011) 'Convergent evolution in plant specialized metabolism', *Annual Review of Plant Biology,* 62(1), pp. 549-566.

Polturak, G., Dippe, M., Stephenson, M. J., Chandra Misra, R., Owen, C., Ramirez-Gonzalez, R. H., Haidoulis, J. F., Schoonbeek, H.-J., Chartrain, L., Borrill, P., Nelson, D. R., Brown, J. K. M., Nicholson, P., Uauy, C. and Osbourn, A. (2022) 'Pathogen-induced biosynthetic pathways encode defense-related molecules in bread wheat', *Proceedings of the National Academy of Sciences,* 119(16), pp. e2123299119.

Pryszcz, L. P. and Gabaldón, T. (2016) 'Redundans: an assembly pipeline for highly heterozygous genomes', *Nucleic Acids Research,* 44(12), pp. e113-e113.

Pustahija, F., Brown, S. C., Bogunić, F., Bašić, N., Muratović, E., Ollier, S., Hidalgo, O., Bourge, M., Stevanović, V. and Siljak-Yakovlev, S. (2013) 'Small genomes dominate in plants growing on serpentine soils in West Balkans, an exhaustive study of 8 habitats covering 308 taxa', *Plant and Soil,* 373(1), pp. 427-453.

Putieva, Z. M., Mzhel'skaya, L. G., Gorovits, T. T., Kondratenko, E. S. and Abubakirov, N. K. (1977) 'Triterpene glycosides of *Acanthophyllum gypsophiloides* V. D-quinovose in acanthophyllosides B and C', *Chemistry of Natural Compounds,* 13(6), pp. 679-682.

Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R. and Osbourn, A. (2004) 'A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants', *Proceedings of the National Academy of Sciences,* 101(21), pp. 8233-8238.

Raguindin, P. F., Itodo, O. A., Stoyanov, J., Dejanovic, G. M., Gamba, M., Asllanaj, E., Minder, B., Bussler, W., Metzger, B., Muka, T. and Glisic, M. and Kern, H. (2021), 'A systematic review of phytochemicals in oat and buckwheat', *Food chemistry*, *338*, pp. 127982.

Rao, X. and Dixon, R. A. (2019) 'Co-expression networks for plant biology: why and how', *Acta Biochimica et Biophysica Sinica,* 51(10), pp. 981-988.

Reed, J., Orme, A., El-Demerdash, A., Owen, C., Martin, L. B. B., Misra, R. C., Kikuchi, S., Rejzek, M., Martin, A. C., Harkess, A., Leebens-Mack, J.,

Louveau, T., Stephenson, M. J. and Osbourn, A. (2023) 'Elucidation of the pathway for biosynthesis of saponin adjuvants from the soapbark tree', *Science,* 379(6638), pp. 1252-1264.

Reed, J. and Osbourn, A. (2018) 'Engineering terpenoid production through transient expression in Nicotiana benthamiana', *Plant Cell Reports*, pp. 1-11.

Reed, J., Stephenson, M. J., Miettinen, K., Brouwer, B., Leveau, A., Brett, P., Goss, R. J., Goossens, A., O'Connell, M. A. and Osbourn, A. (2017) 'A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules', *Metabolic Engineering,* 42, pp. 185-193.

Rees, A. (1819) 'The Cyclopedia or Universal Dictionary', *Philadelphia: Samuel F. Bradford and others, XXXVIII, see varnishing*.

Robert, X. and Gouet, P. (2014) 'Deciphering key features in protein structures with the new ENDscript server', *Nucleic Acids Research,* 42(W1), pp. W320-W324.

Sadowska, B., Budzynska, A., Wieckowska-Szakiel, M., Paszkiewicz, M., Stochmal, A., Moniuszko-Szajwaj, B., Kowalczyk, M. and Rozalska, B. (2014) 'New pharmacological properties of *Medicago sativa* and *Saponaria officinalis* saponin-rich fractions addressed to *Candida albicans*', *Journal of Medical Microbiology,* 63(Pt 8), pp. 1076-86.

Sainsbury, F. and Lomonossoff, G. P. (2014) 'Transient expressions of synthetic biology in plants', *Current Opinion in Plant Biology,* 19, pp. 1-7.

Sainsbury, F., Thuenemann, E. C. and Lomonossoff, G. P. (2009) 'pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants', *Plant Biotechnology Journal,* 7(7), pp. 682-693.

Sasaki, N., Nishizaki, Y., Ozeki, Y. and Miyahara, T. (2014) 'The role of acyl-glucose in anthocyanin modifications', *Molecules,* 19(11), pp. 18747-18766.

Schmid, R., Heuckeroth, S., Korf, A., Smirnov, A., Myers, O., Dyrlund, T. S., Bushuiev, R., Murray, K. J., Hoffmann, N. and Lu, M. (2023) 'Integrative analysis of multimodal mass spectrometry data in MZmine 3', *Nature Biotechnology,* 41(4), pp. 447-449.

Schmidt, J. F., Moore, M. D., Pelcher, L. E. and Covello, P. S. (2007) 'High efficiency *Agrobacterium rhizogenes*-mediated transformation of *Saponaria vaccaria* L.(Caryophyllaceae) using fluorescence selection', *Plant Cell Reports,* 26(9), pp. 1547-1554.

Seppey, M., Manni, M. and Zdobnov, E. M. (2019) 'BUSCO: assessing genome assembly and annotation completeness', *Gene Prediction*: Springer, pp. 227-245.

Shah, B. A., Qazi, G. N. and Taneja, S. C. (2009) 'Boswellic acids: a group of medicinally important compounds', *Natural Product Reports,* 26(1), pp. 72-89.

Sharkey, T. D., Wiberley, A. E. and Donohue, A. R. (2008) 'Isoprene emission from plants: why and how', *Annals of Botany,* 101(1), pp. 5-18.

Shrivastava, S. (2020) 'Introduction to glycoside hydrolases: classification, identification and occurrence', *Industrial Applications of Glycoside Hydrolases*: Springer, pp. 3-84.

Siedenburg, G. and Jendrossek, D. (2011) 'Squalene-hopene cyclases', *Applied and Environmental Microbiology,* 77(12), pp. 3905-3915.

Smulek, W., Zdarta, A., Pacholak, A., Zgola-Grzeskowiak, A., Marczak, L., Jarzebski, M. and Kaczorek, E. (2017) '*Saponaria officinalis* L. extract: surface active properties and impact on environmental bacterial strains', *Colloids and Surfaces B Biointerfaces,* 150, pp. 209-215.

Soneson, C., Love, M. and Robinson, M. (2016) 'Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences', *F1000Research,* 4(1521).

Sparg, S., Light, M. and Van Staden, J. (2004) 'Biological activities and distribution of plant saponins', *Journal of Ethnopharmacology,* 94(2-3), pp. 219-243.

Springob, K. and Kutchan, T. M. (2009) 'Introduction to the different classes of natural products', *Plant-Derived Natural Products: Synthesis, Function, and Application*: Springer, pp. 3-50.

St-Pierre, B. and De Luca, V. (2000) 'Origin and diversification of the BAHD superfamily of acyltransferases involved in secondary metabolism', *Recent Advances in Phytochemistry,* 34(9), pp. 285-315.

Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics,* 30(9), pp. 1312-1313.

Stanke, M. and Morgenstern, B. (2005) 'AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints', *Nucleic Acids Research,* 33(suppl_2), pp. W465-W467.

Stehle, F., Brandt, W., Milkowski, C. and Strack, D. (2006) 'Structure determinants and substrate recognition of serine carboxypeptidase-like acyltransferases from plant secondary metabolism', *FEBS Letters,* 580(27), pp. 6366-6374.

Stephenson, M. J., Reed, J., Brouwer, B. and Osbourn, A. (2018) 'Transient expression in nicotiana benthamiana leaves for triterpene production at a preparative scale', *Journal of Visualized Experiments,* 138, pp. e58169.

Stonik, V. and Elyakov, G. (1988) 'Secondary metabolites from echinoderms as chemotaxonomic markers', *Bioorganic Marine Chemistry*: Springer, pp. 43-86.

Sun, H.-X., Xie, Y. and Ye, Y.-P. (2009) 'Advances in saponin-based adjuvants', *Vaccine,* 27(12), pp. 1787-1796.

Sun, J., Wang, S., Xia, F., Wang, K.-Y., Chen, J.-M. and Tu, P.-F. (2014) 'Five new benzophenone glycosides from the leaves of *Aquilaria sinensis* (Lour.) Gilg', *Chinese Chemical Letters,* 25(12), pp. 1573-1576.

Takahashi, N., Iguchi, T., Kuroda, M., Mishima, M. and Mimaki, Y. (2022) 'Novel Oleanane-type triterpene glycosides from the *Saponaria officinalis* L. seeds and apoptosis-inducing activity via mitochondria', *International Journal of Molecular Sciences,* 23(4), pp. 2047.

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P. and Osbourn, A. (2014) 'Triterpene biosynthesis in plants', *Annual Review of Plant Biolology,* 65, pp. 225-57.

Tholl, D. (2015) 'Biosynthesis and biological functions of terpenoids in plants', *Biotechnology of Isoprenoids*: Springer, pp. 63-106.

Tohge, T. and Fernie, A. R. (2012) 'Co-expression and co-responses: within and beyond transcription', *Frontiers in Plant Science,* 3, pp. 248.

Torrens-Spence, M., Fallon, T. and Weng, J. (2016) 'A workflow for studying specialized metabolism in nonmodel eukaryotic organisms', *Methods in Enzymology*: Elsevier, pp. 69-97.

Van Tunen, A., Koes, R., Spelt, C., Van der Krol, A., Stuitje, A. and Mol, J. (1988) 'Cloning of the two chalcone flavanone isomerase genes from *Petunia hybrida*: coordinate, light-regulated and differential expression of flavonoid genes', *The EMBO Journal,* 7(5), pp. 1257-1263.

Vogt, T. and Jones, P. (2000) 'Glycosyltransferases in plant natural product synthesis: characterization of a supergene family', *Trends in Plant Science,* 5(9), pp. 380-386.

Vranová, E., Coman, D. and Gruissem, W. (2013) 'Network analysis of the MVA and MEP pathways for isoprenoid synthesis', *Annual review of plant biology,* 64, pp. 665-700.

Vuong, T. V. and Wilson, D. B. (2010) 'Glycoside hydrolases: catalytic base/nucleophile diversity', *Biotechnology and Bioengineering,* 107(2), pp. 195-205.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J. and Schatz, M. C. (2017) 'GenomeScope: fast reference-free genome profiling from short reads', *Bioinformatics,* 33(14), pp. 2202-2204.

Wang, P. (2021) 'Natural and synthetic saponins as vaccine adjuvants'. *Vaccines*, 9(3), 222.

Wen, B., Tian, J.-M., Huang, Z.-R., Xiao, S.-J., Yuan, W.-L., Wang, J.-X., Li, H.-L., Xu, X.-K. and Shen, Y.-H. (2020) 'Triterpenoid saponins from the roots of *Psammosilene tunicoides*', *Fitoterapia,* 144, pp. 104596.

Weng, A., Bachran, C., Fuchs, H. and Melzig, M. (2008) 'Soapwort saponins trigger clathrin-mediated endocytosis of saporin, a type I ribosome-inactivating protein', *Chemico-Biological Interactions,* 176(2-3), pp. 204-211.

Weng, J.-K., Philippe, R. N. and Noel, J. P. (2012) 'The rise of chemodiversity in plants', *Science,* 336(6089), pp. 1667-1670.

Weng, J. K. (2014) 'The evolutionary paths towards complexity: a metabolic perspective', *New Phytologist,* 201(4), pp. 1141-9.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G. and Gitzendanner, M. A. (2014) 'Phylotranscriptomic analysis of the origin and early diversification of land plants', *Proceedings of the National Academy of Sciences,* 111(45), pp. E4859-E4868.

Williams, S. (2013) *Transglycosylases*. CAZypedia. Available at: https://www.cazypedia.org/index.php/Transglycosylases (Accessed: March 15 2024).

Withers, S. G., Dombroski, D., Berven, L. A., Kilburn, D. G., Miller Jr, R. C., Warren, R. A. J. and Gilkes, N. R. (1986) 'Direct 1H NMR determination of the stereochemical course of hydrolyses catalysed by glucanase components of the cellulase complex', *Biochemical and Biophysical Research Communications,* 139(2), pp. 487-494.

Wu, T. D., Reeder, J., Lawrence, M., Becker, G. and Brauer, M. J. (2016) 'GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality', *Satistical Genomics: Methods and Protocols,* 1418, pp. 283-334.

Xu, R., Fazio, G. C. and Matsuda, S. P. (2004) 'On the origins of triterpenoid skeletal diversity', *Phytochemistry,* 65(3), pp. 261-291.

Xu, Z., Escamilla-Treviño, L., Zeng, L., Lalgondar, M., Bevan, D., Winkel, B., Mohamed, A., Cheng, C.-L., Shih, M.-C. and Poulton, J. (2004) 'Functional genomic analysis of *Arabidopsis thaliana* glycoside hydrolase family 1', *Plant Molecular Biology,* 55(3), pp. 343-367.

Yu, B. and Sun, J. (2009) 'Current synthesis of triterpene saponins', *Chemistry–An Asian Journal,* 4(5), pp. 642-654.

Zannino, L., Carelli, M., Milanesi, G., Croce, A. C., Biggiogera, M. and Confalonieri, M. (2024) 'Histochemical and ultrastructural localization of triterpene saponins in *Medicago truncatula*', *Microscopy Research and Technique*, 87, pp. 2143-2153.

Zhao, C. L., Cui, X. M., Chen, Y. P. and Liang, Q. (2010) 'Key enzymes of triterpenoid saponin biosynthesis and the induction of their activities and gene expressions in plants', *Natural Product Communications,* 5(7), pp. 1934578X1000500736.

Zhao, S., Guo, Y., Sheng, Q. and Shyr, Y. (2014) 'Heatmap3: an improved heatmap package with more powerful and convenient features', *BMC Bioinformatics,* 15(10), pp. P16.

Zhou, F. and Pichersky, E. (2020) 'More is better: the diversity of terpene metabolism in plants', *Current Opinion in Plant Biology,* 55, pp. 1-10.

Ziegler, J. and Facchini, P. J. (2008) 'Alkaloid biosynthesis: metabolism and trafficking', *Annual Review of Plant Biolology,* 59, pp. 735-769.

A

Miscellaneous

A.1 Primers sequences

Table A.1.1. Primer oligonucleotide sequences. All primers used in this project are listed. Primers used for Gateway cloning are gene specific primers with either attB1 or attB2 adaptor sequences. F, forward primer; R, reverse primer.

Name	Sequence (5′ → 3′)
Sanger Sequencing	
attL1-F	TCGCGTTAACGCTAGCATGGATCTC
attL2-R	ACATCAGAGATTTTGAGACACGGGC
attB1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTA
attB2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTA
SobAS1-Middle	GATTACACCTCTAATCAAACAGCT
SoCSL1-Middle	CGGGTTTTAATCACCATGCTAAAG
Gateway Cloning of candidate genes from *S. officinalis*	
SobAS1-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTGGAGGTTAAAAATAGCAGAAG
SobAS1-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAACTAAGCCTCAAAGGAACATG
SoC28-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAACTCTTCTTCATATGTGGA
SoC28-attb2-F	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGCAGATACAGTTACGGGTTT
SoC28C16-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGCTAATTACCTTACTAAGTG
SoC28C16-attb2-F	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGCGAGGGTGGCGGATT
CYP1-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGCATAGTAATAGTAAGATGGGTA
CYP1-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAACGTCTACGAAACATGAGAG
CYP2-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGCTCATTTTGTCACTACTA
CYP2-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAACAATAAGCAATAGGAATTACATG
CYP3-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGGAGTGTTGCTAGTGCTG
CYP3-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATAGTCGGAAACTCGAATTTTCAA
CYP4-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGATTTCTTAGACATTTTCATACTC
CYP4-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCACCTAGGCAGAATCACTGC
CYP5-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGACGCATTTACTTTATTAATGCT
CYP5-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATTGTTAATTTTGGCAATTATAGGG
CYP6-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGCTAATTACTTTGTTAAGTG
CYP6-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCATCGACGAGCCGACATATG

CYP7-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGATATTCTTGTAGGTTTGCTTTT

(Table A.1.1. continued)

Name	Sequence (5′ → 3′)
CYP7-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAAAAGCTTTCGTTTGCTTTGCG
CSL1-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCACCCCACAACACCTG
CSL1-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGAGCGACCTTTTCTAGCTTT
UGT1-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTAATGAAAACAATACAATTCAAG
UGT1-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGAAGATGAAAGCCACTCAATG
UGT2-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGGAATCAAAGGAGGAAG
UGT2-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAAATTTTTGTAGCACAGCTTTG
UGT3-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAACACAGAAATGGAGATCAC
UGT3-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCATTGTGACGTGCTTTTTAACTC
UGT4-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCTGCCAAAATGTTGCACG
UGT4-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCACTCGACGAGTGCTTGTAAA
UGT5-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGACTCGAATTCTAACAACAAC
UGT5-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAGACTAGTTTTTGTGACAATTCATTT
UGT6-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCCGAAAATAAATTAAAAATATTATT CATA
UGT6-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGCGATTTCATGGGATCCC
UGT7-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGGTTCAAATACAGAAGCAACT
UGT7-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAGCCTTCCTTAACGATCTC
UGT8-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTACCCTTTGTTTGCCATGG
UGT8-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCATCCAATTAAATCCTTCAAATTTTGAA
UGT9-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCAAACCAAGGTGAACAAAAA
UGT9-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTATCTGGTAATATGCGCCACAA
UGT10-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCGGATCAAAATGATAAAAAGGT
UGT10-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAGAAAGATGAAACCCACTCAATAA
UGT11-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGAAGAAACAACTTCATTTAGT
UGT11-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAACCCACAAGTTTTAGCAGATC
UGT12-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAAGTCACCACTAAAGTTGTAC
UGT12-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAATTAGCAACCTTACTCATTTTATC
UGT13-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGGTACTAAAGAGTTACACATAG
UGT13-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTACTTCTCAACAAGATCTTGTAG
AT1-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAGAGCCACAAAACTTAGAA
AT1-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAGGCAACAACAAAAGCACTAAA
AT2-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAATCCTTTGGAAAGAAAGTC
AT2-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATCCAATGACACAAAACTCATAAA
AT3-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGACTTCCAAAAGGTAGAAGT
AT3-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAATTATCGACCCTAAAGCTTG
AT4-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGGAAGTTATAGTAATAATAACAACA
AT4-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATTTAAAATAGAAGAAAAAGGTTTGAC
AT5-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGATGACATTTTGTAAGCACAACA
AT5-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAGAGACTACTATTGATACACGT
AT6-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCTAATAAAAATAATCCCAAATTAAA AATA
AT6-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGAAACATAAGACATAAACTCGTG

AT7-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAACCTTCAAAAATGGAAGTG
AT7-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAAAGTTTGGGGAAGCAAAGG
AT8-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAATTTCCAAAATATAGTAGTCAAAAG
AT8-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAACGTCGACGTGATTATTGTC

(Table A.1.1. continued)

Name	Sequence (5′ → 3′)
AT9-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAGTGAAAATTGTACGTAGG
AT9-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAGCTGGGCGTGGCATATTC
SCPL1-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTTGTCTTTTACCACAACCGA
SCPL1-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTATAGCATATCGGGCAGAGG
SCPL2-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCACTTGCCACACCTCT
SCPL2-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTATAACGGAAAATAATTCACCCATC
SCPL3-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCATACACATTGTTTGTTCACAA
SCPL3-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCATATAGGAATTCCGTCTAGCC
SCPL4-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAAGATGTCAATATACTCTTTGTTT
SCPL4-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTACAAAGGACTCAATGAAAACCAC
SCPL5-attb1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTTGAGGAAGAATAATCATCACAT
SCPL5-attb2-R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAACAGGTTTTGCAGATAAAAACT

A.2 Saponins reported from *S. officinalis*

Table A.2.1. List of saponins previously isolated from *S. officinalis.*

Name	Chemical Formula	m/z [M]	Composition	Extracted Material	Reference
Quillaic acid-based saponins					
Saponarioside A (SpA)	C82H128O45	1832	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)-β-D-4-O-acetylquinovopyranosyl-(1→4)]-β-D-fucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1998)
Saponarioside B (SpB)	C77H120O41	1700	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-4-O-acetylquinovopyranosyl-(1→4)]-β-D-fucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1998)
CAN1	C70H109O37	1542	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN2	C72H111O38	1584	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-glucopyranosyl-(1→3)-[β-D-6-O-acetyl-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN3	C75H117O41	1674	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)

(Table A.2.1. continued)

Name	Chemical Formula	m/z [M]	Composition	Extracted Material	Reference
CAN4	C77H119O42	1716	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-glucopyranosyl-(1→3)-[β-D-6-O-acetyl-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN5	C76H120O42	1704	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-xylopyranosyl-(1→4)-[β-D-glucopyranosyl-(1→3)]-α-L-rhamnopyranosyl-(1→2)-[α-D-galactopyranosyl(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN6	C78H121O43	1746	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-xylopyranosyl-(1→4)-[β-D-6-O-acetyl-glucopyranosyl-(1→3)]-α-L-rhamnopyranosyl-(1→2)-[α-D-galactopyranosyl-(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN7	C81H128O46	1836	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1 →2)-[α-D-galactopyranosyl-(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN8	C76H119O42	1704	3-O-α-L-galactopyranosyl-(1→4)-β-D-xylopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN9	C72H111O38	1584	3-O-β-D-galactopyranosyl-(1→2)-[α-L-arabinopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid-28-O-β-D-xylopyranosyl-(1→4)-[β-D-6-O-acetyl-glucopyranosyl-(1→3)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)

(Table A.2.1. continued)

Name	Chemical Formula	m/z [M]	Composition	Extracted Material	Reference
POL 5	C85H132O47	1904	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-(6-O-acetyl)-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)-(4-O-acetyl)-β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
POL4	C83H129O46	1862	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-β-D-glucopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)-(4-O-acetyl)-β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
POL6	C77H120O41	1700	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)-(4-O-acetyl)-β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
POL7	C78H122O42	1730	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-β-D-glucopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[(4-O-acetyl)-β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
POL8	C80H124O43	1772	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-(6-O-acetyl)-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl(1→4)]-α-L-rhamnopyranosyl-(1→2)-[(4-O-acetyl)-β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
JAP1	C76H120O41	1689	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-[β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Seed (commercial)	(Takahashi *et al.*, 2022)

(Table A.2.1. continued)

Name	Chemical Formula	m/z [M]	Composition	Extracted Material	Reference
JAP2	C78H122O42	1731	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl quillaic acid 28-O-β-D-glucopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-[(4-O-acetyl)-β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Seed (commercial)	(Takahashi *et al.*, 2022)
JAP3	C76H120O41	1689	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid 28-O-β-D-glucopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-O-[β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Seed (commercial)	(Takahashi *et al.*, 2022)
JAP4	C80H124O43	1773	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-quillaic acid 28-O-β-D-glucopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-O-[[3,4-di-O-acetyl-β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranoside	Seed (commercial)	(Takahashi *et al.*, 2022)
Gypsogenin-based saponins					
CAN10	C70H109O36	1526	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-gypsogenin-28-O-β-D-glucopyranosyl-(1→3)[β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN11	C72H111O37	1568	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-gypsogenin-28-O-β-D-6-O-acetyl-glucopyranosyl-(1 →3)-[β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN12	C75H117O40	1658	3-O-β-D-galactopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-gypsogenin-28-O-β-D-glucopyranosyl-(1→3)[β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)

(Table A.2.1. continued)

Name	Chemical Formula	m/z [M]	Composition	Extracted Material	Reference
CAN13	C77H119O41	1700	3-O-β-D-galactopyranosyl-(1→2)-[β-Dxylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-gypsogenin-28-O-β-D-O-acetyl-glucopyranosyl-(1 →3)-[β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)]-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
CAN14	C69H107O35	1496	3-O-β-D-galactopyranosyl(1→2)-[β-D-xylopyranosyl-(1→3)]-β-D-glucuronopyranosyl-gypsogenin-28-O-β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranoside	Roots (commercial, dried)	(Lu *et al.*, 2015)
Gypsogenic acid-based saponins					
Saponarioside C (SpC)	C59H94O29	1266	3-O-β-D-xylopyranosylgypsogenic acid-28-O-α-D-galactopyranosyl-(1→6)-β-D-glucopyranosyl-(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1999)
Saponarioside D (SpD)	C59H94O29	1266	3-O-β-D-xylopyranosyl-gypsogenic acid-28-O-β-D-glucopyranosyl(1→2)-β-D-glucopyranosyl-(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1999)
Saponarioside E (SpE)	C60H96O30	1296	3-O-β-D-glucopyranosyl-gypsogenic acid-28-O-β-D-glucopyranosyl-(1→2)-β-D-glucopyranosyl-(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1999)
Saponarioside H (SpH)	C41H64O14	780	3-O-β-D-xylopyranosylgypsogenic acid-28-O-β-D-glucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1999)
Saponarioside L (SpL)	C53H84O24	1104	3-O-β-D-xylopyranosylgypsogenic acid 28-O-β-D-glucopyranosyl-(1→3)-[β-D-glucopyranosyl-(1→6)]-β-D-glucopyranoside	Whole Plants (air dried)	(Koike, Jia and Nikaido, 1999)
Saponarioside M (SpM)	C53H84O24	1104	3-O-β-D-glucopyranosylgypsogenic acid 28-O-β-D-glucopyranosyl-(1→2)-β-D-glucopyranosyl-(1→6)-β-D-glucopyranoside	Whole Plants (air dried)	(Koike, Jia and Nikaido, 1999)

(Table A.2.1. continued)

Name	Chemical Formula	m/z [M]	Composition	Extracted Material	Reference
POL3	C59H91O28	1248	3-O-β-D-xylopyranosyl-gypsogenic acid-28-O-[β-D-glucopyranosyl-(1→3)]-[6-O-(3-hydroxy-3-methylglutaryl)-β-D-glucopyranosyl-(1→6)]-β-D-glucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
JAP6	C65H102O33	1411	3-O-β-D-xylopyranosyl-gypsogenic acid-28-O-β-D-glucopyranosyl-(1→2)-(6-O-3-hydroxy-3-methylglutaryl-βD-glucopyranosyl)-(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Seed (commercial)	(Takahashi *et al.*, 2022)
16α-Hydroxygypsogenic acid-based saponins					
Saponarioside F (SpF)	C59H94O30	1282	3-O-β-D-xylopyranosyl-16α-hydroxygypsogenic acid-28-O-β-D-glucopyranosyl-(1→2)-β-D-glucopyranosyl-(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1999)
Saponarioside G (SpG)	C53H84O25	1120	3-O-β-D-xylopyranosyl-16α-hydroxygypsogenic acid-28-O-β-D-glucopyranosyl(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Whole Plants (fresh)	(Jia, Koike and Nikaido, 1999)
Saponarioside I (SpI)	C59H94O30	1282	3-O-β-D-xylopyranosyl-16α-hydroxygypsogenic acid 28-O-a-D-galactopyranosyl-(1→6)-β-D-glucopyranosyl-(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Whole Plants (air dried)	(Koike, Jia and Nikaido, 1999)
POL1	C47H73O20	958	3-O-β-D-xylopyranosyl-16α-hydroxygypsogenic acid-28-O-[β-D-glucopyranosyl (1→6)]-β-D-glucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
POL2	C65H103O35	1444	3-O-β-D-xylopyranosyl-16α-hydroxygypsogenic acid-28-O-[β-D-glucopyranosyl-(1→3)]-[α-D-galactopyranosyl-(1→6)-α-D-galactopyranosyl-(1→6)-β-D-glucopyranosyl-(1→6)]-β-D-glucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016)
JAP5	C53H84O25	1121	3-O-β-D-xylopyranosyl-16α-hydroxygypsogenic acid-28-O-β-D-glucopyranosyl-(1→2)-O-β-D-glucopyranosyl-(1→6)-β-D-glucopyranoside	Seed (commercial)	(Takahashi *et al.*, 2022)

(Table A.2.1. continued)

Name	Chemical Formula	m/z [M]	Composition	Extracted Material	Reference
JAP7	C65H102O34	1427	3-O-β-D-xylopyranosyl-16α-hydroxygypsogenic acid-28-O-β-D-glucopyranosyl-(1→2)-O-(6-O-3-hydroxy-3-methylglutaryl-β-D-glucopyranosyl)-(1→6)-O-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Seed (commercial)	(Takahashi *et al.*, 2022)
Olean-11,13(18)-diene-23,24-dioic acid-based saponin					
Saponarioside J	C52H82O24	1102	3-O-β-D-xylopyranosylolean-11,13(18)-diene-23,28-dioic acid 28-O-β-D-glucopyranosyl-(1→3)-[β-D-glucopyranosyl-(1→6)]-β-D-glucopyranoside	Whole Plants (air dried)	(Koike, Jia and Nikaido, 1999)
3,4-Seco-16α-hydroxygypsogenic acid-base saponin					
Saponarioside K	C48H76O21	988	3,4-seco-16α-hydroxygypsogenic acid 28-O-β-D-glucopyranosyl-(1→3)-[β-D-glucopyranosyl-(1→6)]-β-D-glucopyranoside	Whole Plants (air dried)	(Koike, Jia and Nikaido, 1999)
16α-Hydroxyolean-12-ene-23α, 28β-dioic acid-base saponin					
Dianchinenoside B	C41H64O15	796	3β-O-β-D-xylopyranosyl-16α-hydroxyolean-12-ene-23α, 28β-dioic acid 28-O-β-D-glucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016; Li *et al.*, 1993)
3,4-Seco-gypsogenic acid-base saponin					
Vaccaroside D	C54H86O25	1134	3,4-seco-gypsogenic acid-28-O-β-D-glucopranosyl-(1→2)-β-D-glucopyranosyl-(1→6)-[β-D-glucopyranosyl-(1→3)]-β-D-glucopyranoside	Roots (commercial, dried)	(Moniuszko-Szajwaj *et al.*, 2016; Koike, Jia and Nikaido, 1998)

A.3 plantiSMASH output

Table A.3.1. Details of plantiSMASH output of *S. officinalis* genome.

Locus tag	From	To	Strand	Category	Pfam Domains
Cluster 1 - Chr01 - Saccharide					
Saoffv11000760m	12071585	12071902	+	Other genes	n/a
Saoffv11000761m	12072056	12072448	+	Glycosyltransferase	UDPGT_2
Saoffv11000763m	12093560	12094021	+	Glycosyltransferase	UDPGT_2
Saoffv11000764m	12094832	12098061	-	Other genes	n/a
Saoffv11000767m	12237606	12238216	+	Other genes	n/a
Saoffv11000773m	12269980	12271609	+	Other genes	n/a
Saoffv11000775m	12317803	12318306	+	Other genes	n/a
Saoffv11000776m	12321156	12323041	-	Cytochrome 450	p450
Saoffv11000778m	12366870	12369730	-	Other genes	n/a
Cluster 2 - Chr01 - Saccharide					
Saoffv11004374m	165549897	165554551	+	Glycosyltransferase	Glycos_transf_1
Saoffv11004376m	165567162	165570404	+	Other genes	n/a
Saoffv11004378m	165571552	165572061	-	Other genes	n/a
Saoffv11004380m	165604033	165609029	+	BAHD acyltransferase	Transferase
Saoffv11004381m	165623746	165625158	+	Other genes	n/a
Saoffv11004384m	165651384	165656559	+	Methyltransferase	Methyltransf_7
Saoffv11004385m	165791606	165814633	+	Methyltransferase	Methyltransf_7
Saoffv11004386m	165810128	165810967	-	Other genes	n/a
Saoffv11004390m	165837775	165838392	-	Other genes	n/a
Cluster 3 - Chr03 - Saccharide					
Saoffv11015856m	159915209	159915571	+	Other genes	n/a
Saoffv11015857m	159918763	159920917	+	Other genes	n/a
Saoffv11015859m	159923016	159924278	+	Other genes	n/a
Saoffv11015860m	159924592	159926163	-	Other genes	n/a
Saoffv11015861m	159929152	159931071	+	Other genes	n/a
Saoffv11015862m	159931337	159938861	-	Glycosyltransferase	UDPGT_2
Saoffv11015863m	159945587	159946766	-	Glycosyltransferase	UDPGT_2
Saoffv11015868m	159984614	159985216	-	Other genes	n/a
Saoffv11015869m	159992530	159993426	-	Other genes	n/a
Saoffv11015871m	160016001	160023396	-	Glycosyltransferase	UDPGT_2
Saoffv11015872m	160038226	160038384	+	Other genes	n/a
Saoffv11015875m	160043945	160047126	+	(Other) biosynthetic genes	DAHP_synth_2
Saoffv11015876m	160049123	160053313	-	Other genes	n/a
Cluster 4 - Chr03 - Saccharide					
Saoffv11016196m	162620452	162622530	+	Other genes	n/a
Saoffv11016199m	162623692	162625840	-	Glycosyltransferase	Glycos_transf_1
Saoffv11016201m	162629307	162631331	+	Other genes	n/a

(Table A.3.1 continued)

Locus tag	From	To	Strand	Category	Pfam Domains
Saoffv11016203m	162632312	162632803	+	Other genes	n/a
Saoffv11016205m	162636223	162638505	+	Oxidoreductase	n/a
Saoffv11016209m	162644573	162645031	-	Other genes	n/a
Saoffv11016210m	162645688	162648090	+	Oxidoreductase	adh_short
Saoffv11016211m	162650062	162652929	+	Other genes	n/a
Saoffv11016212m	162655347	162658067	+	Other genes	n/a
Saoffv11016216m	162662330	162672371	-	Other genes	n/a
Saoffv11016219m	162679869	162682092	-	BAHD acyltransferase	Transferase
Saoffv11016220m	162685191	162688548	-	Other genes	n/a
Cluster 5 - Chr04 - Saccharide					
Saoffv11018774m	90617308	90617574	+	Other genes	n/a
Saoffv11018775m	90618767	90620230	-	Glycosyltransferase	UDPGT_2
Saoffv11018778m	90629974	90631344	-	Other genes	n/a
Saoffv11018780m	90633290	90639046	-	Methyltransferase	Methyltransf_2
Saoffv11018783m	90750316	90752085	+	Other genes	n/a
Saoffv11018785m	90758728	90761364	-	Dioxygenase	DIOX_N, 2OG-FeII_Oxy
Saoffv11018786m	90761419	90764322	+	Other genes	n/a
Saoffv11018787m	90773169	90774707	-	Other genes	n/a
Saoffv11018791m	90787955	90788407	-	Other genes	n/a
Cluster 6 - Chr05 - Alkaloid					
Saoffv11026782m	138296345	138304222	-	Other genes	n/a
Saoffv11026783m	138313514	138324259	-	Strictosidine synthase-like	Str_synth
Saoffv11026789m	138339588	138341242	+	Other genes	n/a
Saoffv11026790m	138344553	138349927	-	Other genes	n/a
Saoffv11026792m	138352027	138357834	-	Other genes	n/a
Saoffv11026794m	138363400	138364731	+	Other genes	n/a
Saoffv11026796m	138367373	138369893	-	Other genes	n/a
Saoffv11026799m	138376992	138378195	-	Methyltransferase	n/a
Saoffv11026800m	138379271	138381902	+	Other genes	n/a
Saoffv11026801m	138382825	138383828	-	Other genes	n/a
Cluster 7 - Chr08 - Lignan-saccharide					
Saoffv11041753m	113459298	113461298	+	Dioxygenase	DIOX_N, 2OG-FeII_Oxy
Saoffv11041757m	113501123	113501966	-	Other genes	n/a
Saoffv11041758m	113509426	113509816	-	Other genes	n/a
Saoffv11041760m	113581797	113582006	+	Other genes	n/a
Saoffv11041761m	113582103	113583805	+	Dioxygenase	2OG-FeII_Oxy
Saoffv11041764m	113610196	113613011	+	Other genes	n/a
Saoffv11041765m	113631399	113635590	-	Other genes	n/a
Saoffv11041766m	113668349	113673308	-	Other genes	n/a
Saoffv11041769m	113738189	113738620	-	Dirigent enzymes	Dirigent

(Table A.3.1. continued)

Locus tag	From	To	Strand	Category	Pfam Domains
Saoffv11041771m	113788076	113793641	-	Dirigent enzymes	Dirigent
Saoffv11041772m	113809775	113811939	+	Other genes	n/a
Saoffv11041773m	113812676	113817728	-	Other genes	n/a
Saoffv11041776m	113875604	113877284	+	Other genes	n/a
Saoffv11041777m	113878239	113879582	-	Other genes	n/a
Saoffv11041780m	113880772	113884377	-	Glycosyltransferase	UDPGT_2
Saoffv11041782m	114008114	114013535	-	Glycosyltransferase	UDPGT_2
Saoffv11041783m	114016156	114017681	-	Glycosyltransferase	UDPGT_2
Saoffv11041784m	114018448	114019839	-	Glycosyltransferase	UDPGT_2
Saoffv11041785m	114021294	114024308	-	Other genes	n/a
Saoffv11041787m	114056013	114056754	+	Other genes	n/a
Saoffv11041788m	114058759	114080138	+	Other genes	n/a
Cluster 8 - Chr08 - Putative					
Saoffv11042415m	129131319	129134978	-	Other genes	n/a
Saoffv11042418m	129158290	129159868	+	BAHD acyltransferase	Transferase
Saoffv11042422m	129180227	129182078	+	BAHD acyltransferase	Transferase
Saoffv11042423m	129218068	129219429	-	BAHD acyltransferase	Transferase
Saoffv11042424m	129274286	129275162	+	Other genes	n/a
Saoffv11042425m	129282977	129284180	+	BAHD acyltransferase	Transferase
Saoffv11042426m	129290110	129291477	-	BAHD acyltransferase	Transferase
Saoffv11042428m	129316375	129317748	-	BAHD acyltransferase	Transferase
Saoffv11042436m	129591192	129592541	-	BAHD acyltransferase	Transferase
Saoffv11042437m	129599064	129600407	-	BAHD acyltransferase	Transferase
Saoffv11042440m	129617786	129622511	+	Other genes	n/a
Saoffv11042442m	129624758	129625835	+	Other genes	n/a
Saoffv11042443m	129626567	129628864	-	Other genes	n/a
Saoffv11042446m	129684750	129687252	+	Dioxygenase	DIOX_N, 2OG-FeII_Oxy
Saoffv11042447m	129687979	129689082	-	Other genes	n/a
Saoffv11042454m	129711737	129714614	+	Other genes	n/a
Saoffv11042455m	129715760	129720604	-	Other genes	n/a
Saoffv11042456m	129718452	129718676	-	Other genes	n/a
Cluster 9 - Chr09 - Saccharide					
Saoffv11049097m	139661846	139663057	+	Other genes	n/a
Saoffv11049098m	139663665	139665532	+	Other genes	n/a
Saoffv11049101m	139680660	139696753	+	Glycosyltransferase	UDPGT_2
Saoffv11049104m	139742203	139742853	-	Other genes	n/a
Saoffv11049106m	139760003	139761736	+	Other genes	n/a
Saoffv11049108m	139764902	139775892	+	Other genes	n/a

(Table A.3.1. continued)

Locus tag	From	To	Strand	Category	Pfam Domains
Saoffv11049111m	139783673	139785190	+	Glycosyltransferase	UDPGT_2
Saoffv11049112m	139788412	139789899	+	Glycosyltransferase	UDPGT_2
Saoffv11049114m	139790366	139794828	-	Aminotransferase	Aminotran_1_2
Saoffv11049118m	139806469	139812546	+	Other genes	n/a
Cluster 10 - Chr10 - Saccharide					
Saoffv11053421m	119305774	119306475	-	Other genes	n/a
Saoffv11053422m	119307723	119308202	+	Other genes	n/a
Saoffv11053423m	119308612	119310140	-	Glycosyltransferase	UDPGT_2
Saoffv11053424m	119317866	119318508	+	Other genes	n/a
Saoffv11053425m	119319739	119323893	+	Other genes	n/a
Saoffv11053426m	119328449	119329150	+	BAHD acyltransferase	Transferase
Saoffv11053427m	119329431	119329931	+	BAHD acyltransferase	Transferase
Saoffv11053429m	119347456	119350126	-	Other genes	n/a
Cluster 11 - Chr11 - Saccharide					
Saoffv11059653m	128475568	128479293	-	Glycosyltransferase	UDPGT_2
Saoffv11059655m	128489346	128492006	-	Other genes	n/a
Saoffv11059658m	128542434	128543090	-	Other genes	n/a
Saoffv11059659m	128547592	128569274	+	Other genes	n/a
Saoffv11059660m	128592445	128593192	-	Other genes	n/a
Saoffv11059662m	128605786	128606025	+	Other genes	n/a
Saoffv11059667m	128646113	128646564	-	Glycosyltransferase	UDPGT_2
Saoffv11059669m	128652514	128652993	-	Glycosyltransferase	UDPGT_2
Saoffv11059670m	128679322	128682021	+	(Other) biosynthetic genes	DAHP_synth_2
Saoffv11059671m	128683320	128687170	+	Other genes	n/a
Saoffv11059673m	128687846	128692220	-	Other genes	n/a
Saoffv11059675m	128703752	128707071	-	Other genes	n/a
Saoffv11059677m	128708907	128712668	+	Other genes	n/a
Saoffv11059678m	128716324	128716836	-	Other genes	n/a
Cluster 12 - Chr12 - Saccharide					
Saoffv11066305m	134526626	134529077	+	Other genes	n/a
Saoffv11066306m	134531133	134532665	-	Cytochrome 450	p450
Saoffv11066307m	134535856	134539375	-	Other genes	n/a
Saoffv11066308m	134547704	134552744	-	Glycosyltransferase	Glycos_transf_2
Saoffv11066310m	134568100	134572755	-	Other genes	n/a
Saoffv11066311m	134575037	134584659	-	Other genes	n/a
Saoffv11066312m	134587755	134597973	+	Other genes	n/a
Saoffv11066314m	134599709	134599954	-	Other genes	n/a
Saoffv11066315m	134600968	134603956	+	Methyltransferase	Methyltransf_11
Saoffv11066316m	134605383	134610009	+	Other genes	n/a
Saoffv11066317m	134611287	134614956	+	Other genes	n/a

(Table A.3.1. continued)

Locus tag	From	To	Strand	Category	Pfam Domains
Cluster 13 - Chr12 - Saccharide					
Saoffv11066562m	136058059	136060384	+	Other genes	n/a
Saoffv11066564m	136062856	136078161	-	Cytochrome 450	p450
Saoffv11066565m	136082251	136085387	-	Other genes	n/a
Saoffv11066566m	136100160	136104999	+	Other genes	n/a
Saoffv11066567m	136107729	136111820	-	Other genes	n/a
Saoffv11066569m	136129266	136131136	+	Other genes	n/a
Saoffv11066570m	136145869	136150580	+	Other genes	n/a
Saoffv11066571m	136151192	136156074	-	Other genes	n/a
Saoffv11066574m	136167317	136169104	+	Other genes	n/a
Saoffv11066576m	136182853	136192183	-	Glycosyltransferase	Glycos_transf_1
Saoffv11066577m	136194201	136197859	-	Other genes	n/a
Saoffv11066583m	136203860	136207739	+	Other genes	n/a
Cluster 14 - Chr13 - Lignan					
Saoffv11067403m	13799730	13801289	+	Cytochrome 450	p450
Saoffv11067404m	13880547	13882003	+	Other genes	n/a
Saoffv11067409m	13885601	13889964	-	Oxidoreductase	adh_short
Saoffv11067410m	13891998	13898598	-	Other genes	n/a
Saoffv11067413m	13921844	13926374	-	Other genes	n/a
Saoffv11067416m	13929472	13938289	-	Other genes	n/a
Saoffv11067419m	13941795	13944596	-	Other genes	n/a
Saoffv11067421m	13952520	13953674	+	Other genes	n/a
Saoffv11067422m	13954566	13955312	-	Other genes	n/a
Saoffv11067423m	13956365	13956958	+	Dirigent enzymes	Dirigent
Cluster 15 - Chr13 - Polyketide					
Saoffv11070308m	114536129	114539132	+	Other genes	n/a
Saoffv11070309m	114557251	114558072	+	Other genes	n/a
Saoffv11070310m	114558124	114558618	+	Ketosynthase	Chal_sti_synt_C
Saoffv11070314m	114623774	114624367	+	Other genes	n/a
Saoffv11070315m	114639960	114652082	+	Scl acyltransferase	Peptidase_S10
Saoffv11070317m	114677644	114679060	+	Scl acyltransferase	Peptidase_S10
Saoffv11070319m	114697410	114703902	+	Other genes	n/a

B

NMR

B.1 NMR of Saponarioside A

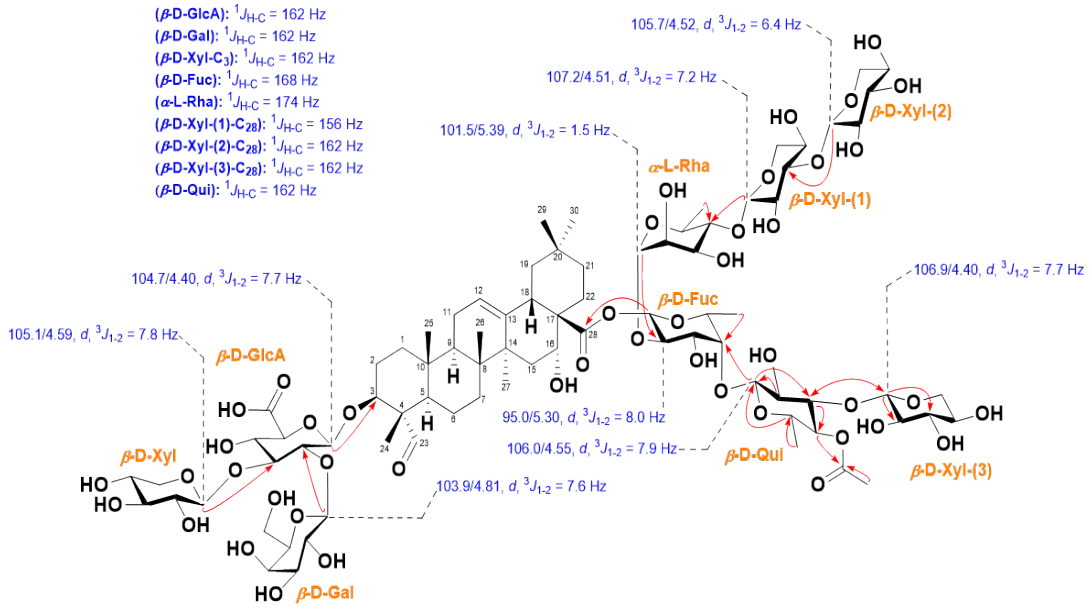


Figure B.1.1. Key HMBC of saponarioside A standard isolated from soapwort leaves. Red arrows indicate H→C. Chemical structure of saponarioside A, 3-*O*-{β-D-xylopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-O-{β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-xylopyranosyl-(1→3)-β-D-4-*O*-acetylquinovopyranosyl-(1→4)]-β-D-fucopyranosyl ester}-quillaic acid

Table B.1.1. ^{1}H, ^{13}C NMR spectroscopic data recorded for saponarioside A standard isolated from soapwort leaves.

No.	δC Type	δH mult (J in Hz)	No.	δC Type	δH mult (J in Hz)
1	39.5, CH2	1.71/1.09, m	C3-Xyl-1	105.1, CH	4.59, d (7.8)
2	25.9, CH2	1.98/1.79, m	C3-Xyl-2	75.5, CH	3.22, m
3	86.4, CH	3.86, m	C3-Xyl-3	78.4, CH	3.29, overlapped
4	56.5, Cq	-	C3-Xyl-4	71.2, CH	3.51, m
5	49.3, CH, overlapped	1.31, m	C3-Xyl-5	67.3, CH2	3.91/3.24, m
6	21.6, CH2	1.47/0.93, m	Fuc-1	95.0, CH	5.30, d (8)
7	33.7, CH2	1.49/1.35, m	Fuc-2	75.2, CH	3.79, m
8	41.2, Cq	-	Fuc-3	77.4, CH	3.75, m
9	48.1, CH	1.73, m	Fuc-4	84.4, CH	3.76, m
10	37.3, Cq	-	Fuc-5	72.2, CH	3.70, m
11	24.6, CH2	1.92/1.92, m	Fuc-6	17.1, CH3	1.24, d (6.3)
12	123.4, CH	5.30, m	Rha-1	101.5, CH	5.39, d (1.5)
13	144.9, Cq	-	Rha-2	71.9, CH	3.92, m
14	43.0, Cq	-	Rha-3	72.5, CH	3.78, m
15	36.7, CH2	1.94/1.44, m	Rha-4	85.3, CH	3.51, m
16	74.6, CH	4.47, m	Rha-5	68.8, CH	3.76, m
17	50.2, Cq	-	Rha-6	18.5, CH3	1.31, d (6.2)
18	42.4, CH	2.94, dd (14.0, 3.8)	C28-Xyl (1)-1	107.2, CH	4.51, d (7.2)
19	48.2, CH2	2.31, t (13.4)/1.05, m	C28-Xyl (1)-2	75.3, CH	3.37, m
20	31.5, Cq	-	C28-Xyl (1)-3	87.5, CH	3.48, m
21	36.8, CH2	1.94/1.17, m	C28-Xyl (1)-4	69.6, CH	3.54, m
22	32.3, CH2	1.92/1.71, m	C28-Xyl (1)-5	67.1, CH2	3.88/3.22, m
23	211.4, CH	9.45, s	C28-Xyl (2)-1	105.7, CH	4.52, d (6.4)
24	11.1, CH3	1.17, s	C28-Xyl (2)-2	75.3, CH	3.37, m
25	16.6, CH3	1.01, s	C28-Xyl (2)-3	77.9, CH	3.38, m
26	17.8, CH3	0.75, s	C28-Xyl (2)-4	71.2, CH	3.57, m
27	27.3, CH3	1.39, s	C28-Xyl (2)-5	67.4, CH2	3.94/3.29, m
28	177.3, Cq	-	Qui-1	106.0, CH	4.45, d (7.9)
29	33.5, CH3	0.88, s	Qui-2	73.7, CH	3.46, m
30	24.9, CH3	0.94, s	Qui-3	85.5, CH	3.68, m
GlcA-1	104.7, CH	4.40, d (7.7)	Qui-4	75.1, CH	4.63, d (9.5)
GlcA-2	78.3, CH	3.64, m	Qui-5	75.2, CH	3.38, m
GlcA-3	86.8, CH	3.68, m	Qui-6	18.2, CH3	1.27, d (6.2)
GlcA-4	71.5, CH	3.56, m	4-OAc	21.3, CH3/172.2, Cq	2.04, s/-

(Table B.1.1. continued)

No.	δC Type	δH mult (J in Hz)	No.	δC Type	δH mult (J in Hz)
GlcA-5	75.2, CH	3.79, m	C28-Xyl (3)-1	106.9, CH	4.40, d (7.7)
GlcA-6	ND	-	C28-Xyl (3)-2	75.8, CH	3.17, m
Gal-1	103.9, CH	4.81, d (7.6)	C28-Xyl (3)-3	77.9, CH	3.38, m
Gal-2	73.6, CH	3.46, m	C28-Xyl (3)-4	71.0, CH	3.51, m
Gal-3	75.3, CH	3.36, m	C28-Xyl (3)-5	67.1, CH2	3.88/3.22, m
Gal-4	70.9, CH	3.81, m			
Gal-5	76.7, CH	3.48, m			
Gal-6	62.3, CH2	3.74/3.74, m			

B.2 NMR of Saponarioside B

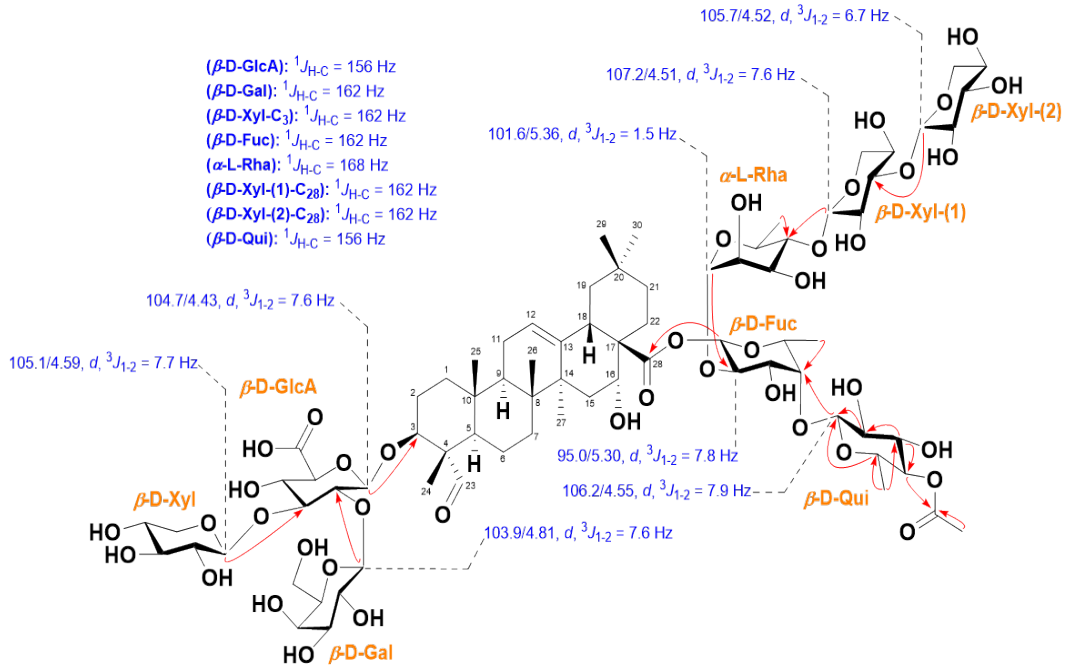


Figure B.2.1. Key HMBC of saponarioside B standard isolated from soapwort leaves. Red arrows indicate H→C. Saponarioside B, 3-*O*-{β-D-xylopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-4-*O*-acetylquinovopyranosyl-(1→4)]-β-D-fucopyranosyl ester}-quillaic acid

Table B.2.1. ^{1}H, ^{13}C NMR spectroscopic data recorded for saponarioside B standard isolated from soapwort leaves.

No.	δC Type	δH mult (*J* in Hz)	No.	δC Type	δH mult (*J* in Hz)
1	39.4, CH2	1.71/1.10, m	Gal-3	75.4, CH	3.37, m
2	25.9, CH2	1.98/1.78, m	Gal-4	70.9, CH	3.82, m
3	86.5, CH	3.87, dd (12.3, 4.8)	Gal-5	76.7, CH	3.49, m
4	56.5, Cq	-	Gal-6	62.3, CH2	3.75/3.75, m
5	49.4, CH, overlapped	1.32, m	C3-Xyl-1	105.1, CH	4.59, d (7.7)
6	21.6, CH2	1.48/0.93, m	C3-Xyl-2	75.5, CH	3.23, m
7	33.7, CH2	1.49/1.35, m	C3-Xyl-3	78.4, CH	3.30, overlapped
8	41.2, Cq	-	C3-Xyl-4	71.2, CH	3.51, m
9	48.2, CH	1.73, m	C3-Xyl-5	67.3, CH2	3.91/3.24, m
10	37.3, Cq	-	Fuc-1	95.0, CH	5.30, d (7.8)
11	24.6, CH2	1.92/1.92, m	Fuc-2	75.5, CH	3.80, m
12	123.4, CH	5.30, m	Fuc-3	77.4, CH	3.75, m
13	144.9, Cq	-	Fuc-4	84.2, CH	3.76, m
14	43.0, Cq	-	Fuc-5	72.2, CH	3.71, m
15	36.7, CH2	1.94/1.44, m	Fuc-6	17.1, CH3	1.24, d (6.4)
16	74.6, CH	4.47, m	Rha-1	101.6, CH	5.36, d (1.5)
17	50.2, Cq	-	Rha-2	71.9, CH	3.92, m
18	42.4, CH	2.94, dd (14.3, 3.9)	Rha-3	72.5, CH	3.79, m
19	48.1, CH2	2.31, t (13.6)/1.05, m	Rha-4	85.3, CH	3.51, m
20	31.5, Cq	-	Rha-5	68.8, CH	3.77, m
21	36.8, CH2	1.94/1.17, m	Rha-6	18.5, CH3	1.31, d (6.2)
22	32.3, CH2	1.93/1.73, m	C28-Xyl (1)-1	107.2, CH	4.51, d (7.6)
23	211.4, CH	9.46, s	C28-Xyl (1)-2	75.3, CH	3.37, m
24	11.1, CH3	1.17, s	C28-Xyl (1)-3	87.5, CH	3.48, m
25	16.6, CH3	1.01, s	C28-Xyl (1)-4	69.6, CH	3.54, m
26	17.8, CH3	0.75, s	C28-Xyl (1)-5	67.1, CH2	3.88/3.22, m
27	27.3, CH3	1.39, s	C28-Xyl (2)-1	105.7, CH	4.52, d (6.7)
28	177.3, Cq	-	C28-Xyl (2)-2	75.3, CH	3.37, m
29	33.5, CH3	0.88, s	C28-Xyl (2)-3	77.9, CH	3.38, m
30	24.9, CH3	0.94, s	C28-Xyl (2)-4	71.2, CH	3.58, m
GlcA-1	104.7, CH	4.43, d (7.6)	C28-Xyl (2)-5	67.4, CH2	3.93/3.28, m
GlcA-2	78.3, CH	3.65, m	Qui-1	106.2, CH	4.55, d (7.9)
GlcA-3	86.8, CH	3.68, m	Qui-2	73.5, CH	3.46, m

(Table B.2.1. continued)

No.	δC Type	δH mult (J in Hz)	No.	δC Type	δH mult (J in Hz)
GlcA-4	71.8, CH	3.56, m	Qui-3	75.2, CH	3.13, m
GlcA-5	75.5, CH	3.79, m	Qui-4	79.2, CH	4.89, d (9.4)
GlcA-6	ND	-	Qui-5	75.2, CH	3.38, m
Gal-1	103.9, CH	4.81, d (7.6)	Qui-6	18.3, CH3	1.27, d (6.1)
Gal-2	73.7, CH	3.47, m	4-O-Ac	20.9/172.8 CH3/Cq	2.11, s/-

B.3 NMR of QA-TriF(Q)RXX

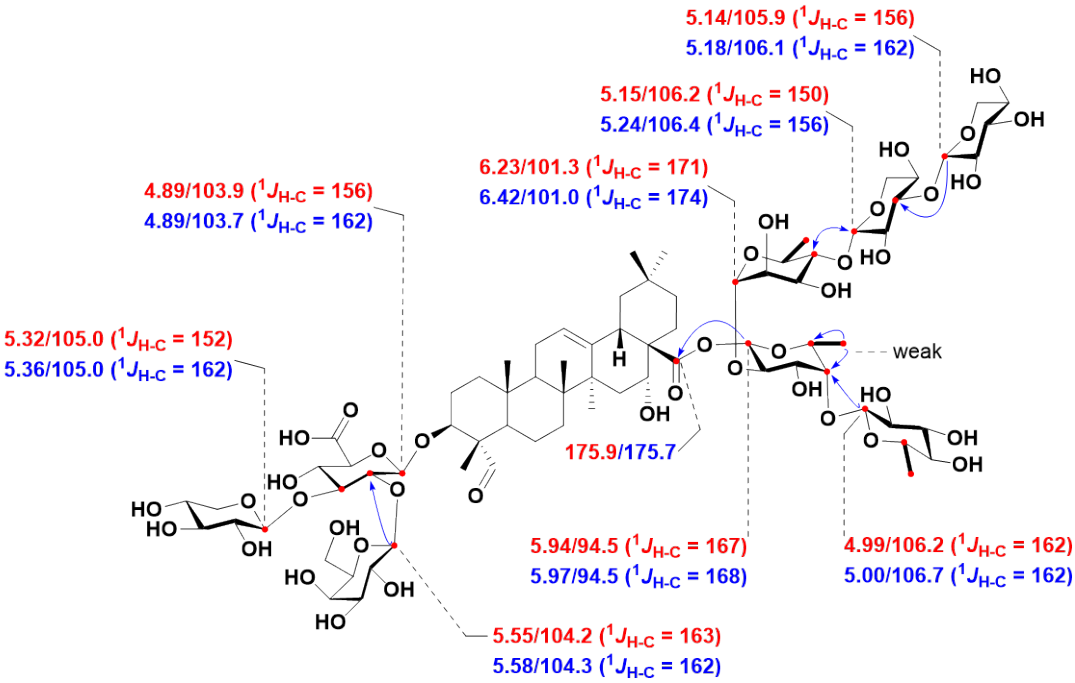


Figure B.3.1. Key HMBC of QA-TriF(Q)RXX produced by transient expression of *SoGH1* in *N. benthamiana*. Blue arrows indicate H→C. QA-TriF(Q)RXX was structurally elucidated as 3-*O*-{β-D-xylopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-*O*-{β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-quinovopyranosyl-(1→4)]-β-D-fucopyranosyl ester}-quillaic acid

Table B.3.1. ^{1}H, ^{13}C NMR spectroscopic data for anomeric protons recorded for QA-TriF(Q)RXX produced by transient expression of *SoGH1* in *N. benthamiana.*

Sugar	Recorded ^{1}H-NMR	^{13}C-NMR	Coupled HSQC ($^{1}J_{H-C}$) Hz	Literature ^{1}H-NMR	^{13}C-NMR	Coupled HSQC ($^{1}J_{H-C}$) Hz
GlcA-1	4.89	103.7	162	4.89	103.9	156
Gal-1	5.58	104.3	162	5.55	104.2	163
Xyl(C-3)-1	5.36	105	162	5.32	105	152
Fuc-1	5.97	94.5	168	5.94	94.5	167
Rha-1	6.42	101	174	6.23	101.3	171
Xyl'(C-28)-1	5.24	106.4	156	5.15	106.2	150
Xyl"(C-28)-1	5.18	106.1	162	5.14	105.9	156
Qui-1	5	106.7	162	4.99	106.2	162

C

Gene discovery and characterization

C.1 Literature sequences used as BLAST queries

Table C.1.1. List of literature OSCs used as BLASTP queries and in phylogenetic analysis of candidate soapwort OSCs. The sequences were retrieved from (Thimmappa *et al.*, 2014), except QsbAS was retrieved from (Reed *et al.*, 2023). Reference for each sequence can be found through their unique GeneBank/Uniprot ID.

Name	GenBank/Uniprot ID	Species
CrCAS	XP_042928216	*Chlamydomonas reinhardtii*
PgPNZ1	AB009031	*Panax ginseng*
LjOSC7	AB244671	*Lotus japonicus*
AtLAS1	AT3G45130	*Arabidopsis thaliana*
OsOSC2	AK121211	*Oryza sativa*
AsCS1	AJ311790	*Avena strigosa*
PgPNX	AB009029	*Panax ginseng*
AtCAS1	AT2G07050	*Arabidopsis thaliana*
PsPSX	D89619	*Pisum sativum*
LjOSC5	AB181246	*Lotus japonicus*
LjOSC3	AB181245	*Lotus japonicus*
GgLUS1	AB116228	*Glycyrrhiza glabra*
AsOXA1	AY836006	*Aster sedifolius*
AaBAS	EU330197	*Artemisia annua*
SlTTS1	HQ266579	*Solanum lycopersicum*
PgPNY1	AB009030	*Panax ginseng*
VhBS	DQ915167	*Saponaria vaccaria*
QsbAS	OQ107256	*Quillaja saponaria*
LjOSC1	AB181244	*Lotus japonicus*
GgbAS1	AB037203	*Glycyrrhiza glabra*
PsPSY	AB034802	*Pisum sativum*
MtbAS1	AJ430607	*Medicago truncatula*

Table C.1.2. List of literature CSLs used as BLASTP queries and in phylogenetic analysis of candidate soapwort CSLs. Literature cellulose synthase (CesA) sequences were also used to build CSL phylogenetic tree. The sequences were retrieved from (Carroll and Specht, 2011; Jozwiak *et al.*, 2020; Chung *et al.*, 2020; Reed *et al.*, 2023). Reference for each sequence can be found through their unique GeneBank/Uniprot ID.

Name	GenBank/UniProt ID	Species
CESA1	O48946	*Arabidopsis thaliana*
CESA2	O48947	*Arabidopsis thaliana*
CESA3	Q941L0	*Arabidopsis thaliana*
CESA4	Q84JA6	*Arabidopsis thaliana*
CESA5	Q8L778	*Arabidopsis thaliana*
CESA6	Q94JQ6	*Arabidopsis thaliana*
CESA7	Q9SWW6	*Arabidopsis thaliana*
CESA8	Q8LPK5	*Arabidopsis thaliana*
CESA9	Q9SJ22	*Arabidopsis thaliana*
CSLB1	O80898	*Arabidopsis thaliana*
CSLB2	O80899	*Arabidopsis thaliana*
CSLD1	O49323	*Arabidopsis thaliana*
CSLD2	Q9LFL0	*Arabidopsis thaliana*
CSLE1	Q8VZK9	*Arabidopsis thaliana*
CSLE2	Q0DXZ1	*Oryza sativa* Japonica
CSLF1	Q6ZF89	*Oryza sativa* Japonica
CSLF2	Q84S11	*Oryza sativa* Japonica
CSLG1	Q570S7	*Arabidopsis thaliana*
CSLG2	Q8VYR4	*Arabidopsis thaliana*
CSLH1	Q339N5	*Oryza sativa* Japonica
CSLH2	Q7PC71	*Oryza sativa* Indica
GmCSLM1	BBN60792	*Glycine max*
SlCSLM	XP_004234035	*Solanum lycopersicum*
GmCSyGT1	BBN60789	*Glycine max*
GuCSyGT	BBN60794	*Glycyrrhiza uralensis*
SOAP5	XP_021842158	*Spinacia oleracea*
LjCSyGT	BBN60795	*Lotus japonicus*
QsCSLM1	WEU75093	*Quillaja saponaria*
VvCSLM	CBI26389	*Vitis vinifera*
SmCESA0A	EFJ17343	*Selaginella moellendorffii*

Table C.1.3. List of literature BAHD ATs used as BLASTP queries to mine for BAHD AT candidates in soapwort. The sequences were retrieved from (Bontpart *et al.*, 2015). Reference for each sequence can be found through their unique GeneBank/Uniprot ID.

Name	GenBank/Uniprot ID	Species
At3AT1	NP171890	*Arabidopsis thaliana*
Gt5AT	BAA74428	*Gentiana triflora*
Dv3MAT	AAO12206	*Dahlia variabilis*
Sc3MaT	AAO38058	*Pericallis cruenta*
NtMAT1	BAD93691	*Nicotiana tabacum*
Glossy2	CAA61258	*Zea mays*
CER2	CAA61258	*Arabidopsis thaliana*
Ss5MaT2	AAR26385	*Salvia splendens*
VAAT	CAC09062	*Fragaria vesca*
SalAT	AAK73661	*Papaver somniferum*
Pun1	AAV66311	*Capsicum annum*
ACT	AAO73071	*Hordeum vulgare*
AtHCT	NP199704	*Arabidopsis thaliana*
CHAT	AAN09797	*Arabidopsis thaliana*
TAT	AAF34254	*Taxus cuspidata*
TpHCT1A	ACI16630	*Trifolium pratense*

Table C.1.4. List of literature SCPL ATs used as BLASTP queries and in phylogenetic analysis of candidate soapwort SCPL ATs. The sequences were retrieved from (Bontpart *et al.*, 2018). Reference for each sequence can be found through their unique GeneBank/Uniprot ID.

Name	GenBank/Uniprot ID	Species
HvSCPI	AAA32940	*Hordeum vulgare*
SCPLe	AAF44708	*Solanum lycopersicum*
SpGAC	AAF64227	*Solanum pennellii*
AtSMT	AAF78760	*Arabidopsis thaliana*
AtSCT	AAK52316	*Arabidopsis thaliana*
BnSCT1	AAQ91191	*Brassica napus*
AtSCPL17	AAS99709	*Arabidopsis thaliana*
AsSCPL1	ACT21078	*Avena sativa*
AtSAT	AEC07395	*Arabidopsis thaliana*
AtSST	AEC07397	*Arabidopsis thaliana*
MtSCP1	AES71587	*Medicago truncatula*
CsSCPL	AIW39897	*Camellia sinensis*
DkSCPL1	BAF56655	*Diospyros kaki*
CtAT1	BAF99695	*Clitoria ternalea*
DkSCPL2	BAH89272	*Diospyros kaki*
DgSCPL1	BAO04182	*Delphinium grandiflorum*
DgSCPL2	BAO04183	*Delphinium grandiflorum*
DgSCPL3	BAO04184	*Delphinium grandiflorum*
HvSCPIII	CAA70817	*Hordeum vulgare*
HvCBPII	CAB59202	*Hordeum vulgare*
BRS1	CAB79779	*Arabidopsis thaliana*
SbHNL1	CAD12888	*Sorghum bicolor*
TaCBPII	CAI64396	*Triticum aestivum*
BnSCT2	CAM91991	*Brassica napus*
ScCPY	NP_014026	*Saccharomyces cerevisae*

C.2 Full GC/MS spectra of heterologous expression experiments

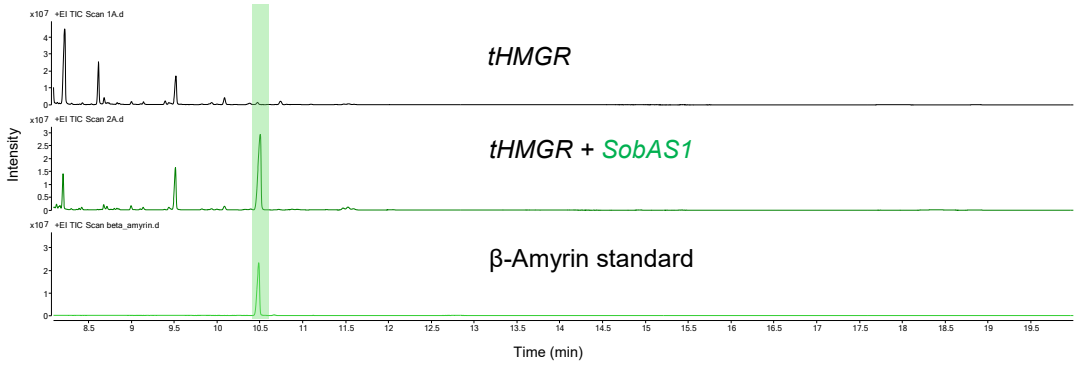


Figure C.2.1. Activity of *SobAS1* transiently expressed in *N. benthamiana*. The full GC-MS total ion chromatograms (TICs) of leaf extracts co-expressing *AstHMGR* (*tHMGR*) and *SobAS1*, along with a control (leaf only expressing *AstHMGR*) and a commercial β-amyrin standard are shown. Highlighted peak corresponds to β-amyrin.

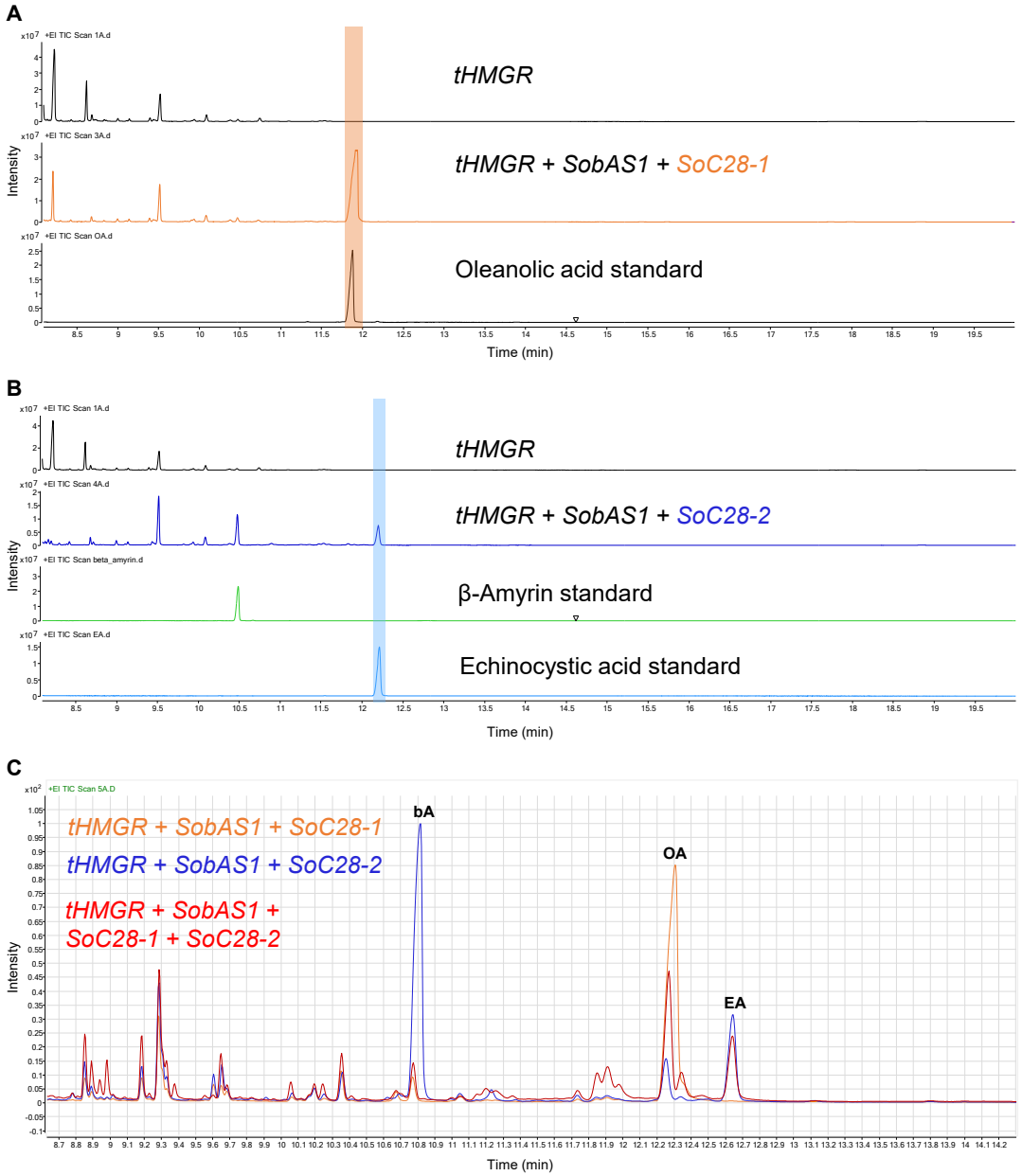


Figure C.2.2. Activity of *SoC28-1* and *SoC28-2* transiently expressed in *N. benthamiana*. The full GC-MS total ion chromatograms (TICs) of leaf extracts co-expressing *AstHMGR* (*tHMGR*), *SobAS1*, and either **(A)** *SoC28-1* or **(B)** *SoC28-2* are shown. A control (leaf only expressing *AstHMGR*) and commercial standards are shown. Highlighted peaks corresponds to oleanolic acid (orange) and echinocystic acid (blue). **(C)** Co-expression of both *SoC28-1* and *SoC28-2* together with *SobAS1* did not lead to increased production of echinocystic acid (EA). bA, β-amyrin; OA, oleanolic acid.

C.3 Full LC/MS spectra of heterologous expression experiments

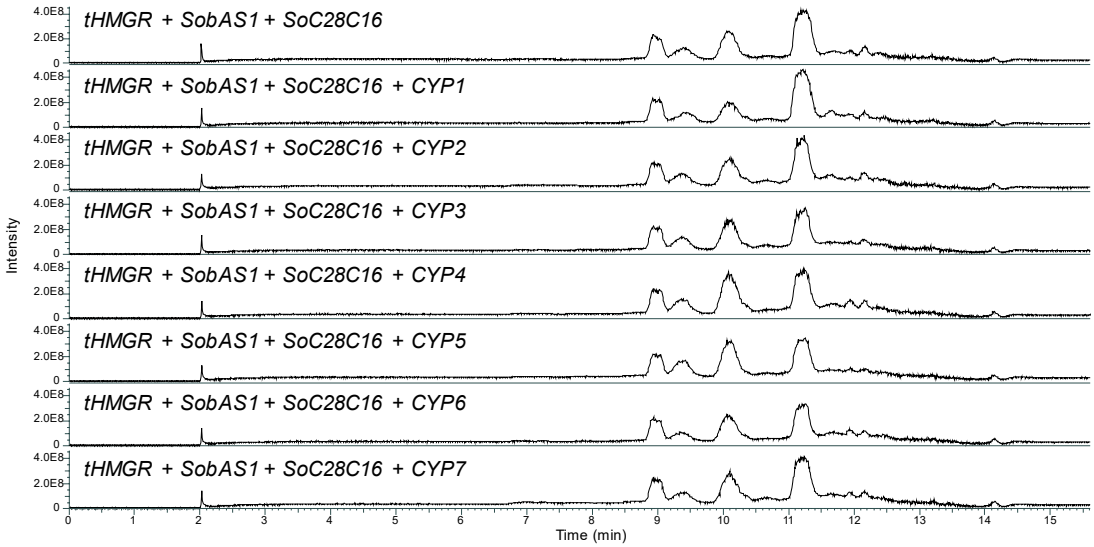


Figure C.3.1. Candidate soapwort CYPs transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing the minimal gene set to produced echinocystic acid (*AstHMGR*, *SobAS1*, *SoC28C16*) and one of soapwort CYP candidates (1-7) are shown. *tHMGR = AstHMGR.*

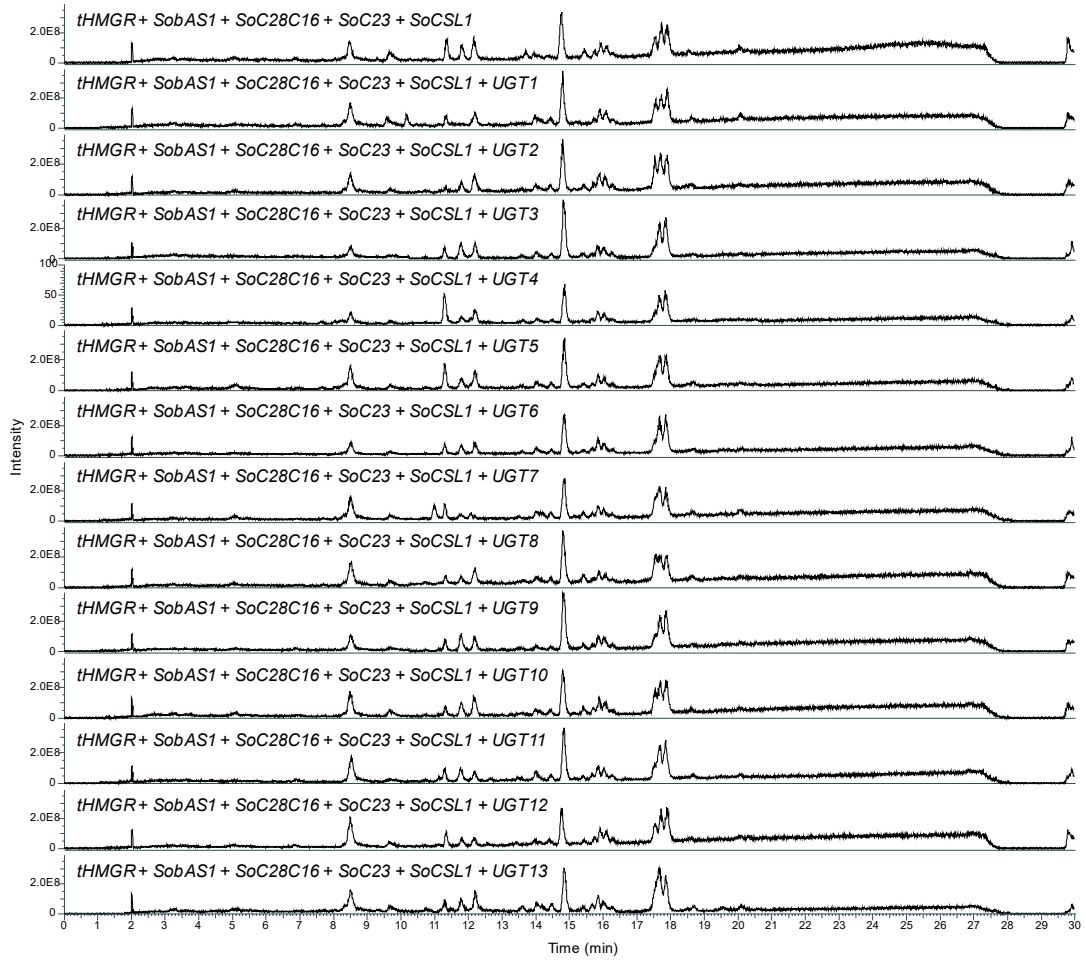


Figure C.3.2. Testing candidate soapwort UGTs for C-3 galactosyltransferase activity. Candidate soapwort UGTs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing genes to produce compound **2** (*AstHMGR*, *SobAS1*, *SoC28C16, SoC23 and SoCSL1*) (QA-Mono) and one of soapwort UGT candidates (1-13) are shown. *tHMGR = AstHMGR.*

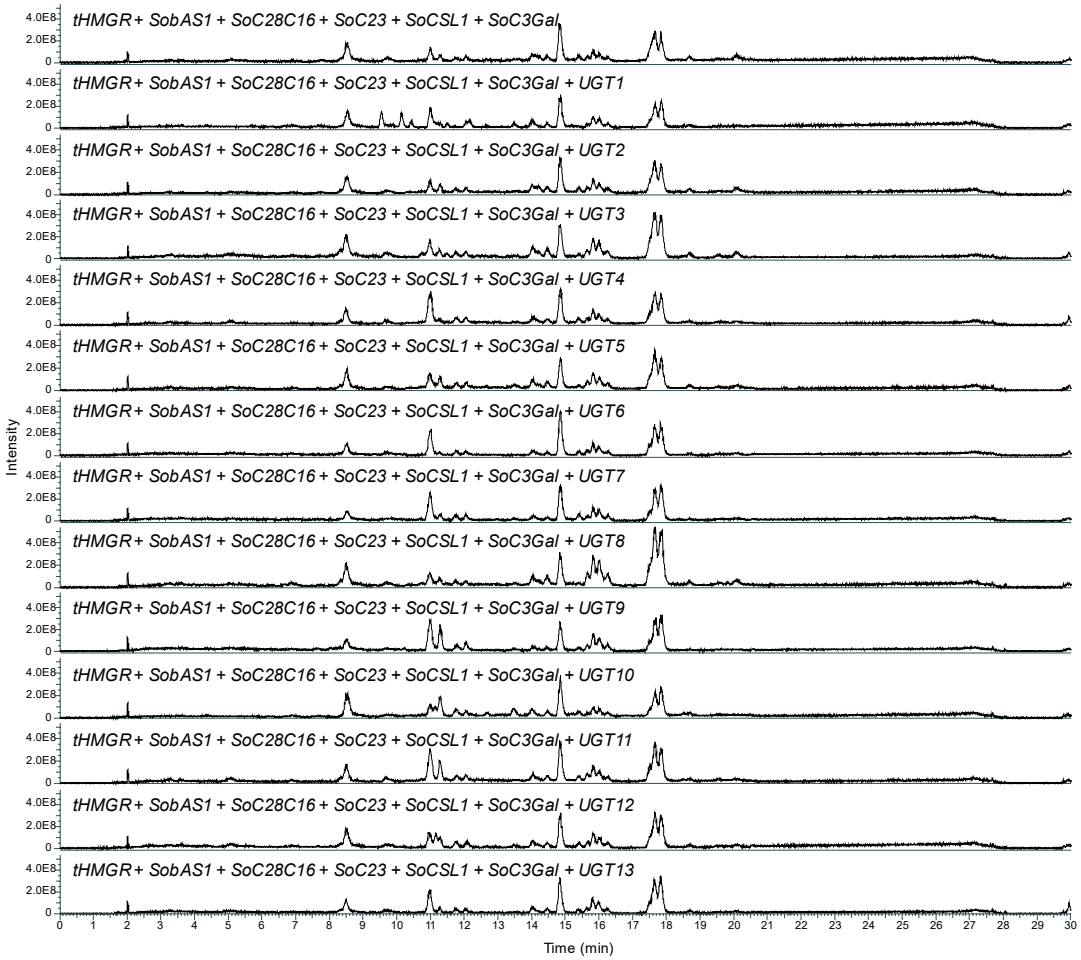


Figure C.3.3. Testing candidate soapwort UGTs for C-3 xylosyltransferase activity. Candidate soapwort UGTs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing genes to produce compound **3** (*AstHMGR*, *SobAS1*, *SoC28C16, SoC23, SoCSL1 and SoC3Gal*) (QA-Di) and one of soapwort UGT candidates (1-13) are shown. *tHMGR* = *AstHMGR.*

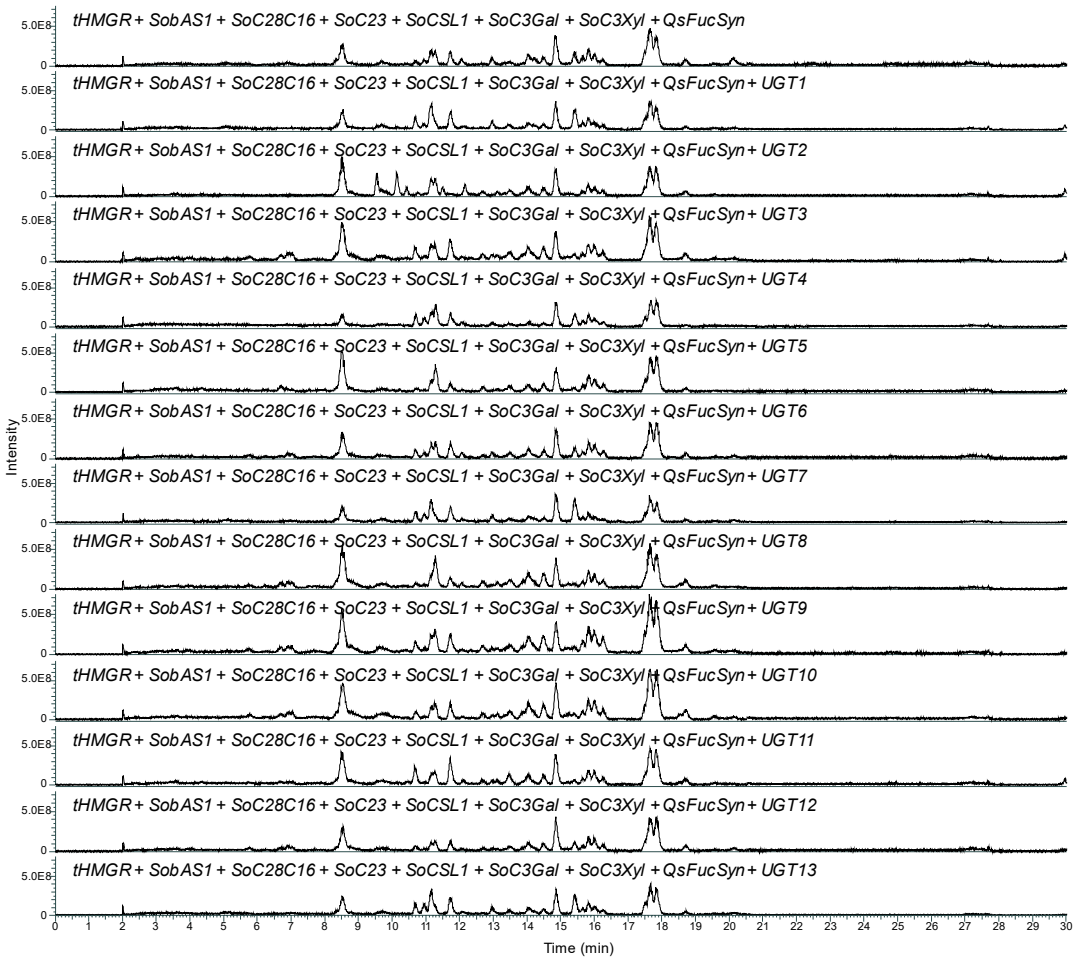


Figure C.3.4. Testing candidate soapwort UGTs for C-28 fucosyltransferase activity. Candidate soapwort UGTs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing genes to produce compound **4** (*AstHMGR*, *SobAS1*, *SoC28C16, SoC23, SoCSL1, SoC3Gal, SoC3Xyl* and *QsFucSyn*) (QA-Tri) and one of soapwort UGT candidates (1-13) are shown. *tHMGR = AstHMGR*. *QsFucSyn* from *Q. saponaria* (Reed *et al.*, 2023) was co-expressed to increase the production of D-fucose in *N. benthamiana*.

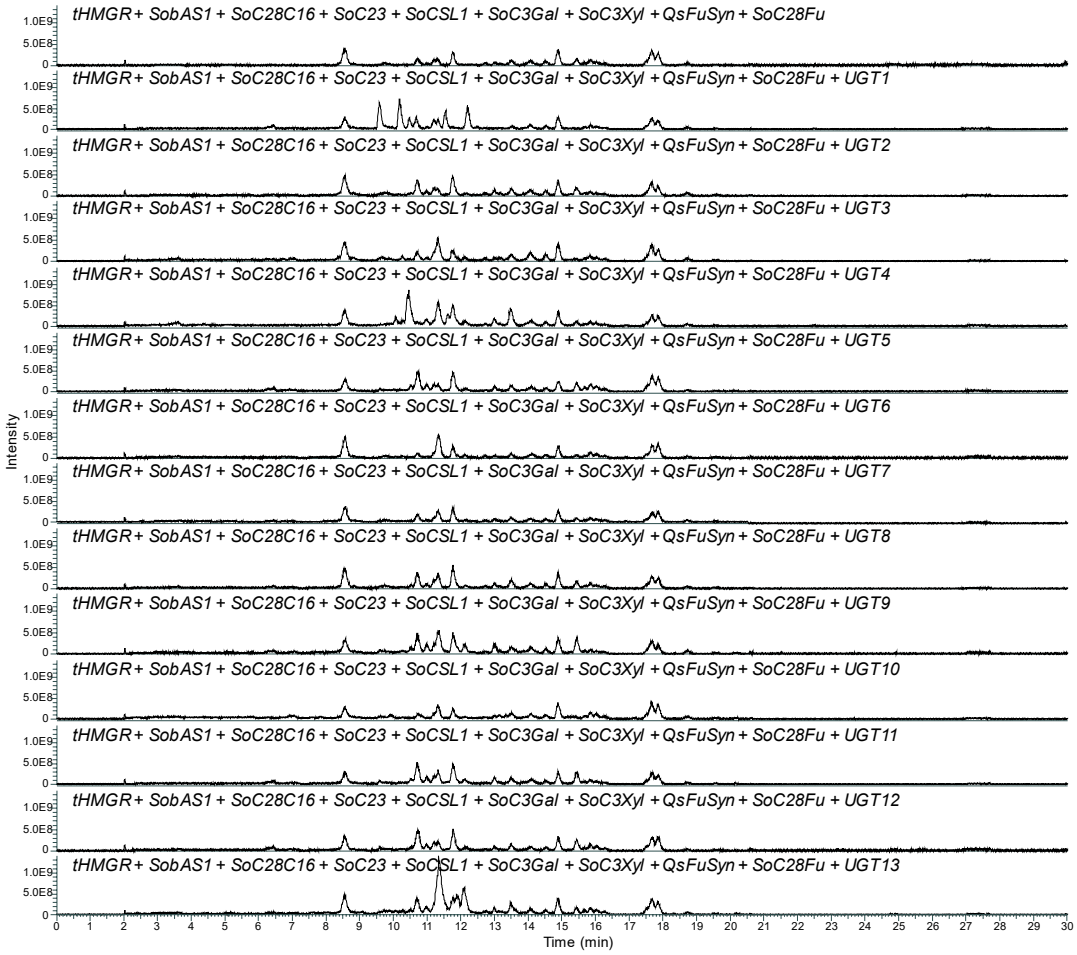


Figure C.3.5. Testing candidate soapwort UGTs for C-28 rhamnosyltransferase activity. Candidate soapwort UGTs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing genes to produce compound **5** (*AstHMGR*, *SobAS1*, *SoC28C16, SoC23, SoCSL1, SoC3Gal, SoC3Xyl, QsFucSyn* and *SoC28Fu*) (QA-TriF) and one of soapwort UGT candidates (1-13) are shown. *tHMGR = AstHMGR*. *QsFucSyn* from *Q. saponaria* (Reed *et al.*, 2023) was co-expressed to increase the production of D-fucose in *N. benthamiana*.

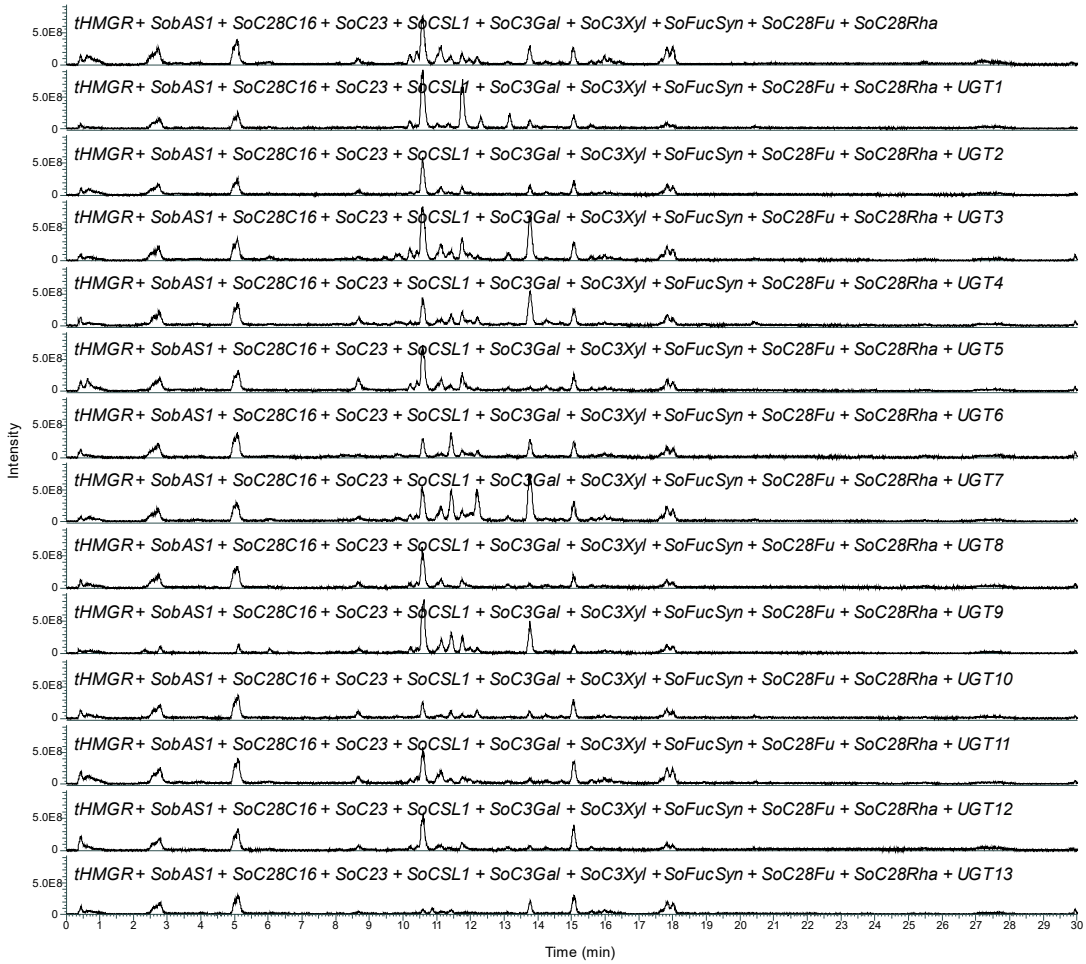


Figure C.3.6. Testing candidate soapwort UGTs for C-28 xylosyltransferase activity. Candidate soapwort UGTs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing genes to produce compound **6** (*AstHMGR*, *SobAS1*, *SoC28C16, SoC23, SoCSL1, SoC3Gal, SoC3Xyl, SoFucSyn, SoC28Fu* and *SoC28Rha*) (QA-TriFR) and one of soapwort UGT candidates (1-13) are shown. *tHMGR = AstHMGR*. *SoSDR1* (*SoFucSyn*) identified from *S. officinalis* was co-expressed to increase the production of D-fucose in *N. benthamiana* instead of *QsFucSyn*.

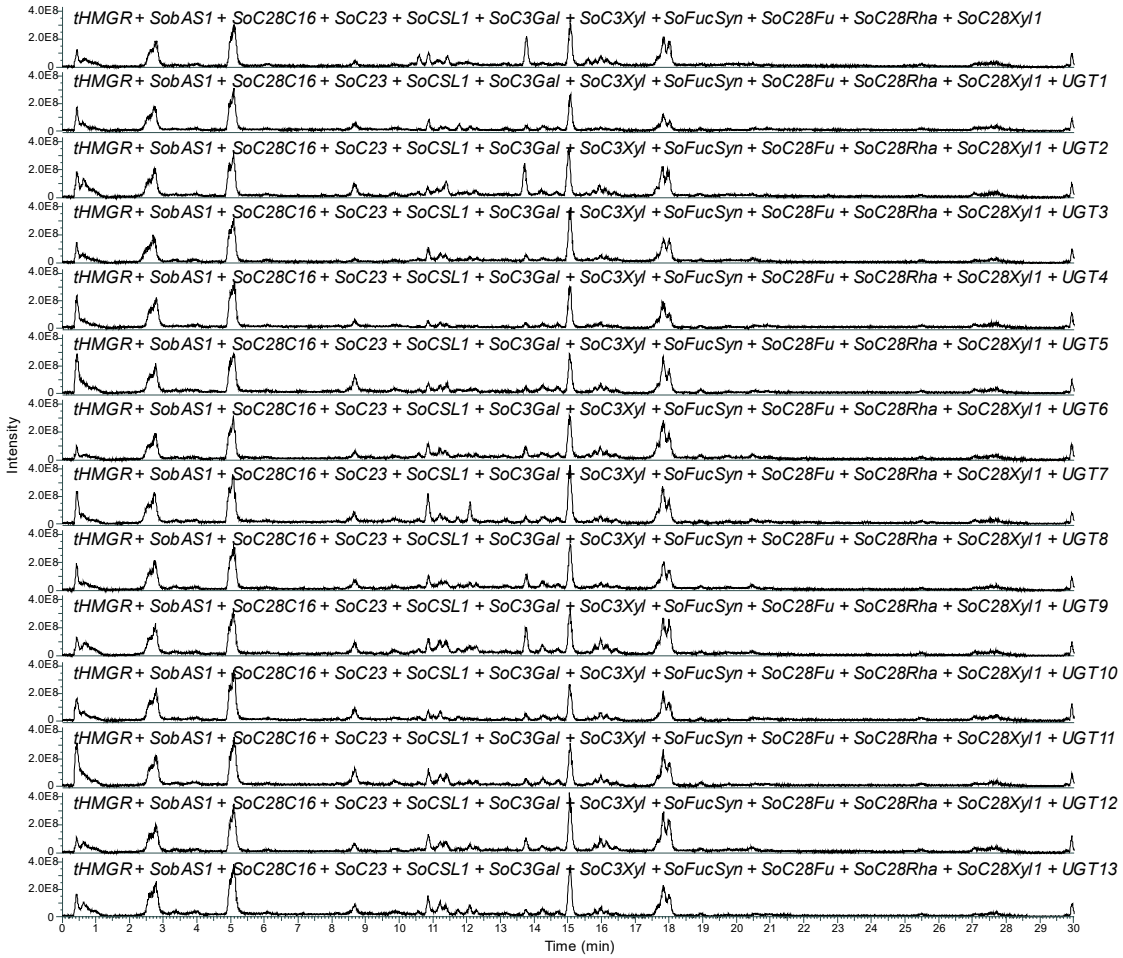


Figure C.3.7. Testing candidate soapwort UGTs for second C-28 xylosyltransferase activity. Candidate soapwort UGTs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing genes to produce compound **7** (*AstHMGR*, *SobAS1*, *SoC28C16, SoC23, SoCSL1, SoC3Gal, SoC3Xyl, SoFucSyn, SoC28Fu*, *SoC28Rha* and *SoC28Xyl1*) (QA-TriFRX) and one of soapwort UGT candidates (1-13) are shown. *tHMGR* = *AstHMGR*. *SoSDR1* (*SoFucSyn*) was co-expressed to increase the production of D-fucose in *N. benthamiana*.

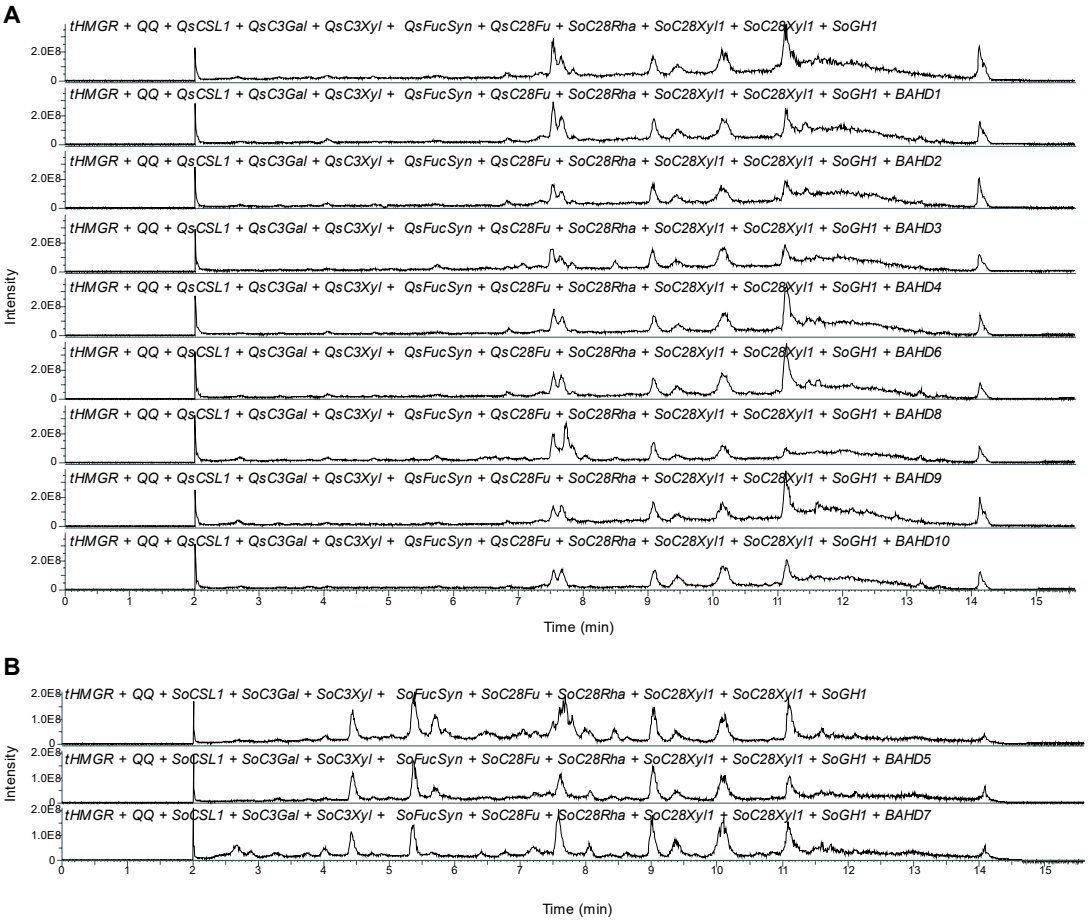


Figure C.3.8. Testing candidate soapwort BAHD ATs for acetyltransferase activity. Candidate soapwort BAHD ATs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing mixture of genes from *S. officinalis* and *Q. saponaria* to produce compound **10** (QA-TriF(Q)RXX) and one of soapwort BAHD AT candidates (1-10) are shown. *tHMGR = AstHMGR*. Experiments **(A)** and **(B)** were performed at separate times and used readily available *A. tumefaciens* strains carrying either *S. officinalis* (*So*) or *Q. saponaria* (*Qs*) genes to produce **10** at the time of the experiment. QQ, goldengate vector harbouring *Q. saponaria* genes required to produce quillaic acid (*QsbAS* + *QsC28* + *QsC16* + *QsC23*) (Reed *et al.*, 2023). *SoSDR1* (*SoFucSyn*) or *QsFucSyn* was co-expressed to increase the production of D-fucose in *N. benthamiana*.

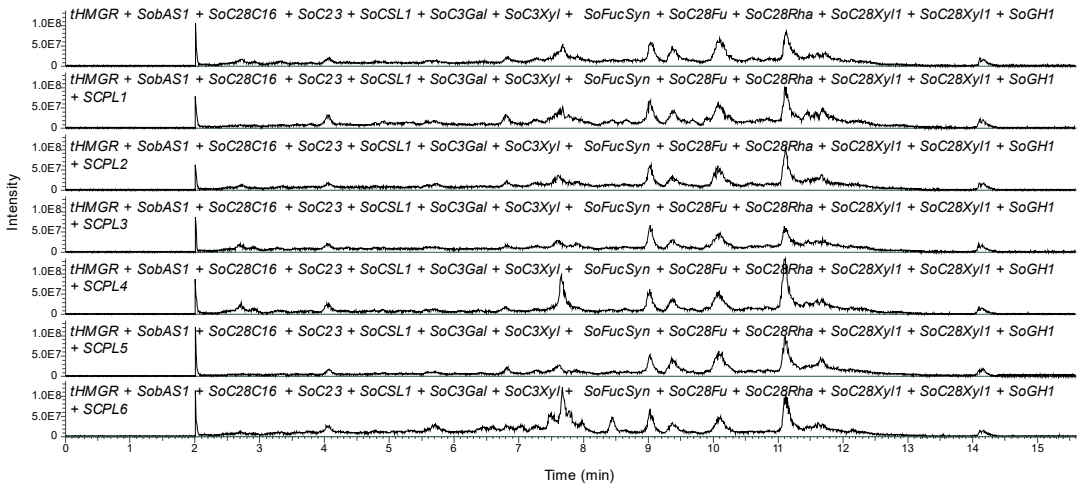


Figure C.3.9. Testing candidate soapwort SCPL ATs for acetyltransferase activity. Candidate soapwort SCPL ATs were transiently expressed in *N. benthamiana.* The full LC-MS total ion chromatograms (TICs) of leaf extracts co-expressing genes to produce compound **10** (*AstHMGR*, *SobAS1*, *SoC28C16, SoC23, SoCSL1, SoC3Gal, SoC3Xyl, SoFucSyn, SoC28Fu*, *SoC28Rha, SoC28Xyl1, SoC28Xyl2* and *SoGH1*) (QA-TriF(Q)RXX) and one of soapwort SCPL AT candidates (1-5) are shown. *tHMGR* = *AstHMGR*. *SoSDR1* (*SoFucSyn*) was co-expressed to increase the production of D-fucose in *N. benthamiana*.

D

Co-expression analysis

D.1 Co-expression analysis of *S. officinalis* candidate genes

Table D.1.1. List of *S. officinalis* genes showing positive correlation (PCC ≥ 0.500) with SobAS1 expression. Genes annotated as either CYP, CSL, UGT or AT based on the human readable description (AHRD) are shown. SoGH1 (TRINITY_DN530_c2_g1) is included for comparison.

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN1084_c0_g4	1.000	Terpene cyclase/mutase family member	IPR018333 (Squalene cyclase), IPR032696 (Squalene cyclase, C-terminal), IPR032697 (Squalene cyclase, N-terminal)
TRINITY_DN645_c1_g2	0.989	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN3796_c0_g1	0.987	Cellulose synthase	IPR005150 (Cellulose synthase)
TRINITY_DN1618_c1_g2	0.986	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN28657_c0_g1	0.981	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN1473_c3_g1	0.980	vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN5570_c0_g3	0.979	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN5701_c1_g1	0.975	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN21519_c0_g1	0.972	Cellulose synthase	IPR005150 (Cellulose synthase)
TRINITY_DN54808_c0_g7	0.972	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN530_c2_g1	0.971	Beta-glucosidase, putative	IPR001360 (Glycoside hydrolase family 1), IPR017853 (Glycoside hydrolase superfamily)
TRINITY_DN651_c0_g3	0.969	Cytochrome P450	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN345366_c0_g1	0.969	Cellulose synthase	IPR005150 (Cellulose synthase), IPR013083 (Zinc finger, RING/FYVE/PHD-type), IPR029044 (Nucleotide-diphospho-sugar transferases)
TRINITY_DN5729_c1_g1	0.967	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN5570_c0_g1	0.961	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN51550_c0_g1	0.960	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN347728_c0_g1	0.956	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN2822_c1_g3	0.954	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN41181_c0_g1	0.954	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN2993_c0_g1	0.954	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN342_c0_g1	0.953	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN5422_c7_g1	0.950	UDP-glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN13626_c1_g2	0.946	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN53369_c0_g1	0.946	Ent-kaurenoic acid oxidase	IPR001128 (Cytochrome P450)
TRINITY_DN69958_c0_g1	0.946	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN3011_c0_g3	0.941	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN14107_c4_g1	0.938	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN58802_c0_g3	0.931	Cytochrome P450 family protein	IPR001128 (Cytochrome P450)
TRINITY_DN3341_c0_g1	0.921	vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN5664_c0_g3	0.921	N-acetyltransferase domain-containing protein	IPR001128 (Cytochrome P450)
TRINITY_DN221488_c0_g1	0.916	omega-hydroxypalmitate O-feruloyl transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN283414_c0_g1	0.916	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN23622_c0_g2	0.915	Cellulose synthase	IPR005150 (Cellulose synthase)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN31287_c0_g2	0.914	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN10898_c0_g1	0.914	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN586_c1_g1	0.909	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN8790_c0_g3	0.909	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN47434_c0_g2	0.908	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN40880_c0_g3	0.908	Transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN47337_c0_g1	0.907	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN111518_c0_g1	0.907	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN349059_c0_g1	0.907	Cytochrome P450 6B7	IPR001128 (Cytochrome P450)
TRINITY_DN234703_c0_g1	0.904	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN4499_c3_g1	0.903	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN5664_c0_g1	0.903	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN171936_c0_g1	0.903	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN26666_c0_g1	0.902	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN46549_c0_g1	0.900	Cellulose synthase	IPR005150 (Cellulose synthase)
TRINITY_DN14375_c0_g2	0.899	Abscisic acid 8'-hydroxylase	IPR001128 (Cytochrome P450)
TRINITY_DN2549_c0_g1	0.898	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN348845_c0_g1	0.898	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN27658_c0_g1	0.897	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN47780_c0_g2	0.895	UDP-glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN11658_c0_g2	0.894	Cellulose synthase	IPR005150 (Cellulose synthase), IPR029044 (Nucleotide-diphospho-sugar transferases)
TRINITY_DN302393_c0_g1	0.892	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN187995_c0_g1	0.891	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN31873_c2_g1	0.891	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN14107_c6_g1	0.890	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)

TRINITY_DN5184_c0_g1	0.889	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN336578_c0_g1	0.889	Cytochrome P450	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN44858_c0_g1	0.886	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN5384_c0_g3	0.886	Vinorine synthase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN2154_c1_g2	0.886	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN10048_c0_g1	0.885	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN54846_c0_g1	0.883	Cytochrome P450 4F22	IPR001128 (Cytochrome P450)
TRINITY_DN29178_c1_g1	0.882	Vinorine synthase-like protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN6661_c0_g1	0.881	UDP-glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN140829_c0_g1	0.881	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN55859_c0_g1	0.881	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN86763_c0_g1	0.880	spermidine hydroxycinnamoyl transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN41684_c0_g4	0.880	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN57970_c0_g1	0.879	Cellulose synthase	IPR005150 (Cellulose synthase), IPR013083 (Zinc finger, RING/FYVE/PHD-type), IPR029044 (Nucleotide-diphospho-sugar transferases)
TRINITY_DN206380_c0_g1	0.878	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN81626_c0_g1	0.876	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN95894_c0_g1	0.876	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN69950_c0_g1	0.872	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN5555_c0_g1	0.870	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN16150_c3_g1	0.867	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN11697_c0_g1	0.866	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)

TRINITY_DN354780_c0_g1	0.865	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN65986_c0_g1	0.865	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN225683_c0_g1	0.864	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN3451_c1_g2	0.864	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN315042_c0_g1	0.863	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN3341_c0_g4	0.863	vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN469_c0_g1	0.861	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN41487_c0_g1	0.861	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN54846_c0_g2	0.861	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN320_c1_g2	0.861	omega-hydroxypalmitate O-feruloyl transferase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN200239_c0_g1	0.857	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN183736_c0_g1	0.856	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN284374_c0_g1	0.856	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN8560_c0_g1	0.855	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN86505_c0_g1	0.855	Cellulose synthase	IPR005150 (Cellulose synthase)
TRINITY_DN24060_c2_g2	0.855	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN64217_c0_g1	0.854	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN823_c1_g1	0.853	Cellulose synthase	IPR005150 (Cellulose synthase), IPR029044 (Nucleotide-diphospho-sugar transferases)
TRINITY_DN135458_c0_g1	0.853	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN19883_c0_g5	0.852	Cellulose synthase	IPR005150 (Cellulose synthase), IPR013083 (Zinc finger, RING/FYVE/PHD-type)
TRINITY_DN21123_c0_g2	0.852	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN71518_c0_g1	0.851	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN2172_c1_g1	0.851	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN101327_c0_g1	0.844	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN9977_c0_g1	0.841	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)

TRINITY_DN507_c0_g1	0.840	Glyco_transf_28 domain-containing protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase), IPR004276 (Glycosyltransferase family 28, N-terminal domain)
TRINITY_DN32485_c0_g4	0.839	UDP-glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN30153_c0_g1	0.839	Geraniol 8-hydroxylase	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN2210_c0_g1	0.836	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN1707_c1_g2	0.835	BAHD acyltransferase DCR	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN316011_c0_g1	0.834	vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN97163_c0_g1	0.832	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN4777_c0_g3	0.826	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN370449_c0_g1	0.825	Vinorine synthase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN101327_c0_g6	0.825	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN70669_c0_g2	0.822	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN200550_c0_g1	0.820	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN18871_c1_g1	0.820	vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN205600_c0_g1	0.819	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN98247_c0_g1	0.815	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN31985_c0_g1	0.812	Vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN133_c0_g2	0.810	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN44187_c6_g1	0.808	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN7831_c0_g3	0.808	Cytochrome P450	IPR001128 (Cytochrome P450)

TRINITY_DN42024_c0_g1	0.808	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN300781_c0_g1	0.808	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN43050_c0_g1	0.807	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN215579_c0_g1	0.806	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN71147_c0_g2	0.805	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN316434_c0_g1	0.803	Cytochrome P450	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN78115_c0_g1	0.800	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN319_c5_g2	0.799	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN4811_c1_g2	0.799	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN145201_c0_g1	0.799	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN5932_c0_g1	0.796	Terpene cyclase/mutase family member	IPR018333 (Squalene cyclase), IPR032696 (Squalene cyclase, C-terminal)
TRINITY_DN3977_c2_g2	0.792	Cellulose synthase, putative	IPR005150 (Cellulose synthase)
TRINITY_DN44344_c0_g1	0.791	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN38479_c0_g1	0.788	Cellulose synthase	IPR005150 (Cellulose synthase), IPR029044 (Nucleotide-diphospho-sugar transferases)
TRINITY_DN112951_c0_g1	0.788	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN49528_c0_g1	0.788	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN7600_c1_g1	0.787	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN63648_c0_g1	0.787	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN227820_c0_g1	0.784	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN153392_c0_g2	0.782	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN30822_c0_g1	0.782	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN26926_c0_g1	0.782	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN4739_c0_g2	0.780	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN6399_c0_g1	0.779	Cytochrome P450-like protein	IPR001128 (Cytochrome P450)
TRINITY_DN27217_c0_g5	0.778	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN15401_c0_g1	0.777	Cytochrome P450, putative	IPR001128 (Cytochrome P450)

TRINITY_DN36550_c1_g3	0.774	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN92303_c0_g1	0.774	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN227534_c0_g1	0.772	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN16060_c0_g2	0.770	UDP-glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN63553_c0_g1	0.766	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN20638_c0_g2	0.765	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN26899_c0_g1	0.762	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN49669_c0_g1	0.762	Cytochrome P450	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN72113_c1_g1	0.762	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN3977_c2_g1	0.761	Cellulose synthase	IPR005150 (Cellulose synthase)
TRINITY_DN2774_c12_g1	0.758	UDP-glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN1501_c0_g1	0.757	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN47437_c0_g1	0.753	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN2641_c0_g1	0.753	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN105863_c0_g1	0.751	Beta-amyrin 28-oxidase	IPR001128 (Cytochrome P450)
TRINITY_DN317510_c0_g1	0.748	Trans-cinnamate 4-monooxygenase	IPR001128 (Cytochrome P450)
TRINITY_DN20518_c0_g1	0.746	Cellulose synthase	IPR005150 (Cellulose synthase)
TRINITY_DN123742_c0_g1	0.746	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN2324_c2_g1	0.746	Flavonoid 3'-hydroxylase	IPR001128 (Cytochrome P450)
TRINITY_DN3352_c6_g1	0.744	Ent-kaurenoic acid oxidase	IPR001128 (Cytochrome P450)
TRINITY_DN69953_c0_g1	0.742	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN61190_c0_g1	0.740	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN71040_c0_g1	0.740	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN30822_c1_g1	0.732	cytochrome P450 CYP749A22-like	IPR001128 (Cytochrome P450)
TRINITY_DN56102_c0_g1	0.732	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN38818_c1_g2	0.731	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN1707_c1_g1	0.731	BAHD acyltransferase DCR-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)

TRINITY_DN44076_c0_g2	0.729	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN30822_c2_g1	0.729	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN2172_c0_g1	0.728	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN81788_c0_g1	0.727	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN46019_c0_g1	0.725	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN186179_c0_g1	0.719	Vinorine synthase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN132712_c0_g1	0.718	Ent-kaurenoic acid oxidase	IPR001128 (Cytochrome P450)
TRINITY_DN285251_c0_g1	0.715	Cytochrome P450	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN823_c0_g2	0.713	Cellulose synthase family protein	IPR005150 (Cellulose synthase), IPR013083 (Zinc finger, RING/FYVE/PHD-type), IPR029044 (Nucleotide-diphospho-sugar transferases)
TRINITY_DN48274_c1_g1	0.709	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN46470_c0_g1	0.708	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN86976_c0_g1	0.708	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN24248_c0_g1	0.707	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN142198_c0_g1	0.704	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN10926_c0_g2	0.702	Glyco_transf_28 domain-containing protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase), IPR004276 (Glycosyltransferase family 28, N-terminal domain)
TRINITY_DN24248_c0_g2	0.701	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN3859_c1_g1	0.699	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN19741_c0_g1	0.697	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN186013_c0_g1	0.697	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN2273_c0_g1	0.694	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN101327_c0_g5	0.694	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN29733_c0_g1	0.691	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN15805_c1_g1	0.690	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN71494_c0_g1	0.690	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)

TRINITY_DN10898_c1_g2	0.688	Shikimate O-hydroxycinnamoyl transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN1501_c1_g1	0.687	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN49017_c0_g1	0.685	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN83629_c0_g1	0.683	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN6632_c0_g3	0.679	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN7322_c3_g1	0.675	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN47434_c0_g1	0.675	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN49_c1_g1	0.673	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN68519_c0_g1	0.671	Cytochrome P450	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN17491_c0_g1	0.668	Cytochrome P450 family ent-kaurenoic acid oxidase	IPR001128 (Cytochrome P450)
TRINITY_DN64217_c1_g1	0.666	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN233747_c0_g1	0.662	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN44601_c0_g1	0.662	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN29117_c0_g3	0.662	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN32758_c0_g2	0.661	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN33242_c0_g1	0.661	Glyco_transf_28 domain-containing protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase), IPR004276 (Glycosyltransferase family 28, N-terminal domain)
TRINITY_DN177126_c0_g1	0.661	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN78857_c0_g1	0.661	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN23881_c0_g1	0.657	Protein ECERIFERUM 26-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN358914_c0_g1	0.656	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN12730_c0_g5	0.656	Shikimate O-hydroxycinnamoyltransferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN26905_c0_g1	0.654	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN12883_c0_g1	0.650	Cytochrome P450	IPR001128 (Cytochrome P450)

TRINITY_DN7835_c0_g4	0.650	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN117399_c0_g2	0.649	Cytochrome P450 83A1	IPR001128 (Cytochrome P450)
TRINITY_DN88851_c0_g1	0.648	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN224660_c0_g1	0.648	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN17940_c0_g1	0.647	Transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN16426_c0_g1	0.647	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN10494_c1_g1	0.647	protein ECERIFERUM 26	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN17675_c0_g1	0.646	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN74148_c0_g1	0.641	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN2210_c0_g2	0.633	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN2822_c1_g2	0.632	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)

(Table D.1.1. continued)

TRINITY ID	**PCC**	**AHRD**	**Interpro ID Description**
TRINITY_DN56469_c0_g1	0.631	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN71494_c0_g2	0.631	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN56642_c0_g1	0.629	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN3075_c0_g1	0.629	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN7654_c0_g1	0.623	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN96413_c0_g1	0.606	vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN72776_c0_g1	0.605	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN37974_c0_g1	0.603	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN56102_c0_g2	0.598	Cytochrome P450 family protein	IPR001128 (Cytochrome P450)
TRINITY_DN124767_c0_g2	0.596	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN32514_c0_g1	0.590	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN3796_c1_g1	0.589	Cellulose synthase	IPR005150 (Cellulose synthase)
TRINITY_DN285107_c0_g1	0.585	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN320_c1_g3	0.584	omega-hydroxypalmitate O-feruloyl transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN54760_c0_g1	0.583	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)

TRINITY_DN31689_c0_g1	0.576	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN51657_c0_g1	0.575	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN12883_c1_g1	0.569	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN290705_c0_g1	0.568	Omega-hydroxypalmitate O-feruloyl transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN52434_c0_g1	0.568	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN30349_c0_g2	0.567	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN26047_c0_g1	0.564	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN3735_c1_g1	0.561	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN27404_c0_g1	0.555	Terpene cyclase/mutase family member	IPR018333 (Squalene cyclase), IPR032696 (Squalene cyclase, C-terminal)
TRINITY_DN4739_c0_g1	0.555	BAHD acyltransferase DCR-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)

(Table D.1.1. continued)

TRINITY ID	**PCC**	**AHRD**	**Interpro ID Description**
TRINITY_DN26561_c0_g3	0.553	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN35843_c0_g1	0.551	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN26526_c0_g1	0.551	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN89084_c0_g1	0.550	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN4419_c0_g1	0.544	Vinorine synthase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN149620_c0_g1	0.540	cytochrome P450 87A3-like	IPR001128 (Cytochrome P450)
TRINITY_DN148101_c0_g1	0.540	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN13035_c0_g4	0.539	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN290761_c0_g1	0.536	HXXXD-type acyl-transferase family protein	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN4609_c0_g3	0.533	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN9000_c0_g2	0.532	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN19633_c0_g1	0.518	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN3663_c2_g1	0.518	Allene oxide synthase, chloroplastic	IPR001128 (Cytochrome P450)

TRINITY_DN107095_c0_g1	0.515	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN29376_c0_g1	0.515	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN195506_c0_g1	0.514	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN135474_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN153729_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN185960_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN201816_c0_g1	0.514	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN211883_c0_g1	0.514	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN213941_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN223273_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN258788_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN269522_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN314343_c0_g1	0.514	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN317992_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN318472_c0_g1	0.514	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN323951_c0_g1	0.514	Cytochrome P450 4g15	IPR001128 (Cytochrome P450)
TRINITY_DN33893_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)

(Table D.1.1. continued)

TRINITY ID	PCC	AHRD	Interpro ID Description
TRINITY_DN346693_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN87586_c0_g1	0.514	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN92095_c0_g1	0.514	Cytochrome P450, putative	IPR001128 (Cytochrome P450)
TRINITY_DN92456_c0_g1	0.514	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN158314_c0_g2	0.514	Cytochrome p450	IPR001128 (Cytochrome P450)
TRINITY_DN276751_c0_g1	0.514	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN325712_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN70834_c0_g1	0.514	Unknown protein	IPR001128 (Cytochrome P450)
TRINITY_DN90413_c0_g1	0.514	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN90607_c0_g1	0.514	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN200575_c0_g1	0.514	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN298018_c0_g1	0.514	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN84385_c0_g1	0.514	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)

TRINITY_DN259489_c0_g1	0.514	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN320648_c0_g1	0.514	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN155264_c0_g1	0.514	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN325565_c0_g1	0.514	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN302722_c0_g1	0.514	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN90523_c0_g1	0.514	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN65959_c0_g1	0.514	Unknown protein	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN205351_c0_g1	0.514	vinorine synthase-like	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN350759_c0_g1	0.514	spermidine hydroxycinnamoyl transferase	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)
TRINITY_DN15179_c0_g2	0.509	Glycosyltransferase	IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase)
TRINITY_DN298411_c0_g1	0.509	Cytochrome P450	IPR001128 (Cytochrome P450)
TRINITY_DN23350_c0_g4	0.500	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltr	IPR003480 (Transferase), IPR023213 (Chloramphenicol acetyltransferase-like domain)

E

Sequences

E.1 Sequences of characterised *S. officinalis* enzymes

>SobAS1_CDS
ATGTGGAGGTTAAAAATAGCAGAAGGTGGAAATGACCCGTATTTGTATAGCACAAACAATTTTGTAGG
ACGTCAAACTTGGGAATTTGATAGCGAGTACGGTACTCCTGAAGCTATAAAAGAAGTAGAAGAAGCTC
GACAAATTTTTTACAAAAATCGATTTCAAGTTAAGCCTTGTGGCGATCTTCTATGGCGTTTTCAGTTC
CTAAGAGAGAAAAACTTCAAGCAAACAATACCGCAAGTGAAGGTGGGTGATGGGGAGGAGGTCACCTA
CGAAGCCGCCTCAACGACGTTAAAGCGTTCCGTCAACTTACTCACGGCCCTGCAGGCCGACGACGGTC
ACTGGCCTGCTGAAATTGCTGGCCCTCAATTTTTCCTCCCTCCTTTGGTGTTTTGCTTGTACATCACC
GGACATCTCAACGTTGTTTTCAATGTTCATCACCGTGAAGAAATTCTTCGTAGCATTTATTATCACCA
GAATGAGGATGGAGGGTGGGGGTTGCACATTGAAGGACACAGCACCATGTTCTGTACGGCGTTGAACT
ACATATGTTTGCGGATGCTAGGAGTCGGTCCTGATGAAGGAGACGACAACGCTTGCCCTAGGGCTCGT
AAATGGATCCTCGACCATGGTAGTGTCACTCATATCCCTTCTTGGGGAAAGACTTGGCTTTCTATACT
CGGTTTGTTTGATTGGTCCGGAAGTAACCCGATGCCACCTGAGTTTTGGATTCTGCCTACTTTCATGC
CTATGTATCCAGCGAAAATGTGGTGTTACTGTCGAATGGTGTACATGCCGATGTCGTACTTATACGGG
AAGAGGTTCGTTGGTCCGATTACACCTCTAATCAAACAGCTCAGAGAGGAACTTTTCAGTGAACCGTT
TGAAGAAATCAAGTGGAAAAAAGTCCGTCATCTGTGTGCACCGGAGGATCTCTACTACCCGCATCCAT
TGATTCAAGACTTAATGTGGGACAGTCTTTACTTATTCACCGAGCCTCTTCTTACTCGCTGGCCGTTC
AACAATTTGATACGACAGAAGGCCTTACAAGTGACGATGGATCATATACATTACGAAGATGAGAACAG
TCGATACATAACCATAGGATGCGTTGAAAAGGTTTTGTGTATGTTGGCCTGTTGGGTTGAAGACCCAA
ATGGTGTTTGTTACAAAAAACATCTTGCTAGAGTTCCCGATTATATATGGATTGCCGAGGATGGCCTT
AAAATGCAGAGTTTTGGAAGTCAACAGTGGGACTGTGGCTTTGCTGTGCAAGCATTACTAGCTTCGAA
TATGAGTCTTGATGAAATCGGACCTGCCCTTAAGAAAGGCCACTTCTTTATCAAAGAGTCTCAGGTGA
AAGATAATCCCTCGGGTGATTTCAAGAGCATGCACCGTCATATCTCGAAGGGATCGTGGACGTTTTCT
GACCAAGATCATGGTTGGCAGGTCTCTGACTGCACTGCAGAAGGCCTTAAGTGCTGCTTGATCTTATC
AACCATGCCGCCAGAAATTGTTGGAGAAAAGATGGACCCTGAGAGGCTCTACGACTCTGTCAATGTCC
TGCTTTCTCTACAGAGTGAAAATGGAGGTCTATCTGCTTGGGAACCAGCTGGAGCACAAGCTTGGTTA
GAGCTTCTAAATCCAACGGAATTCTTCGCAGACATTGTGATCGAGCATGAGTATGTTGAATGTACTGG
TGCATCAATTCAAGCTCTGGTATTATTCAAGAAAATGTACCCTGGTCACCGAAAGAAAGAGATCGAAA
ATTTCATAGCCAAGGCCGCGAAATACCTCGAGGACACCCAATATCCAAACGGCTCTTGGTATGGAAAT
TGGGGTGTGTGTTTCACGTATGGGACGTGGTTTGCGCTAGGAGGGCTAGCGGCAGCGGGCAAAACATA
CGCGAATTGTGCTGCGATGCGAAAAGGTGTTGAATTCCTTCTTAAGTCACAAAAGGAGGACGGTGGGT
GGGGCGAAAGCTATGTTTCATGCCCGAAAAAGGACTTCGTGCCGCTGGAAGGACCATCCAATCTAACT
CAAACCGCATGGGCGTTGATGGGTCTAATTTACGCACGACAGATGGAGAGGGATCCGACACCGCTACA
CCAAGCAGCAAAGCTTTTGATCAATTCACAACTCGAAAACGGAGATTTCCCTCAACAGGAAATAACAG
GAGTATTCATGAAGAATTGCATGCTACACTATCCAATGTACAGGACTATTTATCCACTGTGGGCTATT
GCAGAATATAGGACGCATGTTCCTTTGAGGCTTAGTTAA

>SobAS1_translated
MWRLKIAEGGNDPYLYSTNNFVGRQTWEFDSEYGTPEAIKEVEEARQIFYKNRFQVKPCGDLLWRFQF
LREKNFKQTIPQVKVGDGEEVTYEAASTTLKRSVNLLTALQADDGHWPAEIAGPQFFLPPLVFCLYIT
GHLNVVFNVHHREEILRSIYYHQNEDGGWGLHIEGHSTMFCTALNYICLRMLGVGPDEGDDNACPRAR

KWILDHGSVTHIPSWGKTWLSILGLFDWSGSNPMPPEFWILPTFMPMYPAKMWCYCRMVYMPMSYLYG
KRFVGPITPLIKQLREELFSEPFEEIKWKKVRHLCAPEDLYYPHPLIQDLMWDSLYLFTEPLLTRWPF
NNLIRQKALQVTMDHIHYEDENSRYITIGCVEKVLCMLACWVEDPNGVCYKKHLARVPDYIWIAEDGL
KMQSFGSQQWDCGFAVQALLASNMSLDEIGPALKKGHFFIKESQVKDNPSGDFKSMHRHISKGSWTFS
DQDHGWQVSDCTAEGLKCCLILSTMPPEIVGEKMDPERLYDSVNVLLSLQSENGGLSAWEPAGAQAWL
ELLNPTEFFADIVIEHEYVECTGASIQALVLFKKMYPGHRKKEIENFIAKAAKYLEDTQYPNGSWYGN
WGVCFTYGTWFALGGLAAAGKTYANCAAMRKGVEFLLKSQKEDGGWGESYVSCPKKDFVPLEGPSNLT
QTAWALMGLIYARQMERDPTPLHQAAKLLINSQLENGDFPQQEITGVFMKNCMLHYPMYRTIYPLWAI
AEYRTHVPLRLS

>SoC28_CDS
ATGGAACTCTTCTTCATATGTGGACTAGTACTCTTCTCCACCCTATCACTAATATCCCTCTTCCTCCT
CCACAACCACAGTTCTGCTCGGGGGTACAGGCTGCCCCCGGGCAGAATGGGATGGCCCTTCATAGGCG
AGTCATACGAGTTTTTAGCAAACGGGTGGAAAGGGTACCCGGAAAAGTTTATATTTAGCAGGTTGGCC
AAGTATAAACCGAATCAAGTATTTAAGACGTCGATCCTAGGAGAAAAAGTCGCGGTAATGTGTGGCGC
GACATGTAACAAGTTCTTGTTCTCGAACGAGGGCAAATTAGTAAATGCTTGGTGGCCGAATTCGGTTA
ATAAGATCTTCCCTTCTTCTACTCAAACTTCTTCCAAGGAAGAAGCTAAGAAGATGCGGAAACTTCTC
CCTACATTCTTTAAACCCGAGGCACTACAACGATACATACCCATCATGGACGAAATTGCGATCCGACA
CATGGAGGACGAATGGGAAGGCAAATCCAAAATCGAAGTATTCCCACTCGCAAAACGCTACACATTTT
GGCTAGCGTGCCGTCTATTCCTAAGCATAGACGACCCGGTACACGTAGCCAAATTCGCTGACCCGTTC
AACGACATTGCCTCAGGGATCATATCGATCCCAATAGACCTCCCCGGCACACCATTCAACCGGGGAAT
TAAGGCCTCGAATGTCGTGAGACAGGAATTGAAGACCATAATAAAGCAGAGGAAATTGGACCTGTCCG
ACAACAAGGCGTCCCCGACACAGGATATATTGTCACACATGTTATTAACTCCCGACGAAGACGGGCGG
TATATGAATGAATTGGACATTGCTGATAAAATTCTCGGGTTGTTAATTGGAGGACATGATACTGCAAG
TGCTGCTTGTACTTTTGTTGTGAAGTTTCTTGCTGAACTCCCTCATATTTACGACGGTGTTTACAAAG
AGCAAATGGAGATAGCAAAGTCGAAAAAAGAAGGAGAGCGATTAAATTGGGAGGACATACAAAAGATG
AAATATTCATGGAATGTGGCCTGTGAAGTCATGCGTTTAGCACCTCCTCTTCAAGGCGCTTTTCGTGA
AGCCCTCTCTGATTTTATGTACGCCGGTTTCCAAATTCCCAAGGGTTGGAAGTTATATTGGAGCGCAA
ACTCAACACATAGGAACCCAGAATGCTTCCCAGAGCCGGAAAAATTCGACCCAGCAAGGTTCGATGGG
AGCGGTCCGGCCCCATACACGTACGTACCGTTCGGAGGAGGGCCGAGAATGTGCCCAGGAAAAGAGTA
TGCAAGGCTAGAAATATTGGTGTTCATGCACAACATTGTCAAGAGATTTAAGTGGGAAAAACTTATTC
CTGATGAAACCATTGTTGTTAATCCCATGCCGACCCCGGCTAAAGGCCTACCCGTCCGCCTTCGTCCT
CATTCCAAACCCGTAACTGTATCTGCTTAA

>SoC28_translated
MELFFICGLVLFSTLSLISLFLLHNHSSARGYRLPPGRMGWPFIGESYEFLANGWKGYPEKFIFSRLA
KYKPNQVFKTSILGEKVAVMCGATCNKFLFSNEGKLVNAWWPNSVNKIFPSSTQTSSKEEAKKMRKLL
PTFFKPEALQRYIPIMDEIAIRHMEDEWEGKSKIEVFPLAKRYTFWLACRLFLSIDDPVHVAKFADPF
NDIASGIISIPIDLPGTPFNRGIKASNVVRQELKTIIKQRKLDLSDNKASPTQDILSHMLLTPDEDGR
YMNELDIADKILGLLIGGHDTASAACTFVVKFLAELPHIYDGVYKEQMEIAKSKKEGERLNWEDIQKM
KYSWNVACEVMRLAPPLQGAFREALSDFMYAGFQIPKGWKLYWSANSTHRNPECFPEPEKFDPARFDG
SGPAPYTYVPFGGGPRMCPGKEYARLEILVFMHNIVKRFKWEKLIPDETIVVNPMPTPAKGLPVRLRP
HSKPVTVSA

>SoC28C16_CDS
ATGGAGCTAATTACCTTACTAAGTGCTCTTCTTGTTCTTGCTATAGTGAGTTTATCTACATTTTTCGT
CCTTTACTATAATACTCCTACTAAGGACGGCAAAACTCTCCCTCCCGGTCGTATGGGCTGGCCTTTTA
TAGGCGAGTCCTACGACTTTTTTGCCGCCGGTTGGAAAGGGAACCCGAGAGCTTCATTTTCGACCGGT
TGAAGAAATTTGCTAAGGGGAACCTGAACGGTCAGTTCAGGACGAGCTTGTTTGGGAACAAGTCGATT
GTGGTGGCGGGGGCTGCTGCTAACAAGCTTCTTTTCTCGAATGAAAAGAAGCTTGTTACCATGTGGTG
GCCCCCGTCTATTGATAAGGCCTTCCCGTCGACTGCACAGTTGAGTGCGAACGAGGAGGCCTTATTGA
TGAGGAAGTTTTTTCCTTCTTTTTTGATTAGAAGGGAGGCGCTCCAGCGCTACATCCCTATTATGGAC
GACTGCACCCGTCGTCACTTCGCGACGGGTGCGTGGGGTCCGTCGGACAAGATCGAGGCCTTCAATGT
GACCCAAGACTACACGTTTTGGGTCGCCTGCAGAGTCTTCATGAGCATAGACGCTCAGGAAGACCCTG
AGACGGTAGACTCCCTCTTTAGGCACTTTAACGTGCTTAAAGCGGGAATCTACTCAATGCACATCGAT
CTCCCGTGGACGAACTTCCACCACGCGATGAAGGCGTCCCACGCCATCAGGAGCGCCGTGGAGCAAAT
CGCGAAGAAAAGAAGGGCGGAATTGGCCGAGGGAAAGGCGTTCCCGACACAAGATATGCTGTCTTACA
TGCTCGAAACGCCAATTACATCGGCGGAGGATAGCAAGGACGGGAAAGCGAAGTATTTGAATGACGCC
GATATCGGGACGAAGATACTTGGTCTTCTTGTTGGTGGCCATGACACAAGTAGTACAGTTATTGCCTT
CTTTTTCAAGTTCATGGCTGAAAATCCTCATGTTTATGAGGCTATTTACAAAGAACAAATGGAGGTAG
CGGCCACAAAAGCGCCGGGGGAGCTTCTAAATTGGGATGACTTGCAGAAAATGAAGTACTCGTGGTGT
GCGATTTGCGAGGTTATGCGTTTGACTCCCCCTGTCCAAGGCGCCTTTCGCCAAGCCATCACCGACTT

CACCCATAATGGTTACCTTATTCCCAAGGGTTGGAAGATATACTGGAGTACACACTCAACACACAGAA
ATCCCGAAATCTTCCCACAACCAGAGAAATTCGACCCAACAAGATTCGAAGGAAACGGGCCACCAGCG
TTCTCATTCGTGCCATTCGGAGGAGGCCCGAGAATGTGTCCGGGTAAAGAATATGCAAGGCTACAAGT
GCTTACATTTGTGCACCACATTGTGACCAAATTCAAGTGGGAACAAATTCTACCTAATGAAAAGATCA
TTGTTAGCCCTATGCCGTACCCGGAGAAGAATCTTCCGCTTCGTATGATTGCTCGGTCTGAATCCGCC
ACCCTCGCTTAA

>SoC28C18_translated
MELITLLSALLVLAIVSLSTFFVLYYNTPTKDGKTLPPGRMGWPFIGESYDFFAAGWKGKPESFIFDR
LKKFAKGNLNGQFRTSLFGNKSIVVAGAAANKLLFSNEKKLVTMWWPPSIDKAFPSTAQLSANEEALL
MRKFFPSFLIRREALQRYIPIMDDCTRRHFATGAWGPSDKIEAFNVTQDYTFWVACRVFMSIDAQEDP
ETVDSLFRHFNVLKAGIYSMHIDLPWTNFHHAMKASHAIRSAVEQIAKKRRAELAEGKAFPTQDMLSY
MLETPITSAEDSKDGKAKYLNDADIGTKILGLLVGGHDTSSTVIAFFFKFMAENPHVYEAIYKEQMEV
AATKAPGELLNWDDLQKMKYSWCAICEVMRLTPPVQGAFRQAITDFTHNGYLIPKGWKIYWSTHSTHR
NPEIFPQPEKFDPTRFEGNGPPAFSFVPFGGGPRMCPGKEYARLQVLTFVHHIVTKFKWEQILPNEKI
IVSPMPYPEKNLPLRMIARSESATLA

>SoC23_CDS
ATGGAGTATTTGCCGTACATTGCAACATCAATTGCGTGCATAGTAATACTAAGATGGGCATTGAACAT
GATGCAATGGCTATGGTTCGAACCGAGGCGGTTGGAGAAATTACTTAGAAAACAAGGACTTCAAGGAA
ATTCATATAAGTTTTTATTTGGAGATATGAAGGAAAGTTCTATGTTGAGAAATGAAGCTTTAGCAAAG
CCTATGCCTATGCCTTTTGATAATGACTACTTTCCTCGTATTAATCCTTTTGTTGATCAACTTCTTAA
CAAATATGGTATGAATTGTTTCTTGTGGATGGGGCCTGTTCCGGCTATTCAAATCGGAGAACCAGAGT
TAGTTAGGGAAGCTTTCAACCGGATGCACGAGTTTCAAAAGCCCAAAACTAACCCTTTGAGTGCTTTA
CTCGCCACCGGACTTGTTAGCTACGAGGGCGACAAATGGGCCAAGCACCGCCGCCTTATCAACCCCTC
TTTTCATGTTGAAAAGCTCAAGCTTATGATTCCTGCATTCCGCGAGAGCATTGTGGAGGTGGTCAATC
AATGGGAGAAGAAAGTACCTGAAAACGGCTCTGCTGAAATAGATGTATGGCCGTCTCTTACTAGTTTA
ACCGGAGATGTTATCTCAAGAGCTGCCTTTGGCAGCGTGTATGGCGATGGAAGAAGGATTTTCGAACT
TCTAGCTGTTCAGAAAGAACTCGTTTTAAGTCTGCTCAAGTTTTCGTACATCCCTGGATACACGTATT
TGCCAACAGAGGGAAACAAGAAGATGAAGGCGGTGAACAATGAGATACAAAGACTACTCGAAAACGTG
ATTCAAAACAGAAAGAAGGCGATGGAAGCCGGAGAAGCAGCAAAAGATGATCTGTTGGGTTTACTGAT
GGATTCCAATTACAAGGAGAGTATGCTTGAAGGCGGCGGGAAAAACAAAAAATTGATCATGAGTTTTC
AAGATCTTATTGACGAGTGTAAGCTCTTCTTCTTAGCTGGGCACGAGACGACTGCTGTGTTACTTGTG
TGGACTTTGATTTTGTTGTGTAAGCACCAAGACTGGCAAACCAAAGCTCGCGAAGAAGTTTTGGCTAC
TTTTGGAATGTCGGAACCCACTGATTATGATGCCTTAAACCGTCTCAAGATTGTGACAATGATACTAA
ATGAGGTCCTAAGATTGTACCCACCGGTTGTTTCAACCAACCGAAAACTATTCAAGGGCGAAACAAAA
CTCGGAAACTTGGTAATACCACCAGGTGTCGGTATCTCACTATTAACCATCCAAGCAAACCGTGACCC
GAAAGTTTGGGGGGAGGATGCAAGTGAGTTCCGACCTGATAGATTTGCAGAAGGGCTAGTGAAGGCGA
CTAAGGGCAATGTCGCGTTTTTCCCCTTCGGTTGGGGTCCTAGGATTTGTATTGGCCAAAATTTTGCG
CTGACCGAGTCAAAGATGGCGGTTGCTATGATATTGCAACGCTTCACTTTCGACCTTTCACCGTCTTA
CACTCATGCTCCGTCGGGCCTTATTACTCTTAACCCGCAATATGGGGCTCCTCTCATGTTTCGTAGAC
GTTAA

>SoC23_translated
MEYLPYIATSIACIVILRWALNMMQWLWFEPRRLEKLLRKQGLQGNSYKFLFGDMKESSMLRNEALAK
PMPMPFDNDYFPRINPFVDQLLNKYGMNCFLWMGPVPAIQIGEPELVREAFNRMHEFQKPKTNPLSAL
LATGLVSYEGDKWAKHRRLINPSFHVEKLKLMIPAFRESIVEVVNQWEKKVPENGSAEIDVWPSLTSL
TGDVISRAAFGSVYGDGRRIFELLAVQKELVLSLLKFSYIPGYTYLPTEGNKKMKAVNNEIQRLLENV
IQNRKKAMEAGEAAKDDLLGLLMDSNYKESMLEGGGKNKKLIMSFQDLIDECKLFFLAGHETTAVLLV
WTLILLCKHQDWQTKAREEVLATFGMSEPTDYDALNRLKIVTMILNEVLRLYPPVVSTNRKLFKGETK
LGNLVIPPGVGISLLTIQANRDPKVWGEDASEFRPDRFAEGLVKATKGNVAFFPFGWGPRICIGQNFA
LTESKMAVAMILQRFTFDLSPSYTHAPSGLITLNPQYGAPLMFRRR

>SoCSL1_CDS
ATGTCACCCCACAACACCTGCACTCTACAAATAACCCGAGCCCTCCTCAGCCGCCTCCACATCCTCTT
CCACTCCGCCCTCGTCGCCTCCGTCTTCTACTACCGCTTTTCCAACTTCTCCTCTGGCCCGGCATGGG
CCCTCATGACTTTCGCCGAGCTCACCCTCGCCTTCATCTGGGCCCTCACCCAGGCCTTCCGCTGGCGG
CCCGTCGTCCGGGCCGTCTTCGGGCCCGAGGAGATTGACCCGGCCCAGCTCCCGGGTCTGGACGTGTT
CATATGCACGGCAGACCCGAGGAAGGAGCCGGTGATGGAGGTGATGAACTCGGTGGTGTCGGCATTGG
CGTTGGATTATCCGGCAGAGAAGCTGGCGGTTTACTTGTCGGACGACGGCGGGTCGCCCTTGACTAGG
GAGGTTATTAGGGAGGCTGCCGTGTTTGGGAAGTACTGGGTCGGGTTTTGTGGGAAGTATAATGTTAA
GACGAGGTGTCCTGAGGCCTATTTTAGTTCGTTTTGTGATGGTGAAAGAGTTGATCATAATCAGGATT

ATTTGAACGACGAGCTTTCCGTCAAGTCGAAATTTGAAGCGTTTAAGAAGTATGTGCAAAAAGCAAGT
GAAGACGCCACCAAATGTATTGTTGTCAATGATCGTCCTTCTTGTGTTGAGATTATTCATGACAGCAA
GCAGAACGGAGAGGGTGAAGTGAAAATGCCGCTTCTTGTTTACGTAGCCAGGGAAAAAAGACCGGGTT
TTAATCACCATGCTAAAGCCGGAGCCATTAATACACTTCTTCGAGTGTCGGGTTTACTGAGCAATAGC
CCTTTCTTTTTGGTGTTGGATTGTGATATGTACTGTAATGATCCAACGTCTGCGCGTCAAGCTATGTG
CTTCCATCTTGACCCGAAACTAGCTCCCTCTCTCGCGTTTGTGCAATACCCTCAAATTTTCTACAACA
CCAGCAAAAACGACATCTATGATGGTCAGGCCAGAGCAGCTTTTAAGACTAAATATCAAGGCATGGAT
GGTCTTAGAGGGCCGGTTATGAGTGGCACGGGGTATTTCTTGAAGAGGAAAGCATTGTACGGAAAACC
ACACGACCAAGATGAATTACTCAGGGAGCAGCCAACGAAGGCCTTTGGCTCCTCTAAGATATTCATCG
CGTCCCTTGGTGAAAATACCTGTGTTGCCTTGAAAGGATTGAGTAAAGACGAGTTGTTGCAAGAGACT
CAAAAATTGGCTGCTTGTACATACGAATCAAACACGTTATGGGGTAGCGAGGTTGGATACTCGTACGA
CTGCTTGTTGGAGAGCACATACTGTGGGTACTTATTACACTGCAAAGGATGGATCTCAGTATATCTAT
ACCCGAAAAAGCCGTGTTTCTTGGGGTGTGCAACAGTGGACATGAATGATGCCATGCTTCAGATAATG
AAATGGACTTCTGGATTGATTGGCGTTGGCATATCAAAGTTCAGCCCGTTCACATACGCCATGTCTCG
GATCTCCATTATGCAAAGTCTTTGCTATGCTTACTTCGCTTTTTCGGGCCTATTTGCTGTCTTCTTCT
TGATCTATGGCGTTGTTCTTCCGTATTCCCTCTTGCAGGGTGTTCCGCTCTTCCCCAAGGCAGGAGAT
CCATGGCTTTTGGCATTTGCGGGAGTATTCATATCCTCGCTTCTTCAGCACCTGTACGAGGTTCTCTC
AAGCGGAGAAACAGTGAAAGCGTGGTGGAACGAGCAAAGAATCTGGATCATAAAATCAATCACCGCCT
GTCTGTTTGGTCTTCTGGACGCTATGCTTAACAAAATTGGCGTCTTAAAGGCTAGTTTCAGACTGACA
AACAAGGCTGTCGACAAACAAAAACTCGATAAATACGAGAAGGGCAGGTTCGATTTCCAAGGCGCACA
AATGTTCATGGTCCCTCTCATGATTCTGGTGGTATTCAATTTGGTCTCGTTCTTTGGCGGCTTAAGAA
GAACCGTCATTCATAAAAACTACGAAGACATGTTCGCGCAGCTTTTCCTCTCGTTGTTCATTCTAGCT
CTTAGCTATCCTATCATGGAGGAGATTGTCCGAAAAGCTAGAAAAGGTCGCTCTTAA

>SoCSL1_translated
MSPHNTCTLQITRALLSRLHILFHSALVASVFYYRFSNFSSGPAWALMTFAELTLAFIWALTQAFRWR
PVVRAVFGPEEIDPAQLPGLDVFICTADPRKEPVMEVMNSVVSALALDYPAEKLAVYLSDDGGSPLTR
EVIREAAVFGKYWVGFCGKYNVKTRCPEAYFSSFCDGERVDHNQDYLNDELSVKSKFEAFKKYVQKAS
EDATKCIVVNDRPSCVEIIHDSKQNGEGEVKMPLLVYVAREKRPGFNHHAKAGAINTLLRVSGLLSNS
PFFLVLDCDMYCNDPTSARQAMCFHLDPKLAPSLAFVQYPQIFYNTSKNDIYDGQARAAFKTKYQGMD
GLRGPVMSGTGYFLKRKALYGKPHDQDELLREQPTKAFGSSKIFIASLGENTCVALKGLSKDELLQET
QKLAACTYESNTLWGSEVGYSYDCLLESTYCGYLLHCKGWISVYLYPKKPCFLGCATVDMNDAMLQIM
KWTSGLIGVGISKFSPFTYAMSRISIMQSLCYAYFAFSGLFAVFFLIYGVVLPYSLLQGVPLFPKAGD
PWLLAFAGVFISSLLQHLYEVLSSGETVKAWWNEQRIWIIKSITACLFGLLDAMLNKIGVLKASFRLT
NKAVDKQKLDKYEKGRFDFQGAQMFMVPLMILVVFNLVSFFGGLRRTVIHKNYEDMFAQLFLSLFILA
LSYPIMEEIVRKARKGRS

>SoC3Gal_CDS
ATGGGTTCAAATACAGAAGCAACTGAAATACCCAAAATGCCCTTGAAAATAGTCTTCCTTACACTTCC
TATAGCCGGACACATGCTCCACATTGTAGACACCGCAAGCACATTTGCCATACATGGAGTCGAGTGTA
CCATAATCACTACCCCTGCAAATGTCCCTTTCATCGAAAAATCAATCTCTGCAACCAACACCACAATT
CGACAGTTCCTCAGTATCCGCCTCGTCGATTTCCCCCATGAAGCTGTCGGCCTTCCTCCCGGTGTCGA
AAACTTCAGTGCAGTCACGTGTCCGGATATGAGACCCAAAATATCGAAAGGACTTTCGATCATACAAA
AACCAACTGAAGACTTAATCAAGGAAATATCACCTGATTGTATTGTTTCTGACATGTTTTACCCTTGG
ACTTCTGATTTCGCCCTTGAAATAGGTGTTCCAAGGGTGGTTTTTCGCGGTTGTGGGATGTTTCCCAT
GTGTTGTTGGCATAGTATTAAGTCACATTTACCACATGAGAAGGTTGACAGAGATGATGAAATGATTG
TTCTTCCTACATTGCCTGATCATATAGAGATGAGAAAATCTACATTACCTGATTGGGTAAGGAAACCA
ACTGGGTACAGTTATTTGATGAAGATGATTGATGCGGCCGAATTGAAGAGTTATGGAGTAATTGTTAA
TAGTTTTAGTGATTTAGAGAGGGATTATGAGGAGTATTTTAAGAATGTCACCGGGTTAAAGGTGTGGA
CCGTCGGTCCGATTTCGTTACATGTGGGTCGGAATGAGGAGTTAGAAGGGTCAGATGAGTGGGTCAAA
TGGCTAGATGGGAAAAAACTAGACTCGGTTATTTATGTTAGTTTTGGTGGGGTGGCGAAGTTTCCACC
CCACCAGCTGAGAGAAATCGCGGCCGGATTAGAATCATCTGGCCACGATTTTGTTTGGGTGGTGAGGG
CGAGTGACGAAAATGGCGACCAAGCTGAAGCGGATGAGTGGTCCCTACAAAAATTTAAAGAGAAAATG
AAGAAAACTAACCATGGGTTGGTTATAGAGAGTTGGGTCCCACAACTTATGTTTTTGGAACATAAGGC
TATCGGAGGAATGTTGACACATGTTGGTTGGGGTACAATGTTGGAAGGGATTACAGCGGGTTTACCGT
TGGTGACGTGGCCATTGTATGCCGAGCAGTTTTACAATGAGAGGTTGGTGGTTGATGTGTTGAAGATT
GGAGTTGGTGTTGGGGTGAAAGAGTTCTGTGGGTTGGATGATATTGGCAAGAAGGAGACCATTGGTAG
GGAGAATATCGAGGCATCGGTGAGATTAGTGATGGGCGATGGCGAGGAGGCGGCTGCCATGAGACTGC
GGGTGAAGGAGTTGAGTGAGGCGTCTATGAAGGCGGTTCGAGAAGGTGGTTCATCTAAGGCTAATATA
CACGATTTCCTTAACGAGCTGTCTACGTTGAGATCGTTAAGGCAGGCTTGA

>SoC3Gal_translated
MGSNTEATEIPKMPLKIVFLTLPIAGHMLHIVDTASTFAIHGVECTIITTPANVPFIEKSISATNTTI
RQFLSIRLVDFPHEAVGLPPGVENFSAVTCPDMRPKISKGLSIIQKPTEDLIKEISPDCIVSDMFYPW
TSDFALEIGVPRVVFRGCGMFPMCCWHSIKSHLPHEKVDRDDEMIVLPTLPDHIEMRKSTLPDWVRKP
TGYSYLMKMIDAAELKSYGVIVNSFSDLERDYEEYFKNVTGLKVWTVGPISLHVGRNEELEGSDEWVK
WLDGKKLDSVIYVSFGGVAKFPPHQLREIAAGLESSGHDFVWVVRASDENGDQAEADEWSLQKFKEKM
KKTNHGLVIESWVPQLMFLEHKAIGGMLTHVGWGTMLEGITAGLPLVTWPLYAEQFYNERLVVDVLKI
GVGVGVKEFCGLDDIGKKETIGRENIEASVRLVMGDGEEAAAMRLRVKELSEASMKAVREGGSSKANI
HDFLNELSTLRSLRQA

>SoC3Xyl_CDS
ATGAAGTCACCACTAAAGTTGTACTTCCTGCCATACATATCACCAGGCCATATGATCCCACTTTCCGA
AATGGCTCGGTTATTCGCCAACCAAGGGCACCACGTGACCATCATCACCACCACCTCGAACGCCACCC
TCCTCCAAAAATACACCACCGCCACCCTGTCTCTACATCTTATTCCCCTCCCTACCAAAGAGGCCGGC
CTTCCAGACGGCCTCGAAAACTTCATTTCTGTCAACGATCTTGAAACCGCTGGCAAACTCTACTACGC
TCTTTCCCTCCTGCAACCCGTCATTGAGGAGTTTATCACGTCTAACCCGCCCGATTGTATCGTGTCCG
ACATGTTCTATCCCTGGACTGCGGACCTGGCGTCCCAACTCCAGGTCCCGCGTATGGTCTTTCATGCA
GCGTGTATATTCGCTATGTGCATGAAAGAGTCAATGCGGGGCCCTGACGCCCCGCATCTGAAGGTCAG
CTCTGATTATGAGCTGTTTGAAGTCAAGGGGCTACCGGACCCGGTTTTTATGACCCGGGCCCAGCTCC
CTGACTACGTGCGTACCCCAAACGGGTACACACAGCTCATGGAGATGTGGCGAGAAGCGGAAAAGAAA
AGTTACGGTGTTATGGTTAATAATTTTTACGAACTTGACCCGGCTTATACCGAGCATTATAGTAAGAT
TATGGGCCATAAGGTCTGGAATATTGGGCCTGCGGCCCAAATTCTTCACCGTGGTTCTGGTGATAAAA
TCGAGAGGGTTCACAAAGCCGTTGTTGGTGAAAACCAATGCTTGAGTTGGCTCGACACTAAGGAACCT
AACTCGGTTTTTTACGTCTGCTTTGGGAGCGCGATTAGGTTCCCTGATGATCAGCTCTACGAAATTGC
TAGCGCGCTAGAATCATCTGGCGCGCAGTTTATATGGGCCGTTCTTGGAAAAGACTCGGATAATTCAG
ACTCGAACTCAGACTCAGAATGGCTGCCTGCAGGGTTCGAGGAAAAAATGAAGGAAACGGGTAGAGGG
ATGATAATACGAGGTTGGGCCCCACAGGTGTTGATATTGGACCACCCGTCTGTAGGCGGGTTTATGAC
TCACTGTGGCTGGAACTCGACAATTGAGGGGGTTAGCGCGGGGGTGGGGATGGTGACATGGCCGTTGT
ATGCGGAACAATTTTACAATGAGAAGTTAATAACACAAGTGCTTAAGATAGGGGTGGAGGCCGGGGTG
GAGGAGTGGAACTTGTGGGTGGATGTTGGGAGGAAATTGGTGAAGAGAGAGAAGATCGAGGCGGCAAT
TAGGGCGGTGATGGGTGAGGCCGGGGTGGAGATGAGGAGGAAGGCGAAAGAGTTGAGTGTCAAGGCTA
AGAAGGCGGTGCAGGATGGTGGGTCGTCTCACCGTAATTTAATGGCTTTGATCGAAGATCTGCAGAGG
ATTAGAGATGATAAAATGAGTAAGGTTGCTAATTAG

>SoC3Xyl_translated
MKSPLKLYFLPYISPGHMIPLSEMARLFANQGHHVTIITTTSNATLLQKYTTATLSLHLIPLPTKEAG
LPDGLENFISVNDLETAGKLYYALSLLQPVIEEFITSNPPDCIVSDMFYPWTADLASQLQVPRMVFHA
ACIFAMCMKESMRGPDAPHLKVSSDYELFEVKGLPDPVFMTRAQLPDYVRTPNGYTQLMEMWREAEKK
SYGVMVNNFYELDPAYTEHYSKIMGHKVWNIGPAAQILHRGSGDKIERVHKAVVGENQCLSWLDTKEP
NSVFYVCFGSAIRFPDDQLYEIASALESSGAQFIWAVLGKDSDNSDSNSDSEWLPAGFEEKMKETGRG
MIIRGWAPQVLILDHPSVGGFMTHCGWNSTIEGVSAGVGMVTWPLYAEQFYNEKLITQVLKIGVEAGV
EEWNLWVDVGRKLVKREKIEAAIRAVMGEAGVEMRRKAKELSVKAKKAVQDGGSSHRNLMALIEDLQR
IRDDKMSKVAN

>SoSDR1_CDS
ATGGCTGAAGCATCCTCATTTCTTGCACAGAAAAGGTATGCGGTCGTGACAGGAGCAAACAAAGGACT
AGGACTAGAAATATGCGGACAGCTTGCTTCACAGGGGGTGACGGTACTGCTGACATCCAGAGATGAAA
AACGAGGCTTAGAAGCCATTGAGGAGCTTAAGAAATCGGGGATTAATTCGGAAAATCTTGAATATCAT
CAGCTGGATGTTACTAAGCCAGCTAGTTTCGCTTCTCTGGCCGATTTCATCAAGGCCAAATTTGGCAA
GCTTGATATCCTGGTGAACAATGCAGGGATCAGCGGTGTTATTGTAGATTATGCAGCTTTAATGGAAG
CCATTCGCCGTCGAGGGGCAGAGATCAATTACGATGGAGTGATGAAACAGACCTACGAGCTAGCAGAG
GAATGCTTGCAAACAAATTACTATGGTGTGAAAAGAACCATTAATGCTCTCCTTCCGCTACTTCAGTT
TTCCGATTCACCAAGGATCGTCAATGTTTCCTCCGATGTTGGCCTCCTTAAGAAAATACCCGGCGAGA
GAATCAGAGAAGCCTTAGGCGACGTGGAAAAACTTACGGAAGAAAGCGTGGACGGGATTTTAGACGAG
TTTCTAAGAGATTTCAAGGAAGGCAAGATCGCAGAGAAAGGTTGGCCTACGTTTAAGAGCGCCTATTC
AATCTCAAAGGCGGCGCTCAATTCGTACACGAGGGTTTTAGCACGGAAATACCCGTCGATCATCATCA
ACTGTGTCTGCCCGGGTGTCGTCAAAACCGATATCAATCTTAAAATGGGCCACTTGACGGTTGAAGAA
GGCGCGGCCAGTCCCGTGAGGTTAGCACTCATGCCCCTTGGTTCGCCTTCCGGCCTGTTCTATACTCG
AAACGAAGTAACTCCATTTGAATGA

>SoSDR1_tranlsated
AEASSFLAQKRYAVVTGANKGLGLEICGQLASQGVTVLLTSRDEKRGLEAIEELKKSGINSENLEYHQ
LDVTKPASFASLADFIKAKFGKLDILVNNAGISGVIVDYAALMEAIRRRGAEINYDGVMKQTYELAEE
CLQTNYYGVKRTINALLPLLQFSDSPRIVNVSSDVGLLKKIPGERIREALGDVEKLTEESVDGILDEF
LRDFKEGKIAEKGWPTFKSAYSISKAALNSYTRVLARKYPSIIINCVCPGVVKTDINLKMGHLTVEEG
AASPVRLALMPLGSPSGLFYTRNEVTPFE

>SoC28Fu_CDS
ATGTCGGATCAAAATGATAAAAAGGTCGAAATAATAGTATTTCCATACCATGGCCAAGGTCACATGAA
CACCATGCTACAATTCGCCAAACGAATTGCGTGGAAAAACGCCAAAGTTACAATCGCTACGACATTGT
CCACCACTAATAAAATGAAGTCCAAGGTCGAGAATGCCTGGGGCACTTCTATAACCTTGGACTCCATT
TACGATGACTCTGACGAGTCGCAGATAAAATTCATGGACCGTATGGCCAGGTTTGAGGCTGCTGCAGC
CTCGAGCCTGTCCAAACTCCTGGTCCAGAAAAAAGAAGAAGCTGACAACAAAGTCTTGTTGGTTTACG
ACGGGAATTTGCCGTGGGCGCTGGATATCGCCCACGAGCATGGCGTGCGTGGGGCCGCGTTTTTTCCA
CAGTCGTGTGCGACGGTCGCCACGTACTACTCGTTGTATCAAGAGACGCAGGGGAAGGAGCTAGAGAC
GGAGTTGCCGGCGGTGTTTCCGCCGTTGGAGTTGATACAACGGAATGTACCGAATGTGTTTGGATTGA
AGTTTCCGGAGGCGGTTGTGGCTAAGAATGGGAAGGAGTATAGTCCTTTTGTGTTGTTTGTGTTGAGG
CAGTGTATTAACCTTGAGAAGGCTGATTTGCTGCTTTTCAATCAGTTTGATAAGTTGGTTGAACCTGG
GGAGGTTCTGCAATGGATGTCGAAGATATTCAACGTAAAGACAATCGGACCGACACTTCCATCTTCAT
ACATCGACAAACGAATCAAAGACGACGTGGACTACGGTTTCCACGCATTCAACCTCGACAACAACTCC
TGCATCAATTGGCTTAACTCCAAACCCGCTCGCTCTGTCATCTACATAGCATTTGGGAGCAGCGTCCA
CTACAGCGTTGAGCAAATGACCGAAATAGCCGAGGCCTTAAAGAGCCAACCGAACAATTTCCTTTGGG
CAGTCCGAGAAACCGAACAAAAGAAACTCCCTGAAGACTTCGTCCAACAAACCTCGGAAAAAGGGTTA
ATGCTCTCATGGTGCCCTCAATTAGATGTTTTGGTGCATGAATCAATCAGTTGTTTTGTGACACATTG
TGGTTGGAACTCGATTACAGAGGCACTTAGCTTCGGGGTACCAATGCTGTCAGTGCCACAGTTTTTGG
ACCAGCCTGTTGATGCTCACTTTGTGGAACAGGTTTGGGGTGCTGGAATTACGGTCAAGAGGAGCGAA
GACGGTTTGGTTACTCGAGACGAAATTGTTCGGTGCTTGGAGGTGTTAAATAATGGCGAAAAGGCGGA
GGAAATTAAGGCGAATGTGGCGAGGTGGAAGGTTTTGGCTAAGGAAGCTTTGGATGAAGGTGGTAGTT
CTGATAAGCACATTGACGAAATTATTGAGTGGGTTTCATCTTTCTAA

>SoC28Fu_translated
MSDQNDKKVEIIVFPYHGQGHMNTMLQFAKRIAWKNAKVTIATTLSTTNKMKSKVENAWGTSITLDSI
YDDSDESQIKFMDRMARFEAAAASSLSKLLVQKKEEADNKVLLVYDGNLPWALDIAHEHGVRGAAFFP
QSCATVATYYSLYQETQGKELETELPAVFPPLELIQRNVPNVFGLKFPEAVVAKNGKEYSPFVLFVLR
QCINLEKADLLLFNQFDKLVEPGEVLQWMSKIFNVKTIGPTLPSSYIDKRIKDDVDYGFHAFNLDNNS
CINWLNSKPARSVIYIAFGSSVHYSVEQMTEIAEALKSQPNNFLWAVRETEQKKLPEDFVQQTSEKGL
MLSWCPQLDVLVHESISCFVTHCGWNSITEALSFGVPMLSVPQFLDQPVDAHFVEQVWGAGITVKRSE
DGLVTRDEIVRCLEVLNNGEKAEEIKANVARWKVLAKEALDEGGSSDKHIDEIIEWVSSF

>SoC28Rha_CDS
ATGTCTGCCAAAATGTTGCACGTAGTTATGTACCCATGGTTCGCATACGGTCACATGATCCCATTTTT
ACATTTATCGAACAAATTAGCCGAAACCGGTCACAAAGTCACGTACATACTCCCCCCAAAAGCGCTAA
CCCGCTTACAAAACCTCAACCTAAATCCGACCCAAATCACGTTCCGGACCATCACGGTCCCCCGAGTT
GATGGGTTACCCGCTGGTGCCGAGAACGTGACCGATATTCCGGATATTACTCTGCATACTCATTTGGC
CACGGCGCTGGATCGAACCCGACCCGAATTTGAGACGATTGTCGAGTTGATTAAGCCGGATGTGATAA
TGTATGACGTGGCGTATTGGGTGCCAGAGGTGGCGGTGAAGTATGGGGCGAAGAGTGTTGCGTATAGT
GTGGTGTCGGCGGCAAGTGTGTCGCTGAGTAAGACGGTGGTTGATCGGATGACGCCGTTGGAGAAACC
GATGACGGAGGAGGAGAGGAAGAAGAAGTTTGCTCAGTATCCTCACTTAATTCAGCTTTATGGTCCTT
TTGGTGAAGGTATCACCATGTACGACCGTCTAACAGGCATGCTTAGCAAGTGTGACGCTATAGCTTGT
AGGACCTGCCGTGAGATTGAAGGCAAGTATTGCCAATATTTATCCACTCAATATGAAAAGAAAGTCAC
CCTTACCGGCCCGGTTCTTCCCGAGCCGGAAGTCGGGGCCACACTGGAGGCCCCTTGGTCCGAGTGGC
TTAGTCGGTTCAAGCTTGGTTCGGTTTTATTTTGTGCCTTTGGGAGCCAATTTTACTTGGACAAGGAC
CAGTTCCAGGAAATCATCCTCGGGCTTGAAATGACAAATTTACCCTTTCTGATGGCTGTTCAGCCCCC
TAAGGGTTGCGCCACTATCGAGGAGGCGTACCCTGAGGGGTTTGCTGAGCGGGTCAAGGACCGAGGAG
TCGTGACAAGCCAGTGGGTGCAACAGCTGGTTATACTGGCCCACCCAGCGGTTGGGTGCTTTGTGAAC
CATTGCGCGTTTGGGACAATGTGGGAGGCCTTATTGAGCGAAAAGCAGTTGGTGATGATCCCTCAACT
AGGTGACCAAATACTGAACACCAAAATGTTGGCCGATGAATTGAAAGTCGGGGTTGAAGTCGAGAGAG
GAATCGGTGGGTGGGTGTCTAAGGAGAATTTGTGTAAGGCGATCAAGTCCGTCATGGACGAGGATAGT
GAAATTGGCAAGGACGTGAAACAAAGTCATGAAAAATGGAGGGCGACTTTGTCGAGCAAAGATTTAAT
GTCGACTTATATTGATAGTTTCATCAAAGATTTACAAGCACTCGTCGAGTGA

>SoC28Rha_translated
MSAKMLHVVMYPWFAYGHMIPFLHLSNKLAETGHKVTYILPPKALTRLQNLNLNPTQITFRTITVPRV
DGLPAGAENVTDIPDITLHTHLATALDRTRPEFETIVELIKPDVIMYDVAYWVPEVAVKYGAKSVAYS
VVSAASVSLSKTVVDRMTPLEKPMTEEERKKKFAQYPHLIQLYGPFGEGITMYDRLTGMLSKCDAIAC
RTCREIEGKYCQYLSTQYEKKVTLTGPVLPEPEVGATLEAPWSEWLSRFKLGSVLFCAFGSQFYLDKD
QFQEIILGLEMTNLPFLMAVQPPKGCATIEEAYPEGFAERVKDRGVVTSQWVQQLVILAHPAVGCFVN
HCAFGTMWEALLSEKQLVMIPQLGDQILNTKMLADELKVGVEVERGIGGWVSKENLCKAIKSVMDEDS
EIGKDVKQSHEKWRATLSSKDLMSTYIDSFIKDLQALVE

>SoC28Xyl1_CDS
ATGGGTACTAAAGAGTTACACATAGTAATGTACCCATGGCTAGCATTTGGTCATTTCATACCATACCT
TCATCTCTCTAACAAACTCGCTCAAAAAGGCCATAAAATCACTTTCTTACTTCCTCATAGAGCCAAAC
TTCAACTTGACTCCCAAAATTTATATCCCTCACTTATTACCCTCGTACCAATTACCGTCCCACAGGTC
GACACCCTTCCTCTCGGGGCCGAATCGACTGCTGATATCCCCCTTAGTCAGCACGGTGACCTCTCCAT
CGCCATGGACCGTACTCGACCCGAGATTGAGTCTATCTTGTCTAAACTTGACCCAAAACCGGACCTGA
TTTTCTTCGATATGGCGCAGTGGGTGCCTGTCATAGCGTCTAAGCTTGGGATCAAGTCTGTTTCGTAT
AATATCGTTTGCGCCATTTCGTTGGACCTTGTTCGAGATTGGTATAAGAAGGATGATGGAAGTAATGT
GCCTAGTTGGACATTGAAGCATGACAAGTCATCCCATTTCGGGGAGAATATTAGTATTCTCGAGCGAG
CGCTGATTGCGCTCGGGACGCCTGATGCCATAGGCATCAGGTCGTGTCGGGAGATAGAGGGGGAGTAC
TGTGACAGCATAGCGGAACGATTTAAGAAACCGGTCTTACTAAGCGGGACGACCTTACCTGAACCATC
CGACGACCCACTTGACCCAAAATGGGTCAAGTGGCTCGGAAAGTTCGAGGAAGGTTCGGTTATTTTTT
GCTGCCTAGGGAGTCAGCACGTGTTAGACAAGCCCCAGCTCCAGGAGCTGGCGCTGGGGCTTGAAATG
ACGGGGTTGCCATTCTTCCTAGCGATTAAACCACCGCTAGGATACGCAACCCTAGACGAGGTACTACC
CGAGGGGTTTTCAGAACGGGTTCGAGATCGAGGGGTGGCTCATGGGGGATGGGTACAACAGCCTCAGA
TGCTGGCACACCCTTCTGTAGGGTGCTTTTTGTGTCACTGTGGGTCGTCGTCGATGTGGGAGGCATTA
GTGAGTGATACGCAGCTCGTATTGTTTCCTCAAATACCAGATCAAGCTCTAAACGCGGTTTTAATGGC
GGATAAACTTAAGGTCGGGGTGAAGGTCGAGAGAGAGGACGACGGAGGGGTGTCGAAAGAGGTTTGGA
GTAGAGCAATAAAGAGTGTGATGGATAAGGAGAGTGAAATTGCTGCGGAAGTGAAGAAGAATCATACT
AAGTGGAGAGATATGTTGATTAATGAAGAATTTGTGAATGGGTACATTGACAGTTTCATTAAGGATCT
ACAAGATCTTGTTGAGAAGTAG

>SoC28Xyl1_translated
MGTKELHIVMYPWLAFGHFIPYLHLSNKLAQKGHKITFLLPHRAKLQLDSQNLYPSLITLVPITVPQV
DTLPLGAESTADIPLSQHGDLSIAMDRTRPEIESILSKLDPKPDLIFFDMAQWVPVIASKLGIKSVSY
NIVCAISLDLVRDWYKKDDGSNVPSWTLKHDKSSHFGENISILERALIALGTPDAIGIRSCREIEGEY
CDSIAERFKKPVLLSGTTLPEPSDDPLDPKWVKWLGKFEEGSVIFCCLGSQHVLDKPQLQELALGLEM
TGLPFFLAIKPPLGYATLDEVLPEGFSERVRDRGVAHGGWVQQPQMLAHPSVGCFLCHCGSSSMWEAL
VSDTQLVLFPQIPDQALNAVLMADKLKVGVKVEREDDGGVSKEVWSRAIKSVMDKESEIAAEVKKNHT
KWRDMLINEEFVNGYIDSFIKDLQDLVEK

>SoC28Xyl2_CDS
ATGGAGGAATCAAAGGAGGAAGTACATGTAGCATTCTTCCCATTCATGACACCAGGTCACTCAATCCC
AATGCTAGACTTGGTACGTTTGTTCATTGCTCGTGGTGTCAAAACTACTGTCTTCACTACTCCTCTTA
ATGCTCCTAATATTTCCAAATACCTCAACATTATCCAAGATTCCTCATCAAACAAAAACACCATTTAT
GTAACTCCTTTTCCTTCTAAAGAAGCCGGTTTACCGGAAGGTGTGGAAAGCCAGGATAGTACCACTTC
CCCCGAAATGACCCTCAAGTTCTTTGTTGCTATGGAATTACTTCAAGACCCCCTTGATGTTTTTTTAA
AAGAAACCAAACCTCATTGTCTTGTTGCTGATAATTTCTTCCCTTACGCCACCGACATCGCTTCTAAG
TATGGCATTCCTAGGTTTGTTTTTCAGTTCACTGGCTTCTTTCCTATGTCTGTCATGATGGCCTTAAA
TCGTTTCCACCCTCAAAACTCTGTATCATCTGATGACGACCCCTTTCTTGTTCCCAGTTTACCCCATG
ACATCAAATTGACTAAGTCACAATTGCAACGAGAGTACGAGGGTAGTGATGGTATTGACACCGCTCTT
TCTAGGCTCTGTAATGGCGCCGGTAGAGCTTTGTTTACTAGTTATGGTGTCATTTTTAACAGCTTCTA
CCAACTCGAACCTGATTATGTTGATTATTATACCAACACCATGGGGAAACGATCCAGGGTTTGGCATG
TGGGCCCAGTGTCGTTATGCAACCGTCGACACGTGGAGGGTAAATCTGGTAGGGGGAGAAGTGCTTCA
ATTAGTGAGCATTTGTGCTTAGAGTGGCTCAATGCCAAAGAACCAAATTCAGTGATATATGTATGTTT
TGGTAGTCTCACATGTTTCTCCAATGAGCAACTCAAAGAAATCGCAACCGCCTTAGAAAGGTGTGAAG
AGTATTTTATATGGGTGTTGAAGGGTGGCAAAGATAATGAGCAAGAGTGGTTGCCACAAGGGTTTGAA
GAGAGGGTTGAAGGGAAAGGACTAATCATACGGGGGTGGGCCCCACAAGTGTTGATTTTAGACCATGA
AGCCATAGGCGGGTTTGTGACACACTGTGGTTGGAACTCGACACTAGAAAGTATATCAGCGGGGGTGC
CCATGGTGACATGGCCCATATATGCAGAGCAATTTTATAATGAGAAATTGGTGACGGATGTACTGAAG
GTGGGGGTTAAAGTAGGGTCAATGAAGTGGAGTGAGACGACGGGGGCGACTCATTTAAAGCATGAGGA
AATAGAAAAAGCATTGAAGCAAATAATGGTGGGAGAAGAGGTGTTAGAGATGAGAAAAAGAGCAAGTA

AGTTGAAAGAGATGGCTTATAATGCTGTTGAAGAAGGAGGCTCTTCTTATTCTCACCTCACTTCCTTA
ATCGACGACCTTATGGCTTCCAAAGCTGTGCTACAAAAATTTTGA

>SoC28Xyl2_translated
MEESKEEVHVAFFPFMTPGHSIPMLDLVRLFIARGVKTTVFTTPLNAPNISKYLNIIQDSSSNKNTIY
VTPFPSKEAGLPEGVESQDSTTSPEMTLKFFVAMELLQDPLDVFLKETKPHCLVADNFFPYATDIASK
YGIPRFVFQFTGFFPMSVMMALNRFHPQNSVSSDDDPFLVPSLPHDIKLTKSQLQREYEGSDGIDTAL
SRLCNGAGRALFTSYGVIFNSFYQLEPDYVDYYTNTMGKRSRVWHVGPVSLCNRRHVEGKSGRGRSAS
ISEHLCLEWLNAKEPNSVIYVCFGSLTCFSNEQLKEIATALERCEEYFIWVLKGGKDNEQEWLPQGFE
ERVEGKGLIIRGWAPQVLILDHEAIGGFVTHCGWNSTLESISAGVPMVTWPIYAEQFYNEKLVTDVLK
VGVKVGSMKWSETTGATHLKHEEIEKALKQIMVGEEVLEMRKRASKLKEMAYNAVEEGGSSYSHLTSL
IDDLMASKAVLQKF

>SoGH1_CDS
ATGGTTCTTAGTCGATTGGATTTTCCGTCCGATTTCATTTTTGGCTCCGGCACGTCAGCTTCTCAGGT
AGAAGGTGCAGCACTAGAGGATGGGAAGACTTCGACTGCATTTGAAGGATTCTTAACTCGCATGAGTG
GAAATGATTTGAGCAAAGGAGTTGAAGGCTACTACAAATACAAGGAAGACGTCCAGTTAATGGTGCAA
ACAGGACTAGATGCATACAGATTCTCCATTTCATGGTCAAGACTAATTCCCGGTGGAAAAGGACCCGT
CAACCCAAAAGGTTTACAATATTATAATAACTTTATCGACGAACTCATCAAAAATGGAATACAACCGC
ACGTTACTCTGCTGCATTTCGACATACCGGACACACTTATGACTGCTTATAATGGATTGAAGGGTCAA
GAATTTGTGGAAGATTTCACGGCATTTGCTGACGTGTGCTTCAAGGAATTTGGTGACCGAGTTTTGTA
TTGGACGACGGTCAATGAAGCAAATAATTTTGCAAGTCTAACACTCGATGAGGGCAATTTTATGCCGT
CTACTGAACCGTACATTAGAGGTCACAATATCATTCTTGCTCATGCATCCGCGGTAAAACTATACCGA
GAAAAATATAAGAAAACCCAAAATGGATTCATAGGCTTGAATTTATATGCAAGCTGGTATTTTCCCGA
GACCGATGACGAACAAGATTCAATTGCCGCTCAAAGAGCCATTGATTTTACTATTGGATGGATAATGC
AACCATTGATATACGGAGAATATCCAGAAACATTGAAGAAACAAGTGGGAGAAAGACTGCCAACATTT
ACAAAAGAAGAGTCAACGTTCGTTAAAAATTCGTTTGACTTCATTGGAGTGAATTGCTACGTCGGCAC
TGCTGTTAAGGATGACCCTGACAGCTGTAACAGTAAAAATAAAACTATTATTACTGACATGTCTGCTA
AACTTTCTCCTAAAGGTGAACTAGGAGGAGCGTATATGAAGGGATTGTTGGAATACTTCAAAAGAGAT
TACGGCAATCCGCCAATTTACATTCAAGAAAATGGTTATTGGACACCGCGTGAATTAGGAGTGAACGA
TGCGTCAAGGATCGAATACCATACTGCTTCTCTTGCTAGCATGCACGATGCTATGAAGAATGGGGCAA
ATGTAAAGGGATATTTCCAATGGTCATTTTTGGATCTCTTGGAGGTGTTCAAATACAGCTATGGCCTC
TACCATGTCGATTTGGAAGACCCGACCCGAGAAAGACGACCCAAGGCATCCGCCAATTGGTACGCGGA
GTTCTTGAAGGGTTGCGCTACTTCTAACGGGAATGCTAAAGTTGAAACTCCGTTGTAA

>SoGH1_translated
MVLSRLDFPSDFIFGSGTSASQVEGAALEDGKTSTAFEGFLTRMSGNDLSKGVEGYYKYKEDVQLMVQ
TGLDAYRFSISWSRLIPGGKGPVNPKGLQYYNNFIDELIKNGIQPHVTLLHFDIPDTLMTAYNGLKGQ
EFVEDFTAFADVCFKEFGDRVLYWTTVNEANNFASLTLDEGNFMPSTEPYIRGHNIILAHASAVKLYR
EKYKKTQNGFIGLNLYASWYFPETDDEQDSIAAQRAIDFTIGWIMQPLIYGEYPETLKKQVGERLPTF
TKEESTFVKNSFDFIGVNCYVGTAVKDDPDSCNSKNKTIITDMSAKLSPKGELGGAYMKGLLEYFKRD
YGNPPIYIQENGYWTPRELGVNDASRIEYHTASLASMHDAMKNGANVKGYFQWSFLDLLEVFKYSYGL
YHVDLEDPTRERRPKASANWYAEFLKGCATSNGNAKVETPL

>SoBAHD1_CDS
ATGGAACCTTCAAAAATGGAAGTGAAAATAATATCGTCCGAAACCATCAAACCGTCATCTCCGACACC
ATCCCACCTTCGAAAATATACACTTTCTTTGCTCGACCAAAAATACACGCCTATCGTTGTTCCGGCCA
TTCTATTCTATGAGCGCCCACAAGGGGTGGCGCCATTGGATATGGACCGTCTCAGAACATGCCTCTCA
CAGACACTTACCGCGTTTTACCCTTTAGCCGGACGAGCTGAATCTCGAGACGTTATAATATGTAATGA
CGAAGGTATCCCCTTCGTTGAGGCTCATGTCGATTGTGAACTTTCGAGTGTTGTTAAGTCGCTTTCGT
CCCTAGGGAGTGATTTGCGGTCTTTTTACCCGCCTAGGGACGGTTTACTCGAGGGGGGAATTCAGTTT
GCTATTCAGATGAATGTGTTTAGTTGTGGCGGGTTTGCGTTCGCGTGGTATTGCACGCATAACGTTAC
TGACGGGACCTCGACTGCTAACTTTTTTAGGTATTGGACTGCGCTGTATGCTCAACGTAGTGAGTACG
CAGTCCAAGACCTAATGGATTTCAATTCCGTCGTCACTGCCTTTCCCCCTGTGCCGCCCCGTGTACCG
CAGGAGGAAAAACCGGTGACAACGGAATTGAAACCCGAGAAACAAGAGGGACAAGAAAAGGAGGAAAA
GAAAAAATCGTCATTTAATTTCAGTTTTCAATCTCACATCGTGGCGAGGAGTTTCTTGATAAAGAGCA
AGGCGGTCGCAGAGTTGAAGGCCAAGTCGGTAAGCGAGGAAGTGCCATATCCGAGTCGGTTCGAGGCC
GTGTCGGCTTTCCTATGGAAATCGATAGTGTCAAGCTCGACAACAGAAGGGAAGACGATGATCAATAT
GCCCGTAAACTTGAGACCACGGGTGGACCCGCCATTACCCTTGGACTCCGTAGGTAACATTTTCGAAA
ATGCACTCGTACAGTCCGAGAAAAAAGCGGAGCTCCACGAATTCGTTGCAAGGATCCGTGGATCAATC
TCGAAAATGAAAGATTTTGCCACGGAATATCAAGGCGAAAAGCGGGAAGAAGCTAAGGACGCACATTG
GAAAAGATTCATAAAAGCGGTTATCGAGTGTAAGGGGAAAGACGCCTACGTAATTTCGCCTTGGTATA

AGTCGTCCGGGTTTACGGACATAGATTTCGGGTTTGGGACCCCGATACGGGTCGTACCCATGGACGAT
GTCGTAAATCATAATCAAAGGAACACGATAATGTTGATGGAGTTTGTTGATTCCGACGGTGATGGATT
TGAAGCTTGGATGTTCCTGGAGGAGGAATGTATCAAGTTTTTGGAGTCCAACCCGGAATTTCTTGCCT
TTGCTTCCCCAAACTTTTAA

>SoBAHD1_translated
MEPSKMEVKIISSETIKPSSPTPSHLRKYTLSLLDQKYTPIVVPAILFYERPQGVAPLDMDRLRTCLS
QTLTAFYPLAGRAESRDVIICNDEGIPFVEAHVDCELSSVVKSLSSLGSDLRSFYPPRDGLLEGGIQF
AIQMNVFSCGGFAFAWYCTHNVTDGTSTANFFRYWTALYAQRSEYAVQDLMDFNSVVTAFPPVPPRVP
QEEKPVTTELKPEKQEGQEKEEKKKSSFNFSFQSHIVARSFLIKSKAVAELKAKSVSEEVPYPSRFEA
VSAFLWKSIVSSSTTEGKTMINMPVNLRPRVDPPLPLDSVGNIFENALVQSEKKAELHEFVARIRGSI
SKMKDFATEYQGEKREEAKDAHWKRFIKAVIECKGKDAYVISPWYKSSGFTDIDFGFGTPIRVVPMDD
VVNHNQRNTIMLMEFVDSDGDGFEAWMFLEEECIKFLESNPEFLAFASPNF