The Biosynthesis of Aescin

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

Eliot Jan-Smith

March 2022

Contents

Abstract: 5

Acknowledgements: 6

List of Abbreviations: 7

1 – General Introduction: 8

1.1 Saponins within wider plant specialised metabolism: 8

1.2 The Biosynthesis of triterpenoid saponins: 10

1.3 Aescin, a triterpenoid saponin from *Aesculus spp.*: 15

1.4 The pharmaceutical properties of aescin: 18

1.5 Natural role and other applications of aescin: 20

1.6 Similar saponins to aescin: 21

1.7 Bio-engineering of triterpenoid saponins: 22

Project aims: 23

2 – Database Mining and Triterpenoid Acylation: 25

2.1 Introduction: 25

2.2 Methods: 27

2.3 The distribution of acylation on triterpenoid scaffolds: 28

2.4 The frequency of acyl groups involved in beta-amyrin acylation: 34

2.5 Analysis of reported ACTs: 36

2.6 Predicting a biosynthetic pathway to aescin: 40

2.7 Conclusions: 42

3 – Evaluating the triterpenoid toolkit for reconstruction of the aescin pathway: 45

3.1 Introdution: 45

3.2 Methods: 51

3.3 Validation of the biosynthesis of beta-amyrin in *N. benthamiana*: 52

3.4 Evaluation of the activity of CYP 72A69 in *N. benthamiana*: 54

3.5 Evaluation of the activity of CYP 716A12 in *N. benthamiana*: 55

3.6 Evaluation of the activity of CYP 93E1 in *N. benthamiana*: 56

3.7 Evaluation of the activity of CYP 716-2012090 in *N. benthamiana*: 58

3.8 Combinatorial trials of toolkit CYPs: 60

3.9 Cloning and testing of a reported erythrodiol synthase: 63

3.10 Conclusions: 65

4 – Identification and Cloning of Candidate Biosynthetic Enzymes from *A. pavia*: 66

4.1 Introduction: 66

4.2 Methods: 68

4.3 Identification and cloning of candidates from *A. pavia*: 69

4.4 Characterisation of the activity of ApCYP1: 71

4.5 Evaluation of the activity of ApCYP1 in combination with toolkit CYPs: 73

4.6 Isolation of 21β-OH beta-amyrin and kudzusapogenol C: 76

4.7 Conclusions: 81

5 – Metabolic Profiling and Transcriptomic Analysis of *Aesculus*: 82

5.1 Introduction: 82

5.2 Methods: 84

5.3 LC-MS analysis of aescin standard: 86

5.4 Metabolic profiling of *Aesculus* leaves: 88

5.5 Metabolic profiling of *Aesculus* flowers: 90

5.6 Metabolic profiling of *Aesculus* fruits: 91

5.7 RNAseq experiments in *A. pavia* and *A. hippocastanum*: 97

5.8 Transcriptome assembly: 97

5.9 Conclusions: 101

6 – Identification and Cloning of OSC and CYP Candidate Genes Using *de novo* Transcriptome Data: 102

6.1 Introduction: 102

6.2 Methods: 105

6.3 Differential expression and co-expression analysis of *A. pavia* transcriptome: 106

6.4 Identification of an active OSC from *A. hippocastanum*: 110

6.5 Cloning and testing of AhCYP1, an *A. hippocastanum* homologue of ApCYP1: 112

6.6 Identification of AhCYP2 from the *A. hippocastanum* transcriptome: 113

6.7 Identification and characterisation of AhCYP3: 118

6.8 Isolation and structural determination of aescin intermediates: 119

6.9 Combinatorial biosynthesis of protoaescigenin in *N. benthamiana*: 127

6.10 Conclusions: 129

7 – Acylation and Glycosylation of the Aescin Aglycone: 131

7.1 Introduction: 131

7.2 Methods: 135

7.3 Identification and characterisation of AhBAHD1: 135

7.4 Identification and characterisation of AhCSL1: 142

7.5 Characterisation of ApUGT2, a D-galactosyltransferase: 149

7.6 Identification and characterisation of AhUGT1: 151

7.7 Identification and characterisation of AhUGT2: 153

7.8 Conclusion: 157

8 – General Discussion and Conclusions: 159

8.1 Aescin biosynthetic pathway elucidation using publicly available bioinformatic resources: 159

8.2 Characterisation of aescin biosynthetic enzymes from *A. hippocastanum*: 160

8.3 Progress along the aescin biosynthetic pathway: 162

8.4 Relevance of findings to bio-engineering of saponins: 164

8.5 Prospects for future work: 165

Concluding remarks: 166

9 – General Materials and Methods: 167

9.1 Molecular biology methods: 167

9.2 Transient expression in *N. benthamiana* methods: 175

9.3 Extraction and analysis of infiltrated leaves methods: 176

10 -Bibliography: 180

Appendix A – Reported ACTs: 201

Appendix B – Protein sequences of cloned enzymes: 206

Appendix C – Nucleotide sequences of synthetic genes: 215

Appendix D – NMR of isolated compounds: 218

Abstract

Aescin is a mixture of triterpenoid saponins found in horse chestnuts (*Aesculus spp.*) which has elicited interest from the chemical industry for a range of applications, from medicinal use to inclusion in household cleaning products. Wider development of products containing aescin would be greatly aided by cheaper, more environmentally-friendly supplies – one potential means to achieve this could be production through heterologous expression of biosynthetic enzymes. This work therefore seeks to elucidate the biosynthetic pathway to aescin and reconstitute it in a heterologous plant host. To this end a biosynthetic pathway was predicted by database mining, and previously reported triterpenoid biosynthetic enzymes from other species were used to reconstitute early steps by transient expression in *Nicotiana benthamiana*. An existing transcriptome assembly for *A. pavia* was used to identify the first active enzyme from the aescin pathway, a beta-amyrin C-21β hydroxylase. Together with enzymes from other species, this was used to isolate a product not previously obtained in N. benthamiana: kudzusapogenol C.

Novel transcriptomic resources for *A. pavia* and *A. hippocastanum* were subsequently generated and used to discover further aescin pathway enzymes involved in formation and oxidation of the aescin beta-amyrin aglycone: AhOSC1, AhCYP1, AhCYP2, and AhCYP3. The activities of the cytochrome P450s (CYPs) identified have been validated by scale-up, purification and NMR, allowing isolation of novel aescin intermediates up to 21β,22α,24,28-OH beta-amyrin. Identification of a cellulose synthase-like enzyme (AhCSL1) and a UDP-dependent sugar transferase (AhUGT1) capable of adding D-glucuronic acid and D-galactose, respectively, along with partial characterisation of the activities of a BAHD and a second UGT, led to detection of intermediates only two transformations removed from aescin. The enzymes identified through this work therefore provide an excellent platform for further work on bioengineering of aescin and analogues thereof. The novel activities of some of the enzymes characterised here may also enable production of new-to-nature acylated and glycosylated triterpenoids.

Acknowledgments

First and foremost, I would like to thank my supervisor Prof Anne Osbourn for the opportunity to carry out this project, and for all her help and guidance over these four and a half years. Her advice and encouragement has been incredibly valuable. I would also like to thank my secondary supervisors Prof Rob Field and Prof George Lomonossoff, and my industrial supervisor Sarah Hosking for their advice.

I am deeply grateful to all the members of the Osbourn group over my time here. The support I have received from the group has been genuinely touching, and I am hugely indebted to all of you. I have learnt so much from my time in this group. I would particularly like to thank Thomas Louveau for supervision during the first half of my project and for his patience in teaching me so much, and Michael Stephenson for supervision during the later half of my PhD.

The staff at the John Innes Centre have been vital to this project, especially those in horticultural services, lab support, and metabolomics. I am very grateful to Dr Lionel Hill and Dr Paul Brett for their support with LC-MS and GC-MS.

To all my friends, in Norwich, Birmingham, Derby, or elsewhere – thank you.

Lastly, I would like to thank my family, without which this project would not have been possible. Jim and Chantal, my parents, whose love and support has helped me through all the long days in the lab, and my sister Alix. I could not imagine having completed this PhD without you.

This thesis is dedicated to the memory of my grandmother Jean, with whom I first collected conkers as a boy.

List of Abbreviations

ACT - Acyltransferase

BAHD – (BEAT, AHCT, HCBT, DAT) Acyltransferase

CoA – Coenzyme A

COSY – Correlation spectroscopy

CSL – Cellulose synthase-like

CYP – Cytochrome P450

EI – Earlham Institute

GC-MS – Gas chromatography – mass spectrometry

GFP – Green fluorescent protein

HMBC – Heteronuclear multiple-bond correlation spectroscopy

HSQC – Heteronuclear single-quantum correlation spectroscopy

JIC – John Innes Centre

LC-MS – Liquid chromatography – mass spectrometry

MS – Mass spectrometry

OneKP – One thousand plants

OSC – Oxido-squalene cyclase

PCR – Polymerase chain reaction

ppm – parts per million

NBI – Norwich Biosciences Institutes

NMR – Nuclear magnetic resonance

SCPL – Serine carboxypeptidase-like

tHMGR – truncated HMG-CoA-Reductase

TMS - Trimethylsilyl

UGT – Uridine diphosphate dependent glycosyltransferase

Chapter 1 – General Introduction

1.1 – Saponins within wider plant specialised metabolism

The plant kingdom produces a great variety of chemical products. These may broadly be split into two categories: the products of primary metabolism, and those of secondary, or specialised, metabolism. Although the distinction between these is not always entirely clear, primary metabolites are typically defined as those deemed essential to basic physiological 'housekeeping' processes, e.g. respiration, photosynthesis, or growth. Primary metabolites are highly conserved evolutionarily across the vast majority of plants (Weng, Philippe et al. 2012, Maeda 2019). Loss or change in function of the genes involved in primary metabolism typically has deleterious effects. In contrast, secondary (or specialised) metabolites are generally regarded as being dispensable for growth and survival under laboratory conditions, but have important roles in interactions between plants and other organisms in natural environments (Pichersky and Gang 2000, Weng, Philippe et al. 2012). The ability to produce particular types of specialised metabolites is usually restricted to narrow taxonomic subgroups of plants, e.g. steroidal glycoalkaloids in the Solanaceae (Friedman and McDonald 1999); or limonoids in the Rutaceae and Meliaceae (Hasegawa 1999), most likely as a consequence of adaptation to particular environmental niches (Pichersky and Gang 2000, Pichersky and Lewinsohn 2011). This has resulted in the emergence of enormous chemical diversity within higher plants. It has been estimated that the plant kingdom collectively may be capable of production of over one million different compounds (Afendi, Okada et al. 2012).

Specialised metabolism has evolved from aspects of primary metabolism. The feedstock for most specialised metabolic pathways is drawn from the primary metabolome (the totality of primary metabolic products in a plant species), and many of the enzymes implicated in specialised metabolic pathways are likely to have originated from gene duplication events of primary metabolic enzymes, followed by subsequent gain of functionality (Ober 2010, Pichersky and Lewinsohn 2011). This origin is thought to play a part in the observed chemical diversity – gene duplication of primary metabolic enzymes eases the selection pressure on those enzymes, which then tend to evolution for increased substrate promiscuity (Pichersky and Gang 2000, Aharoni, Gaidukov et al. 2005, Weng, Philippe et al. 2012). The resultant diversity of specialised metabolism includes such classes of compounds as the polyketides, isoprenoids, shikimates, and many more.

The saponins constitute a large group of structurally complex specialised metabolites found across a range of plant species. They are a subset of the broader class of isoprenoids (Holstein and Hohl 2004). Isoprenoids are formed by concatenation of the five-carbon molecules isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP), and named according to the number of these molecules that are employed in their synthesis (Geissman and Grout 1969). Molecules formed from six of these molecules, for a thirty-carbon structure, are termed triterpenoids, or, if degraded through subsequent loss of carbons, steroids (Patterson and Nes 1991, Mann, Davidson et al. 1994). Saponins have either a triterpenoid or a steroid scaffold (the aglycone or sapogenol) that has been modified by addition of one or more sugar chains (Hostettmann and Marston 1995). Saponins with a single sugar chain are termed monodesmosidic, the sugar chain typically attached at the C-3 position of the scaffold. Addition of a further sugar chain can also occur, commonly at the C-28 position, to give bidesmosidic saponins (Kasai, Yamasaki et al. 1999). In plants, steroidal saponins are more closely associated with the monocots, and triterpenoid saponins with the dicots, although exceptions to this rule have been found (Springob and Kutchan 2009).

Saponins take their name from their erstwhile use as the active principle of soaps and detergents, a use which in some cases is still ongoing (Kunatsa and Katerere 2021). The use of saponin-containing plants as soaps is often indicated by a species' names – either common or scientific – for example soapwort (*Saponaria officinalis*), soapbark (*Quillaja saponaria*), or soapberry (*Sapindus saponaria*). Laundry detergents made from saponin-containing plants have a long history, with reports of the ashes of horse chestnuts being used to clean laundry dating back to the early 19th century (Duhamel du Monceau 1819). The presence of saponins in a plant was previously determined by the formation of a long-lived foam or lather upon agitation of aqueous extracts of the plants. The detergent activity of saponins is a function of their bipartite structure, with a highly hydrophilic moiety (the sugar chain) and a significantly less polar hydrophobic moiety (the aglycone). The marked difference in polarity between these two moieties leads to surfactant activity and micelle formation (Hostettmann and Marston 1995). This structure also has implications for other activities, notably haemolysis (red blood cell lysis), molluscicidal, and piscicidal activity, common to many saponins.

Haemolysis has historically been used as a means of quantification of saponin concentration, many of the better-studied saponins having reported 'haemolytic indices' (Voutquenne, Lavaud et al. 2002). Haemolytic activity is observed to be greater in

monodesmosidic saponins than in those with multiple sugar chains (Springob and Kutchan 2009). Although this widespread haemolytic activity might be expected to lead to saponins being highly toxic to mammals, this is not generally the case in practice; the toxicity of saponins administered intravenously is high, but saponins tend to be very poorly absorbed in the gut when ingested orally (Navarro Del Hierro, Herrera et al. 2018). Interestingly, toxicity is significantly greater in gill-breathing organisms, where saponins attack the membranes of the gills. This has led to their use as piscicides, as concentrations sufficient to kill fish remain well below levels toxic to those eating the fish (Cannon, Burton et al. 2004). For example, the de-fatted ground seeds of a species of tea (*Camellia oleifera*) are still used in Chinese aquaculture to kill fish, for which the tea saponins are thought to be responsible (Chen, Yang et al. 2010). Not all saponins can be said to have low toxicity, however. Glycyrrhizin, a triterpenoid saponin from liquorice, can be highly toxic (even fatal) with excessive consumption (Omar, Komarova et al. 2012).

The bitter taste of many saponins points to their potential roles in the plants of origin as anti-feedant or insecticidal agents that deter herbivory (Koul 2008). However, there is considerable interest in the pharmaceutical and other potentially useful properties of saponins. Many of these compounds are now being investigated for a range of applications, from anti-tumour, antiviral, and anti-inflammatory, to uses in the cosmetics industry as bio-surfactants (Sparg, Light et al. 2004, Asres, Seyoum et al. 2005, Bezerra, Rufino et al. 2018). Examples of some of these applications include the use of QS-21 from *Quillaja saponaria*, in addition to other saponins, as an adjuvant to increase the immune response caused by vaccines (Barr, Sjölander et al. 1998), or the use of saponins from soapwort (*Saponaria officinalis*) as foaming agents and emulsifiers in the food industry (Jurado Gonzalez and Sörensen 2020). This interest in saponins has spurred research into alternative means of obtaining industrial-scale quantities of molecules of interest for commercial development.

1.2 – The biosynthesis of triterpenoid saponins

As mentioned previously, the triterpenoid scaffolds from which saponins are made are isoprenoids, with metabolic origins in the mevalonate pathway (MVA pathway). The MVA pathway transforms units of acetyl-CoA (an ubiquitous product of primary metabolism) into five-carbon molecules of isopentenyl pyrophosphate (IPP) or dimethylallyl pyrophosphate (DMAPP), as illustrated in figure 1.2.1 (Miziorko 2011). This pathway is well characterised, and in plants the rate-limiting step is the action of HMG-CoA-Reductase

(HMGR), which self-regulates according to the concentration of metabolites, limiting the overall accumulation of MVA pathway products (Polakowski, Stahl et al. 1998).

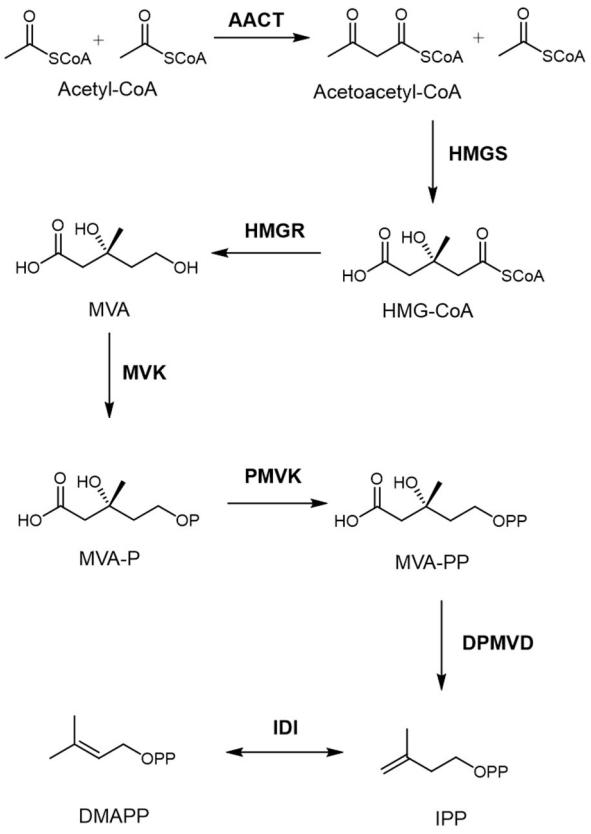


Figure 1.2.1 – The first part of the mevalonate (MVA) pathway, which leads to biosynthesis of triterpenoids in plants. Starting from acetyl-CoA, this pathway forms five-carbon products isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP), which go on to be used in the biosynthesis of mono-, di-, and triterpenoids. The rate-limiting step is the action of HMG-CoA-Reductase (HMGR), which regulates the accumulation of products.

The products IPP and DMAPP are then combined into progressively longer and longer carbon chains, as depicted in figure 1.2.2, to give the thirty-carbon linear compound 2,3-oxidosqualene, which is the immediate precursor for triterpenoid biosynthesis (Nicholas 1963, Do, Kiss et al. 2009)..

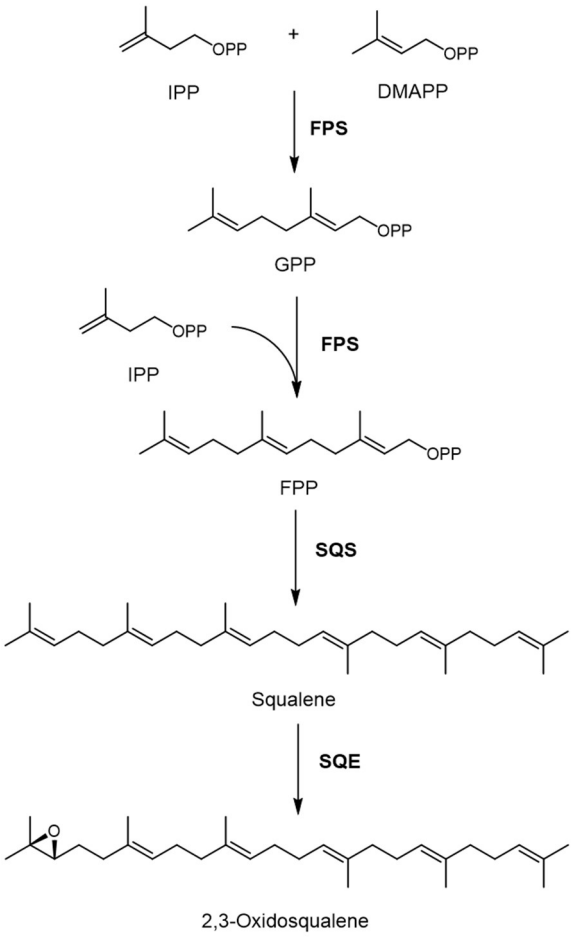


Figure 1.2.2 – The second part of the biosynthesis of triterpenoid precursors in plants. Squalene is formed by the condensation of two units of farnesyl pyrophosphate (FPP) end to end by squalene synthase (SQS), which is then epoxidised by squalene epoxidase (SQE). The oxygen from the 2,3-epoxy group becomes the C-3 hydroxyl of the triterpenoid upon cyclisation by an OSC.

The pathway so far is common to the biosynthesis of all the triterpenoids. Structural diversity starts to emerge with cyclisation of 2,3-oxidosqualene by oxidosqualene cyclases (OSCs). These attack the epoxide of folded 2,3-oxidosqualene, which initiates a cascade of ring formations using the C-C double bonds, leading to cyclic structures (Christianson 2006, Thimmappa, Geisler et al. 2014). The products formed may have a portion of the carbon chain left uncyclised, or may be completely cyclised into a pentacyclic triterpenoid. Despite using the same substrate and relying on the conformation of folded 2,3-oxidosqualene to form different products, the OSCs show marked product specificity (Wendt, Schulz et al. 2000). To demonstrate their structural diversity, some OSC products from plants are depicted in figure 1.2.3, including cycloartenol, the precursor to the steroids in plants (Hostettmann and Marston 1995).

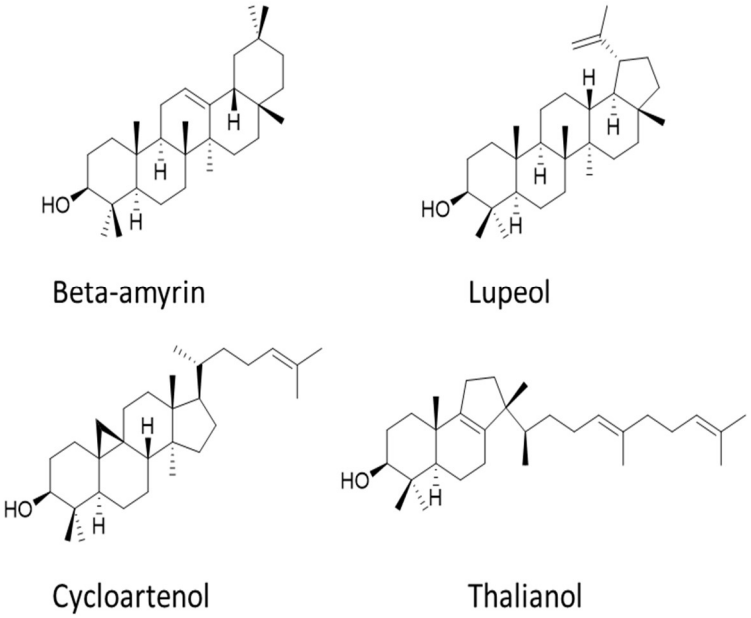


Figure 1.2.3 – Chemical structures of four triterpenoids formed by the cyclisation of 2,3-oxidosqualene by OSCs. These structures are very diverse, despite originating from the same molecule. Of note are beta-amyrin, one of the most frequently encountered OSC products in plants, and cycloartenol, the precursor to plant steroids (animals use lanosterol). Like beta-amyrin, lupeol is also widely distributed across taxa, whereas thalianol is much more taxonomically restricted (Liu, Cheema et al. 2020).

The cyclisation products generated by OSC enzymes can then undergo modification by a range of tailoring enzymes. One of the most frequent modifications is oxidation. This can take the form of hydroxylation, epoxidation, or introduction of a carbonyl function (a ketone, aldehyde, or carboxylic acid group). A number of enzymes from the cytochrome P450 (CYP) family that oxidise triterpenoids have been characterised (Seki, Tamura et al. 2015, Ghosh 2017, Miettinen, Pollier et al. 2017). The TriForC database (Miettinen, Iñigo et al. 2018)(see section 1.7) lists 90 CYPs active as part of triterpenoid biosynthetic pathways. CYPs are implicated in biosynthetic pathways for many classes of plant secondary metabolites, and consequentially have been studied extensively (Nelson and Werck-Reichhart 2011, Hamberger and Bak 2013, Xu, Wang et al. 2015). Oxidation of triterpenoids is closely linked to their biological and physical properties: addition of a number of hydroxyl groups to a triterpenoid aglycone of a saponin, for instance, will increase the polarity of the aglycone moiety, altering the surfactant and haemolytic properties of the saponin (Voutquenne, Lavaud et al. 2002). Hydroxylation is also required for some of the further downstream modifications of triterpenoids, including acylation, and also glycosylation at positions other than the C-3 carbon.

Acylation – ie. the addition of an acyl (sometimes termed ester) group – to a saponin can occur on either the aglycone or glycosidic moiety of a saponin, and has been shown to influence bioactivity and bioavailability (Chan 2007, Podolak, Galanty et al. 2010, Zhang, Zhang et al. 2019, Alqathama, Shao et al. 2020). As well as affecting the polarity and distribution of hydrogen bond acceptors in the molecule, acylated saponins can also act as pro-drugs, where the acyl group masks a hydroxyl group relevant to bioactivity until the former is cleaved (Yang, Zhao et al. 2004, Zhang, Zhang et al. 2019). In plant specialised metabolism, acylation is typically carried out by acyltransferases (ACTs) belonging to one of two families: the BAHDs (named for the first four enzymes of the class to be discovered: BEAT, AHCT, HCBT, and DAT) and the serine carboxypeptidase-like ACTs (SCPLs) (Milkowski and Strack 2004, D'Auria 2006, Bontpart, Cheynier et al. 2015). These two families of ACTs differ in the types of acyl donors that they use (BAHDs use acyl-CoA thioesters as an acyl donor, while SCPLs use acyl-sugar derivatives) and in their subcellular localisation (BAHDs are cytosolic, while SCPLs are vacuolar) (Bontpart, Cheynier et al. 2015). In contrast to the OSCs and the CYPs, far fewer ACTs from either family have been characterised from triterpenoid biosynthetic pathways. The TriForC database (Miettinen, Iñigo et al. 2018) lists only a single ACT from triterpenoids: SAD7 from oats, an SCPL responsible for the addition of N-methyl anthranilate in the biosynthesis of the antimicrobial saponin avenacin A-1 (Mugford, Qi et al. 2009, Mugford and Osbourn 2010). The *N*-methyl anthranilate D-glucose acyl donor used by SAD7 is furnished by two other pathway enzymes, the methyltransferase SAD9 and the glucosyltransferase SAD10 (Mugford, Louveau et al. 2013). More recently BAHDs involved in the acylation of thalianol and other *Arabidopsis thaliana* triterpenes have been reported. Unlike SAD7, these enzymes can be expressed in both the heterologous plant host *N. benthamiana* and in yeast without the need for introduction of additional auxiliary enzymes. Thus the acyl-CoA thioester acyl donors used by these BAHDs are presumably available as part of endogenous metabolism (Huang, Jiang et al. 2019, Bai, Fernández-Calvo et al. 2021).

Glycosylation of the triterpenoid aglycone is necessary for formation of a saponin. As previously mentioned, the resultant saponins can be monodesmosidic (a single sugar chain, typically at C-3), bidesmosidic (two sugar chains, typically C-3 and either C-28 or C-30) or even tridesmosidic (three sugar chains, only rarely encountered) (Majinda 2012). Glycosylation is essential for surfactant and haemolytic activity, due to the resultant amphiphilic nature of the saponin. Surfactant activity is usually only observed when at least three monosaccharide units are attached to the aglycone (Kasai, Yamasaki et al. 1999).

Triterpenoid glycosylation is most frequently carried out by uridine diphosphate-glycosyltransferases (UGTs). The UGT enzyme family has been implicated in the biosynthesis of many glycosylated specialised metabolites from a wide range of plant species, including triterpenoid saponins (Gachon, Langlois-Meurinne et al. 2005, Thimmappa, Geisler et al. 2014, De Bruyn, Maertens et al. 2015, Seki, Tamura et al. 2015, Tiwari, Sangwan et al. 2016, Rahimi, Kim et al. 2019). Examples of UGTs that transfer different sugars onto triterpenoids have been characterised, including for addition of D-glucose, D-rhamnose, D-galactose, L-arabinose, and D-xylose (Thimmappa, Geisler et al. 2014).

Other enzyme families have also been implicated in triterpenoid biosynthesis, for instance methyltransferases (Cárdenas, Almeida et al. 2019), epoxide hydrolases (Itkin, Davidovich-Rikanati et al. 2016), and cellulose synthase-like enzymes (recently reported as responsible for glycosylation of triterpenoids with D-glucuronic acid) (Chung, Seki et al. 2020, Jozwiak, Sonawane et al. 2020).

1.3 – Aescin, a triterpenoid saponin from *Aesculus spp.*

A triterpenoid saponin that is of considerable interest to the pharmaceutical and chemical industries is aescin (alternatively spelt escin or aescine), which is derived from extracts of the conkers (seeds) of horse chestnuts (trees belonging to the genus *Aesculus*). The horse chestnut genus consists of between 12 and 19 species distributed across the northern hemisphere, *Aesculus hippocastanum* being the only species native to Europe (Harris, Xiang et al. 2009, Thomas, Alhamd et al. 2019) (figures 1.3.1, 1.3.2.). *A. hippocastanum* is originally native to the Balkan mountains (Avtzis, Avtzis et al. 2007), but has been introduced across much of Europe, including the British Isles, possibly by the Romans (Bradshaw 2004), and then again during the early 17th Century (Thomas, Alhamd et al. 2019). Since then, it has become a very popular ornamental tree for planting in public parks, and the seeds are used in the children's game conkers. The timber of horse chestnuts is unsuitable for most applications (Howard 1945), although niche industrial applications for *A. hippocastanum* have been found, for instance the production of acetone for explosives manufacturing during the first world war by fermentation of the seeds (Jones and Woods 1986). The principal use of horse chestnuts is as medicine.

Figure 1.3.1 – The horse chestnut (*Aesculus hippocastanum*), engraving from Evelyn (1776). The leaves are palmate, and the flowers form a large panicle, having white petals with a red, pink, or yellow centre. The seeds (which contain the saponin aescin) are contained in a spiky shell. The tree typically grows to about 80ft tall, though the tallest in Britain reached 36m (118ft) tall (Thomas, Alhamd et al. 2019). Picture: John Innes Historical Collections, courtesy of the John Innes Foundation.

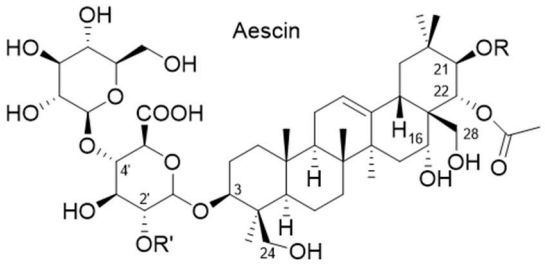
The medicinal use of horse chestnuts dates back to ancient times. The trees take their name from their use by Greeks and Turks to treat various ailments of horses (Thomas, Alhamd et al. 2019). Not long after their introduction to Britain, medicinal effects were already being ascribed to their use. In 1662, John Evelyn presented to the Royal Society a paper suggesting a flour of chestnuts mixed with honey as an electuary to be an effective remedy against “spitting of blood” (Evelyn 1776). An engraving of horse chestnuts from a later edition of his book “Sylva” is shown in figure 1.3.1.

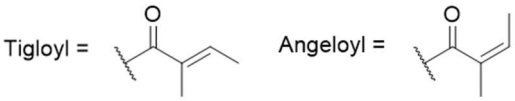
Later reports by Henri-Louis Duhamel du Monceau (see figure 1.3.2) noted that a flour made from conkers foamed upon agitation in water (highly indicative of the presence of saponins), and that the resultant paste is an effective detergent (Duhamel du Monceau 1819). He succeeded in separating the constituents of the conkers, noting the presence of a large portion of an “unbearably bitter extract”. This extract is likely to contain the saponin aescin. The medicinal uses mentioned include as a replacement for the anti-malarial quinine (Duhamel du Monceau 1819). Interestingly, the footnotes of his book describe what may be the first extraction of saponins from *A. hippocastanum*: a Mr Bon of Montpellier is said to have removed the bitterness from horse chestnut seeds in 1720 by soaking in water with lime for 24 hours, before a further eight days in pure water.

Figure 1.3.2 – *Aesculus hippocastanum* after Duhamel du Monceau (1819). The French term "marronier d'Inde" means Indian Chestnut, although these trees are not native to India, originating instead from the Balkans. Duhamel du Monceau recognised the horse chestnut as a potential source of starch, even higher-yielding than potatoes. However, extraction of the starch from conkers is very labour-intensive. Picture: John Innes Historical Collections, courtesy of the John Innes Foundation.

In modern times, horse chestnut seed extract has been reported to have vasoprotective effects, finding use in the treatment of chronic venous insufficiency, varicose veins, and related disorders of the peripheral vascular system (Pittler and Ernst 1998, Underland, Sæterdal et al. 2012, Dudek-Makuch and Studzińska-Sroka 2015). The active component of these extracts has been found to be aescin (Sirtori 2001, Gallelli 2019, Lichota, Gwozdzinski et al. 2019). The term aescin is subject to some confusion, since it refers to a mixture of compounds extracted from the seeds of the horse chestnut. Hostettmann and Marston describe aescin as a mixture of more than thirty individual saponins, containing: β-aescin (a mixture of saponins with acylation at C-21β and C-22α); kryptoaescin (a mixture with acylation at C-21β and C-28); α-aescin (a 4:6 mixture of β-aescin and kryptoaescin, the remnant of crystallisation of β-aescin from solution) and assorted artefacts resulting from hydrolysis of the main saponins (Hostettmann and Marston 1995). Other sources include isoaescins and aesculosides (Zhang, Li et al. 2010, Colson, Decroo et al. 2019). Of the horse chestnut saponins, the β-aescins are those with

the greatest haemolytic index, and are the most thoroughly investigated. In *A. hippocastanum* there are six principal components of β-aescin, as shown in figure 1.3.3.





SAPONIN	R	R'
AESCIN I A	Tigloyl	D-Glucose
AESCIN I B	Angeloyl	D-Glucose
AESCIN II A	Tigloyl	D-Xylose
AESCIN II B	Angeloyl	D-Xylose
AESCIN III A	Tigloyl	D-Galactose
AESCIN III B	Angeloyl	D-Galactose

Figure 1.3.3 – The structures of the six main components of β-aescin found in *A. hippocastanum*. All are based on beta-amyrin, with hydroxylation at C-16α, C-24, and C-28; acylation at C-22α with acetyl and C-21β with either tigloyl (aescin A) or angeloyl (aescin B); and glycosylation at C-3. The sugar chain comprises a D-glucuronic acid, D-glucose at C-4', and another sugar at C-2'. In aescin I this is D-glucose, in aescin II this is D-xylose, and in aescin III this is D-galactose.

First elucidated in the 1960s, the structures of these main aescins are based around beta-amyrin, one of the most common triterpenoid scaffolds. This is hydroxylated at C-16α, C-24, and C-28, acetylated at C-22α, and acylated at C-21β with either tigloyl or angeloyl (very similar small aliphatic acyl groups differing only in their configuration about the double C=C bond) (Wulff and Tschesche 1969). It is thought that the acylation at C-21β and C-22α is particularly significant for the bioactivity of aescin, and that of related saponins which share this pattern of acylation (Chan 2007, Zhang and Li 2007, Podolak, Galanty et al. 2010), as is hydroxylation at C-16α (Schlösser and Wulff 1969). The glycosidic moiety of aescin consists of a D-glucuronic acid at C-3, to which are added D-glucose at C-4' and either D-glucose, D-xylose, or D-galactose at C-2'. Aescin I (C-2' glucose) is the most abundant in *A. hippocastanum* extracts, followed by aescin II (C-2' xylose), with aescin III forming only a very small portion of β-aescin (Price, Johnson et al. 1987).

1.4 – The pharmaceutical properties of aescin

Aescin is not widely used in a purified form. Rather, most available pharmaceutical preparations are sold in the form of horse chestnut seed extract, often standardised for aescin concentration. The use of purified aescin (typically β-aescin) in research is more

common, though the composition of the aescin used is not always clear. Comparatively few studies into aescin use a single purified saponin, instead using what is likely to be a complex mixture of saponins.

Research into the treatment of vascular disorders constitutes the bulk of aescin research. Aescin has been investigated as a topical ointment applied to the affected area, as a cosmeto-textile that administers aescin through the skin (Cravotto, Beltramo et al. 2011), and as an oral medicine (Gallelli 2019), though some sources cite low oral bio-availability (Hostettmann and Marston 1995). The use of aescin to treat localised oedema is well established. It is thought that aescin acts to decrease the permeability of small blood vessels, which can lessen oedema (Sirtori 2001, Zhang, Li et al. 2010). Interestingly, a study in both mice and rats compared the different saponins in aescin, and found that aescins Ib, IIa, and IIb were the most effective at reducing the symptoms of oedema, whilst des-acyl aescins were ineffective (Matsuda, Li et al. 1997). Related vascular disorders for which aescin has been investigated include varicose veins (Lichota, Gwozdzinski et al. 2019), haemorrhoids (Sirtori 2001), chronic venous insufficiency (Yu and Su 2013), and general inflammation (Wang, Fu et al. 2009).

It is thought that the venotonic and anti-inflammatory activity of aescin towards blood vessel walls may be related to inhibition of the NFκB pathway, which is induced by the interaction of aescin with cell membrane-bound cholesterol (Domanski, Zegrocka-Stendel et al. 2016). Aescin, in common with many other saponins, is able to complex cholesterol (De Geyter, Swevers et al. 2012), and experiments with radio-labelled cholesterol have shown that aescin can both liberate cell-membrane cholesterol and inhibit uptake of cholesterol into the membrane (Böttger and Melzig 2013). The effect of this is to significantly alter membrane permeability. Experiments with model membranes have demonstrated that the incorporation of aescin into membranes results in formation of aescin-cholesterol complexes in the membrane, which could lead to pore formation and altered membrane structure (Sreij, Dargel et al. 2019).

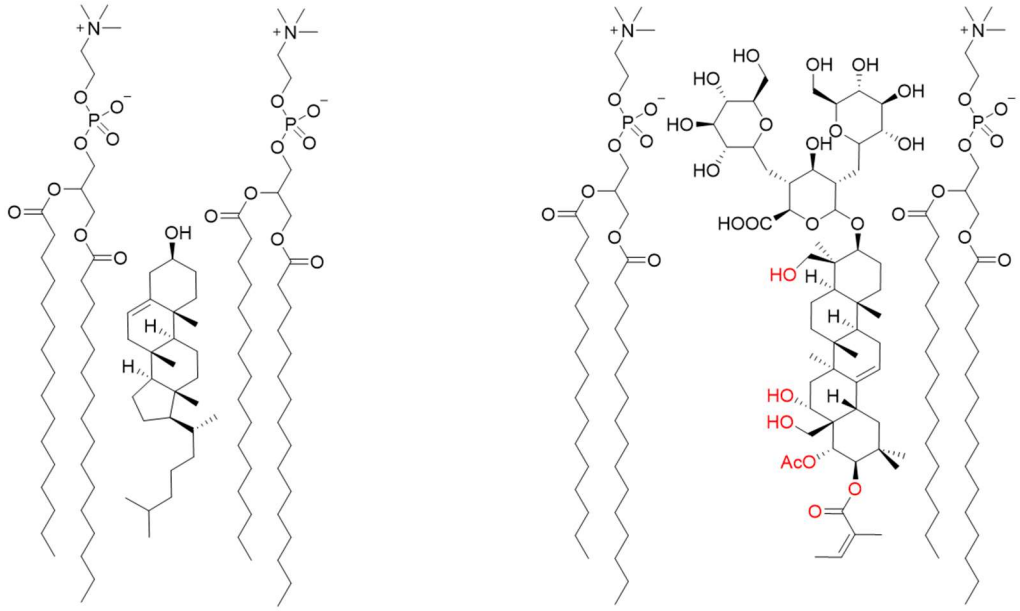


Figure 1.4.1 – Aescin, in common with many saponins, displaces cholesterol from cell membranes, altering their permeability (Böttger and Melzig 2013). Left: cholesterol in the cell membrane, stabilising the hydrophobic portion of the phospholipids. Right: replacement of cholesterol with aescin. The amphiphilic nature of aescin leads it to be readily incorporated into the cell membrane. The aglycone, however, has a number of hydrophilic groups (in red) clustered on the same side of the molecule, which may interact with the hydrophobic moieties of the phospholipids differently.

The use of aescin has also been investigated for the treatment of other types of disorder, including liver damage (Jiang, Xin et al. 2011, Singh, Sidhu et al. 2017), male infertility (Fang, Zhao et al. 2010), cystic fibrosis (Mutyam, Du et al. 2016), and even covid (Gallelli, Zhang et al. 2020). There have also been numerous studies on its potential as an anti-tumour agent (Konoshima and Lee 1986, Zhang and Li 2007, Cheong, Arfuso et al. 2018, Yang, Long et al. 2019, Zhang, Zhang et al. 2019). Aescin does not appear to be particularly toxic – cases of aescin poisoning, though recorded, are very rare (Edem, Kahyaoğlu et al. 2016).

1.5 – Natural role and other applications of aescin

In the horse chestnut, aescin is likely to have evolved to act as an anti-feedant or insecticide, as with many other saponins – consequentially, there has been some research into the use of aescin for plant protection. Studies have shown aescin to be an effective repellent to the horse chestnut leaf miner moth (Ferracini, Curir et al. 2010), and it has been found to active bitter taste receptors in *Drosophila* (Freeman, Wisotsky et al. 2014). Furthermore, aescin has been found to have anti-fungal properties, and to elicit plant defence pathways (Trdá, Janda et al. 2019).

Historically, horse chestnuts have been used as sources of laundry soap, charcoal, mattress stuffing, and as a clarifier for candle tallow (Duhamel du Monceau 1819). More recently, there has been interest in aescin related to its ability as a biosurfactant to form micelles, of interest to the cosmetic and chemical industries (Geisler, Dargel et al. 2019, Geisler, Prévost et al. 2020, Tucker, Burley et al. 2021, Tucker, Burley et al. 2021). Because these micelles are self-assembling and can incorporate other molecules, they have been investigated as drug delivery systems (Liao, Li et al. 2021). These properties have also been investigated for potential bio-remediation applications, eg. cleaning up after crude oil spills (Pekdemir, Ishigami et al. 1999). Aescin has also been investigated as a biosurfactant for removal of cadmium and lead from contaminated soils (Hong, Choi et al. 1998).

The potential of aescin's detergent properties has also been considered in relation to household cleaning products. Along with related triterpenoid saponins, aescin has been found to be a safe and effective surface cleaner for household use, with antibacterial activity (Fink and Filip 2022). Interestingly, some studies have found that horse chestnut seed extract actually performs better at microbial control than purified aescin (Fink, Potočnik et al. 2020). The ability of aescin-containing horse chestnut extracts to form dense, long-lasting foams has also seen investigation into their use as a foaming agent in the food industry, for example in foam mat drying to improve the qualities of various powdered foods (Aktas and Tontul 2021).

1.6 – Similar saponins to aescin

Aesculus hippocastanum belongs to the Sapindaceae family, in the order Sapindales, which, as their names might suggest, contain a number of saponin-producing species including lychee (Kilari and Putta 2016) and soapberry (Huang, Wu et al. 2008). Some of these saponins share structural features in common with aescin. For instance, saponins isolated from *Harpullia* species have the same pattern of hydroxylation of the beta-amyrin scaffold at C-16α, C-24, and C-28, and are also acylated at C-21β and C-22α with angeloyl groups (Dizes, Gerald et al. 1998, Voutquenne, Lavaud et al. 1998, Voutquenne, Guinot et al. 2005). These saponins also share the first sugar of the glycosidic moiety with aescin: D-glucuronic acid (Voutquenne-Nazabadioko 2010).

Saponins with similar structures to aescin are also found in more distantly related species. For instance, antonioside I from *Antonia ovata* (an asterid, so very distantly related to *Aesculus hippocastanum*, a rosid) has an aglycone nearly identical to that of aescin Ia, the sole difference being hydroxylation at C-23 instead of C-24 (Alabdul Magid, Lalun et al.

2012). The saponins from tea (*Camellia sinensis*) also have similar aglycones to aescin. A great number of saponins have been isolated from tea, including the camelliasaponins, theasaponins, and assamsaponins (Murakami, Nakamura et al. 1999, Murakami, Nakamura et al. 2000, Toshio Morikawa, Ning Li et al. 2006, Yoshikawa, Morikawa et al. 2007). Many of these saponins are based on similar aglycones – barringtogenol C (16α,21β,22α,28-OH beta-amyrin) or theasapogenol A (16α,21β,22α,23,28-OH beta-amrin), acylated at C-21β or C-22α with acetyl or angeloyl. The core aglycones of these saponins are very close to the core aescin aglycone protoaescigenin (16α,21β,22α,24,28-OH beta-amyrin) (Jatczak and Grynkiewicz 2014). Figure 1.6.1 demonstrates how closely related these aglycones are.

Aglycone	R_1	R_2	R_3	R_4
Harpuloside	Ang	Ang	OH	H
Antonioside I	Tig	Ac	H	OH
Barringtogenol C	H	H	H	H
Theasapogenol A	H	H	H	OH
Protoaescigenin	H	H	OH	H
Aescin A	Tig	Ac	OH	H
Aescin B	Ang	Ac	OH	H

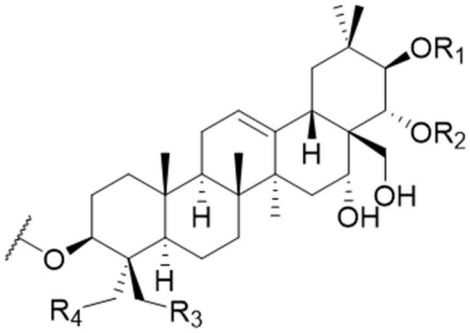


Figure 1.6.1 – Saponins from many different species, some only very distantly related, have similar aglycones to aescin. The saponins from *Harpullia spp.* have aglycones with the same pattern of hydroxylation, differing in the acylation at C-21β and C-22α. Saponins from tea are mostly based on the sapogenins barringtogenol C and theasapogenol A, which are very similar to protoaescigenin, the basis of aescin.

1.7 – Bio-engineering triterpenoid saponins

As interest in saponins as potential drugs or speciality chemicals grows, it will be advantageous to develop methods for production of target saponins and analogues thereof at scale. Many saponins only accumulate at low levels in the producing plant, and are often localised to a particular organ. Problems of access may be further confounded if the plants are slow-growing or not particularly abundant. Saponin-producing plant species that have received the most interest, such as *Quillaja saponaria*, which produces the saponin adjuvant QS-21, are vulnerable to over-exploitation in their native habitats (Szakiel, Pączkowski et al. 2011, Delporte, Rodríguez-Díaz et al. 2021). The *A. hippocastanum* trees that produce aescin take many years to reach maturity and are vulnerable to pests and diseases (Ćalić-Dragosavac, Zdravković-Korać et al. 2010, Percival, Barrow et al. 2011). If aescin is to find wider industrial use, larger-scale, cheaper supplies will need to be

obtained. Previous investigations have focussed on tissue culture of *A. hippocastanum* to maximise aescin production (Ćalić-Dragosavac, Zdravković-Korać et al. 2010, Zdravković-Korać, Milojević et al. 2022), which has the advantage of requiring less space than trees, along with a much shorter maturation time and year-round production. However, *A. hippocastanum* tissue cultures still produce a complex mixture of saponin products, which is undesirable from a pharmaceutical perspective. An approach that would allow production of individual saponins would be preferable. This will require a thorough understanding of the enzymes required for aescin biosynthesis.

Pathway elucidation and bioengineering has been applied to other saponins of pharmaceutical or agronomic importance. The biosynthetic pathway for the antimicrobial oat saponin avenacin, for instance, has been elucidated and reconstituted in tobacco through transient expression (Li, Leveau et al. 2021). Heterologous expression of biosynthetic genes also enables synthesis of derivatives or intermediates that might otherwise be very difficult to obtain, especially through combinatorial biosynthesis (co-expression of enzymes from different biosynthetic pathways to make novel products) (Zhao and Li 2018, Beyraghdar Kashkooli, Van Der Krol et al. 2019). Of the hosts available for heterologous expression, tobacco and yeast are amongst the most popular (Ikram, Zhan et al. 2015, Tatsis and O'Connor 2016, Huang and Osbourn 2019). Both have been employed to investigate saponin biosynthesis, and metabolic engineering platforms have been developed to increase the yields obtained from heterologous expression (Reed and Osbourn 2018). In *Nicotiana benthamiana*, gram-scale yields of triterpenoid products have been achieved by transient plant expression (Reed, Stephenson et al. 2017). Bio-informatic resources to aid research into triterpenoid biosynthesis are also available, for example the TriForC database of characterised triterpenoid biosynthetic enzymes (Miettinen, Iñigo et al. 2018).

Project Aims

Aescin, a mixture of saponins from *A. hippocastanum* and related species, is of significant interest to the pharmaceutical and chemical industries due to its biological properties. However, current sources of aescin (extraction from horse chestnut seeds) would not be commercially viable at larger scales, which disincentivises further research into wider use of aescin. Obtaining a new supply of aescin, preferably one which enabled production of specific saponins rather than a mixture, is therefore highly desirable. The success of bio-engineering approaches to biosynthetic pathway elucidation for saponins from other

species has informed design of a similar approach for aescin, with the aim of identification and characterisation of the enzymes involved in the biosynthesis of aescin. These enzymes could then be deployed for large-scale production of aescin and analogues thereof, providing an alternative to the costly, inconvenient, and environmentally-unfriendly harvesting of horse chestnuts for the chemical industry. These enzymes could also find wider use in future combinatorial biosynthesis experiments to produce novel saponins.

Chapter 2 – Database Mining & Triterpenoid Acylation

Synopsis

Of all the aspects of triterpenoid saponin biosynthesis, acylation is perhaps one of the least well understood; this is despite many of the most commercially important saponins being acylated. Aescin is no exception, being doubly acylated at C-21β and C-22α, which is thought to be a key determinant of biological activity. Elucidation of the biosynthetic pathway of aescin is therefore likely to require a better understanding of the nature of triterpenoid acylation. To this end, the chemical database Reaxys was used to investigate characterised acylated beta-amyrin derivatives – it was found that the two positions acylated in aescin (C-21 and C-22) were by some margin the most frequently acylated. It is hypothesised that steric effects may be partially responsible for this. Furthermore, the acyl groups most frequently encountered in acylated beta-amyrin derivatives are those found in aescin, making aescin an excellent model to investigate triterpenoid acylation. Analysis of the substrate specificities of reported acyltransferases (ACTs) found that these acyl groups were more likely to be transferred by a BAHD than an SCPL acyltransferase. Further investigation using Reaxys into triterpenoids isolated from *Aesculus* enabled the prediction of a putative biosynthetic pathway, which hypothesises a minimum scaffold required for testing of ACTs; together with a focus on BAHDs, it is hoped this will aid identification and testing of candidate aescin biosynthetic enzymes.

Introduction

2.1 – Introduction

As discussed in the previous chapter, acylation is employed in many triterpenoid biosynthetic pathways, altering biochemical properties (for instance solubility), masking reactive hydroxyl groups with a labile protecting group, and providing a route towards increased structural diversity of a common scaffold. The acylation of saponins at certain positions has been found to be closely linked to biological activities (Voutquenne, Lavaud et al. 2002, Chan 2007, Podolak, Galanty et al. 2010); it is also thought that the variance in potency between aescins Ia – IIb may be partially due to the different acyl groups present (Matsuda, Li et al. 1997). However, when compared to oxidation or glycosylation, acylation remains one of the lesser-known facets of triterpenoid biosynthesis; for instance, the TriForC database of triterpenoid biosynthetic enzymes (Miettinen, Iñigo et al. 2018) lists

only a single ACT. A more developed understanding of acylation is highly desirable, not only generally, but also in informing the strategy taken to investigate the biosynthesis of aescin.

Acylation in plant specialised metabolism is typically carried out by ACTs from one of two families: the BAHDs and the SCPLs (D'Auria 2006, Bontpart, Cheynier et al. 2015). As discussed in chapter 1, these differ in their sub-cellular localisation and acyl donor requirements – importantly, the pre-requisites of SCPLs mean that heterologous expression of an SCPL is likely to require co-expression with one or more additional biosynthetic enzymes to be active, as with SAD7 in oats, which requires co-expression with a glycosyltransferase and a methyltransferase to add N-methylanthranilate to avenacin pathway intermediates (Mugford, Louveau et al. 2013). Testing candidate SCPLs for activity would therefore require a different design of experiment to BAHDs, as such, it would be helpful to the project to predict which family is more likely to be involved in aescin biosynthesis.

Aescin is acylated at two positions on the aglycone: at C-21β with either tigloyl (aescin a) or angeloyl (aescin b), and at C-22α with acetyl. Triterpenoids with a similar pattern of acylation have previously been reported from other species, for instance the harpulosides (Dizes, Gerald et al. 1998) or the saponins from tea (Murakami, Nakamura et al. 2000). It is interesting that this same pattern of acylation should be seen in saponins from two species so distantly related as conkers and tea. Little is known as to whether there exist any trends or patterns in triterpenoid acylation which might explain this – consequentially, a wider investigation of the prevalence of acylated triterpenoids would be highly relevant to work on aescin. Such work should focus on the prevalence of acylation on different triterpenoid scaffolds, the frequency with which each position on those scaffolds is acylated, and which acyl groups are most commonly encountered in acylated triterpenoids. Any apparent preferences for acylation onto particular triterpenoid scaffolds or at specific positions on scaffolds could prove revealing when considering the activity of ACTs. The relative frequencies with which different acyl groups are encountered could also prove of significant interest; these may be indicative of the presence of precursor molecules across plants, or of the acyl donor specificities of ACTs themselves.

Chemical databases can be used for this investigation – there exist several publicly available which can search the literature for particular chemical structures (Menon, Krdzavac et al. 2019). Examples of these include PubChem (Kim, Thiessen et al. 2015), ChemSpider (Pence and Williams 2010), and Reaxys (Menon, Krdzavac et al. 2019). One

advantage that Reaxys has over the others is that it enables searches to be limited to compounds isolated as natural products from plants, including restricting the search to a particular species or taxon. This is particularly relevant to this work, as it would exclude any acylated triterpenoids which have been reported as synthetic products, perhaps by derivatisation of a natural compound. Reaxys also allows for literature searches by chemical structure which leave aspects of the structure undefined, which would enable, for instance, a search for triterpenoids with a particular modification at, say, C-21, but including all the results which are additionally substituted elsewhere on the scaffold.

Searching chemical databases could also have wider utility to this project – searching for triterpenoids isolated from *A. hippocastanum* may enable prediction of a biosynthetic pathway to aescin. If pathway intermediates have been isolated and reported, these could be searched for, and a likely order of transformations deduced. This could be very helpful to experimental design for testing of candidate biosynthetic enzymes.

Methods

2.2 – Methods

Triterpenoid scaffolds beta-amyrin, alpha-amyrin, and dammarane were selected for analysis of the distribution of acylation. Query structures were drawn for each position on the scaffold (acylation of sugar chains was not considered) using a MarvinJS software plugin, leaving all other positions on the scaffold available for substitution. The nature of the acyl group was left undetermined. To capture the large proportion of database entries without fully-defined stereochemistry, all stereochemistry was left undefined, including C-23/-24 and C-29/-30 methyl groups, which were treated together in queries. Query structures used to investigate frequency of different acyl groups left the position of acylation undefined, but specified the nature of the acyl group. An example query structure is depicted in figure 2.2.1, here searching for beta-amyrin derivatives acylated at C-21.

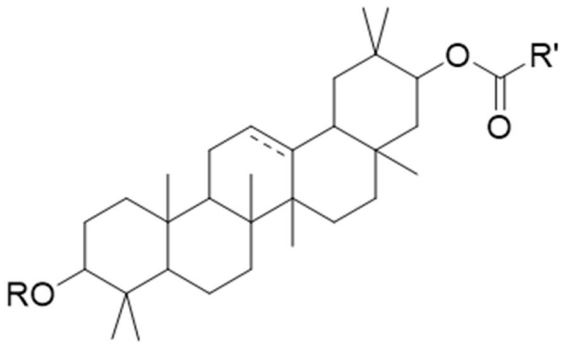


Figure 2.2.2 – Example query structure used for searching the database Reaxys. This structure was used to identify beta-amyrin derivatives acylated at C-21. Stereochemistry is deliberately left undefined. The C-12 – C-13 double bond was not required to account for possible derivatives such as epoxides. Substitution was enabled at every non-quaternary position.

Reaxys searches were carried out in December 2017 using these query structures, including the search field IS "isolated from natural source". The results were ordered by molecular weight and exported as PDF files. The total number of results for each query was noted and used for analysis. Reaxys searches for aescin pathway prediction in 2.6 were carried out using the IS "isolated from natural source" with the specific terms: "horse chestnut" "*Aesculus*" and "*A. hippocastanum*".

A full list of the BAHDs and SCPLs used in 2.5 can be found in appendices A1 and A2, respectively.

Results & Discussion

2.3 – The distribution of acylation on triterpenoid scaffolds

Analysis of the distribution of acyl groups on three of the most common triterpenoid scaffolds in plants (alpha-amyrin, beta-amyrin, and dammarane) was carried out. The structures of these scaffolds are depicted in figure 2.3.1.

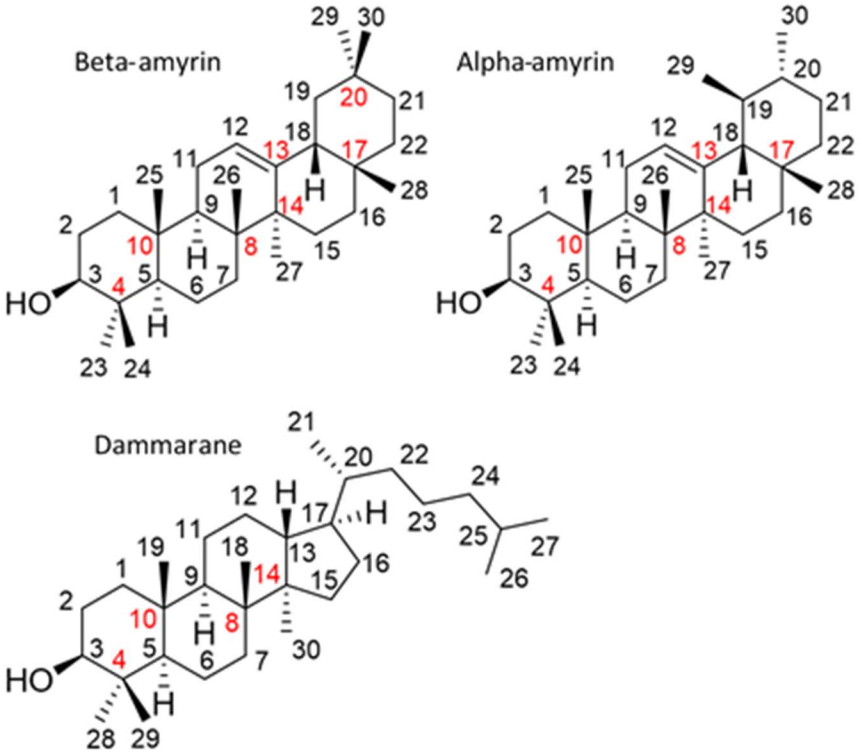


Figure 2.3.1 – The numbering scheme of the scaffolds investigated here: beta-amyrin, alpha-amyrin, and dammarane. Quaternary carbons (those which are already fully substituted, and cannot be modified without breaking of an existing bond) are numbered in red. Alpha- and beta-amryin are almost identical, differing only in the placement of the methyl groups 29 and 30

A query structure for acylation at each position (except for quaternary carbons) was constructed according to 2.2. An example query structure is shown in figure 2.2.1. These structures allowed for substitution at any position, but not for breaking of carbon-carbon bonds. It was noticed that many database entries were not annotated with fully-defined stereochemistry; the query structures were thus drawn without defined stereochemistry to better capture these entries. The Reaxys database was searched (December 2017) using these query structures, specifying that entries must have been reported as isolated from a natural source – this could theoretically include organisms other that plants, though manual assessment of returned entries did not find any non-plant organisms reported. The

number of database entries returned for each search was recorded – the results for the dammarane scaffold are shown in figure 2.3.2.

POSITION	ENTRIES	POSITION	ENTRIES
1	2	**17**	None
2	None	**18**	None
3	160	**19**	4
5	None	**20**	None
6	1	**21**	None
7	25	**22**	1
9	None	**23**	42
11	5	**24**	15
12	21	**25**	None
13	None	**26/27**	1
15	2	**28/29**	None
16	2	**30**	2

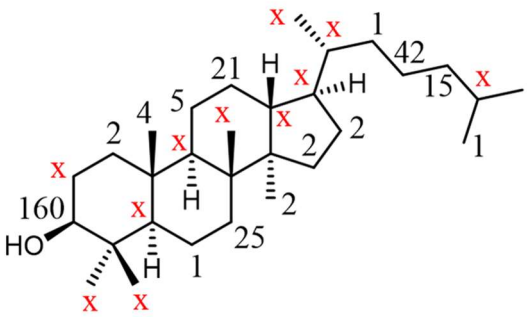


Figure 2.3.2 – The results of Reaxys searches for acylated dammarane derivatives isolated from natural products. The structure on the left shows the number of database entries returned for acylation at each position on the scaffold. Some positions – eg. C-3, C-7, C-12, and C-23, are acylated much more frequently than the others.

The results showed C-3 to be the position for which the most acylated derivatives were returned. This was to be expected – acylation requires hydroxylation, and C-3 is the only position in all three of the scaffolds investigated to already have a hydroxy group. Acylation at other positions returned fewer entries, though some positions were more represented in the database than others, such as C-7, C-12, and C-23. Interestingly, these positions were distributed across the scaffold. Positions for which the most acylated derivatives were found did not seem to be clustered in any part of the molecule.

Similar searches were carried out for alpha-amyrin; the results of these searches are shown in figure 2.3.3.

POSITION	ENTRIES	POSITION	ENTRIES
1	4	**19**	8
2	33	**20**	2
3	27	**21**	2
5	1	**22**	11
6	3	**23/24**	36
7	None	**25**	None
9	None	**26**	None
11	4	**27**	7
12	1	**28**	8
15	2	**29**	None
16	1	**30**	2
18	None		

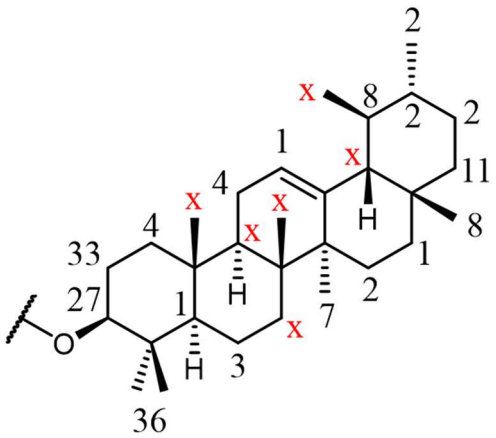


Figure 2.3.3 – The results of Reaxys searches for acylated alpha-amyrin derivatives isolated from natural products. The structure on the left shows the number of database entries returned for acylation at each position on the scaffold. The positions which returned the greatest number of acylated derivatives appear to be clustered, namely C-2, C-3, and the methyl groups C-23 / C-24.

The results obtained with alpha-amyrin scaffolds were slightly different to those with dammarane – this time, C-3 was not the position returning the most database entries, though it was still the third highest. Interesting, the only two positions for which more acylated derivatives were found were C-2 and C-23/24, the two available positions directly flanking C-3. These appear to cluster together on the scaffold, with very few acylated derivatives being found for positions elsewhere on alpha-amyrin. Alpha-amyrin is very similar structurally to beta-amyrin, the scaffold found in aescin (the only difference being the placement of the methyl group C-29) – and here the two positions found acylated in aescin, C-21 and C-22, only returned 2 and 11 entries respectively. To see if this apparent cluster around C-3 was reflected in beta-amyrin, the same searches were carried out with beta-amyrin scaffolds, the results of which are shown in figure 2.3.4.

POSITION	ENTRIES	POSITION	ENTRIES
1	1	**18**	None
2	35	**19**	1
3	32	**21**	623
5	None	**22**	579
6	2	**23/24**	88
7	None	**25**	1
9	None	**26**	None
11	2	**27**	25
12	11	**28**	145
15	7	**29/30**	13
16	141		

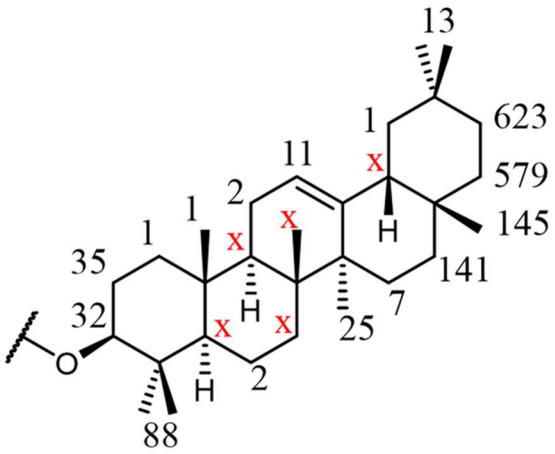


Figure 2.3.4 – The results of Reaxys searches for acylated beta-amyrin derivatives isolated from natural products. The structure on the left shows the number of database entries returned for acylation at each position on the scaffold. Many more entries were returned for beta-amyrin than for dammarane or alpha-amyrin. These results show a very apparent cluster of positions for which the most entries were returned: C-21, C-22, C-28 and C-16. A second cluster akin to that seen with alpha-amyrin around C-3 is also seen.

Many more database entries were returned for searches with beta-amyrin than searches with dammarane or alpha-amyrin. The distribution of positions with the greatest number of acylated derivatives is even more striking than for alpha-amyrin, with two positions (C-21 and C-22) returning significantly more entries than any others. Looking at these positions on the structure, it can be seen that there are two apparent clusters of positions for which acylation is more frequent – around C-3, as seen for alpha-amyrin, and on the opposite end of the molecule (C-21, C-22, C-28, and C-16). Interestingly, the positions in the centre of the molecule (rings B and C) returned very few entries when compared to those at the ends of the molecule (rings A and E). It can also be seen that positions at ring junctures (C-5, C-9, C-18) were not found to be acylated. These observations might be partially explained by the steric effects of acylation at each position – the apparently favoured positions on the A and E rings are those with the least steric hindrance, whereas the positions towards the centre of the molecule, especially those at ring junctures, are heavily constrained by the steric effects of the rigid pentacyclic structure. A side-on view of the structure of beta-amyrin shown in figure 2.3.5 depicts this, especially the reduced steric hindrance of positions C-21 and C-22.

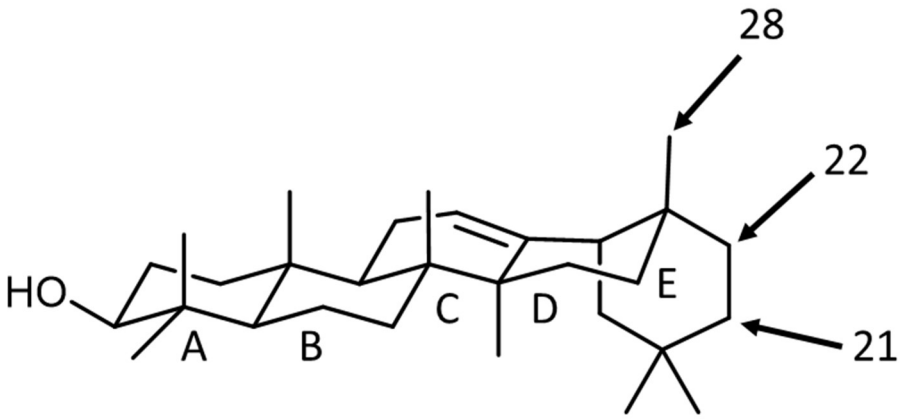


Figure 2.3.5 – The structure of beta-amyrin with the rings labelled. The E ring is out of plane with the rest of the molecule, which significantly reduces the steric hindrance at positions C-21 and C-22. C-28 also appears to be less influenced by neighbouring positions, with the other methyl groups being on the opposite side of the molecule (rings A and B).

This shows the E ring to be out of plane with the rest of the molecule, which might go some way to explaining the much greater number of acylated derivatives identified for positions on this side of the molecule. The pattern observed here is consistent with parallel work carried out by Michael Stephenson in the Osbourn group, which used similar database searching to investigate the distribution of oxidations (hydroxylation, epoxidation, or oxidation to a ketone, an aldehyde, or a carboxylic acid) around the beta-amyrin scaffold. That work found a similar cluster of positions between the E and D rings for which many more oxidised derivatives were identified – given that hydroxylation of a position is required for acylation to occur, this pattern of oxidation may well be the root cause of the observed pattern of acylation. The most commonly oxidised position found through those searches was C-28 rather than C-21, though this apparent discrepancy is readily rationalised by the nature of C-28 oxidation, which is most frequently found oxidised up to a carboxylic acid (and therefore not capable of undergoing acylation), and only much less frequently hydroxylated.

The two positions that returned the greatest number of acylated derivatives (C-21 and C-22) just so happen to be the two positions found acylated in the aescin aglycone. These were more commonly acylated than any other position by some margin. This could potentially make the acylation of aescin intermediates a good model for wider study of triterpenoid acylation.

2.4 – Frequency of acyl groups involved in beta-amyrin acylation

Further investigation into beta-amyrin acylation was carried out. Rather than the position on the beta-amyrin scaffold, this work focussed on the nature of the acyl group present in reported acylated derivatives. Query structures were built which allowed for acylation at any position but specified the nature of the acyl group – these were constructed for a range of acyl groups commonly encountered in specialised metabolism. These query structures were used to search Reaxys as previously, and the results are shown in figure 2.4.1.

Acyl group	Number of entries returned
Acetyl	965
Angeloyl / Tigloyl	535
Anthranilate (inc. derivatives)	21
Benzoyl	53
Caffeate	20
Cinnamoyl	55
p-Coumaroyl	24
Ferrulate	9
Formyl	1
Isobutyrate	14
Isovalerate	21
Malonyl	7
alpha-Methylbutyrate	91
Palmitoyl	4
Propionate	5

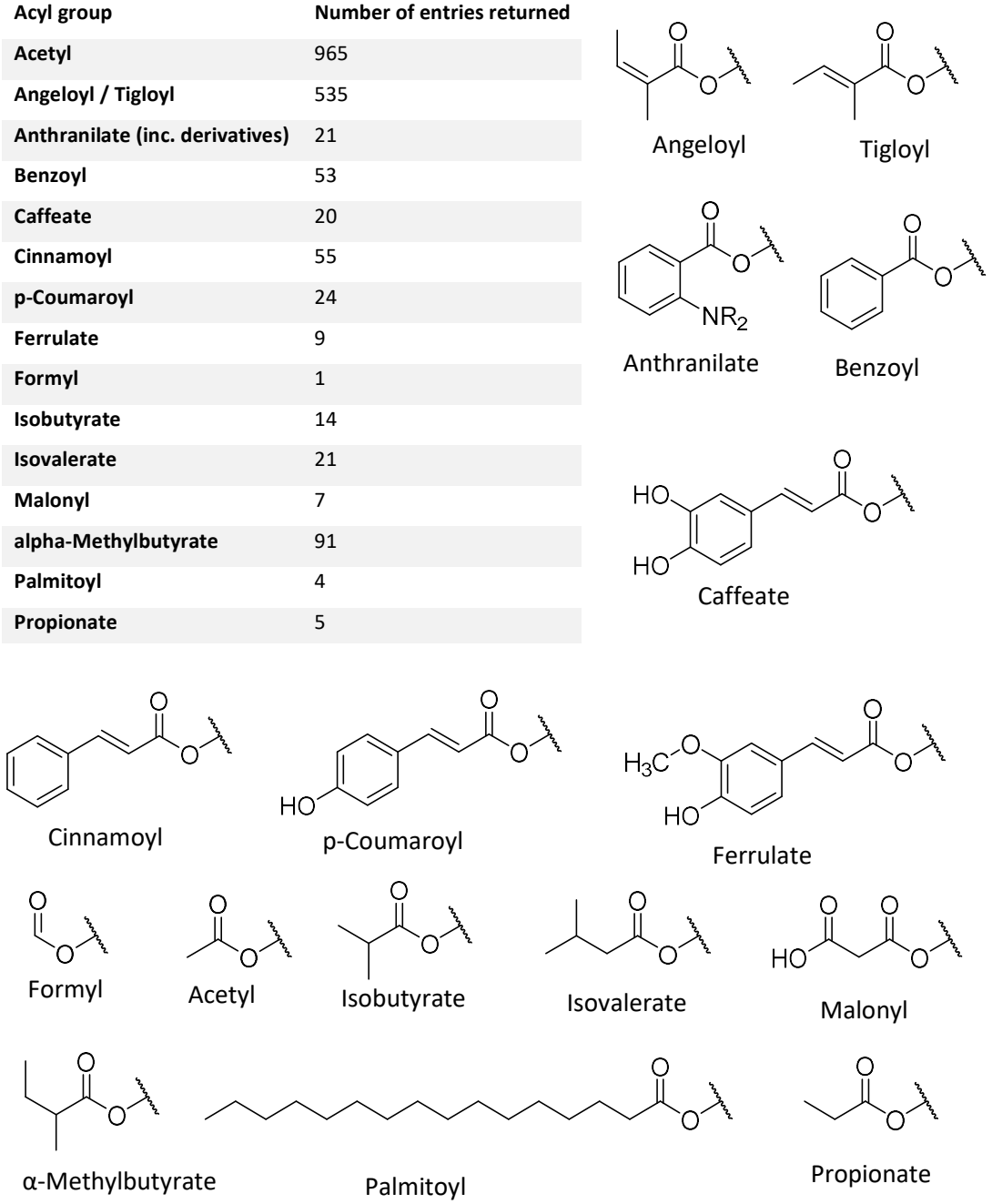


Figure 2.4.1 – The number of database entries returned for Reaxys searches for beta-amyrin derivatives acylated with particular acyl groups, along with the structures of the acyl groups searched for. The acyl groups angeloyl and tigloyl were combined in searches using a scaffold with undefined E/Z geometry. Searches for acetylated and angleoylated/tigloylated derivatives returned the most entries by some margin.

Of the searches carried out, two returned many more entries than the others – those for acetylated beta-amyrin derivatives and those for angeloylated or tigloylated beta-amyrin derivatives (these two were considered together, as many database entries did not properly annotate the geometry about the double bond that distinguish them). It was expected that more acetylated compounds should be reported than many other types of

acyl group – acetyl is one of the simplest acyl groups, and acetyl-CoA (the acyl donor employed by BAHD ACTs) is a primary metabolite ubiquitous to plant cells (Shi and Tu 2015). The number of beta-amyrin derivatives with angeloyl or tigloyl groups had not been expected, though this is particularly relevant to aescin biosynthesis as these are the acyl groups which are found at C-21β of the aescin aglycone. After these two acyl groups, α-methylbutyrate returned the most entries; this molecule can be thought of as tigloyl or angeloyl without the C=C double bond. The CoA esters of tigloyl and α-methylbutyrate are also products of primary metabolism and so are likely widespread and readily incorporated into specialised metabolic pathways (Robinson, Bachhawat et al. 1956).

These results are particularly interesting when compared with the results of beta-amyrin searches in 2.3 and the structure of the aescin aglycone – aescin is acylated at the two most commonly acylated positions (C-21 and C-22) with the most common acyl groups (22α-acetyl and either 21β-tigloyl (aescin a) or 21β-angeloyl (aescin b)). This indicates that the pattern of acylation observed in aescin is widespread across the reported acylated oleananes, which validates aescin as an excellent model for the study of triterpenoid acylation. Comparison of these results with published ACTs is harder – only a single ACT has been reported as active on beta-amyrin derivatives, namely SAD7, an SCPL from oats (Mugford, Qi et al. 2009, Mugford, Louveau et al. 2013). This is active on C-21, the most commonly acylated position, to which it transfers N-methylanthranilate, an anthranilate derivative (these returned 21 database entries, far fewer than acetyl or tigloyl). Prediction of which ACT family is most likely to be implicated in aescin biosynthesis will therefore warrant a wider assessment of published ACTs.

2.5 – Analysis of reported ACTs

The previous chapter introduced the two main classes of acyltransferases, the BAHDs and the SCPLs, and discussed their differing substrate requirements (BAHDs using acyl-CoA thioester donors and the SCPLs using acyl-sugars), along with their sub-cellular localisation and other relevant differences. When identifying potential candidates in the aescin pathway, it would be helpful to predict which of the two types of acyltransferases might be more likely to be involved – one way of investigating this might be to see whether an ACT family is more closely associated with beta-amyrin acylation, or of transfer of a particular acyl group found in aescin. Unfortunately, only a single ACT is characterised as active on the beta-amyrin scaffold. As such, a wider look at the broad class of metabolite the ACT families are active on could be informative. For this, the literature was searched

and a list of reported BAHDs and SCPLs was compiled, together with the substrate they were active on and the acyl group(s) they were found to transfer. The full lists can be found in appendix A1 (BAHD) and appendix A2 (SCPL).

A previous review in 2006 found 46 BAHDs, and a subsequent review in 2015 found 22 SCPLs and 69 BAHDs (D'Auria 2006, Bontpart, Cheynier et al. 2015). Here, 105 BAHDs and 23 SCPLs were itemised (see appendices for details). A summary of the number of ACTs by family reported as using selected classes of metabolites as substrates is given in table 2.5.1.

SUBSTRATE	NO. ACTS	NO. BAHDS	NO. SCPLS
ALKALOIDS	4	4	0
ANTHOCYANINS	17	11	6
ALIPHATICS	5	4	1
SIMPLE AMINES	16	16	0
FLAVONOIDS	18	13	5
MONOTERPENES	4	4	0
TRITERPENOIDS	8	7	1
SHIKIMATES	18	17	1
COMPLEX POLYMERS	7	6	1
CHLOROGENIC ACID	4	4	0
CHOLINE	3	0	3

Table 2.5.1 – Selected classes of plant metabolites, with the numbers of ACTs, BAHDs, and SCPLs reported as using these as substrates. In total, eight ACTs were found to be active on triterpenoids, of which seven were BAHDs and one SCPL.

Though the only ACT active on beta-amyrin derived aglycones (SAD7) is an SCPL, seven other ACTs were found to be active on triterpenoids, all of which were BAHDs. One of these, SOAP10 from spinach, is also active on beta-amyrin derived saponins, though it acts to acetylate the sugar chain rather than the aglycone, so is likely to demonstrate markedly different substrate specificity to other triterpenoid ACTs (Jozwiak, Sonawane et al. 2020). The other seven triterpenoid ACTs act on aglycones – a summary of the activity of these is provided in figure 2.5.1.

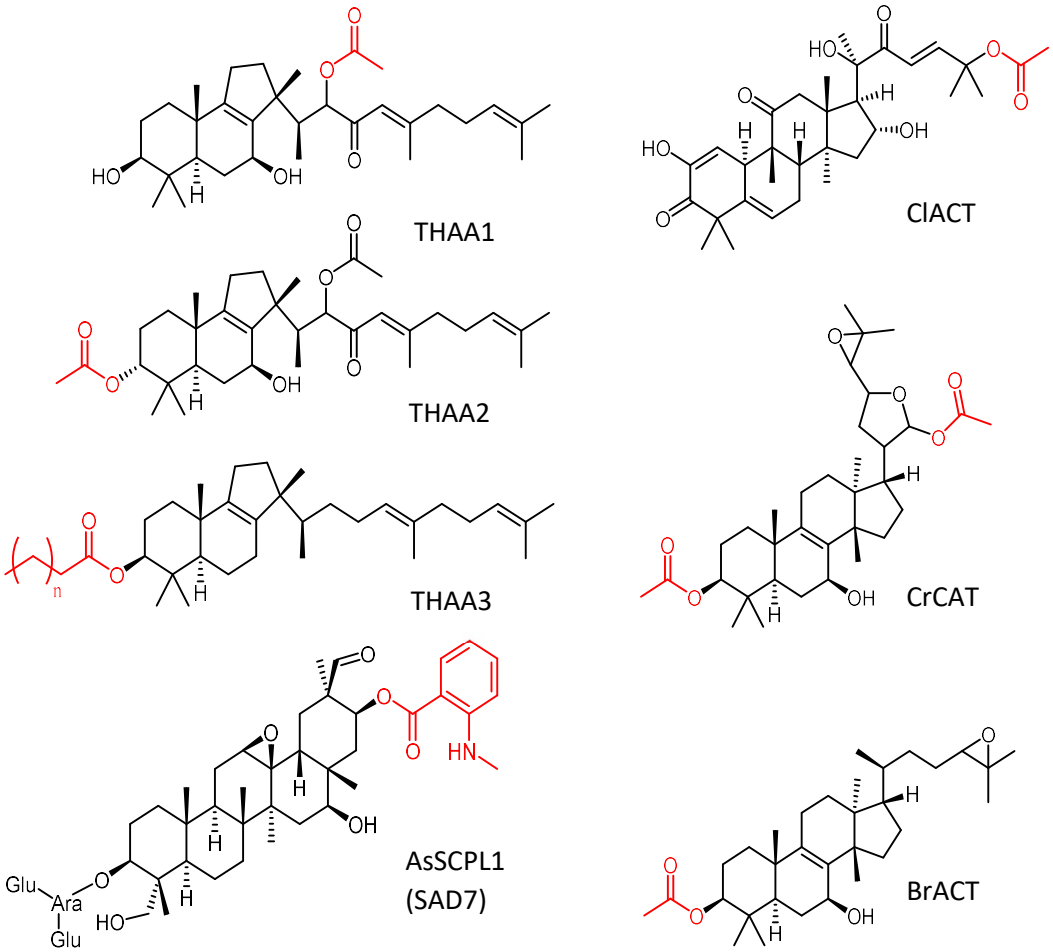


Figure 2.5.1 – The reported products formed by the triterpenoid ACTs active on aglycones found in this work, with the acyl group added marked in red. THAA1, THAA2, and THAA3 from A. thaliana *(Huang, Jiang et al. 2019), ClACT from* Citrullus lanatus *(Zhou, Ma et al. 2016), CrACT from* C. rubella *(Liu, Suarez Duran et al. 2020), BrACT from* B. rapa *(Liu, Suarez Duran et al. 2020), & SAD7 from* Avena strigosa *(Mugford, Qi et al. 2009).*

Further analysis of the listed ACTs was undertaken by looking at the acyl groups they transferred. Some of the ACTs, especially the BAHDs, were found to display a degree of promiscuity in this regard, capable of the transfer of different acyl groups, which is consistent with previous reports (D'Auria 2006, Kruse, Weigle et al. 2020). A table summarising the acyl groups transferred by BAHDs and SCPLs is shown in table 2.5.2.

ACYL GROUP	ENTRIES (2.4)	ACTS	BAHDS	SCPLS	TRITERP ACTS
ACETYL	965	22	22	0	6
ANGELOYL/TIGLOYL	535	1	1	0	-
ANTHRANILATE	21	2	1	1	1
BENZOYL	53	15	11	4	1
CAFFEATE	20	23	23	0	-
CINNAMOYL	55	4	4	0	-
COUMAROYL	24	50	47	3	-
FERRULATE	9	11	11	0	-
FORMYL	1	0	0	0	-
ISO-BUTYRATE	14	1	0	1	-
ISO-VALERATE	21	0	0	0	-
MALONYL	7	21	20	1	-
A-METHYLBUTYRYL	91	0	0	0	-
PALMITOYL	4	1	1	0	1
PROPIONYL	5	0	0	0	-
SINAPOYL	-	9	2	7	-
GALLOYL	-	6	0	6	-

Table 2.5.2 – The number of ACTs, BAHDs, SCPLs, and ACTs active on triterpenoids found transferring each acyl group, along with the number of database entries returned for beta-amyrin for that group (see 2.4 – sinapoylated and galloylated beta-amyrin derivatives were not searched for). The SCPLs are mostly involved in transfer of aromatic acyl groups, whereas the BAHDs appear to be active on a wider range of acyl groups.

The results of this show that the SCPLs appear to be mostly involved in the transfer of aromatic acyl groups, which has previously been noted (Bontpart, Cheynier et al. 2015). The BAHDs, by comparison, are implicated in acylation with a wider range of acyl groups, both aromatic and non-aromatic. When it comes to acetyl and angeloyl/tigloyl, the two acyl groups found in aescin, all characterised acetyltransferases and tigloyltransferases are BAHDs, which is particularly important when considering candidates for aescin biosynthesis. Though it is not possible to prioritise one ACT family over the other based on their involvement in triterpenoid biosynthesis, it seems reasonable to do so based on their acyl group preferences. Acetylation appears to be carried out exclusively by BAHDs, and although only one tigloyltransferase has been found, it is expected that BAHDs are also favoured here – as SCPLs are so greatly implicated in acylation with aromatic groups. As such, it is proposed to focus on BAHD candidates for acylation of aescin intermediates.

2.6 – Predicting a biosynthetic pathway to aescin

The chemical database Reaxys was used to predict a putative biosynthetic pathway to aescin. Searches were restricted to compounds isolated from "*Aesculus*", "*Aesculus hippocastanum*", and "horse chestnut". Over 100 triterpenoid structures were found. Query structures were built from the basic beta-amyrin scaffold as in 2.3, but with substitution at certain positions prevented. These were used to investigate the reported isolation of putative aescin intermediates. First, investigation of hydroxylation was carried out. Query structures were built with substitution prevented at each of the five positions hydroxylated in aescin. These were used to search Reaxys, and the results shown in figure 2.6.1.

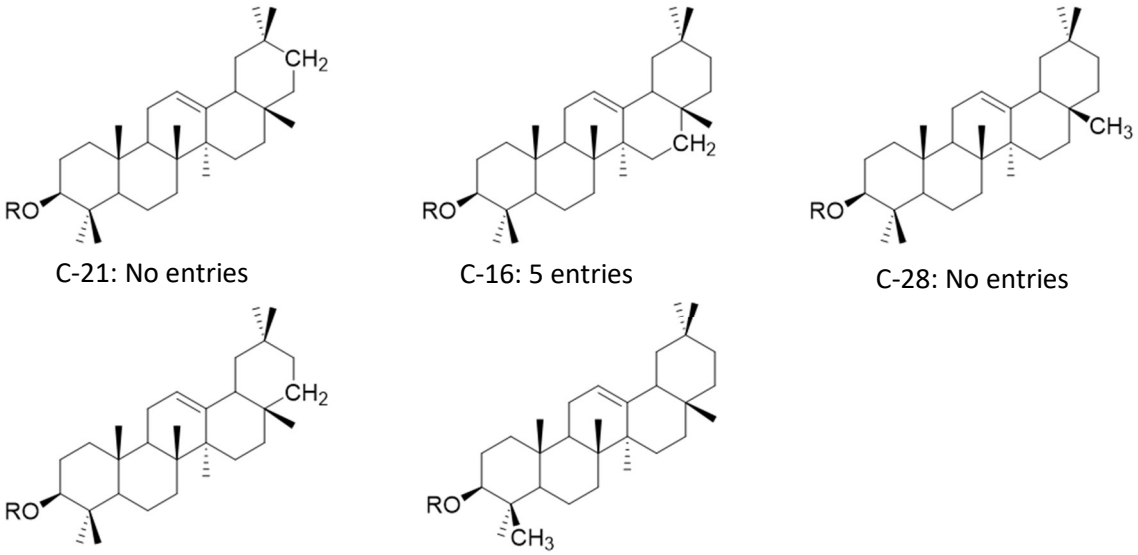
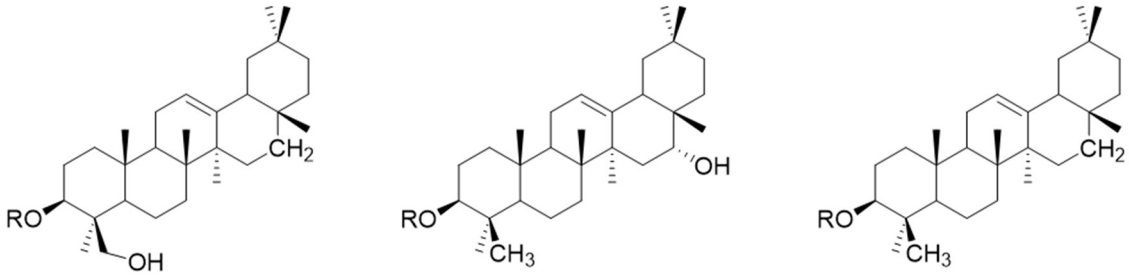


Figure 2.6.1 – Results of Reaxys searches for compounds isolated from Aesculus *lacking hydroxylation at one of the five positions hydroxylated in aescin. Two positions, C-16 and C-24, returned entries. No compounds were reported lacking hydroxylation at C-21, C-22, or C-28.*

The results of these searches found that no compounds had been isolated lacking hydroxylation at C-21, C-22, or C-28. However, compounds were reported lacking C-16 or C-24 hydroxylation. This was investigated further, using query scaffolds combining the two positions. The results of this are shown in figure 2.6.2.



24-OH but no C-16: No entries

16α-OH but no C-24: 53 entries

No C-16 or C-24: 3 entries

Figure 2.6.2 – Results of Reaxys searches for compounds isolated from Aesculus *with missing hydroxylation from C-16 or C-24. No compounds with 24-OH but lacking C-16 hydroxylation were found. Of the 56 compounds lacking C-24 hydroxylation (2.6.1) all but three had C-16α hydroxylation.*

The results of these searches demonstrate that no compounds lacking C-16 hydroxylation but with C-24 hydroxylation were found. However, the inverse (lacking C-24 hydroxylation but with 16α-OH) returned 53 database entries. This would appear to indicate that C-24 hydroxylation might be conditional upon C-16 hydroxylation. That no compounds were found lacking C-21, C-22, or C-28 hydroxylation (2.6.1) might suggest that these hydroxylations are very early steps in the pathway, and means it will not be possible to predict the order in which these hydroxylations might occur.

Similar work was carried out to attempt to determine the order in which acylation and glycosylation of the scaffold might occur. These searches were inconclusive – it was not possible to determine whether acylation is dependent on glycosylation, or vice versa. Both acylated structures lacking glycosylation, and glycosylated structures lacking acylation were found. However, it must be noted that the lability of both sugar chains and acyl groups means that artefact formation during extraction is possible, and so database entries may be des-acyl or deglycosyl artefacts rather than true intermediates. It was, however, possible to conclude that both acylation and glycosylation are not dependent on C-16α or C-24 hydroxylation, as both acylated and glycosylated beta-amyrin derivatives have been reported from *Aesculus* lacking these hydroxylations (Yuan, Wang et al. 2013). This is significant, as it implies that these two hydroxylations may not be required for testing of ACT and glycosyltransferase candidates. By contrast, the absence of reported acylated or glycosylated compounds lacking C-21β, C-22α, or C-28 hydroxylation means that these transformations will be required to test downstream candidates.

Using these results, it is possible to predict parts of the biosynthetic pathway to aescin, as shown in figure 2.6.3.

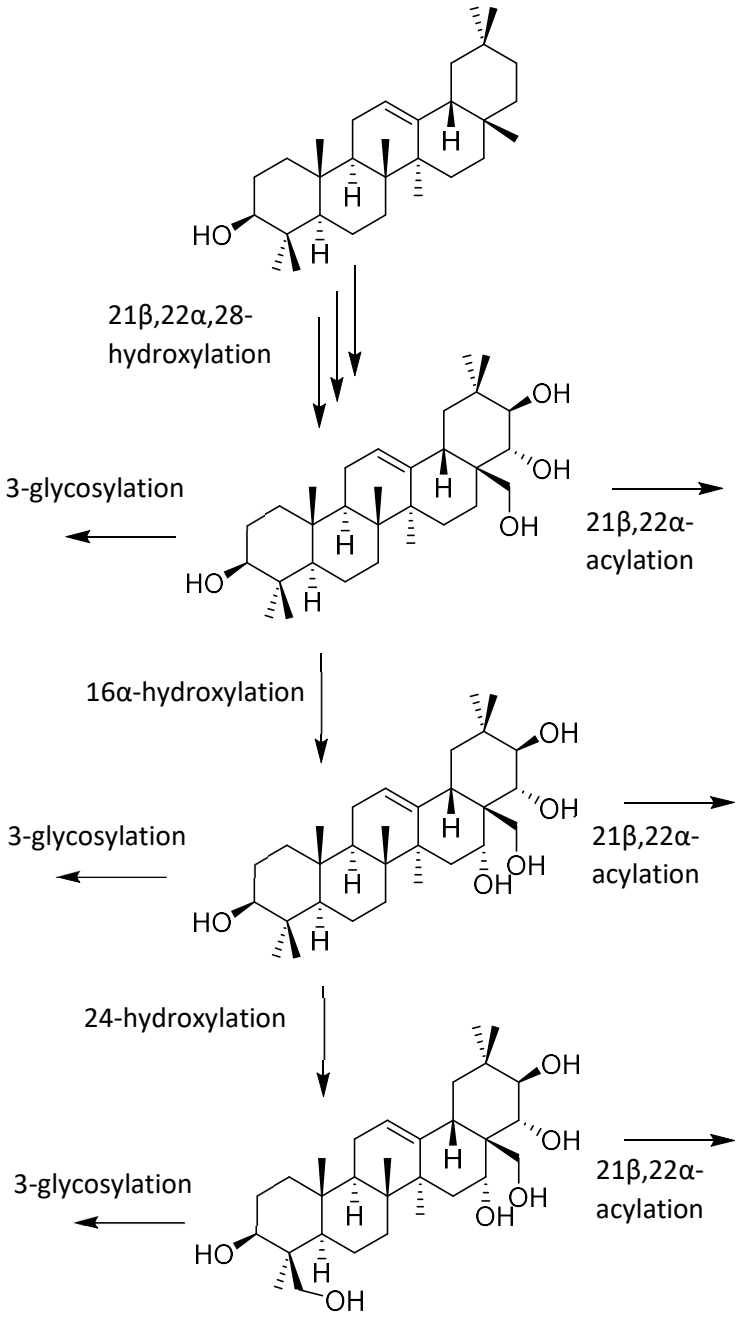


Figure 2.6.3 – A putative biosynthetic pathway based on the results of Reaxys searches for aescin intermediates. It was not possible to separate the order of 21β, 22α, or 28 hydroxylations, but it seems likely that these steps are required for 16α hydroxylation, which is in turn likely required for 24 hydroxylation. Acylation and glycosylation do not appear to be dependent on 16α- or 24-OH, nor are they dependent upon one another. The 21β,22α,28-OH beta-amyrin scaffold is likely to be important in planning future experiments.

2.7 – Conclusions

The database mining carried out in this chapter identified an apparent pattern to the acylation of alpha- and beta-amyrin derived triterpenoids. Acylated beta-amyrin

derivatives reported in the Reaxys database were most commonly acylated at C-21 and C-22. Further database mining found that these triterpenoids were most commonly acylated with acetyl, angeloyl, or tigloyl groups. These findings are significant to work on aescin biosynthesis, as aescin is acylated with these three groups at C-21β and C-22α. This makes aescin a good model for study of triterpenoid acylation.

Analysis of the reported ACTs by their substrate specificity found that very few had been reported as active on triterpenoids, making prioritisation of one ACT family over the other on this basis unwise. However, analysis on the basis of acyl donor specificity found a clear difference between BAHDs and SCPLs – SCPLs were primarily implicated in transfer of aromatic acyl groups, whereas BAHDs were involved in a wider range of acylations. Furthermore, the acyl groups found in aescin have only been reported to be transferred by BAHDs. This justifies a focus on BAHD acyltransferases over SCPLs when selecting candidate biosynthetic enzymes, as carried out in chapters 4 and 7.

The Reaxys database was also used to investigate reported aescin intermediates isolated from *Aesculus* species. This work found that 21β-, 22α-, and 28- hydroxylation of the beta-amyrin scaffold was likely to be the first steps of the pathway, followed by hydroxylation at C-16α, then C-24. The order in which acylation and glycosylation occur could not be determined – it may be the case that the enzymes responsible for these steps are promiscuous in their substrate specificity, resulting in a non-linear biosynthetic pathway. However, no acylated or glycosylated compounds could be detected with a hydroxylated beta-amyrin scaffold lacking C-21β, C-22α, or C-28 hydroxylations. This means that this scaffold (21β,22α,28-OH beta-amyrin, see figure 2.7.1) is to be treated as the minimum scaffold for testing of acyltransferases and glycosyltransferases in subsequent work – testing of these enzymes on a less complex scaffold lacking one of these hydroxylations will not enable inactivity to discount this enzyme. The identification of a minimum scaffold for testing of these enzymes is an important milestone to this project, which will inform the work carried out in subsequent chapters.

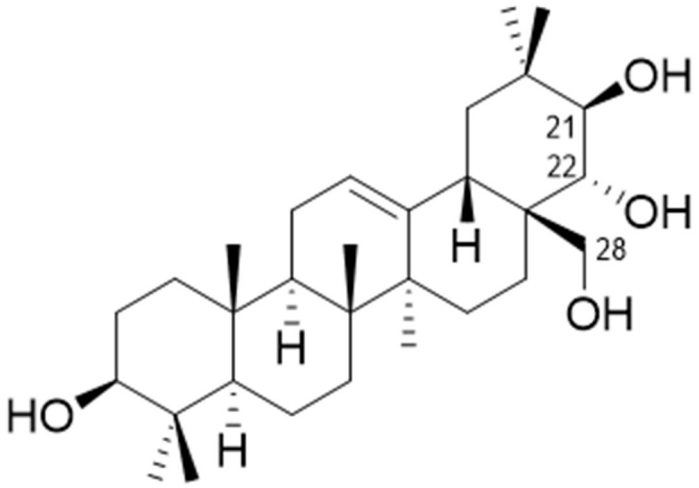


Figure 2.7.1 – The structure of 21β,22α,28-OH beta-amyrin, which is identified in this work as the minimum required scaffold for testing of acyltransferases and glycosyltransferases. Without testing on a scaffold with all three of these hydroxylations, it will not be possible to rule enzymes out of the biosynthetic pathway, as it is not known if this scaffold is required for these enzymes to be active. Work in subsequent chapters ought therefore to focus on production of this scaffold in N. benthamiana.

Chapter 3 – Evaluating the Triterpenoid Toolkit for Reconstruction of the Aescin Pathway

Synopsis

A large number of characterised triterpenoid biosynthetic enzymes have been reported in the literature – one database (TriForC (Miettinen, Iñigo et al. 2018)) lists 125 OSCs alone as of 2022, with many more CYPs, ACTs, UGTs and others. The Osbourn group has established a collection of genes encoding triterpenoid biosynthetic enzymes: the triterpenoid toolkit. This collection offers a potential means of identifying enzymes from other plant species that may enable triterpenoid modifications relevant to the study of aescin, so allowing the testing of *Aesculus* candidate enzymes against a wider range of relevant scaffolds. Co-expression of a beta-amyrin synthase from the avenacin pathway in oat with selected characterised CYPs from the toolkit enabled some of the transformations relevant to the biosynthesis of aescin to be recapitulated, specifically hydroxylation at C-16α, C-21β, and C-24, along with formation of oleanolic acid. Furthermore, combinatorial expression of these CYPs enabled formation of hydroxylated oleanolic acid derivatives. CYPs that carry out other relevant modifications (C-22α and C-28 hydroxylases) were not represented in the toolkit and will need to be accessed from *Aesculus*.

Introduction

3.1 – Introduction to the Triterpenoid Toolkit

The previous chapter introduced the likely minimum scaffold needed for evaluation of candidate downstream acyltransferases and glycosyltransferases implicated in aescin biosynthesis, and set out a putative biosynthetic pathway hypothesising the order of hydroxylation events. It is conceivable that some, or even all, of the aescin pathway enzymes may exhibit substrate specificity, such that the enzymes that carry out particular steps require appropriate modifications to the scaffold before they can act. A strict order of transformations has been reported for other triterpenoid biosynthetic pathways, for instance those of the mogrosides (Itkin, Davidovich-Rikanati et al. 2016) and avenacins (Leveau, Reed et al. 2019). This poses a challenge for the identification and testing of candidate enzymes from *Aesculus*, since enzymes may not show activity unless tested on the correct biosynthetic pathway intermediate that serves as their natural substrate. However, this hurdle can in part be overcome by reconstituting as many steps as possible using enzymes from other species.

Pathway reconstruction through heterologous expression in an appropriate host offers a strategy to accomplish this. Though production of terpenoids and other secondary metabolites is possible through alternative methods, including chemical synthesis or plant cell culture (Corey and Lee 1993, Matsuura, Malik et al. 2018), heterologous expression confers many advantages (Zhao and Li 2018, Huang and Osbourn 2019). The ability to readily combine enzymes from multiple species (allowing for combinatorial biosynthesis), and the speed with which new constructs can be made and tested, provides a powerful approach for investigation of biosynthetic enzymes and pathways (Beyraghdar Kashkooli, Van Der Krol et al. 2019). The heterologous hosts that are most commonly used for triterpene metabolic pathway engineering are yeast (Moses, Pollier et al. 2014, Ikram, Zhan et al. 2015, Tatsis and O'Connor 2016) and *Nicotiana benthamiana* (Reed and Osbourn 2018, Calgaro-Kozina, Vuu et al. 2020, Eljounaidi and Lichman 2020). *N. benthamiana* is the host of choice for the experiments carried out here, since many of the common precursors and co-enzymes required by the metabolic pathways are already present, and expression of plant enzymes in plant cells better replicates the conditions in which the enzymes are naturally active, eg. sub-cellular localisation of the enzymes and their substrates (Owen, Patron et al. 2017). The transient plant expression system is well established for triterpene metabolic engineering, and readily allows both combinatorial biosynthesis for enzyme characterisation and preparative-scale production for compound purification and structural elucidation (Reed, Stephenson et al. 2017, Stephenson, Reed et al. 2018).

A collection of strains of *A. tumefaciens* harbouring expression constructs for enzymes from a range of triterpenoid biosynthetic pathways (the Triterpenoid Toolkit) has been established in the Osbourn laboratory. Using these strains, it is possible to express a variety of biosynthetic enzymes in combination and produce a range of triterpenoids, including novel ones. The toolkit contains enzymes that carry out similar transformations to those seen in aescin biosynthesis. It therefore offers the potential to make some of the intermediates predicted to act as potential substrates for evaluation of biosynthetic gene candidates from *Aesculus*.

A key component of the enzyme toolkit is a truncated form of HMG-CoA Reductase (tHMGR). HMGR is required for the biosynthesis of triterpenoid precursors via the MVA pathway. tHMGR is a feedback-insensitive variant of HMGR that that lacks the portion of the enzyme responsible for self regulation, but retains catalytic activity (Polakowski, Stahl et al. 1998). Co-expression of tHMGR along with other triterpene biosynthetic enzymes boosts precursor supply, and hence results in significantly increased triterpene yields.

The scaffold for aescin biosynthesis is beta-amyrin, a pentacyclic triterpenoid shown in figure 3.1.1. Triterpenoids are made through cyclisation of the linear precursor, 2,3-oxidosqualene by enzymes known as oxidosqualene cyclases (OSCs), of which many have previously been characterised (Xue, Duan et al. 2012, Thimmappa, Geisler et al. 2014, Xue, Tan et al. 2018) – the TriForC database (Miettinen, Iñigo et al. 2018) lists 125 of these. The triterpenoid toolkit established in the Osbourn group contains 24 OSC genes for the biosynthesis of diverse triterpenoid scaffolds, including 5 beta-amyrin synthases. One of these, SAD1, is an OSC from oat (*Avena strigosa*) that catalyses the first committed step in the biosynthetic pathway to the avenacins (Haralampidis, Bryan et al. 2001). SAD1 has previously been shown to generate high levels of beta-amyrin when transiently expressed in *N. benthamiana* (Reed, Stephenson et al. 2017). In the absence of an *Aesculus* beta-amyrin synthase this therefore represents a good starting point for investigation of the biosynthesis of aescin. In this chapter SAD1 was used in combinatorial expression experiments with other toolkit enzymes to generate predicted intermediates of the aescin pathway.

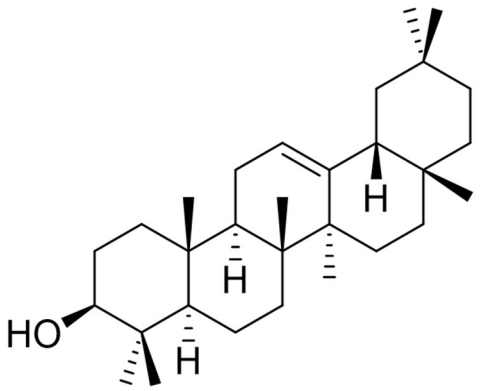


Figure 3.1.1 – The structure of beta-amyrin, the starting point of aescin biosynthesis. A pentacyclic triterpenoid, it is formed by cyclisation of 2,3-oxidosqualene by OSC enzymes known as beta-amyrin synthases.

The putative biosynthetic pathway proposed in the last chapter identified hydroxylation of beta-amyrin as the next steps in aescin biosynthesis, specifically at positions C-21β, -22α, and -28. Hydroxylation of triterpenoids and other plant secondary metabolites is most commonly carried out by Cytochrome P450s (CYPs) (Morant, Bak et al. 2003, Hamberger and Bak 2013, Seki, Tamura et al. 2015, Xu, Wang et al. 2015). This is a large family of genes – for instance, 244 CYPs have been found in the *Arabidopsis* genome alone (Bak, Beisson et al. 2011, Nelson and Werck-Reichhart 2011). Approximately 100 CYPs active on triterpenoids have been characterised - the TriForC database (Miettinen,

Iñigo et al. 2018) lists 90 - with those from the subfamilies 51, 71, 72, and especially 716 most commonly implicated (Ghosh 2017, Miettinen, Pollier et al. 2017). These have been identified from a diverse variety of plant species, and can readily be combined in co-expression experiments, along with a triterpene scaffold-generating OSC, using heterologous expression in *N. benthamiana* as discussed earlier (Reed, Stephenson et al. 2017, Beyraghdar Kashkooli, Van Der Krol et al. 2019). A list of those CYPS represented in the Osbourn lab toolkit that are active on beta-amyrin and derivatives is provided in table 3.1.1.

NAME	POSITION OXIDISED	OXIDATION	SPECIES
ASCYP51H10	12,13β & 16β	Epoxy & Hydroxyl	*Avena strigosa*
ASCYP51H10 A354L	16β	Hydroxyl	*Avena strigosa*
ASCYP51H10 I471M	12,13β	Epoxy	*Avena strigosa*
MTCYP716A12	28	Carboxylic acid	*Medicago trunculata*
PGCYP716A141	16β	Hydroxyl	*Platycodon grandiflorus*
AACYP716A14V2	3	Ketone	*Artemesia annua*
CACYP716C11	2α	Hydroxyl	*Centella asiatica*
SLCYP716E26	6β	Hydroxyl	*Solanum lycopersicon*
PGCYP716S5	12,13α	Epoxy	*Platycodon grandiflorus*
BFCYP716Y1	16α	Hydroxyl	*Bupleurum falcatum*
GUCYP72A154	30	Hydroxyl	*Glycyrrhiza uralensis*
QSCYP716-2073932	28	Carboxylic acid	*Quillaja saponaria*
QSCYP716-2012090(QS16A)	16α	Hydroxyl	*Quillaja saponaria*
ASCYP72A475	21β	Hydroxyl	*Avena strigosa*
ASCYP72A476	30	Hydroxyl	*Avena strigosa*
MTCYP72A61V2	24	Hydroxyl	*Medicago trunculata*
MTCYP72A63	30	Hydroxyl	*Medicago trunculata*
MTCYP72A67V2	2β	Hydroxyl	*Medicago trunculata*
MTCYP72A68V2	23	Carboxylic acid	*Medicago trunculata*
GMCYP72A69	21	Hydroxyl	*Glycine max*
CACYP714E19	23	Hydroxyl	*Centella asiatica*
GUCYP88D6	11	Ketone	*Glycyrrhiza uralensis*
GMCYP93E1	24	Hydroxyl	*Glycine max*
ASCYP94D65	23	Hydroxyl	*Avena strigosa*

Table 3.1.1 –CYPs from the Triterpenoid Toolkit that are active upon beta-amyrin or its derivatives, showing the position acted upon, the nature of the oxidation, and the species which the sequence was cloned from. Most of these CYPs are responsible for the addition of a hydroxy group, though more complex oxidations are known. Multiple activities are also reported, for instance the activity of AsCYP51H10.

As can be seen from Table 3.1.1, CYPs that modify a variety of different positions on the beta-amyrin scaffold have been identified. Comparison of the activity of these CYPs with the hydroxylations required for synthesis of aescin suggests that some of these modifications may be possible using the toolkit. The structure of protoaescigenin, the aglycone core of aescin without any acyl groups, in shown in Fig. 3.1.2. Protoaescigenin has hydroxyl groups at positions C-16α, -21β, -22α, -24, and -28. Enzymes that hydroxylate the C-16α, -21β, and -24 of beta-amyrin are represented in the toolkit. A CYP that oxidises the C-28 position is also available, although this forms a carboxylic acid group rather than an alcohol. These CYPs originate from different triterpene biosynthetic pathways and so may not all work in combination with one another, or may indeed require other modifications to their substrates that are not present in aescin. Although some CYPs display a degree of promiscuity in their substrate requirements, for instance CYP 93E1 (Shibuya, Hoshino et al. 2006), CYP 71E1 (Kahn, Fahrendorf et al. 1999), or CrAO (Huang, Li et al. 2012), it will be necessary to empirically test the ability of the toolkit CYPs to carry out the relevant modifications to the beta-amyrin scaffold by combinatorial biosynthesis.

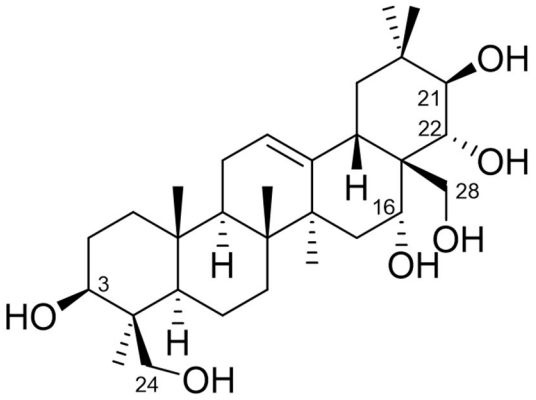


Figure 3.1.2 – The structure of protoaescigenin, the unacylated aglycone of aescin, with the hydroxyl groups numbered. The hydroxyl group at C-3 is present in beta-amyrin, but the other five must be added by CYPs. The presence of a hydroxyl group at C-28 (erythrodiol-type) is slightly more unusual for triterpenoids of this type – a carboxylic acid (oleanolic acid-type) is more common.

The characterised activities of the relevant CYPs from toolkit are shown in fig. 3.1.3. CYP72A69 is a C-21β hydroxylase from the soyasapogenol pathway in soybean (*Glycine max*) (Yano, Takagi et al. 2017), where it forms soyasapogenol A (21β,22β,24-OH beta-amyrin). Although its substrate does not quite correspond to an aescin precursor (the C-22 hydroxylation is 22β rather than 22α), this enzyme is reported to be somewhat promiscuous in the substrates that is is able to use. CYP93E1 is also from the soyasapogenol pathway. This CYP adds an alcohol group to the C-24 position of either beta-amyrin or sophoradiol (22β-OH beta-amyrin) (Shibuya, Hoshino et al. 2006). Whilst this position was not identified as a priority from the putative biosynthetic pathway of the previous chapter, it would still be a relevant transformation to include as a scaffold for testing of candidates. CYP716A12 is a C-28 oxidase from *Medicago truncatula*, a species of

lucerne (Carelli, Biazzi et al. 2011, Fukushima, Seki et al. 2011). This enzyme is reported to be active on a range of closely related triterpenoid scaffolds, where it adds a carboxylic acid group. Though this type of modification is not present in aescin, which has a C-28 hydroxyl group instead, CYP716A12 could still be valuable in testing *Aesculus* candidates; C-28 acid intermediates have been reported from *Aesculus*, and the enzymes carrying out other transformations in the aescin pathway may accept a 28-acid as a valid substrate in place of 28-OH. Finally, CYP 716-2012090 (Qs16a) is a C-16α hydroxylase from *Quillaja saponaria*, which has been identified as active on both beta-amyrin and oleanolic acid (Osbourn and Reed 2019).

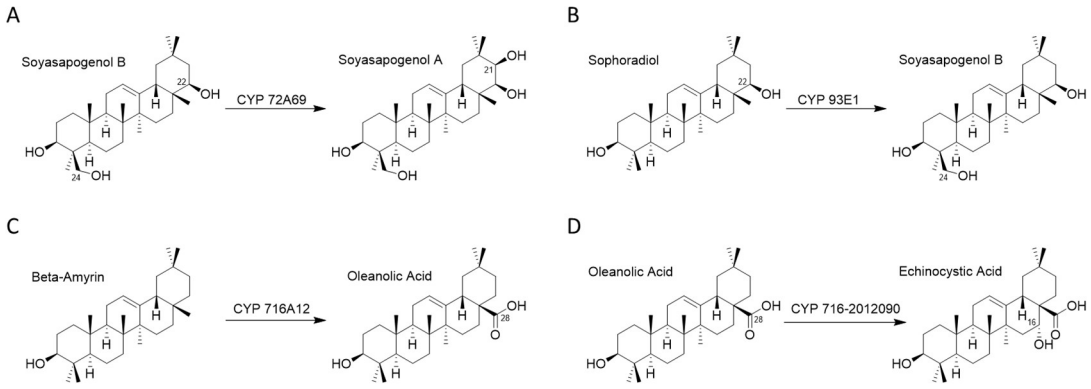


Figure 3.1.3 – The reported activities of toolkit CYPs relevant to the aescin biosynthetic pathway. A: CYP 72A69 – 21β hydroxylation of soyasapogenol B. B: CYP 93E1 – 24 hydroxylation of sophoradiol. C: CYP 716A12 – oxidation of C-28 to form oleanolic acid from beta-amyrin. D: CYP 716-2012090 – 16α hydroxylation of oleanolic acid.

A C-28 hydroxylase is not represented in the toolkit. This makes testing of *Aesculus* candidate enzymes more challenging, as this is likely to be a key step in the biosynthetic pathway and could be a determinant in the substrate specificity of downstream biosynthetic enzymes. Two CYPs from tomato (CYP716A44 and CYP716A46) have been reported to synthesise erythrodiol (28-OH beta-amryin, see fig. 3.1.4) from beta-aymrin when expressed in yeast (Yasumoto, Seki et al. 2017). Both of these CYPs are reported to achieve efficient conversion of beta-amyrin into erythrodiol, with additional formation of oleanolic acid as a secondary product. If active in combination with the toolkit CYPs, these enzymes could be employed to form the erythrodiol derivatives. As these enzymes were tested in yeast, it will first be necessary to validate their activity in *N. benthamiana*.

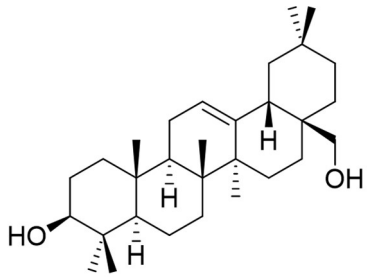


Figure 3.1.4 – The structure of erythrodiol (28-OH beta-amyrin). Erythrodiol is of some interest in pharmaceutical studies into triterpenoids. Though a route to chemical synthesis (Corey and Lee 1993) has been published, it is yet to be synthesised through heterologous plant expression.

Methods

3.2 – Methods used in this chapter

Toolkit enzymes were obtained as glycerol stocks of *Agrobacterium tumefaciens* LBA4404 strains transformed with the relevant constructs, maintained at -70 ^{o}C. These were infiltrated into leaves of *N. benthamiana* as described in section 9.2.1. A strain expressing GFP was used as a control, and to maintain equal OD_{600} between combinations with fewer strains (see 9.2.1).

Tomato root tissue (cultivar: Microtom) stored at -70 ^{o}C was obtained from Hugo Ombredane (formerly of the Osbourn Lab, JIC), from which RNA was extracted according to 9.1.1. From this RNA cDNA was synthesised as described in 9.1.2. Primers were designed for CYP 716A44 and CYP 716A46 (listed in 9.1.3), and these were used for cloning PCR according to 9.1.3 using iProof (Bio-Rad). CYP 716A44 was cloned into expression vector pEAQ-HT-DEST1 (Sainsbury, Thuenemann et al. 2009) per 9.1.4 and subsequently transformed into LBA4404 as set out in 9.1.7. The resultant transformed *Agrobacterium tumefaciens* strain was incorporated into transient expression experiments as with the toolkit enzymes above.

Harvested *N. benthamiana* leaves from co-expression experiments were extracted and derivatised as set out in 9.3.1, with the exception that 50 ppm internal standard (coprostanol) was employed instead of 100 ppm. The resultant samples were then analysed through GC-MS according to the protocol described in 9.3.4. Commercial standards of oleanolic acid and erythrodiol were obtained from Extrasynthese.

3.3 – Validation of the Biosynthesis of beta-Amyrin in *N. benthamiana*

Evaluation of the expression of the toolkit beta-amyrin synthase SAD1 in *N. benthamiana* was carried out both in the absence and in the presence of tHMGR. As can be seen from the GC traces in figure 3.3.1, expression of tHMGR leads to accumulation of the triterpenoid precursor molecule oxidosqualene in the leaf tissue, which, on co-expression with SAD1, is almost entirely consumed to form beta-amyrin. Expression of SAD1 without tHMGR results in a much reduced accumulation of beta-amyrin. This is consistent with the results of previous experiments, which demonstrated a four-fold increase in the production of beta-amyrin by SAD1 through co-expression with tHMGR (Reed and Osbourn 2018). The yields achieved through combination of these enzymes are high, which will allow use of SAD1 as the OSC in combinatorial trials involving downstream enzymes. Reported yields in the literature are also high, with the combination of SAD1 and tHMGR having previously been employed to obtain gram-scale quantities of beta-amyrin through transient plant expression (Reed, Stephenson et al. 2017). The mass spectrum in figure 3.3.1 is highly characteristic of beta-amyrin, with two peaks corresponding to fragments from opposite sides of the molecule. The GC traces and mass spectrum are those of derivatised leaf extracts – reaction of the extracts with TMS-imidazole adds the TMS group to any free -OH groups on the molecules. This changes the retention times and mass spectra of the products, and increases the visibility of some compounds in GC. Care must be taken when comparing traces or spectra that the same derivatising agent has been employed.

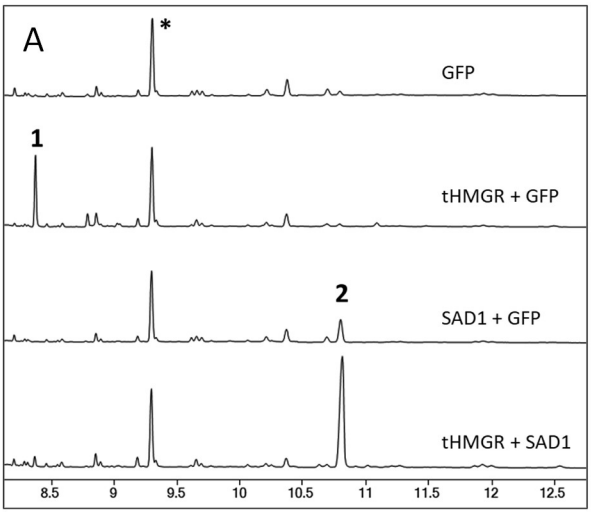
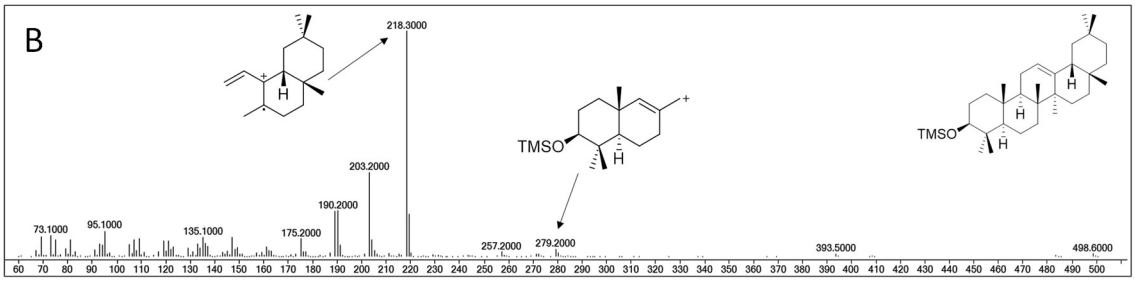


Figure 3.3.1 – The amount of beta-amyrin produced by SAD1 is increased when co-expressed with tHMGR.

*A: GC traces (TMS-derivatised) showing the activity of tHMGR and SAD1. * Internal standard (coprostanol) visible at 9.3 mins. Peak (1): oxidosqualene. Peak (2): beta-amyrin. Accumulation of beta-amyrin is greatly increased when tHMGR is used.*

B: Mass spectrum of peak (2). The peaks at m/z 218 and 279 are characteristic of (derivatised) beta-amyrin, and correspond to the CDE and the ABC* ring fragments respectively.*



The fragmentation pattern seen in the mass spectrum of beta-amyrin is the result of a ring-opening reaction in the centre of the molecule. Ring C, with its double bond, can undergo a retro-Diels-Alder reaction when ionised in the mass spectrometer, resulting in the formation of two fragments: the ABC* and the CDE* ring fragments (figure 3.3.2). For any one ionised molecule, only one fragment will retain the charge which enables detection through MS. Though both fragments are seen, the CDE* ring fragment is detected with greater intensity. Fragmentation of this nature is very useful when analysing the mass spectra of beta-amyrin derivatives, as the absence of one of these peaks could suggest modification on that side of the molecule (Geisler, Hughes et al. 2013, Moses, Pollier et al. 2015). Often new peaks are observed, corresponding to those ring fragments with the modifications attached.

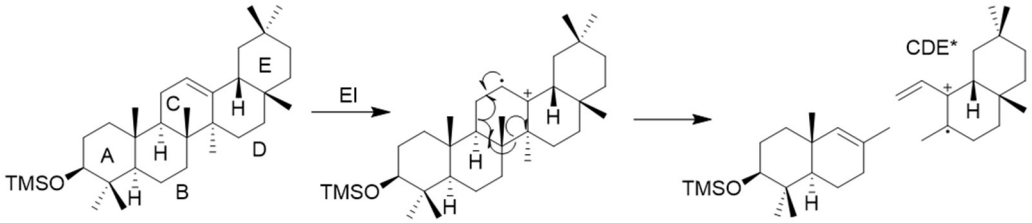


Figure 3.3.2 – The mechanism of formation of the CDE ring fragment through a retro-Diels-Alder fragmentation of TMS-beta-amyrin upon electron ionisation in a mass spectrometer. The CDE* ring fragment (m/z = 218) is the most frequently observed, though a rearrangement of the ionised intermediate leads to formation of the ABC* ring fragment (m/z = 279) with a much lesser intensity in the spectrum.*

3.4 – Evaluation of the Activity of CYP 72A69 in *N. benthamiana*

To establish the value of toolkit CYPs in generating potential triterpenoid intermediates relevant to the biosynthesis of aescin, CYP 72A69 was first tested for its ability to hydroxylate the C-21β position of beta-amyrin (which differs from its reported substrate, soyasapogenol B). Though a beta-amyrin hydroxylating ability for this CYP is speculated upon in the literature, it has not been experimentally characterised (Yano, Takagi et al. 2017). tHMGR, SAD1, and CYP 72A69 were co-expressed in *N. benthamiana*, and analysis of leaf extracts showed accumulation of a new product through GC (see figure 3.4.1). The mass spectrum of this new peak supports addition of a hydroxyl group to the CDE* ring fragment side of the molecule, as shown by the absence of the beta-amyrin CDE* peak and the appearance of a new peak corresponding to addition of OTMS to the CDE* fragment. A new peak is also seen corresponding to loss of a hydroxyl group from the molecule during ionisation. These fragmentation peaks in the mass spectrum are consistent with those reported in the literature for 21β-OH beta-amyrin (Zhou, Zhu et al. 2019). The conversion of beta-amyrin into 21β-OH beta-amyrin is not complete, but may be sufficient to obtain products in combination with other CYPs.

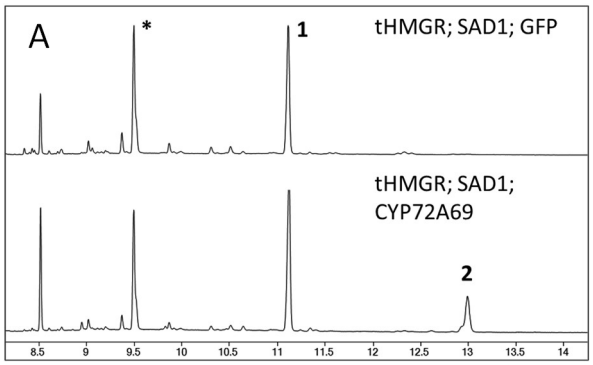
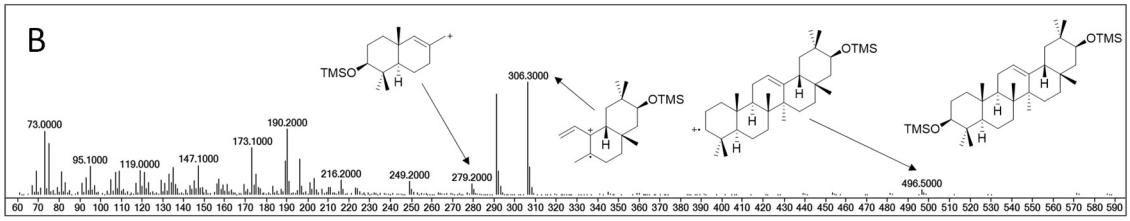


Figure 3.4.1 – CYP 72A69 is active on beta-amyrin.

*A: GC traces (TMS derivatised) showing activity of 72A69 when co-expressed with SAD1 and tHMGR. * Internal standard (coprostanol) visible at 9.5 mins. Peak (1): beta-amyrin. Peak (2): 21β-OH beta-amyrin, as predicted from mass spectra.*

B: Mass spectrum of peak (2). The absence of a peak at m/z 218, and the presence of a new peak at 306, suggests addition of -OTMS to the CDE ring fragment.*



3.5 – Evaluation of the Activity of CYP 716A12 in *N. benthamiana*

Though a C-28 acid is not one of the modifications required for the biosynthesis of aescin, oleanolic acid-derived intermediates have been isolated from *Aesculus*, so a CYP capable of oxidation here could form part of the broader biosynthetic pathway. To this end, the activity of a known C-28 oxidase, CYP 716A12 from *M. trunculata*, was evaluated through transient plant expression experiments. Co-expression with tHMGR and SAD1 results in good conversion of beta-amyrin into oleanolic acid (identity confirmed by characteristic peaks in its mass spectrum and comparison against a commercial standard). Figure 3.5.1 shows the GC traces of leaf extracts expressing CYP 716A12. Given that the C-28 oxidising activity of CYP 716A12 is in fact a series of three successive oxidations, going from methyl to hydroxy, aldehyde, and then acid, it had been

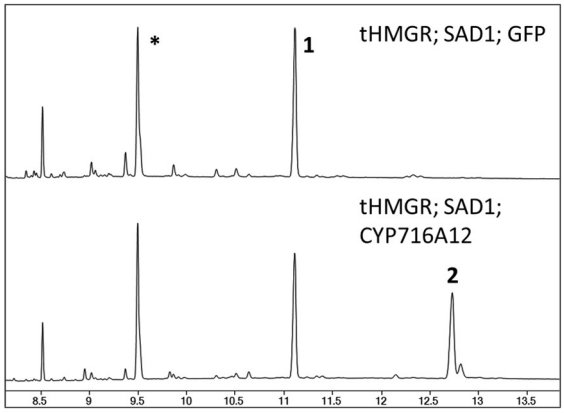


Figure 3.5.1: CYP716A12 is active on beta-amyrin.

GC traces (TMS-derivatised) showing the activity of CYP 716A12 on beta-amyrin in N. benthamiana*. * Internal standard (coprostanol) visible at 9.5 mins. Peak (1) beta-amyrin. Peak (2) corresponds to oleanolic acid, by comparison with known standards and through its mass spectrum.*

hypothesised that some erythrodiol might be observed in the leaf extracts. Indeed, expression of CYP 716A12 in yeast has been reported to result in accumulation of erythrodiol as a side product (Fukushima, Seki et al. 2011). However, this was not observed in these experiments in *N. benthamiana*. The mass spectra in figure 3.5.2 serve to compare the leaf extracts expressing CYP 716A12 with a commercial sample of oleanolic acid. The same peaks and fragmentation patterns are observed, peaks which support oxidation of a methyl group to a carboxylic acid. A new CDE* ring fragment peak is observed corresponding to addition of the two oxygens to C-28, along with peaks resulting from decarboxylation upon ionisation (the C-28 methyl now readily lost as carbon dioxide when converted into a carboxylic acid. The good conversion of beta-amyrin into oleanolic acid by CYP 716A12, and subsequent high yield of oleanolic acid in leaf tissue, make this a promising candidate for combinatorial experiments with further CYPs and other downstream enzymes.

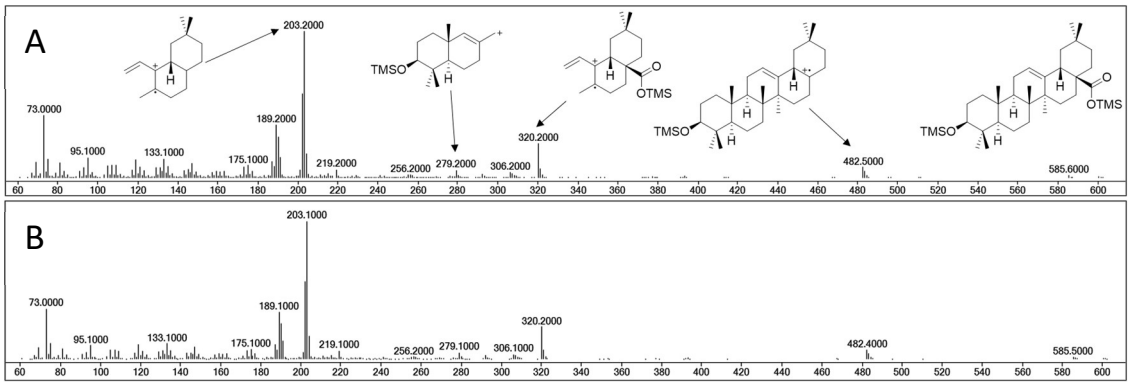


Figure 3.5.2: Mass spectra demonstrating oleanolic acid synthase activity of CYP 716A12.

A: Mass spectrum of peak (2) in figure 3.5.1. The ABC ring fragment is retained, but a new CDE* ring fragment corresponding to addition of a carboxylic acid is seen. Decarboxylation peaks are also visible, supporting conversion of a methyl group into a carboxylic acid.*

B: Mass spectrum of a commercial standard of oleanolic acid. The two spectra are functionally identical, and all the relevant fragmentation peaks are visible in both.

3.6 – Evaluation of the Activity of CYP 93E1 in *N. benthamiana*

C-24 hydroxylation is another transformation relevant to the biosynthesis of aescin, and the toolkit CYP evaluated to carry this out here is CYP 93E1, whose reported substrate in the soyasaponin pathway is sophoradiol (22β-OH beta-amyrin) (Shibuya, Hoshino et al. 2006). Co-expression with tHMGR and SAD1 was carried out in *N. benthamiana*, and GC-MS analysis of leaf extracts (figure 3.6.1) showed near-complete consumption of beta-amyrin and the appearance of a new product. However, the peak for the new product was small;

much smaller than might be expected from complete conversion of beta-amyrin. A second product peak was also observed. The mass spectrum for the first peak appeared to match that of 24-OH beta-amyrin in the literature (Reed, Stephenson et al. 2017), with retention of the unmodified CDE* ring fragment, and new peaks corresponding to addition of a hydroxyl group to the ABC* ring fragment. The second new product peak could correspond to boswellic acid (24-carboxylic acid beta-amyrin) – the mass spectrum also shows an unmodified CDE* fragment, but this time has a peak which matches that of an additional oxidation on the ABC* fragment. These results are consistent with previous investigations into CYP93E1 (Reed, Stephenson et al. 2017), with the alcohol, aldehyde, and acid products all having been detected. It could also go some way to explaining the apparent smaller peaks in the GC traces, as acid products are likely to be less visible through GC due to their increased polarity and lessened volatility. C-24 hydroxylated intermediates may not be in the earlier steps of the putative biosynthetic pathway postulated earlier, however, CYP 93E1 could still be included in wider trials of candidate enzymes for thoroughness.

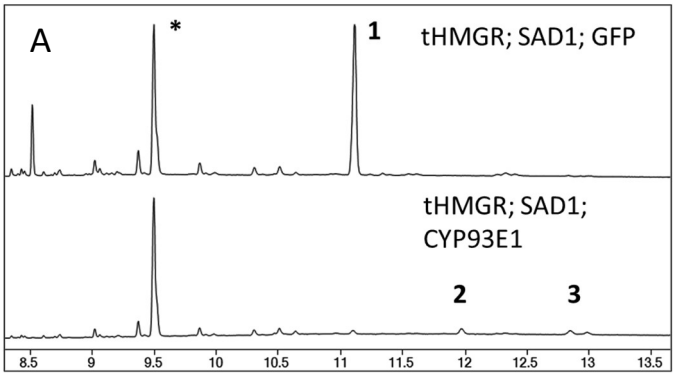
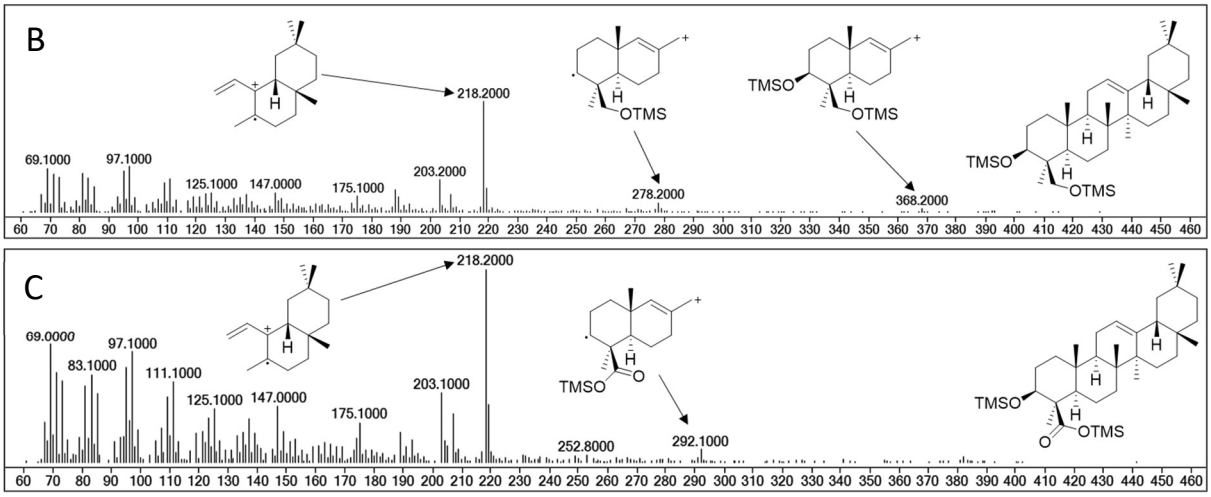


Figure 3.6.1 – CYP 93E1 exhibits oxidising activity against beta-amyrin.

*A: GC traces (TMS-derivatised) showing the activity of CYP 93E1. * Internal standard (coprostanol) visible at 9.5 mins. Peak (1) corresponds to beta-amyrin. Peak (2) corresponds to 24-OH beta-amyrin, according to mass spectra. Peak (3) appears to match that of boswellic acid.*

B: Mass spectrum of peak (2). The CDE ring fragment is retained, but a new ABC* ring fragment corresponding to addition of a hydroxyl group is seen. The peaks observed here are the same as those previously reported in the literature (Reed, Stephenson et al. 2017) for 24-OH beta-amyrin.*

C: Mass spectrum of peak (3). Again, the retention of the CDE ring fragment suggests no modifications on that side of the molecule. The ABC* ring fragments seen in B are absent, replaced with a new peak matching a ABC* fragment with an acid group. This peak is also observed in the literature for boswellic acid – 24-carboxylic acid beta-amyrin (Reed, Stephenson et al. 2017).*



3.7 – Evaluation of the Activity of CYP 716-2012090 in *N. benthamiana*

The final toolkit CYP to be investigated here is CYP 716-2012090 (or Qs16a), a C-16α hydroxylase from *Quillaja saponaria* characterised within the Osbourn Group (Osbourn and Reed 2019). Co-expression of Qs16a with tHMGR and SAD1 was undertaken in *N. benthamiana*. After analysis of leaf extracts through GC-MS, a new peak was detected (see figure 3.7.1). This peak had a mass spectrum similar to that of the product from CYP 72A69 – with a peak at m/z 306 representing addition of an alcohol group to the CDE* ring fragment. The mass spectrum of the new product corresponds very closely to that of 16α-OH beta-amyrin as reported in the literature (Moses, Pollier et al. 2015). This will allow inclusion of CYP 716-2012090 in further trials along with the other CYPs.

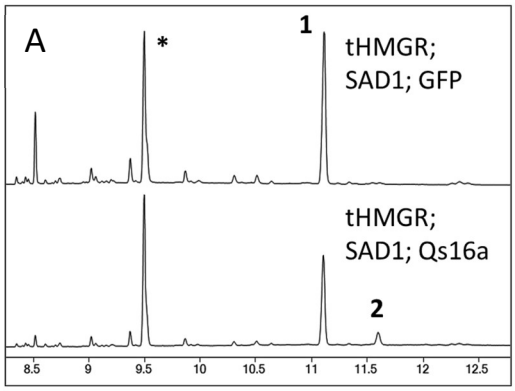


Figure 3.7.1 – Qs16a hydroxylates beta-amyrin.

*A: GC traces (TMS-derivatised) showing the activity of Qs16a. * Internal standard (coprostanol) visible at 9.5 mins. Peak (1) corresponds to beta-amyrin. Peak (2) corresponds to 16α-OH beta-amyrin, according to its mass spectrum.*

Below: Mass spectrum of peak (2). The unmodified ABC ring fragment is retained, but that of the CDE* ring fragment is absent. Instead, a new peak at m/z 306 corresponds to the CDE* ring fragment with addition of a derivatised hydroxyl group. The mass spectrum matches that reported in the literature for 16α-OH beta-amyrin (Moses, Pollier et al. 2015).*

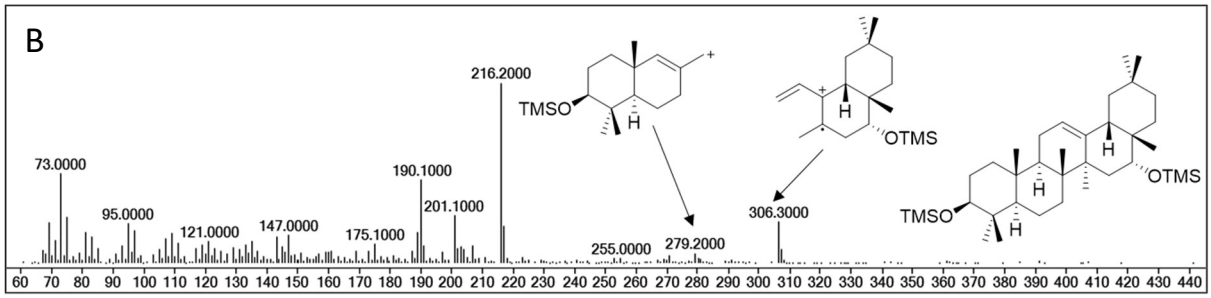


Figure 3.7.2 shows the predicted structures of the products made using each of the four toolkit CYPs, 72A69, 716A12, 93E1, and Qs16a, when tested individually. These enzymes enable introduction of hydroxyl groups to the beta-amyrin scaffold at positions C-21β, C-24, and C-16α, along with oxidation to an acid at C-28 and C-24. Their products having now been validated as potential scaffolds for candidate testing in *N. benthamiana*. These CYPs will therefore be incorporated into the experiments to identify *Aesculus* candidates. In order to widen the range of scaffolds available, combinations of toolkit CYPs will next be evaluated for activity through co-expression.

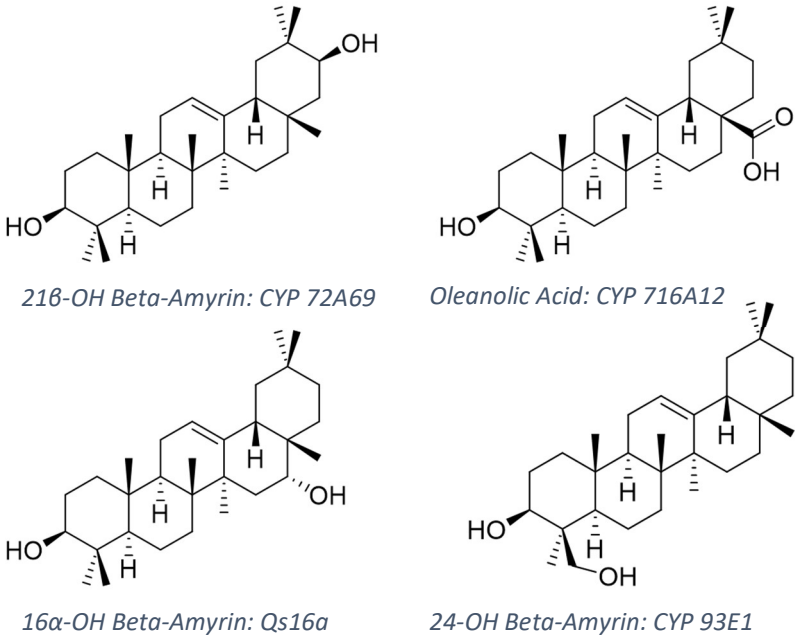


Figure 3.7.2 – The products made by the toolkit CYPs tested here. Clockwise from top-left: 21β-OH beta-amyrin (72A69), oleanolic acid (716A12), 24-OH beta-amyrin (93E1), and 16α-OH beta-amyrin (Qs16a).

3.8 – Combinatorial Trials of Toolkit CYPs

The reported promiscuity of some of the tested CYPs (Fukushima, Seki et al. 2011) raises the possibility that they may be capable of forming new products when co-expressed. Indeed, this has already been reported for one combination: CYP 716A12 and CYP 93E1, which together can form 24-OH oleanolic acid, otherwise known as 4-epi-hederagenin (Reed, Stephenson et al. 2017). These two enzymes were co-expressed with tHMGR and SAD1 to demonstrate this, with the resultant analysis of leaf extracts shown in figure 3.8.1. Interestingly, though the product peaks in GC traces resultant from CYP 93E1 on its own were small, it appeared that the new peak seen on co-expression with CYP 716A12 was larger. The mass spectrum of this new peak could be compared to that reported for 4-epi-hederagenin (Fukushima, Seki et al. 2013), and was found to match, confirming the identity of this product.

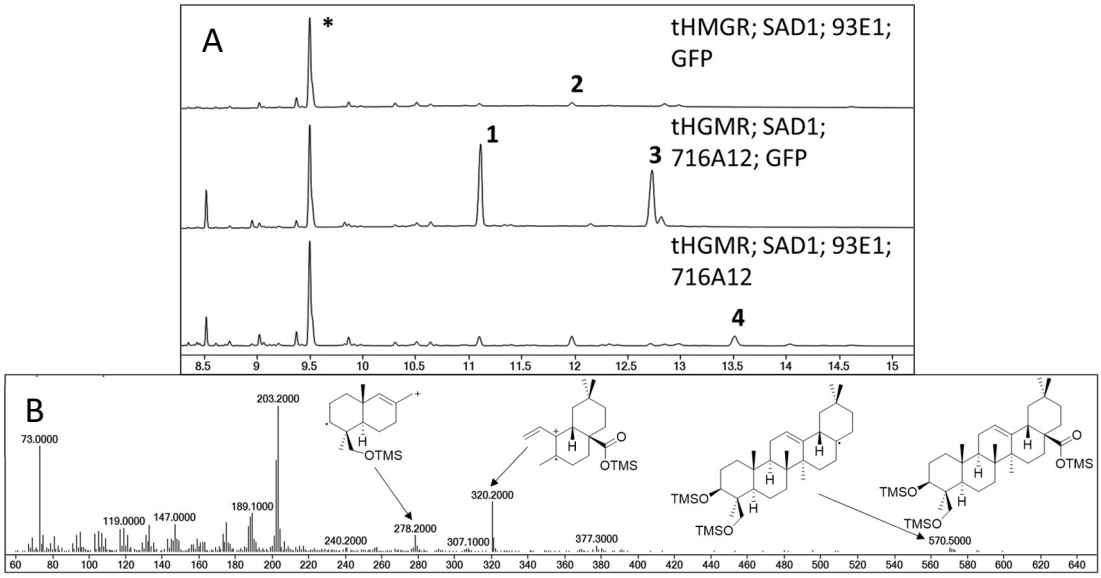


*Figure 3.8.1 – When co-expressed, 93E1 and 716A12 form 4-epi-hederagenin. A: GC traces (TMS-derivatised) showing the result of co-expression of 93E1 and 716A12. * Internal standard (coprostanol) visible at 9.5 mins. Peak (1) corresponds to beta-amyrin, peak (2) to 24-OH beta-amyrin, and peak (3) to oleanolic acid. The new product peak (4) has a mass spectrum which matches that of 4-epi-hederagenin (Fukushima, Seki et al. 2013). B: Mass spectrum of peak (4). Peaks characteristic of decarboxylation indicate C-28 oxidation, whilst an ABC* fragment peak at 278 m/z rather than 279 suggests the presence of a modification on this side of the molecule.*

The other possible pairwise combinations of the four CYPs were also tested in this manner, along with the combinations of three or all four CYPs together. Two further combinations displayed activity on co-expression in *N. benthamiana*: CYP 72A69 and Qs16a, both in combination with CYP 716A12. The combination of CYP 72A69 and CYP 716A12, co-expressed with tHMGR and SAD1, resulted in the appearance of a new peak in GC traces of leaf extracts (figure 3.8.2). The mass spectrum for this peak is consistent with the predicted structure of 21β-OH oleanolic acid, with a peak suggestive of a hydroxylated product undergoing decarboxylation upon ionisation. This molecule could prove a particularly important addition to trials of *Aesculus* candidate enzymes, as C-21β hydroxylation was identified as a potential very early step in aescin biosynthesis in chapter 2.

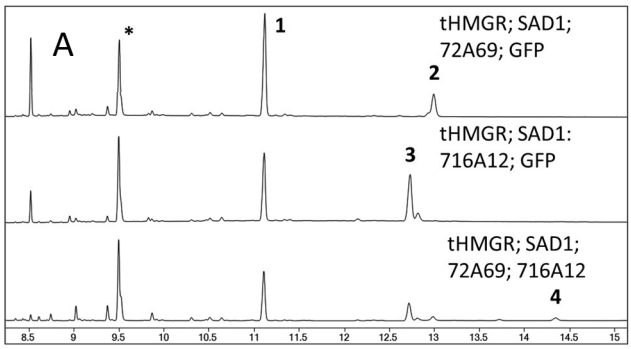
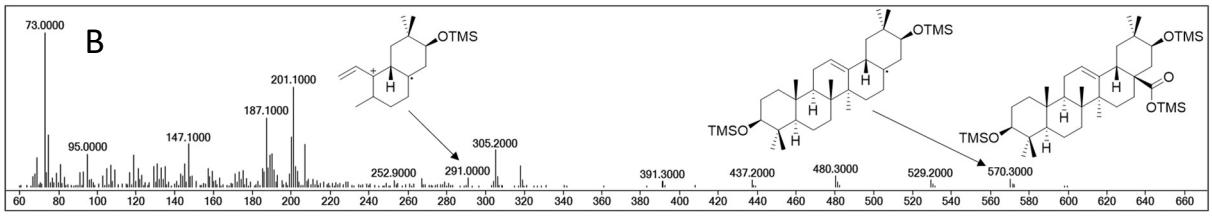


*Figure 3.8.2 – When co-expressed, 72A69 and 716A12 form a new product. A: GC traces (TMS-derivatised) showing the result of co-expression of 72A69 and 716A12. * Internal standard (coprostanol) visible at 9.5 mins. Peak (1) corresponds to beta-amyrin, peak (2) to 21-OH beta-amyrin, and peak (3) to oleanolic acid. A new product is seen at peak (4). B: Mass spectrum of peak (4). A peak characteristic of decarboxylation indicates C-28 oxidation of a hydroxylated scaffold, which is sited on the CDE* fragment. This is consistent with 21β-OH oleanolic acid.*



When co-expressed in *N. benthamiana* with SAD1 and tHMGR, Qs16a and CYP 716A12 appear to form echinocystic acid (16α-OH oleanolic acid). The GC traces of leaf extracts in figure 3.8.3 show the formation of a new peak, the mass spectrum of which matches that reported in the literature for echinocystic acid (Jæger, Ndi et al. 2017), with characteristic peaks showing loss of a hydroxyl group and decarboxylation, along with peaks showing both these modifications are on the CDE* ring fragment.

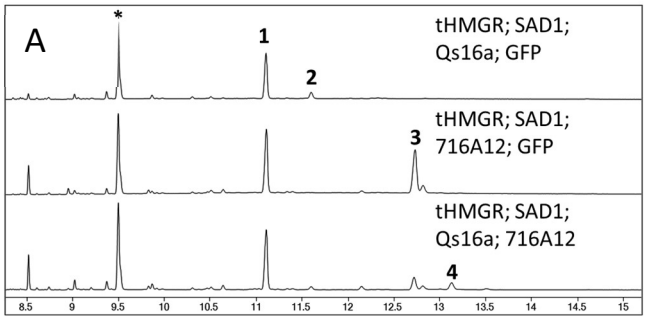
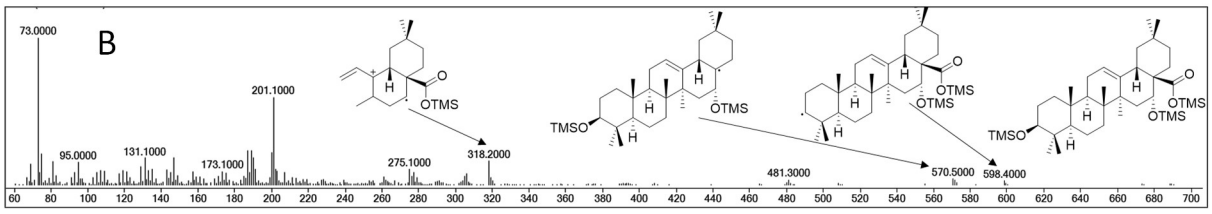
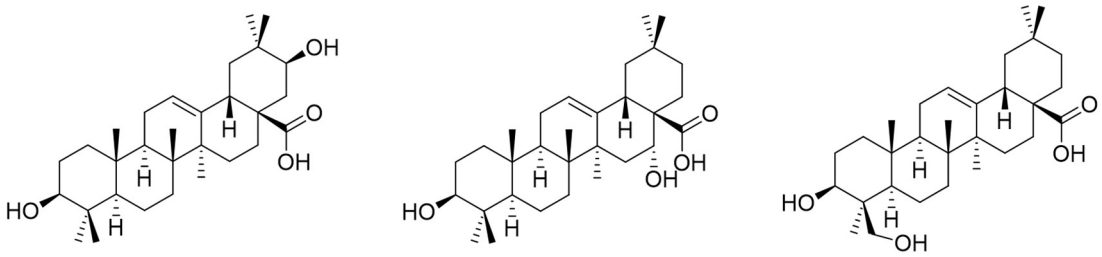


*Figure 3.8.3 – When co-expressed, Qs16a and 716A12 form echinocystic acid. A: GC traces (TMS-derivatised) showing the result of co-expression of Qs16a and 716A12. * Internal standard (coprostanol) visible at 9.5 mins. Peak (1) corresponds to beta-amyrin, peak (2) to 16-OH beta-amyrin, and peak (3) to oleanolic acid. The new product at peak (4) appears to be echinocystic acid, based on its mass spectrum. B: Mass spectrum of peak (4). A peak at 598 m/z corresponds to loss of a hydroxyl group, and decarboxylation and CDE* fragment peaks are also visible. This spectrum matches that reported in the literature for echinocystic acid (Jæger, Ndi et al. 2017).*



With these experiments a further three scaffolds are validated on which *Aesculus* candidates could be tested: 4-epi-hederagenin, 21β-OH oleanolic acid, and echinocystic acid (figure 3.8.4). This now provides a robust platform to test these candidates. However,

the inclusion of further scaffolds, especially those with C-22α or C-28 hydroxylation, would be highly desirable. No CYPs carrying out these transformations are represented in the toolkit, so any reported in the wider literature will have to be considered instead.



21-OH Oleanolic Acid *Echinocystic Acid* *4-epi-hederagenin*

Figure 3.8.4 – The structures of the products from combinatorial trials with CYP 716A12 and other toolkit CYPs. Left to right: 21β-OH oleanolic acid (72A69 and 716A12); echinocystic acid (Qs16α and 716A12); and 4-epi-hederagenin (93E1 and 716A12).

3.9 – Cloning and Testing of a Reported Erythrodiol Synthase

Hydroxylation at the C-28 of beta-amyrin (forming erythrodiol and its derivatives) was identified in the previous chapter as a crucial, and likely early, step in the biosynthetic pathway of aescin. However, thus far a CYP that hydroxylates beta-amyrin at the C-28 position to form erythrodiol as its main product yet to be validated in *N. benthamiana*. This makes testing of *Aesculus* candidate enzymes more challenging, as this is likely to be a key step in the biosynthetic pathway and could be a determinant in the substrate specificity of downstream biosynthetic enzymes.

Two CYPs from tomato (CYP716A44 and CYP716A46) have been reported to synthesise erythrodiol from beta-amyrin when expressed in yeast (Yasumoto, Seki et al. 2017). Both these CYPs are reported to achieve efficient conversion of beta-amyrin into erythrodiol, with additional formation of oleanolic acid as a secondary product. However, the activity of biosynthetic enzymes in yeast and in *N. benthamiana* are not always reported to be identical (Khakimov, Kuzina et al. 2015). One of these CYPs (CYP716A44) was cloned as per section 3.2 and transformed into an appropriate expression vector, before co-expression with SAD1 and tHMGR in *N. benthamiana*. Figure 3.9.1 shows GC traces of CYP 716A44 in combination with SAD1 – the activity is identical to that of CYP 716A12. The mass spectra of these traces (figure 3.9.2) confirm this: the product observed on expression

of CYP 716A44 has an identical spectrum to a commercial sample of oleanolic acid. Oleanolic acid is produced, but erythrodiol was not detected.

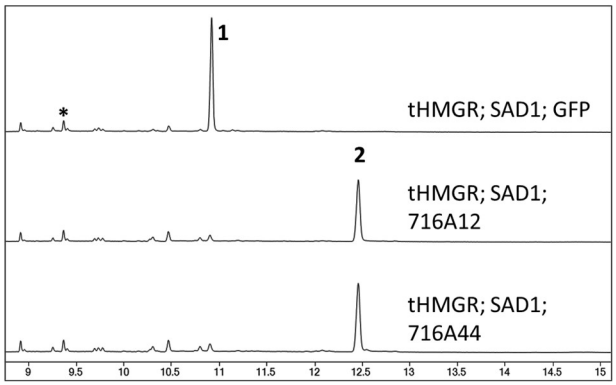


Figure 3.9.1 – Co-expression of CYP 716A44 with tHMGR and SAD1 in N. benthamiana *results in formation of oleanolic acid.*

*GC traces (TMS-derivatised) comparing the expression of 716A12 and 716A44. * Internal standard (coprostanol) visible at 9.3 mins. Peak (1) corresponds to beta-amyrin, peak (2) to oleanolic acid. The activity of 716A44 appears identical to CYP 716A12.*

The reported production of erythrodiol by CYP 716A44 in yeast is not reflective of its activity in plants – this is not unusual, as protein folding and sub-cellular localisation of both proteins and metabolites may differ in yeast and in plants (Yao, Weng et al. 2015, Owen, Patron et al. 2017). Since CYP 716A44 has the same activity as CYP 716A12 in *N. benthamiana* it was not included in further experiments, leaving C-28 hydroxylation of beta-amyrin amongst the transformations unaccounted for in the toolkit.

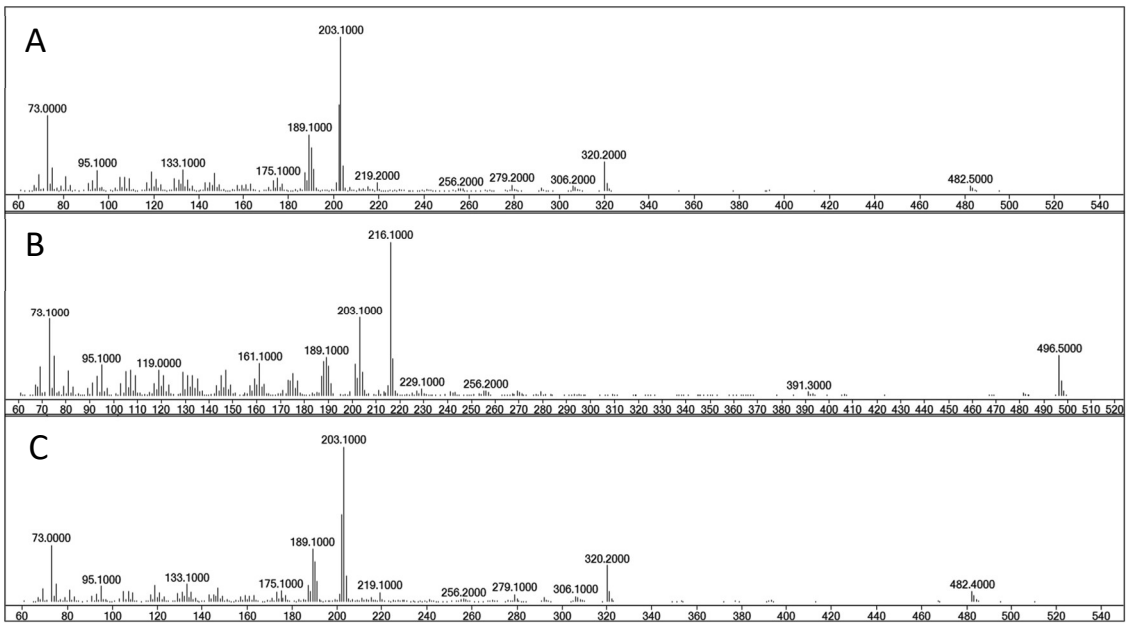


Figure 3.9.2 – The observed product of 716A44 is oleanolic acid. Mass spectra: A: peak (2) in figure 3.9.1 (716A44, SAD1, tHMGR); B: commercial sample of erythrodiol (28-OH beta-amyrin); C: commercial sample of oleanolic acid. The spectra of A and C are identical.

3.10 – Conclusions

The triterpenoid toolkit has been assessed for enzymes that may aid reconstruction of the aescin biosynthetic pathway in *N. benthamiana*, and has been found to provide an excellent starting point for the testing of *Aesculus* candidates. The production of beta-amyrin in *N. benthamiana* through infiltration with a combination of tHMGR and SAD1 is well-validated, and provides a scaffold for the testing of CYP candidates. CYPs from the toolkit are capable of producing three hydroxylated beta-amyrin derivatives relevant to aescin biosynthesis (16α-OH, 21β-OH, and 24-OH beta-amyrin), along with oleanolic acid. These CYPs can be employed in combination to obtain three further scaffolds – hydroxylated oleanolic acid derivatives (21β-OH oleanolic acid, echinocystic acid, and 4-epi-hederagenin), all of which can also be included in assays of *Aesculus* candidate enzymes. 21β-OH beta-amyrin and 21β-OH oleanolic acid in particular are of the most interest, the C-21β hydroxylation having been previously identified as a likely early step in the aescin biosynthetic pathway. Evaluation of a reported erythrodiol synthase failed to recreate C-28 hydroxylation in *N. benthamiana*, so this transformation will need to be effected by enzymes from *Aesculus*. With these scaffolds now validated and established, work will now move on to identification and testing of candidates from *Aesculus pavia*, starting with OSCs and CYPs.

Chapter 4 – Identification and Cloning of Candidate Biosynthetic Enzymes from *Aesculus pavia*

Synopsis

The triterpenoid toolkit furnished the enzymes for beta-amyrin C-16α, C-21β, and C-24 hydroxylation from heterologous plant species, along with a C-28 oxidase capable of forming oleanolic acid derivatives. However, several key pathway steps necessary for reconstructing the aescin pathway were still missing. A focussed approach to identify these missing steps from *Aesculus* was therefore initiated. Publicly available transcriptome data has been generated for *Aesculus pavia* as part of the OneKP project (Leebens-Mack, Barker et al. 2019). Although these transcriptome resources are limited, lacking expression levels in different tissues, they nevertheless enabled the identification of candidate genes for aescin biosynthesis by searching for genes with sequence similarity to characterised biosynthetic enzymes from other species. This led to the identification of a CYP (ApCYP1) that carried out 21β-hydroxylation of the beta-amyrin scaffold, a critical transformation in the biosynthetic pathway. Although a C21β-hydroxylase (CYP 72A69) from the triterpenoid toolkit was able to carry out partial conversion of beta-amyrin to 21β-OH beta-amyrin, ApCYP1 achieves significantly increased conversion over its toolkit equivalent. Furthermore, ApCYP1 allows identification of novel products not previously detected, for instance 21β,24-OH beta-amyrin. ApCYP1 represents the first active aescin biosynthetic enzyme discovered, a significant point in this project.

Introduction

4.1 – Introduction to *Aesculus pavia* and the OneKP Project

The previous chapter identified and validated enzymes from the triterpenoid toolkit that are capable of carrying out transformations relevant to the biosynthesis of aescin. It further provided access to a series of compounds that could be made using these enzymes as viable scaffolds on which to test candidate biosynthetic enzymes from *Aesculus*; a summary of these compounds is provided in figure 4.1.1. Some important transformations were not represented in the toolkit, however. Thus, in order to reconstruct the aescin biosynthetic pathway, it will be necessary to identify and characterise novel enzymes that catalyse these missing steps. The genus *Aesculus* provides the most obvious source of candidates for these steps. Aescin has been reported in the literature to have been isolated

from several different species of *Aesculus*, most notably *A. hippocastanum*, but also *A. indica*, *A. pavia*, *A. chinensis* and others (Hostettmann and Marston 1995, Dinda, Debnath et al. 2010). Whilst any one of these species could be exploited to identify candidates, *A. pavia* already has a publicly available transcriptome, part of the OneKP project.

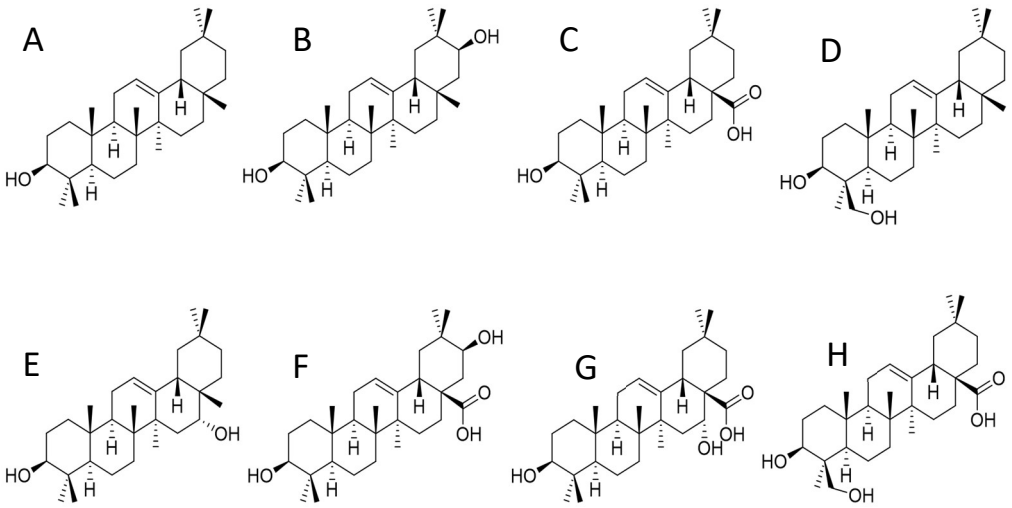


Figure 4.1.1 – The structures of the compounds formed using the triterpenoid toolkit in chapter 3. A: beta-amyrin (SAD1). B: 21ß-OH beta-amyrin (72A69 & SAD1). C: oleanolic acid (716A12 & SAD1). D: 24-OH beta-amyrin (93E1 & SAD1). E: 16α-OH beta-amyrin (Qs16a & SAD1). F: 21ß-OH oleanolic acid (72A69, 716A12, & SAD1). G: 16α-OH oleanolic acid (Qs16a, 716A12, & SAD1). H: 24-OH oleanolic acid (93E1, 716A12, & SAD1).

The OneKP (one thousand plants) project is an endeavour to sequence and make publicly available the transcriptomes of over one thousand plants, primarily intended to enable phylogenetic approaches in investigating the evolution of plants (Matasci, Hung et al. 2014, Carpenter, Matasci et al. 2019). A total of 1124 species has been sequenced thus far, encompassing a very broad range of plants species (Leebens-Mack, Barker et al. 2019) . Across the order Sapindales (the order to which *Aesculus spp.* belong) 15 species are listed, two of which are in the Sapindaceae (soapberry) family – *Acer negundo* (ash-leaved maple) and *Aesculus pavia* (red buckeye). The *Acer* genus is amongst the most closely related to the horse-chestnuts, but aescin is not reported to have been isolated from these trees. *Aesculus pavia*, however, is widely reported as a source of aescin and closely related saponins (Zhang, Li et al. 2006, Zhang and Li 2007, Lanzotti, Termolino et al. 2012), so this transcriptome has the potential to contain biosynthetic genes from the aescin pathway.

The *A. pavia* transcriptome generated by the OneKP project (code HBHB on the database) is reported as having been generated from RNA extracted from “very young leaves,

possibly some very young inflorescence". It is also reported that extraction of RNA was challenging and had to be repeated at least once due to poor quality isolated material. The quality of the RNA should therefore be borne in mind since this could pose an issue for identification of candidate genes.

The distribution of metabolites in different tissues of *A. pavia* has not been thoroughly evaluated, but across the broader *Aesculus* genus aescin is generally found in the fruits (conkers) (Dinda, Debnath et al. 2010, Glensk, Wlodarczyk et al. 2011). There is some uncertainty as to whether aescin is produced in leaf tissue – the presence of saponins in leaves has been reported in *A. pavia*, but it was also demonstrated that *Aesculus* plants were capable of translocating exogenously applied saponins from roots to the leaf tissue (Ferracini, Curir et al. 2010). The use of only one *A. pavia* tissue type for transcriptomics in the OneKP project precludes identification of candidates through differential gene expression. However, it should still be possible to search for candidate genes by sequence similarity with known biosynthetic enzymes from other species – the enzymes from the triterpenoid toolkit provide an excellent source of query sequences with which to conduct these searches.

Methods

4.2 – Chapter-Specific Methods

Candidates were identified through BLASTx or tBLASTn searches (Camacho, Coulouris et al. 2009) of the *A. pavia* transcriptome from the OneKP project, using the OneKP BLAST interface hosted by the China National GeneBank Database using the following query sequences (with accession numbers): *SAD1* (ABG88962.1) for OSCs, *CYP716A12* (FN995113.1), *CYP72A69* (NM_001354945.1), *CYP93E1* (NM_001249225.2), and *CYP87D16* (KF318735.1) for CYPs. Query sequences for BAHD candidate searching were *THAA1, THAA2, NtMAT1, Pf3AT, Ss5MaT1, SalAT, VAAT, HMT/HLT*, and *AMAT*, as listed in Appendix A1. BAHD candidates were manually assessed for the presence of conserved domains HxxxDG and DFGWG before inclusion. Candidate selection and primer design for UGT candidates was carried out by Thomas Louveau (formerly Osbourn Lab, JIC).

Primers were designed for Gateway cloning of selected candidates (see 9.1.3) – these were designed for two-step PCR, with the exception of *HBHB2033231*, *HBHB2008407*, and *HBHB2015618*, which were designed for one-step PCR (9.1.3). *A. pavia*

leaf material was obtained from a potted tree maintained in NBI glasshouses which had been obtained by Thomas Louveau (Pavia 3 – see 5.2), this had been determined to be outside the scope of the Nagoya Protocol according to a self-assessment using the guidelines on www.gov.uk/guidance/abs. RNA was extracted from this leaf material according to 9.1.1, and cDNA synthesised as set out in 9.1.2. Cloning was carried out either by one-step or two-step PCR using iProof (9.1.3), and amplified sequences cloned into expression vector pEAQ-HT-DEST1 (Sainsbury, Thuenemann et al. 2009) according to 9.1.4, before transformation into *Agrobacterium tumefaciens* LBA 4404 (9.1.7). Protein sequences of cloned candidates are listed in Appendix B.

Heterologous expression of toolkit and *Aesculus* enzymes in *N. benthamiana* was carried out according to 9.2.1, with harvested leaves extracted according to 9.3.1 and analysed through GC-MS as set out in 9.3.4. Estimation of the accumulation of 21β-OH beta-amyrin in leaves of *N. benthamiana* was carried out using the protocol in 9.3.6. Large-scale transient expression in *N. benthamiana* followed the protocol 9.2.2 using 100 plants, with subsequent extraction according to 9.3.3 and isolation of product according to 9.3.7 – purification of 21β,24-OH carried out with reverse phase flash column chromatography. NMR analysis of 21β-OH beta-amyrin and 21β,24-OH beta-amyrin was undertaken in $CDCl_3$ and pyridine-D5, respectively, according to the general considerations set out in 9.3.9 – NMR results are found in Appendix D1 and D2.

Results and Discussion

4.3 – Identification and cloning of candidates from *A. pavia*

OSC and CYP enzymes from the triterpenoid toolkit which have been characterised as carrying out transformations relevant to aescin biosynthesis were selected as query sequences, these are detailed in 4.2. Furthermore, 9 BAHDs identified in 2.5 were selected as query sequences; these were chosen across different clades (using previous phylogenetic studies (D'Auria 2006) to capture a broad range of candidates). Using these query sequences BLAST searches were carried out with the OneKP *A. pavia* transcriptome as set out in 4.2.

These searches identified 20 candidate sequences, comprising 1 OSC, 4 CYPs, and 15 BAHDs. The OSC and 3 CYP candidates were successfully cloned. The OSC and CYP

candidates are listed in table 4.3.1 along with the query sequences with which they were identified.

CANDIDATE	OSC/CYP	CLONED	QUERY SEQUENCE	NOTES
HBHB2033231	OSC	✓	49% aa sequence similarity with SAD1	Putative beta-amyrin synthase
HBHB2020895	CYP	✓	69% aa sequence similarity with CYP716A12	Active as ApCYP1, a 21β hydroxylase
HBHB2029750	CYP		61% aa sequence similarity with CYP716A12	
HBHB2008407	CYP	✓	42% aa sequence similarity with CYP93E1	
HBHB2015618	CYP	✓	41% aa sequence similarity with CYP72A69	

Table 4.3.1 – A list of the OSC and CYP candidate sequences identified through BLAST searches of the OneKP A. pavia *transcriptome. One of these, ApCYP1 or* HBHB2020895*, was found to be active, hydroxylating beta-amyrin at C-21β. This activity is identical to that found for CYP72A69 (see 3.4).*

Of the 15 BAHD candidates identified, two were successfully cloned: *HBHB2027681* and *HBHB2000297* – these did not demonstrate any activity on co-expression with toolkit enzymes for scaffolds 21β-OH beta-amyrin and 21β-OH oleanolic acid. Parallel work carried out by Thomas Louveau identified 19 putative UGT candidates from the OneKP transcriptome – primers designed by Thomas Louveau were included in these cloning experiments, leading to 1 candidate being successfully cloned: *HBHB2027528* or ApUGT2. This UGT candidate, though not found to be active in the work presented in this chapter, went on to demonstrate activity as a D-galactosyltransferase in later work (see 7.5).

The OSC candidate identified, *HBHB2033231*, contained the conserved DCTAE domain associated with the catalytic site and a number of the QW structural motifs (Guo, Xiong et al. 2021), along with the MWCYCR motif, the W of which has previously been closely linked to beta-amyrin product specificity (Kushiro, Shibuya et al. 2000). To this end, it was cloned and co-expressed in *N. benthamiana* with tHMGR, but no activity was observed. The three CYP candidates that were successfully cloned were co-expressed in *N. benthamiana* with tHMGR and SAD1 – one of these, HBHB2020895 or ApCYP1, demonstrated activity.

4.4 – Characterisation of the activity of ApCYP1

When co-expressed in *N. benthamiana* with tHMGR and SAD1, HBHB2020895 (ApCYP1) generated a product with a GC retention time of 12.9 min. The retention time of this product was the same as that detected when CYP 72A69, a characterised beta-amyrin C-21β hydroxylase from soybean (Yano, Takagi et al. 2017) (see 3.4), was co-expressed with tHMGR and SAD1 under the same conditions (figure 4.4.1A). The mass spectra of these two peaks confirmed that they were identical (figure 4.4.1B & C), and that they match those reported in the literature for 21β-OH beta-amyrin (Zhou, Zhu et al. 2019). CYP 72A69 gave only partial conversion of beta-amyrin (1) to 21β-OH beta-amyrin (2), while ApCYP1 gave near full conversion. Thus the *A. pavia* ApCYP1 enzyme appears more effective than the toolkit enzyme and may be an authentic aescin biosynthetic enzyme.

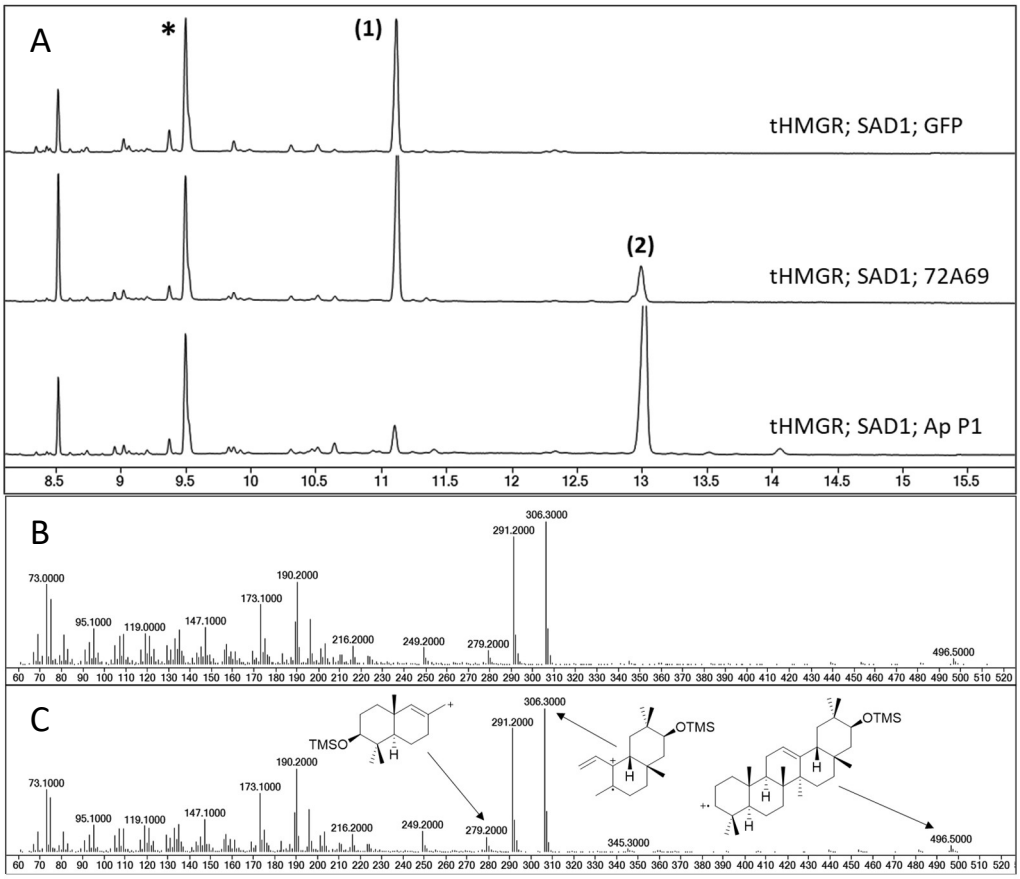


*Figure 4.4.1 – ApCYP1 has identical activity to CYP72A69. A: **GC** traces (TMS-derivatised) comparing the activity of ApCYP1 and CYP72A69 on beta-amyrin. * Internal standard (coprostanol) visible at 9.5 min. (1) beta-amyrin (2) 21β-OH beta-amyrin. Peak (2) is significantly larger in trace for ApCYP1 than for CYP72A69. B: Mass spectrum of peak (2) in trace for CYP72A69. C: Mass spectrum of peak (2) in trace for ApCYP1, with fragmentation peaks characteristic of 21β-OH beta-amyrin annotated. The mass spectra in B and C appear identical, indicating that the product of CYP72A69 and ApCYP1 is the same, namely 21β-OH beta-amyrin.*

ApP1 was identified through amino acid sequence similarity with *CYP716A12* (a characterised beta-amyrin C-28 oxidase from lucerne (Carelli, Biazzi et al. 2011, Fukushima, Seki et al. 2011)), and is also predicted to belong to the CYP 716 subfamily. This family being so frequently implicated in triterpenoid biosynthesis (Miettinen, Pollier et al. 2017), it had been identified as a promising candidate, however its activity as a 21β-OH beta-amyrin hydroxylase had not been predicted. In fact, many of the most closely related sequences to ApP1 in the NCBI GenBank database (Benson, Cavanaugh et al. 2012) are annotated as predicted beta-amyrin C-28 mono-oxygenases. The CYP 716 family has indeed been commonly found to oxidise beta-amyrin derivatives at C-28 (Carelli, Biazzi et al. 2011, Tamura, Seki et al. 2017, Yasumoto, Seki et al. 2017), but it is clear that a much greater variety of transformations are possible using CYPs of this family (Miettinen, Pollier et al. 2017). As such, the discovery of a CYP 716 with beta-amyrin C-21β hydroxylating activity is consistent with previous reports of this subfamily's diverse activities.

The discovery of the first active enzyme from *Aesculus* relevant to the aescin biosynthetic pathway is a major advance for this work. ApCYP1 enables near-complete conversion of beta-amyrin to 21β-OH beta-amyrin, so greatly enabling endeavours to reconstitute the full aescin pathway. It carries out a key step in the aescin pathway, by oxidising the beta-amyrin scaffold at a position that is subsequently acylated. The putative biosynthetic pathway set out in 2.6 identified hydroxylation of this position as likely to be one the earliest steps, making it crucial to the testing of subsequent candidates – furthermore, C-21 was identified as the most frequently acylated position on the beta-amyrin scaffold according to the Reaxys database (see 2.3). Identification of the first active enzyme from the pathway could greatly aid identification of further candidates, such as through co-expression analysis of transcriptomic resources (though this would require a transcriptome with measurements of expression levels across multiple tissues). The apparent increased conversion of beta-amyrin into 21β-OH beta-amyrin by ApCYP1 over CYP 72A69 is also significant for aescin pathway reconstitution: as an early step in a large (possibly 11 steps) pathway, accumulation of intermediates needs to be high to allow for sufficient substrate for the subsequent steps.

A further transient plant expression experiment was then carried out to compare the abilities of CYP72A69 and AsCYP1 to produce 21β-OH beta-amyrin. The two enzymes were each co-expressed with tHMGR and SAD1, with 10 biological replicates for each, and leaf extracts analysed by GC-MS. The area of the product peak in the total-ion chromatogram (TIC) for each trace was integrated, and normalised relative to an internal

standard (200 ppm coprostanol). This enabled estimation of the concentration of 21β-OH beta-amyrin in each extract, and thus the accumulation in mg per gram of dry leaf tissue. The results are shown in figure 4.4.2.

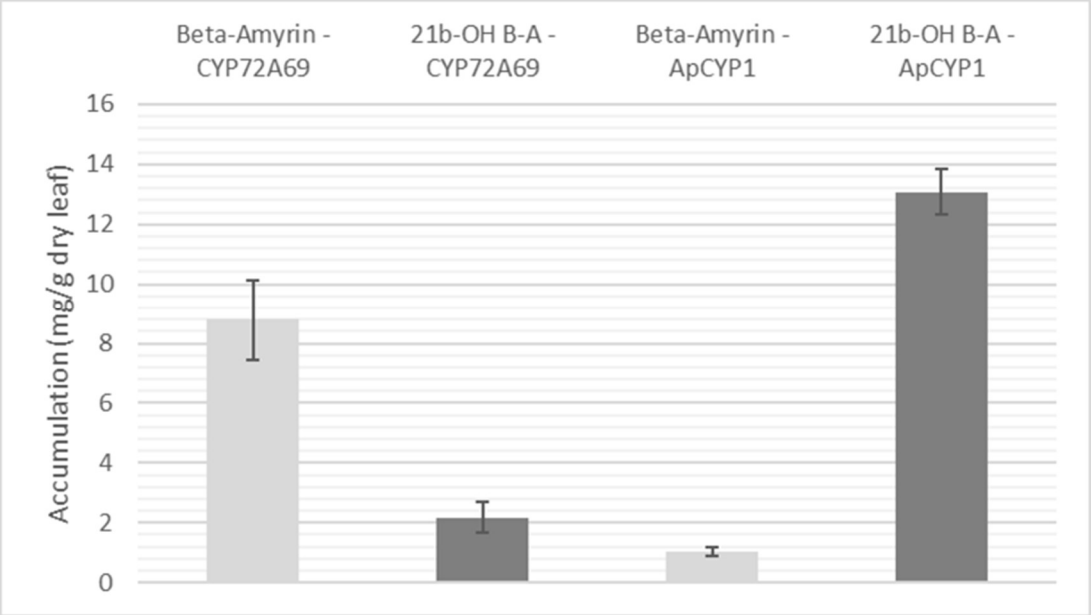


Figure 4.4.2 – ApCYP1 results in significantly increased conversion from beta-amyrin to 21ß-OH beta-amyrin over CYP 72A69. Chart showing the average estimated accumulation (mg/g dry leaf material) of beta-amyrin (light grey) and 21ß-OH beta-amyrin (dark grey) observed by co-expression with tHMGR and SAD1 in N. benthamiana*. An average accumulation of 8.80 mg/g beta-amyrin and 2.18 mg/g 21ß-OH beta-amyrin was observed with CYP 72A69, but ApCYP1 achieved 13.08 mg/g 21ß-OH beta-amryin and only 1.034 mg/g beta-amyrin. 10 biological replicates were used for each condition, error bars indicate standard errors.*

These results show an almost six-fold increase in accumulation of 21β-OH beta-amyrin in the leaf when compared to CYP 72A69 – 13.08 mg/g dry leaf using ApCYP1 against 2.18 mg/g using CYP 72A69. This represents near-complete conversion of substrate (beta-amyrin), and is consistent with some of the higher previously reported estimated accumulation of oxidised beta-amyrin derivatives in *N. benthamiana*, for instance estimated accumulation of approximately 15 mg/g of 12,13β-epoxy-16β-OH beta-amyrin (Reed, Stephenson et al. 2017), though it must be noted that actual isolated yields are typically lower (Reed and Osbourn 2018). This is promising for use of ApCYP1 in further co-expression experiments.

4.5 – Evaluation of the activity of ApCYP1 in combination with toolkit CYPs

The discovery of a new beta-amyrin C-21β hydroxylase warrants investigation of its activity on other beta-amyrin derived scaffolds. ApCYP1 may be more promiscuous in its

substrate specificity, or its increased conversion of beta-amyrin into 21β-OH beta-amyrin may enable identification of new products not previously detected using CYP 72A69 due to very low levels of accumulation in *N. benthamiana*. To this end, ApCYP1 was co-expressed with tHMGR, SAD1, and toolkit CYPs in combinatorial trials, testing its activity on the range of scaffolds depicted in figure 4.1.1. The results of co-expression with tHMGR, SAD1, and CYP 716A12 (forming oleanolic acid) are shown in figure 4.5.1.

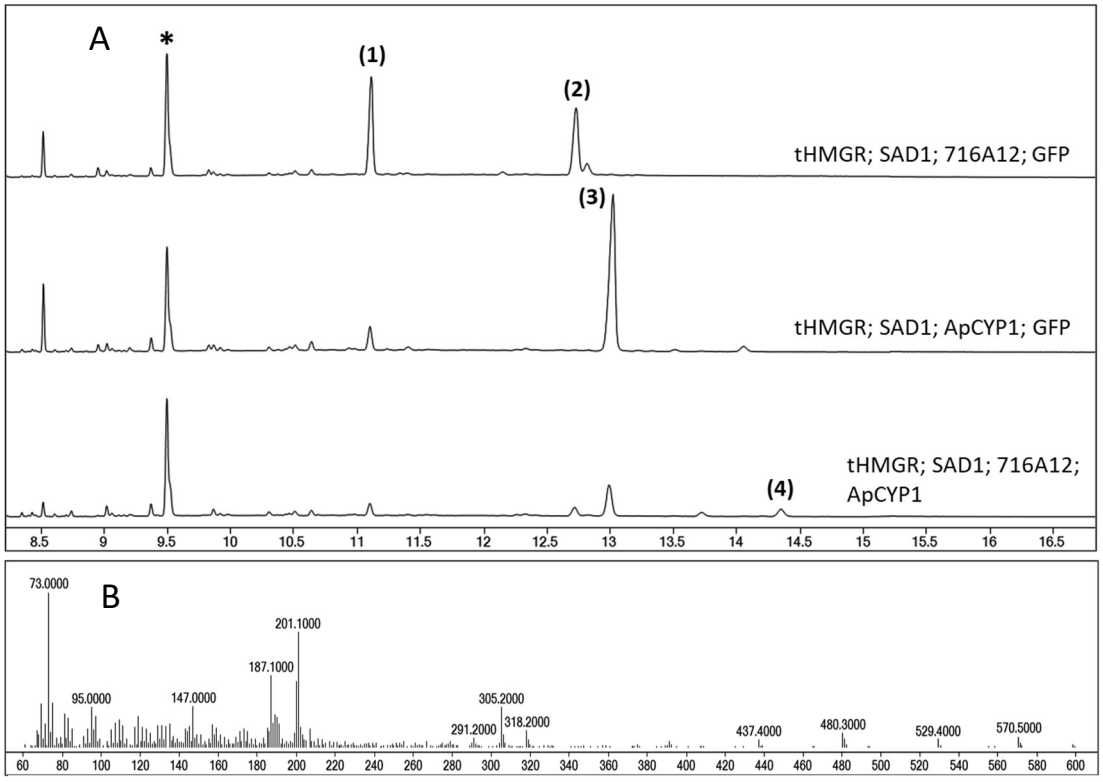


*Figure 4.5.1 – ApCYP1 is active on oleanolic acid. A: **GC** traces (TMS derivatised) demonstrating co-expression of ApCYP1 with tHGMR, SAD1, and CYP716A12. * Internal standard (coprostanol) visible at 9.5 min. (1) beta-amyrin (2) oleanolic acid (3) 21β-OH beta-amyrin (4) New product – this matches 21β-OH oleanolic acid formed by co-expression of CYP72A69 with tHMGR, SAD1, and CYP716A12 (see 3.8.2). This is only visible as a small peak. B: Mass spectrum of peak (4). This matches the mass spectrum of 21β-OH oleanolic acid made with CYP72A69 (3.8.2 B).*

ApCYP1 was active in combination with CYP716A12, forming a new product consistent with 21β-OH oleanolic acid. 21β-OH oleanolic acid, or machaerinic acid, has been previously identified as the core of the aglycone of saponins isolated from various plant species, some with the C-21β hydroxyl group free (Yokosuka, Kawakami et al. 2008, Yokosuka, Okabe et al. 2016), others with acylation at this position (Englert, Weniger et al. 1995). Its bioactivity against cancer cell lines has also been assessed (Yokosuka, Okabe et al. 2016), with it found to be only weakly cytotoxic against HL-60 cells (though more so than the saponins for which this forms the aglycone). Though the traces in figure 4.5.1A show only a very small peak for 21β-OH oleanolic acid, the peaks for the presumed substrates

21β-OH beta-amyrin (peak 3) and oleanolic acid (peak 2) are mostly consumed. It may be the case that 21β-OH oleanolic acid is only poorly visible through GC, possibly due to its increased polarity making it less volatile. The oxidation of C-28 up to an acid, rather than a methyl alcohol, makes this unlikely to be a direct intermediate of aescin, however it will still be employed as a testing scaffold for the evaluation of *Aesculus* candidates.

ApCYP1 was also co-expressed with tHMGR, SAD1, and CYP 93E1 (a characterised C-24 hydroxylase from soybean (Shibuya, Hoshino et al. 2006)) in *N. benthamiana*, the results of which are shown in figure 4.5.2.

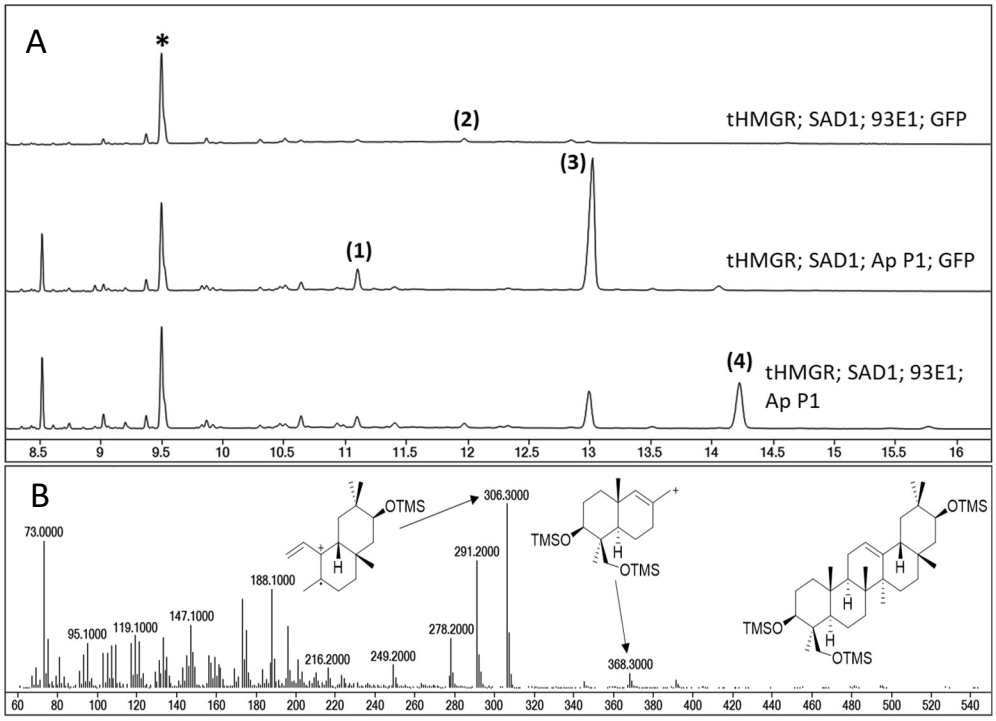


*Figure 4.5.2 – A new product is detected when ApCYP1 and 93E1 are co-expressed together with tHGMR and SAD1 – this is predicted to be 21β,24-OH beta-amryin. A: **GC** traces (TMS-derivatised) demonstrating co-expression of CYP93E1 and ApCYP1 on beta-amyrin. * Internal standard (coprostanol) visible at 9.5 mins. (1) beta-amyrin (2) 24-OH beta-amyrin (3) 21β-OH beta-amyrin (4) new product formed by both CYPs acting together. B: Mass spectrum of peak (4), with structures of possible fragmentation peaks. Characteristic peaks at m/z 368 and 306 suggest addition of OTMS to both the ABC* and the CDE* ring fragments. These peaks are analogous to those reported in the literature for kudzusapogenol C (Kinjo, Miyamoto et al. 1985).*

Together, CYP 93E1 and ApCYP1 appear to form a new product, which had not previously been observed by co-expression of CYP 93E1 and the toolkit C-21β hydroxylase CYP 72A69. The mass spectrum of this new product contains characteristic beta-amyrin fragmentation peaks, suggestive of modification of both the ABC* and CDE* ring fragments through addition of -OTMS groups. These are consistent with the reported mass spectrum for 21β,24-OH beta-amyrin (or kudzusapogenol C), the expected product given the known

activities of the two enzymes (Kinjo, Miyamoto et al. 1985). Kudzusapogenol C was not identified in the putative aescin biosynthetic pathway (2.6) as a likely aescin intermediate – C-24 hydroxylation was predicted to be a later step, possibly requiring hydroxylation at other positions beforehand. CYP 93E1 might therefore be exhibiting a differing substrate specificity than that of the endogenous C-24 hydroxylase from *A. pavia*. The reason behind kudzusapogenol C only being observed with use of ApCYP1 and not CYP 72A69 is unclear: either CYP 72A69 is not able to use 24-OH beta-amyrin as a substrate, or the limited conversion of the two toolkit enzymes together only led to slight accumulation of product, which went undetected. Detection of this new product may provide an additional scaffold for subsequent testing of further candidate enzymes.

4.6 – Isolation of 21β-OH beta-amyrin & kudzusapogenol C

To confirm the identity of the new product observed in figure 4.5.2 (thought to be 21β,24-OH beta-amyrin), analysis through NMR was carried out. To produce sufficient quantity of the new product, a large-scale infiltration of *N. benthamiana* was conducted according to reported methods (Stephenson, Reed et al. 2018), co-expressing tHMGR, SAD1, ApCYP1, and CYP 93E1 in 100 plants. Pressurised solvent extraction with EtOAc was used to extract the dried leaves, from which were purified 7.1 mg 21β-OH beta-amyrin and 3.0 mg kudzusapogenol C.

The ^{1}H NMR for 21β-OH beta-amyrin has previously been reported in the literature (Zhou, Zhu et al. 2019). A ^{1}H NMR of the product obtained here was taken in $CDCl_3$ at 400 MHz (provided in Appendix D1), and the key peaks compared. These were found to match those reported exactly. Of note were the peaks for H-12 (app t) (J = 3.6 Hz) at δ5.22 ppm; for H-21 (dd) (J = 12.0, 4.7 Hz) at δ3.52 ppm; H-3 (dd) (J = 10.9, 4.8 Hz) at δ3.22 ppm; and eight singlet peaks with an integral of 3H (methyl signals H-23, H-24, H-25, H-26, H-27, H-28, H-29, and H-30) – these were observed at δ = 0.79, 0.86, 0.87, 0.94, 0.96, 0.96, 1.00, and 1.13 ppm, identical to those reported (Zhou, Zhu et al. 2019). The match of the coupling constants for H-21 give confidence that the stereochemistry of the product seen here is identical to that reported, namely 21β-hydroxylation. This confirms the identity of the ApCYP1 product as 21β-OH beta-amyrin, as had been suspected by comparison of GC-MS data with the products of a known C-21β hydroxylase from the toolkit (CYP 72A69). The structure of 21β-OH beta-amyrin is shown in figure 4.6.1, with the positions corresponding to the key peaks described here numbered.

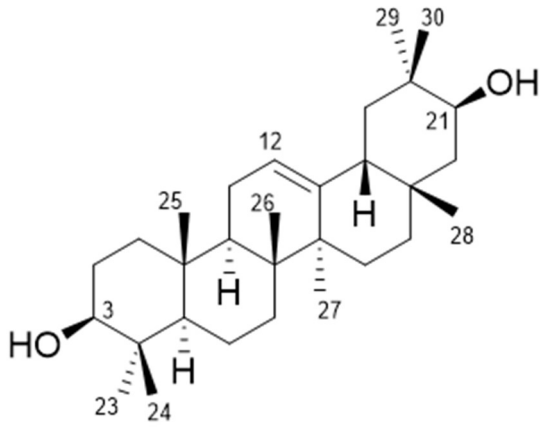


Figure 4.6.1 – The structure of 21β-OH beta-amyrin. The positions numbered are those whose ^{1}H NMR signals were compared to reported values (Zhou, Zhu et al. 2019) to confirm the identity of this product.

NMR spectra of 21β,24-OH beta-amyrin (kudzusapogenol C) have previously been reported, though ^{1}H spectra are only reported as a triacetate derivative (Kinjo, Miyamoto et al. 1985, Ohtani, Ogawa et al. 1992). The ^{13}C spectra are, however, reported for the underivatised molecule, so can serve as a useful comparison. Table 4.6.1 shows the assignments for the NMR spectra of 21β,24-OH beta-amyrin measured in pyridine D-5, alongside the ^{13}C spectra reported. The numbering is shown on figure 4.6.2. The very close match of the ^{13}C peaks serves to validate the identity of the product. In comparison to the ^{1}H spectrum for 21β-OH beta-amyrin, the presence of only 7 CH_3 signals indicates modification of one of the methyl groups. Additionally, two new proton signals are observed at 3.48 and 4.29 ppm, in the region characteristic of alcohol groups. These two signals are the result of two protons at the same position, in this case the C-24; the protons are diastereotopic, so are not in the same chemical environment despite being on the same carbon.

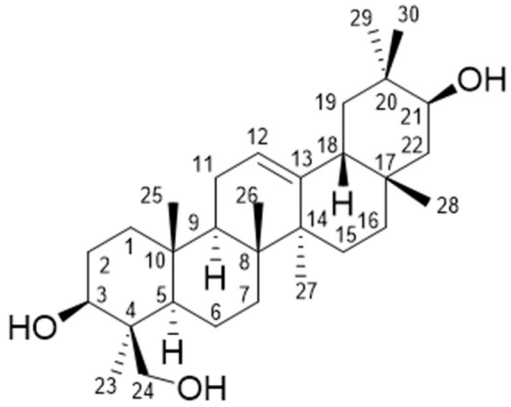


Figure 4.6.2 – The structure of 21β,24-OH beta-amyrin, or kudzusapogenol C, with all positions numbered.

The assignment of the ^{1}H spectrum made use of 2D-NMR experiments, namely COSY, HSQC, and HMBC. COSY indicates proton-proton coupling, which aids in determining

which protons are adjacent to one another. HSQC shows proton-carbon coupling through a single bond, allowing the ^{13}C and ^{1}H environments to be matched to one another, in addition to showing if a carbon environment is associated with more than one proton environment. For instance, in the HSQC spectrum for 21β,24-OH beta-amyrin the ^{13}C signal for C-24 (64.4 ppm) is associated with the two ^{1}H peaks mentioned above, as shown in figure 4.6.3. HMBC shows proton-carbon coupling through multiple bonds – typically two or three. This is extremely useful in structural determination, especially when, as is the case for beta-amyrin derivatives, many of the proton signals are all in a similar region (between 0.7 and 2.0 ppm). Some of the key HMBC and COSY couplings used in assignment of the NMR spectra for 21β,24-OH beta-amyrin are displayed in figure 4.6.4.

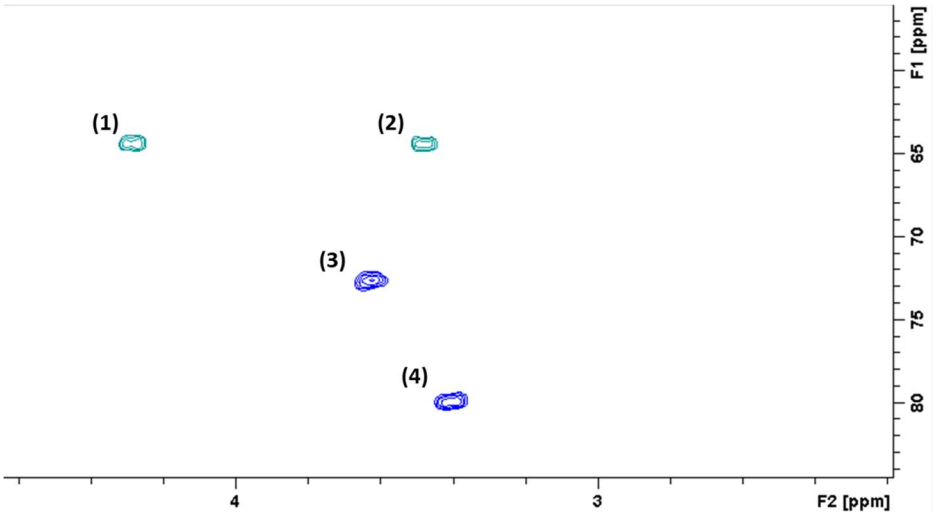
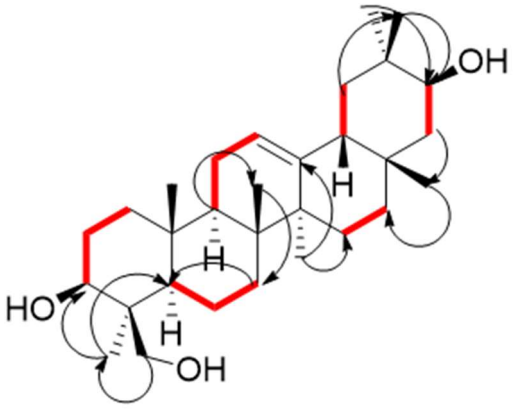


Figure 4.6.3 – Detail from the HSQC spectrum of 21β,24-OH beta-amyrin, showing two ^{1}H signals in the alcohol range to be linked to the same ^{13}C environment. (1) and (2) are the signals for H-24 at δ 4.29 and 3.48 ppm, linked to C-24 at δ 64.4 ppm. (3) H-21 at δ 3.63 ppm is linked to C-21 at δ 72.6 ppm. (4) H-3 at δ 3.41 ppm is linked to C-3 at δ 80.0 ppm.

Using these 2D spectra, assignment started at the olefin (C-12 and C-13), as this is the most distinct environment in the molecule. One proton environment is found in this region (a very characteristic apparent triplet at 5.06 ppm), which can be linked to its carbon peak (122.5 ppm) through HSQC. COSY was then used to identify the signals for H-11, and these then used to find H-9. HMBC can be used to identify H-27 from C-13, in addition to H-25 and H-26 from C-9, and the process is repeated working outwards from the olefin until all positions are accounted for. The assignment is close to those for 21β-OH beta-amyrin and 24-OH beta-amyrin (Shibuya, Hoshino et al. 2006, Zhou, Zhu et al. 2019), and gives confidence that the stereochemistry is as shown.

Figure 4.6.4 – Structure of kudzusapogenol C with selected COSY and HMBC couplings used in assigning the ^{1}H spectrum. COSY are marked with bold red bonds, and HMBC with arrows.



Position	δ ^{13}C per (Ohtani, Ogawa et al. 1992) (ppm)	δ ^{13}C observed (ppm)	δ ^{1}H observed (ppm) (*J* in Hz)
1	38.9	38.6	0.75 (m), 1.35 (m)
2	28.4	28.2	1.77 (m)
3	80.1	80.0	3.41 (dd) (J = 4.6, 11.6)
4	43.2	43.0	-
5	56.3	56.1	0.69 (m)
6	19.1	18.9	1.49 (m), 1.22 (m)
7	33.3	33.1	1.05 (m), 1.25 (m)
8	40.1	39.9	-
9	48.1	47.9	1.43 (dd) (J = 2.9, 9.6)
10	37.0	36.8	-
11	24.1	23.9	1.65 (m), 1.69 (m)
12	122.7	122.5	5.06 (app t) (J = 3.6)
13	144.3	144.1	-
14	41.9	41.7	-
15	26.5	26.3	0.69 (m), 1.49 (m)
16	29.9	28.4	1.85 (m)
17	35.1	34.9	-
18	47.2	47.0	1.95 (dd) (J = 3.8, 14.0)
19	46.5	47.5	1.75 (m), 1.06 (m)
20	36.9	36.8	-
21	72.8	72.6	3.63 (dd) (J = 7.0, 9.7)
22	47.7	46.4	2H, 1.54 (m)
23	23.5	23.4	3H, 1.33 (s)
24	64.5	64.4	3.48 (d) (J = 11.0), 4.29 (d) (J = 11.0)
25	16.2	16.0	3H, 0.70 (s)
26	16.9	16.7	3H, 0.73 (s)
27	26.0	25.8	3H, 0.98 (s)
28	28.7	28.5	3H, 0.72 (s)
29	29.9	29.8	3H, 1.00 (s)
30	17.7	17.5	3H, 0.92 (s)

Table 4.6.1 – Assignment of the ^{1}H and ^{13}C NMR spectra for 21β,24-OH beta-amyrin, recorded in pyridine D-5 at 400 MHz and 100 MHz respectively, alongside the ^{13}C assignments previously reported. The ^{13}C spectrum obtained here matches that reported very closely. The proton spectra could not be directly compared, as only derivatised proton spectra are reported for this molecule, but the proton spectrum obtained here is fully assigned and confirms the predicted structure.

4.7 – Conclusions

Using the OneKP transcriptome of *Aesculus pavia*, the first investigations into the aescin biosynthetic pathway in *Aesculus* have been carried out. This led to identification of the first active enzyme from the aescin pathway, a beta-amyrin 21β-hydroxylase (ApCYP1, a putative CYP 716), demonstrated to achieve significantly increased conversion of beta-amyrin to product over the existing beta-amyrin 21β-hydroxylase in the toolkit. ApCYP1 was active in combination with either CYP 716A12 and CYP 93E1 (an activity not previously observed using CYP 72A69), leading to the detection of product not hitherto formed through transient expression, which was identified as kudzusapogenol C (21β,24-OH beta-amyrin). Large-scale infiltration led to isolation of this new product, which enabled structural determination through NMR for full confirmation of its identity, including the stereochemistry. The detection of new products in through use of ApCYP1 may be a function of its increased activity.

The identification of the first enzyme from the aescin pathway is likely to aid significantly in selection of further candidates, since it can be used in pathway reconstruction experiments in *N. benthamiana* to more effectively generate the 21β-OH beta-amyrin pathway intermediate. The *ApCYP1* gene could also be used as bait to identify other co-expressed *A. pavia* aescin pathway genes. However the OneKP transcriptome resources for *A. pavia* are limited, since only one tissue type is represented. The availability of transcriptome data for multiple *Aesculus* tissues that differ in their levels of aescin would enable better correlations to be made between expression of candidate genes and the likely roles of these is aescin biosynthesis. Furthermore, issues with a large number of apparently truncated sequences and with cloning success rates raise doubts over the validity of the sequences obtained (although cloning problems could be in part due to differences between the *A. pavia* accession used in the OneKP project and the accession used here for experimental work). Further investigation into aescin biosynthesis in *Aesculus* therefore warrants generation of a new in-house transcriptome. This forms the basis of the next chapter.

Chapter 5 – Metabolic Profiling and Transcriptomic Analysis of Aesculus

Synopsis

Having identified the need for improved bio-informatic resources to aid in aescin biosynthetic enzyme candidate selection, the benefits of a genomic or a transcriptomic approach were compared, and the generation of a new transcriptome was considered. To inform this, a survey of the metabolomic profile of tissues from various *Aesculus* species was carried out. This found significant variation in saponin accumulation between tissues, which provided the basis for the design of RNAseq experiments – the species *A. pavia* and *A. hippocastanum* chosen for investigation. Analysis focussed on obtaining a contrast between producing and non-producing tissues. RNAseq was carried out on five samples representing three tissues from *A. pavia* and two from *A. hippocastanum*, and from this two transcriptomes were assembled. Analysis of these showed potential for identification of a much greater range of candidates, which led to the characterisation of active enzymes in chapters 6 and 7.

Introduction

5.1 – Introduction to the Generation of Novel Transcriptomic Resources for *Aesculus*

The previous chapter reported on the use of the OneKP transcriptome data for *Aesculus pavia* for identification of candidate aescin biosynthetic enzymes. From this emerged the first active enzyme – ApCYP1, a CYP active on beta-amyrin and its derivatives, which was shown to be a 21β-hydroxylase. However, the OneKP data for *A. pavia* has limitations for the discovery of further pathway genes because of the suboptimal quality of the transcriptome assembly and because data are only available for a single tissue (reported as “very young leaves, possibly some very young inflorescence”). The discovery of the genes encoding natural product biosynthetic pathways in other plants has been greatly facilitated by carrying out metabolite analysis and transcriptomics on different plants tissues (Bhambhani, Lakhwani et al. 2017, Srivastava, Shukla et al. 2018). This enables correlations to be made between presence of metabolites and expression of likely candidate genes. The availability of high-quality transcriptome resources for multiple different tissues of *A. pavia* that differ in aescin content would pave the way for the discovery of new aescin pathway genes.

Two principal approaches were considered – a genomic and a transcriptomic approach – each enabling differing methods of candidate identification. The elucidation of biosynthetic pathways have been reported using both of these resources (Hodgson, De La Peña et al. 2019, Jozwiak, Sonawane et al. 2020). Whilst either approach enables candidate identification through sequence similarity, a genome offers the additional ability to search for candidates on the basis of clustering into biosynthetic gene clusters (BCGs). Biosynthetic genes for natural product pathways are often found clustered in close proximity to one another in plant genomes, with three or more together being known as a BGC, so searching for instances of multiple genes in triterpenoid biosynthesis-relevant families associated with one another can aid in candidate identification (Boycheva, Daviet et al. 2014, Nützmann and Osbourn 2014, Polturak and Osbourn 2021). Genome mining on this basis has been employed in the investigation of a number of biosynthetic pathways (Shimura, Okada et al. 2007, Winzer, Gazda et al. 2012, Jozwiak, Sonawane et al. 2020, Liu, Cheema et al. 2020), and bio-informatic tools have been developed for this purpose (Kautsar, Hernando et al. 2017, Schläpfer, Zhang et al. 2017, Töpfer, Fuchs et al. 2017). However, not all biosynthetic pathways are encoded by BGCs. In this chapter the generation of *de novo* transcriptome data for *A. pavia* was prioritised as a first step towards discovery of further aescin pathway genes.

Differential expression of sequences in metabolite-producing and non-producing tissues has previously been employed to identify and prioritise candidate genes for plant natural product biosynthetic pathways (Guo, Ma et al. 2014, Liu, Yang et al. 2015, Jain, Srivastava et al. 2016), as has co-expression of genes with that of known biosynthetic enzymes (Sonawane, Pollier et al. 2017, Zhou, Huang et al. 2017, Jozwiak, Sonawane et al. 2020). Co-expression with a known sequence has strong potential when investigating aescin biosynthesis, as an active enzyme from the pathway has already been found (*ApCYP1*) to serve as bait. To establish whether differential expression might be an option for *Aesculus*, it will first be necessary to investigate the levels of aescin in different tissues.

Aescin is well documented to be present in the fruits of various *Aesculus* species (Hostettmann and Marston 1995, Wei, Ma et al. 2004, Dinda, Debnath et al. 2010, Idris, Mishra et al. 2020), but there also reports of trace accumulation of saponins in the bark, roots (Ćalić-Dragosavac, Zdravković-Korać et al. 2010), and leaves (Ferracini, Curir et al. 2010). Within the fruits, saponin accumulation is highly localised to the seed itself, gradually increasing with storage of the seeds until reaching the zygotic embryo stage, whereupon saponins can form up to 6% of the seed by dry weight (Hostettmann and

Marston 1995, Ćalić-Dragosavac, Zdravković-Korać et al. 2010). Aescin production has also been achieved through tissue culture of somatic embryos (Zdravković-Korać, Milojević et al. 2022). The saponin profiles for different *Aesculus* species show some variation, for example in the nature and positions of the acyl groups (Wei, Ma et al. 2004) (Yang, Long et al. 2019), and in sugar chain substituents (Zhang and Li 2007). However, the basic oxidised scaffold of the aglycone is shared across the genus, and the most abundant saponins (aescins I, II, and III) are common to most species (Zhang, Li et al. 2010). In order to identify candidate aescin biosynthetic genes by differential gene expression analysis, it will first be necessary to determine the saponin composition of different *Aesculus* tissues. This will enable gene expression in producing and non-producing tissues to be compared.

Methods

5.2 – Chapter-Specific Methods

Collection of plant material was carried out in spring and summer of 2018 at the following sites: Cambridge University Botanic Gardens (Pavia 2, Hippo 2, Indica, Californica), the Norwich Biosciences Institute grounds (Hippo 1), The Avenues, Norwich (Pavia 1), and a potted *A. pavia* obtained by Thomas Louveau (formerly Osbourn Lab, JIC) maintained in NBI glasshouses (Pavia 3). All material was determined to be outside the scope of the Nagoya Protocol as a result of a self-assessment using the guidelines on www.gov.uk/guidance/abs. Table 5.2.1 shows the tissues collected from each tree. Tissue collected for metabolomic analysis was treated in the same manner as harvested *N. benthamiana* leaves in 9.2.1 – frozen at -70 ^{o}C and freeze-dried. Tissue for transcriptomic analysis from pavia 1 and Hippo 1 was stored at -70 ^{o}C without drying.

TREE	LEAF	LEAF VEIN	FLOWER	SHELL	OUTER CONKER	INNER CONKER
PAVIA 1	PL	PLV	PF	PS	POC	PIC
PAVIA 2	P2L	P2LV		P2S	P2IC	P2IC
PAVIA 3	P3L					
HIPPO 1	HL	HLV	HF	HS	HOC	HIC
HIPPO 2	H2L	H2LV		H2S		H2IC
INDICA	IL	ILV	IF			
CALIFORNICA	CL	CLV	CF			

Table 5.2.1 – The tissues collected from each tree used, with the codes for the traces shown in this chapter. Not all tissue types were obtainable from all trees.

Dried plant material was prepared for metabolic analysis according to the protocols in 9.3.1 (for GC-MS, using 100 ppm internal standard) and 9.3.2 (for LC-MS, using 20 mg plant material instead of 10 mg). GC-MS analysis was carried out according to 9.3.4, and LC-MS according to 9.3.5 using LC gradient A. Commercial samples of aescin were obtained from Sigma-Aldrich, and alpha- and beta-amyrin from Extrasynthese.

RNA extraction was carried out according to 9.1.1. The quality of the extracted RNA was determined by gel electrophoresis, optical measurement on a Thermo NanoDrop spectrophotometer, and by the Earlham Institute using a Perkin-Elmer GX II to generate RNA quality scores. RNAseq was carried out by the Earlham Institute on an Illumina NovaSeq 6000 SP, with all samples uniquely barcoded and pooled into a single lane of sequencing. Transcriptome assembly and annotation was carried out by the Earlham Institute using the software Trinity (v2.8.5) (Haas, Papanicolaou et al. 2013), TransDecoder (v5.5.0), AHRD (v3.3.3), Salmon (0.15.0), and Centrifuge (v1.0.3). Local BLAST databases were made using the *A. pavia* and *A. hippocastanum* transcriptomes using BLAST+ (v2.9.0); these were used to search for *ApCYP1* homologues through blastn.

5.3 – LC-MS Analysis of Aescin Standard

A commercial sample of aescin was obtained from Sigma-Aldrich (reported 95% purity). This was obtained as a white powder, sparingly soluble in pure water. Analysis of a 250 ppm methanolic solution of this by LC-MS was carried out using gradient A set out in 9.3.5 (figure 5.3.1). It is evident that this aescin preparation contains multiple distinct molecules, principally those compounds forming the mixture known as beta-aescin (Hostettmann and Marston 1995), but also others. The two largest peaks in the LC trace, at 14.4 and 14.7 mins, are predicted to be aescin Ib and aescin Ia respectively, by comparison of their relative abundance and retention times with the literature (Colson, Decroo et al. 2019). The smaller peaks are likely to be aescin IIIa and various isoaescins (acetyl group on C-28 rather than C-22), which are difficult to distinguish on the basis of LC traces or mass spectra. Aescins IIa and IIb (C'2-xylose) were not identified in this sample. The mass spectrum shown in figure 5.3.2 for aescin Ia matches those reported (Colson, Decroo et al. 2019, Yang, Long et al. 2019), with several characteristic fragmentation peaks that aid identification. The beta-amyrin retro-Diels-Alder fragmentation peaks seen in previous chapters are observed here too, such as B[1], the result of the CDE* ring fragment with multiple substituents. An aglycone peak is also seen at B[2], along with peaks for loss of a sugar (B[3]), loss of an acyl group (B[4]), loss of an alcohol, and a proton adduct peak. Loss of acyl groups from the aescin aglycone moiety has been reported to be the result of McLafferty re-arrangements (Colson, Decroo et al. 2019). Comparison of the mass spectra of metabolites in *Aesculus* tissues with this spectrum will help to clarify ambiguous LC traces.

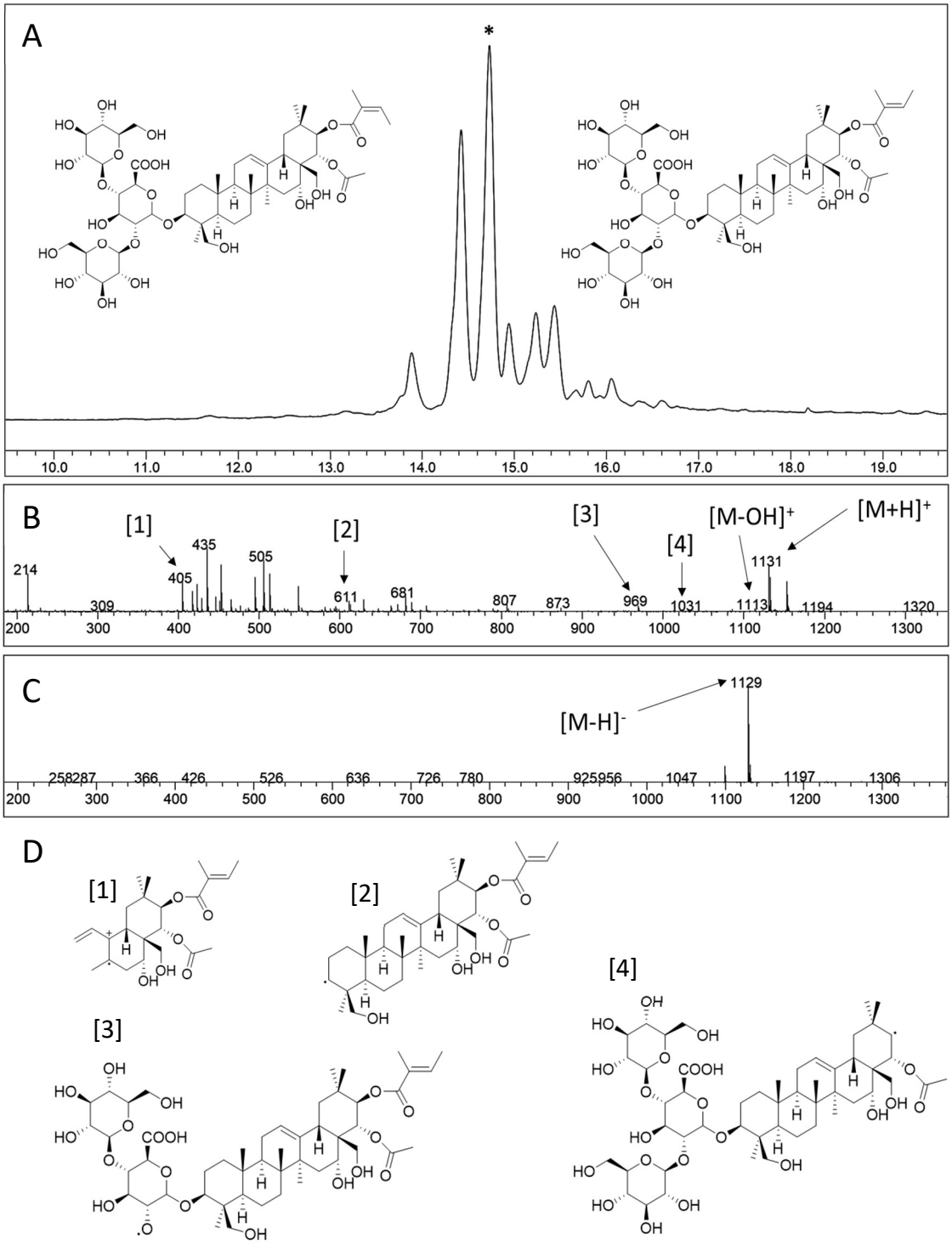


Figure 5.3.1 – LC-MS analysis of a commercial sample of aescin. A: **LC** trace of aescin, showing that aescin is comprised of a mixture of saponins. The two largest peaks are for aescin Ia (* right) and aescin Ib (left). B: Mass spectrum (positive mode) of aescin Ia (peak at 14.7 mins in A). A number of fragmentation peaks are observed, characteristic of the modifications to the scaffold seen in aescin. C: Negative mode mass spectrum of same; a quasi-molecular peak is apparent, and is highly characteristic of saponins with a glucuronic acid. D: Predicted structures of fragmentation peaks seen in B. [1] CDE* ring fragment of aescin, [2] aglycone, loss of sugar moiety, [3] loss of glucose from sugar chain, [4] loss of tigloyl substituent.

5.4 – Metabolomic Profiling of *Aesculus* Leaves

Leaf samples were obtained from four different *Aesculus* species: *A. pavia*, *A. hippocastanum*, *A. indica*, and *A. californica*, as detailed in table 5.2.1. Samples were separated into leaf tissue and leaf vein tissue and analysed separately. Extracts of these were analysed by both LC-MS and GC-MS to investigate the distribution of aescin, related saponins, and intermediates and the resulting profiles compared with those for a commercial sample of aescin. LC-MS analysis of leaf tissue extracts is shown in figure 5.4.1. It is evident that there is some variation in the metabolite profiles for leaves from different species. Although two peaks were observed in the trace for *A. pavia* with similar retention times as aescins Ia and Ib, the mass spectra of these peaks (B & C) established that they were not aescins Ia or Ib. The mass spectra of other peaks in these traces did not suggest the presence of derivatives or other saponins closely related to aescin, due to the absence of key fragmentation peaks characteristic of protoaescigenin- and barringtogenol C-derived triterpenoids. The overall profile of the leaf vein samples was very similar to the leaf tissue.

Analysis of leaf extracts was also carried out by GC-MS of TMS-derivatised EtOAc extracts to identify any aescin aglycone intermediates that may not be detected by LC. GC traces of samples of *A. hippocastanum*, *A. indica*, and *A. californica* did not contain peaks determined to be likely oleanane-type triterpenoid compounds. However, one sample from *A. pavia* 3 did appear to contain small quantities of both beta- and alpha-amyrin, as determined by comparison of retention times and mass spectra against known commercial samples. Figure 5.4.2 shows GC traces of leaf samples from *A. pavia* trees 1, 2, and 3. Interestingly, these triterpenoid peaks could not be detected in samples from the other two *A. pavia* trees (Norwich and Cambridge), despite the overall metabolic profiles being similar. Other peaks observed in these traces did not appear to be oleanolic in origin, according to their mass spectra.

The absence of saponins in the leaves of *Aesculus* is in agreement with most of the literature, which reports aescin and related compounds as being primarily in the seeds (Hostettmann and Marston 1995, Dinda, Debnath et al. 2010). A report of unspecified saponins in the leaves of *A. pavia* (Ferracini, Curir et al. 2010) could not be confirmed here. However, the presence of beta-amyrin, a key intermediate in the biosynthesis of aescin, in the leaves of one *A. pavia* tree might not exclude some biosynthetic activity from leaves altogether, especially as saponins are reported to be readily translocated in *Aesculus* (Ferracini, Curir et al. 2010).

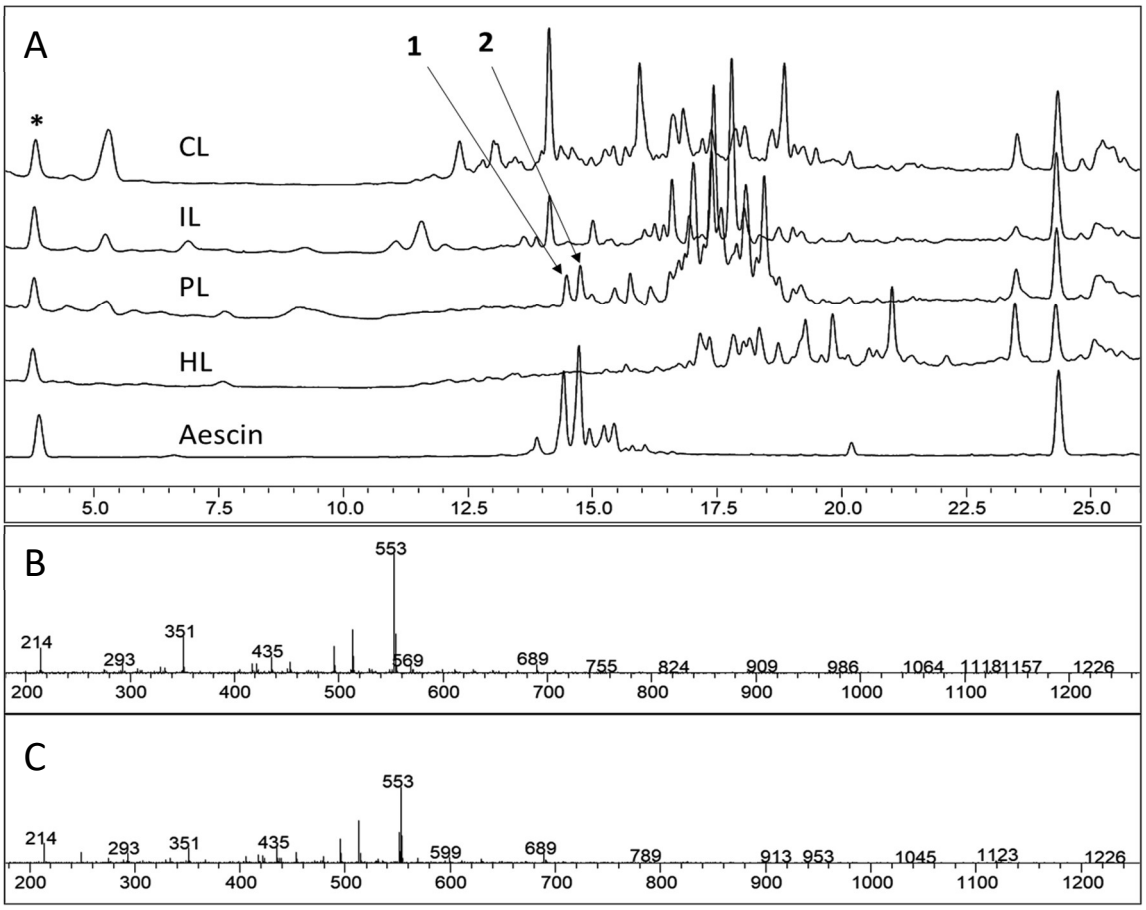


Figure 5.4.1 – LC-MS analysis of leaf extracts from four different *Aesculus* species. A: LC traces comparing leaf extracts from *A. californica*, *A. indica*, *A. pavia*, and *A. hippocastanum* to aescin (250 ppm). * Internal standard (digoxin 20 ppm) seen at 3.9 mins. Two peaks in the trace for *A. pavia* overlap with peaks for aescin. B & C: mass spectra (positive mode) for peaks 1 and 2. Comparison of these to the spectrum of aescin in figure 5.3.1 shows that these peaks are not aescin.

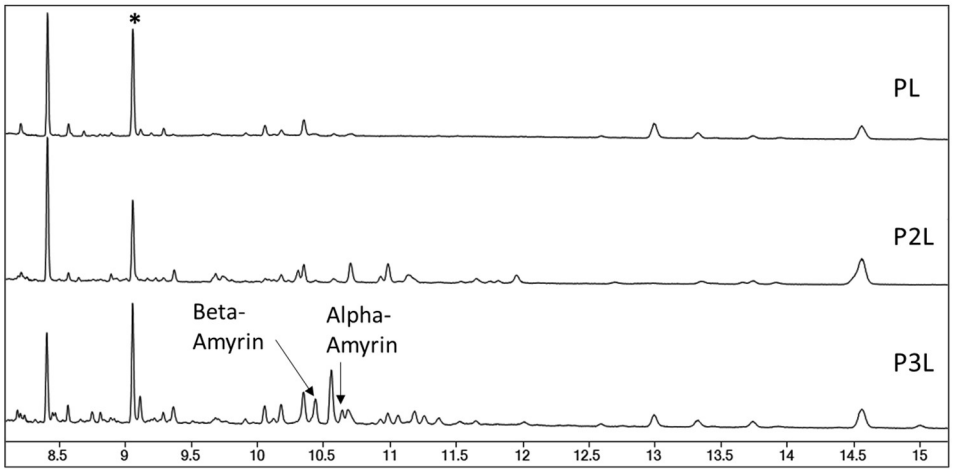


Figure 5.4.2 – GC traces (TMS derivatised) comparing metabolomic profiles of leaf extracts from three different *A. pavia* trees. * Internal standard (coprostanol 100 ppm) visible at 9.1 mins. Peaks likely to be triterpenoids are not detected in samples from pavia 1 and pavia 2, but the trace of pavia 3 contains small peaks which correspond to beta-amyrin (at 10.4 mins) and alpha-amyrin (at 10.7 mins). The mass spectra of these match those of commercial samples exactly.

5.5 – Metabolic Profiling of *Aesculus* Flowers

Flower samples were collected from the same four species, and extracts of these also analysed by LC-MS and GC-MS. LC traces for flower samples are shown in figure 5.5.1. The trace for *A. pavia* flowers contains peaks with similar retention times to those of the aescins. Analysis of the mass spectra of these peaks shows that, whilst they do not match the observed spectra for aescin I, they do closely match the reported spectra of aescin II (Colson, Decroo et al. 2019). The mass spectra provided for one of these peaks A(**1**) shows a highly characteristic peak in the negative mode at m/z = 1099, which corresponds to a peak for $[M\text{-}H]^-$. Other peaks in the positive mode support this, for instance loss of a pentose at [3] (aescin II being C'2-xylose instead of C'2-glucose), a CDE* ring fragment peak at [1] identical to that for aescin I, or an ABC* ring fragment peak with loss of a hexose at [2]. This assignment cannot be compared directly against a standard, as aescin II was not identified in the commercial sample used. However the overlap of the putative aescin II peak with the aescin Ia peak in the standard could suggest the relatively lower-abundance aescin II might be concealed in the trace of the standard. Indeed, the retention times of aescin Ia and IIa (likewise Ib and IIb) have previously been reported as near-identical (Colson, Decroo et al. 2019).

GC-MS analysis of flower tissue extracts did not reveal the presence of aescin intermediates or other known triterpenoids. This is consistent with earlier literature; although some metabolomic investigation of *Aesculus* flowers has previously been carried

out, and carotenoids and flavonoids identified (Deli, Matus et al. 2000, Dudek-Makuch and Matławska 2011), saponins have not been reported.

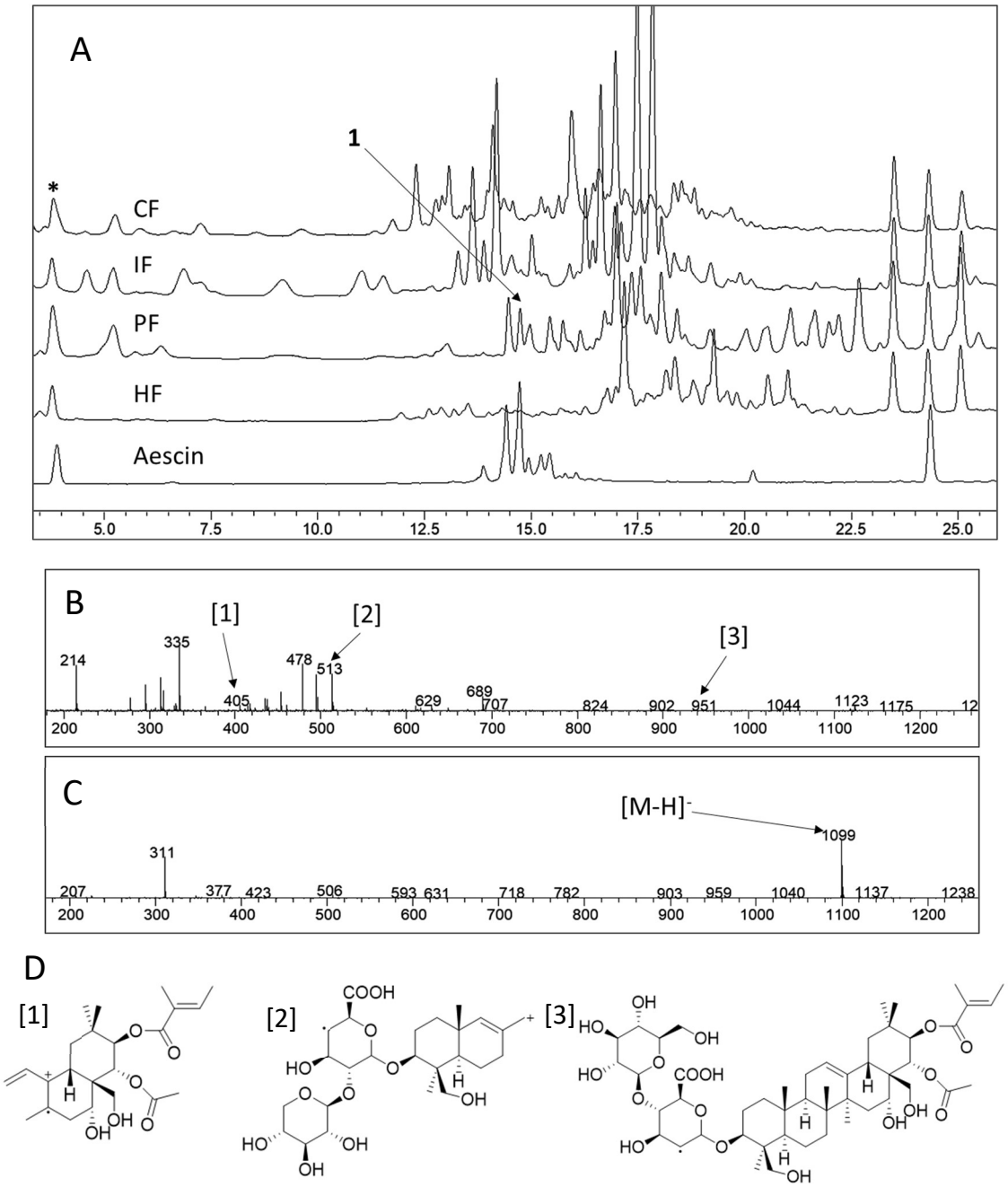


Figure 5.5.1 – LC-MS analysis of flower extracts from four *Aesculus* species detects possible aescin II in *A. pavia* flowers. A: LC traces of four flower samples compared to an aescin standard. * Internal standard (digoxin 20 ppm) seen at 3.9 mins. A peak at **1** appears to be an aescin type saponin. B: Mass spectrum (positive mode) of peak 1 in A. Fragmentation peaks are observed characteristic of aescin derivatives, structures set out in D. C: Negative mode mass spectrum of same; the obvious $[M-H]^-$ peak is seen at m/z 1099, indicative of aescin II saponins, which have lesser mass due to replacement of a hexose with xylose (a pentose). D: Predicted structures of fragmentation peaks seen in B. [1] CDE ring fragment of aescin aglycone, [2] ABC ring fragment with loss of hexose, [3] loss of pentose from aescin II.

5.6 – Metabolomic Profiling of *Aesculus* Fruits

Immature fruits (conkers) were collected from *A. pavia* and *A. hippocastanum* for analysis – immature fruits were used, as mature fruits partially dry out, become lignified, and are likely to be less metabolically active. Fruits were split into three tissue types: the shell or capsule, the outer conker or seed coat, and the inner conker or seed. Comparison between these three tissue types from the same fruit revealed significant differences in their overall metabolic profiles, as shown in LC traces in figure 5.6.1 of extracts from *A. pavia* fruit.

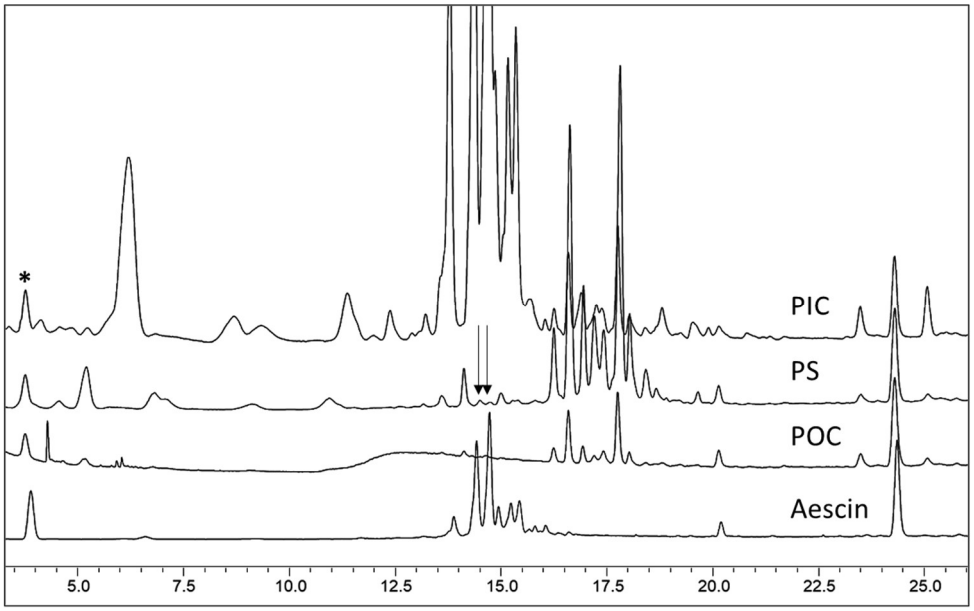


Figure 5.6.1 – LC traces of extracts from different tissues of the fruit of *A. pavia* show varied metabolic profiles. * Internal standard (digoxin 20 ppm) visible at 3.9 mins. The inner conker PIC shows very high concentrations of saponins which exceed the maximum limit of the CAD detector employed, whilst none is detected in the outer conker POC. Small peaks in the trace of the shell extract PS (marked with arrows) appear to correspond to aescin – see figure 5.6.2.

These traces show that the inner conker of *A. pavia* contains very large quantities of aescin-like saponins. By comparison, only very small quantities of aescin-type saponins are detected in the shell, and none in the outer conker. The scale of the difference between the metabolic profiles of these tissues was not previously suspected – though the seeds of *Aesculus* are by a long way the most frequently studied, separate profiling of the different tissues within the fruit has not previously been reported. The apparent detection of very small aescin-like peaks in the shell warranted further investigation – comparison of shell tissue between two different *A. pavia* trees was carried out, as in figure 5.6.2. These small saponin peaks are present in extracts of both trees. Mass spectra of these peaks,

though towards the limit of detection of the instrument, is able to assign these as aescin-type saponins, with a clear aglycone peak the same as aescin. Fragmentation peaks in one of these spectra points to the possible presence of a pentose in the sugar chain of these saponins – potentially making these aescin II-type molecules. Interestingly, analysis of samples from *A. hippocastanum* shells failed to detect any saponins like these observed in *A. pavia*.

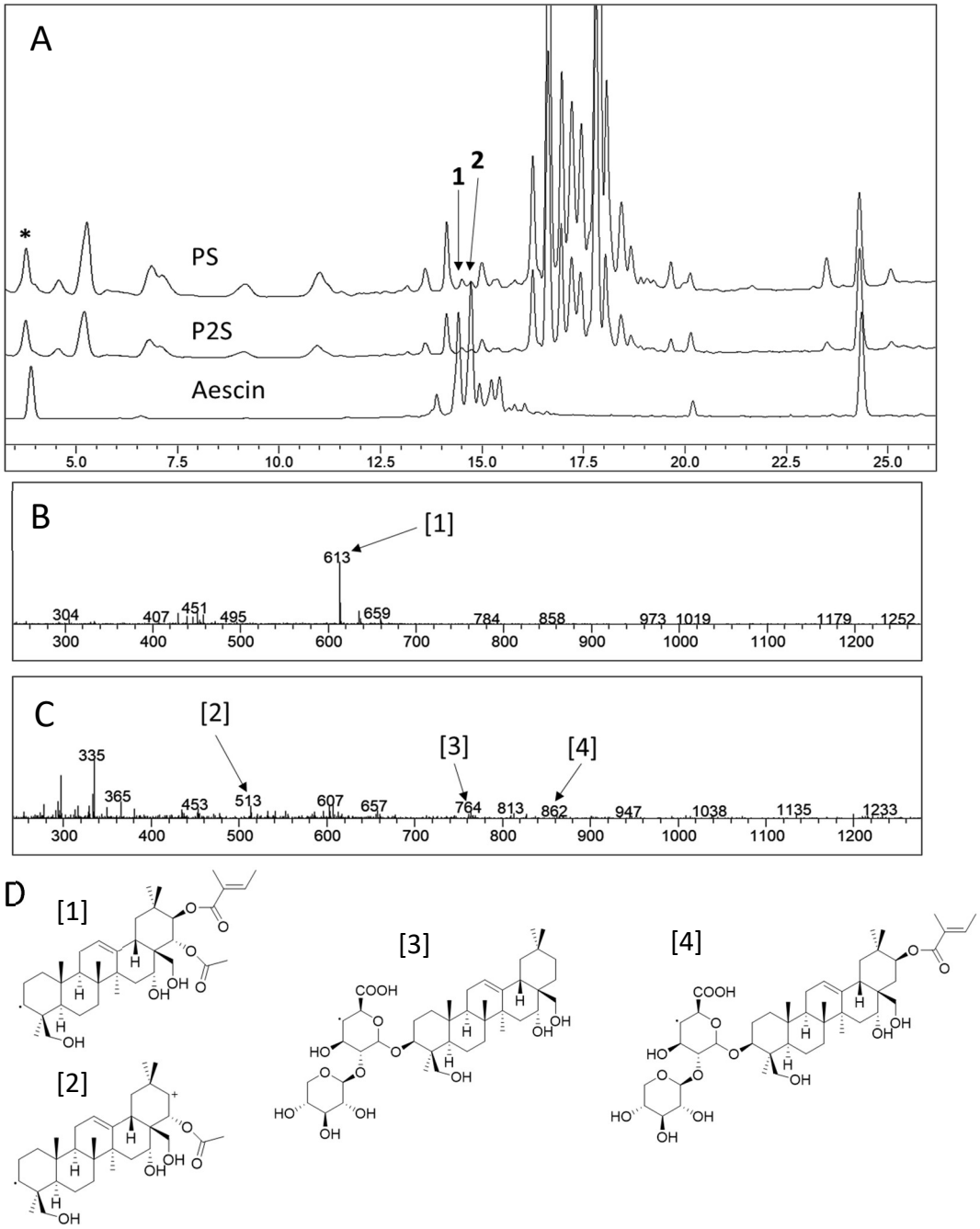


Figure 5.6.2 – Small quantities of aescin-type saponins are detected in the shells of *A. pavia*. A: LC traces of extracts of shell tissue from two *A. pavia* trees. * Internal standard (digoxin 20 ppm) visible at 3.9 mins. Peaks at 1 and 2 correspond to aescin-type saponins, possibly aescin II. B & C: mass spectra (positive mode) of peaks 1 and 2 respectively. A clear aglycone peak at [1] in B supports an aescin-related compound. Peaks [3] and [4] in C suggest a sugar chain with a pentose, as in aescin II. D: possible structures for the fragmentation peaks seen in B and C. [1] aescin aglycone [2] aescin aglycone with loss of tigloyl/angeloyl [3] aescin II with loss of hexose and acyl groups [4] aescin II with loss of hexose and acetyl.

As seen in figure 5.6.1, the saponin content of the inner conkers of *A. pavia* is so high that it exceeds the upper limit of detection of the CAD detector used. A similar result is seen with inner conker extracts of *A. hippocastanum*. To better compare the two, extracts

were diluted tenfold. LC traces are shown in figure 5.6.3. These show very high levels of accumulation of saponins in both species, with the mass spectra showing the same key peaks observed in the spectra of aescin Ia in figure 5.3.1. Interestingly, the negative mode mass spectrum of peak X in the trace of the inner conker of *A. hippocastanum* shows the $[M-H]^-$ peaks of both aescin Ia and aescin IIa (m/z = 1129 and 1099 respectively), here apparently co-eluting. There appears to be a slightly greater diversity of saponins in *A. hippocastanum* than in *A. pavia*, but the overall profile of both is fairly consistent with that of the aescin standard. The very high levels of accumulation here are very much in accordance with the literature, where it is described as forming up to 6- 7 % of the seed by dry weight (Hostettmann and Marston 1995, Ćalić-Dragosavac, Zdravković-Korać et al. 2010).

Analysis of *Aesculus* fruit tissues was also carried out by GC-MS of EtOAc extracts, again detecting very small traces of beta-amyrin in both *A. pavia* and *A. hippocastanum* inner conker samples. Neither beta- nor alpha-amyrin could be detected in shell samples. The detection of beta-amyrin in these tissues is not surprising – though it does not appear to have previously been reported, all saponins isolated from across the genus are derived from beta-amyrin as a scaffold (Zhang, Li et al. 2010), so it is likely to be an intermediate in aescin biosynthesis.

The clear variation in metabolic profiles between the tissues analysed is highly informative to the design of transcriptome sequencing experiments. Given that within the same species some tissues are devoid of saponins (leaf tissues, leaf veins, outer conker), some accumulate only very small quantities of a limited variety of saponins (*A. pavia* flowers, shell tissue), and the inner conker tissues accumulate large quantities of a wide range of aescin-like saponins, it would appear that aescin biosynthesis is highly localised to certain organs in *Aesculus*. These findings now provide the basis for selecting non-producing and producing tissues for transcriptome analysis.

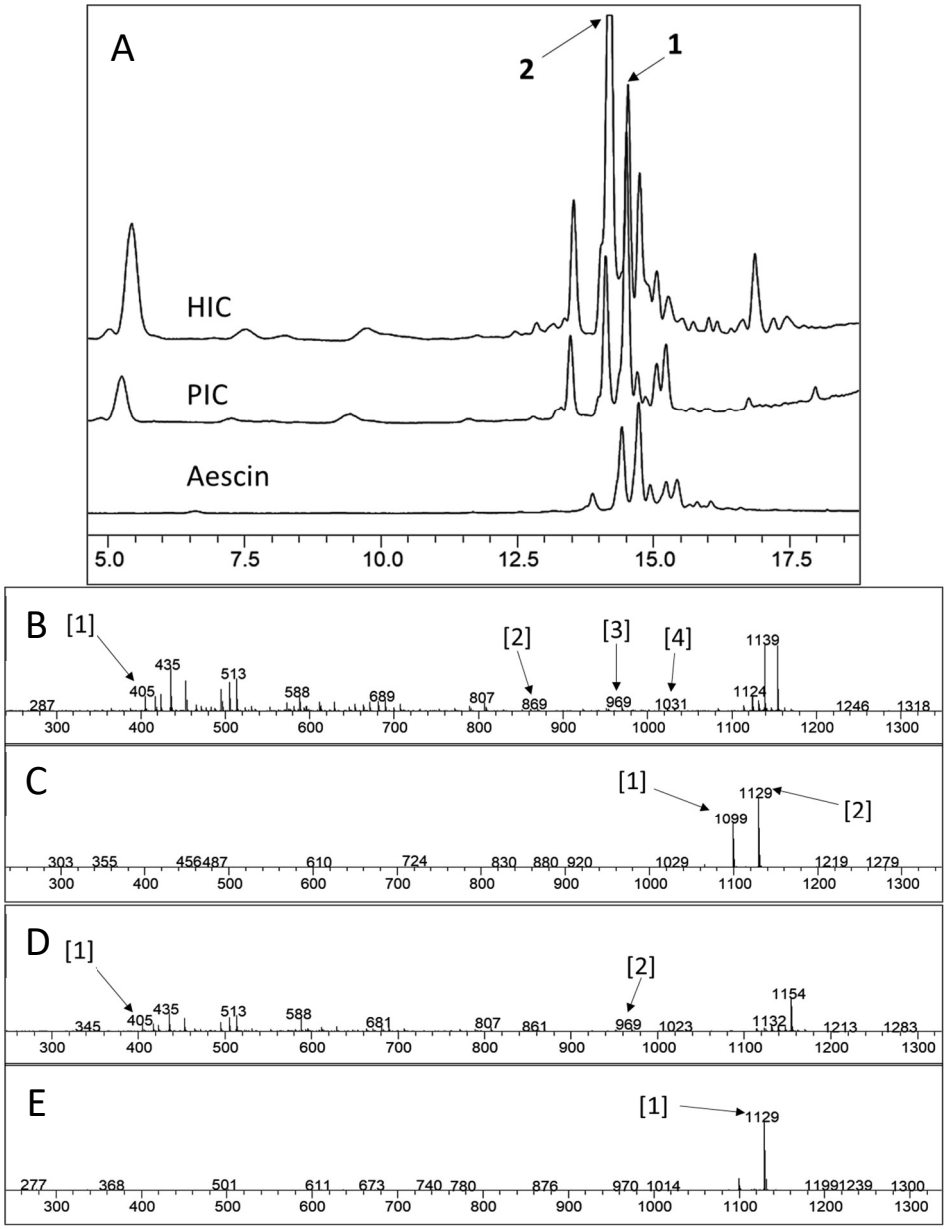


Figure 5.6.3 – Saponins are highly accumulated in the inner conker of both *A. pavia* and *A. hippocastanum*, with a similar profile of aescins to that of commercial standards. A: Detail from LC traces of diluted inner conker extracts of *A. hippocastanum*, *A. pavia*, and an aescin standard (250 ppm). Peaks 1 and 2 correspond to aescins Ia and Ib respectively. B & C: Mass spectra (positive & negative modes) of peak 1 from *A. hippocastanum* extract in A. Key peaks: B[1] CDE* ring fragment of aglycone; B[2] aescin I loss of hexose and tigloyl; B[3] aescin I loss of hexose; B[4] aescin I loss of tigloyl; C[1] aescin II M-H; C[2] aescin I M-H. D & E: Mass spectra (positive & negative modes) of peak 1 from *A. pavia* extract in A. Key peaks: D[1] CDE* ring fragment of aglycone; D[2] aescin I loss of hexose; E[1] aescin I M-H.

5.7 – RNAseq Experiments in *A. pavia* and *A. hippocastanum*

The metabolic profiling reported in the previous sections identified marked differences in accumulation of aescins between different tissues of *Aesculus* species. This opens the way to identify candidate aescin pathway genes based on differential expression in producing and non-producing tissues. The availability of these transcriptome resources would also allow identification of genes that are co-expressed with the previously identified active biosynthetic gene *ApCYP1* using co-expression analysis.

Five tissues were selected for sequencing: *A. pavia* leaf (no accumulation), shell (very slight accumulation of some saponins), inner conker (high levels of accumulation of saponins), and *A. hippocastanum* leaf (no accumulation) and inner conker (high levels of accumulation). RNA was extracted from the five samples, using the trees Pavia 1 (Norwich) and Hippo 1 (NBI site). The conkers and shell proved recalcitrant to RNA extraction using several methods, including Qiagen RNeasy, Invitrogen, and Trizol extraction, but were successfully extracted using the Sigma Spectrum protocol (see 9.1.1) designed for tissues with high secondary metabolite content. This extraction included treatment with DNase to reduce DNA contamination of samples. RNA integrity was assessed by EI through generation of virtual electropherograms and RNA quality scores before samples were barcoded and pooled for sequencing on a single lane of a NovaSeq 6000 SP flow cell with 150 PE reads. The number of reads generated for each sample is listed in table 5.7.1 – the coverage across the two species is broadly comparable.

SPECIES & TISSUE SAMPLE	NO. OF READS
A. pavia – Inner conker	100,282,719
A. pavia – Shell	125,465,446
A. pavia – Leaf	96,709,349
A. hippocastanum – Inner conker	109,828,628
A. hippocastanum – Leaf	93,021,617

Table 5.7.1 – The number of paired-end reads generated through RNAseq carried out by EI of each of the five samples. Fewer reads are generated for leaf tissue than for inner conker or shell tissue, coverage across the two species is similar.

5.8 – Transcriptome Assembly

Two separate transcriptomes were assembled by the Earlham Institute from these reads – one for *A. pavia* and one for *A. hippocastanum*. The assembly was carried out using Trinity (v2.8.5) (Haas, Papanicolaou et al. 2013). Summary statistics for the two assemblies

are given in table 5.8.1 below. The transcriptome assembly of *A. pavia* had a larger total number of contigs than that for *A. hippocastanum*, though the *A. pavia* assembly was based on more samples from more tissues. Due to the large difference in assembly size, the L50 figures are not directly comparable, though comparison of the N20, N50, and N80 figures suggests the distribution of contig lengths to be broadly similar between the two species, with slightly longer contigs seemingly observed in the assembly for *A. hippocastanum*.

	Aesculus hippocastanum	*Aesculus pavia*
Total No. Contigs	219308	391268
L50	38739	67150
Min Length (bp)	179	180
N80 (bp)	762	574
N50 (bp)	1964	1591
N20 (bp)	3509	3100
Max Length (bp)	18828	15740

Table 5.8.1 – Summary statistics of the two transcriptome assemblies carried out by the Earlham Institute using Trinity (v2.8.5). A larger number of contigs were identified for *A. pavia*, but the overall distribution of contig lengths (as indicated by N80, N50, and N20 figures) was comparable.

The assembled transcriptomes were then analysed with TransDecoder (v5.5.0) to identify candidate coding regions of the contigs, restricting each transcript to a single open reading frame. Table 5.8.2 shows the summary statistics of this analysis. As with the contigs, there are more predicted CDS in the assembly for *A. pavia* than for *A. hippocastanum*, but the lengths of these CDS are very similar in both assemblies (as seen from N20, N50, and N80, along with min and max length).

	Aesculus hippocastanum	*Aesculus pavia*
Total No. CDS	85502	129428
L50	20445	30136
Min Length (bp)	255	255
N80 (bp)	726	663
N50 (bp)	1389	1329
N20 (bp)	2469	2385
Max Length (bp)	16383	15324

Table 5.8.2 – Summary statistics of the predicted candidate coding regions of the two transcriptome assemblies. Identification of CDS was carried out by EI using TransDecoder (v5.5.0) restricted to one ORF per transcript. More CDS are found in *A. pavia*, though the distribution of CDS length between the two species is very close.

Predicted coding regions were then functionally annotated using AHRD (v3.3.3) to classify sequences by putative protein family. Transcripts from both assemblies were also analysed using Salmon (v0.15.0) (Patro, Duggal et al. 2017) to estimate abundance in each sample, providing expression levels. This was carried out both with and without cross-species normalisation. Further analysis of the transcriptomes was also carried out using Centrifuge (v1.0.3) (Kim, Song et al. 2016) to identify potential contamination of samples with RNA from other species. Centrifuge classifies sequences according to the species of their closest relative in the reference database (here an internal Earlham Institute database), the results of which are charted on the Krona plots in figure 5.8.1. From these it can be seen that the single largest order is the Sapindales, the order to which *Aesculus* belongs. Over three quarters of sequences are identified as from the Viridiplantae, indicating that contamination from bacterial or fungal sequences appears to be low.

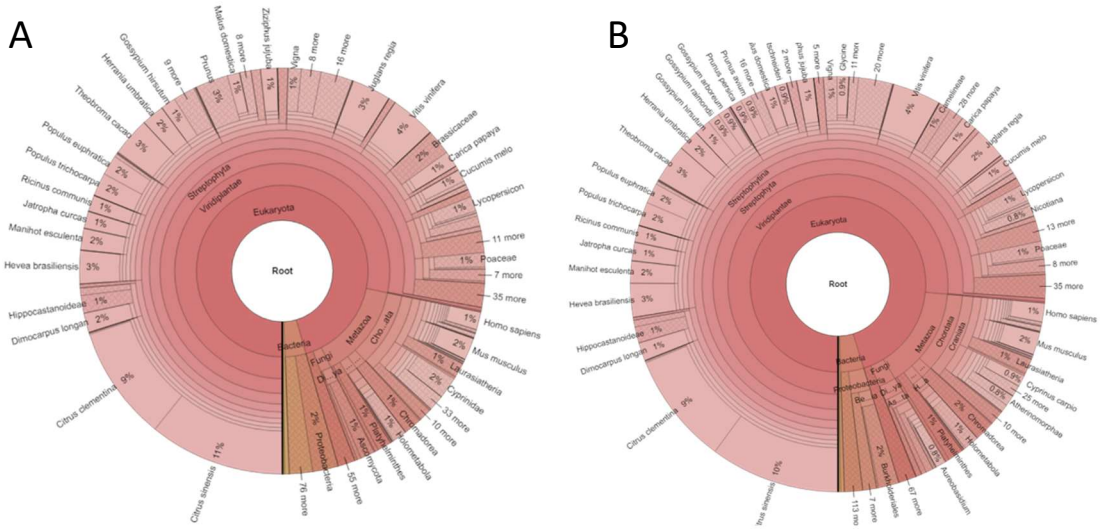


Figure 5.8.1 – Krona plots of the results of taxonomic analyses of sequences from transcriptomic assemblies for *A. hippocastanum* (A) and *A. pavia* (B), carried out by EI using Centrifuge (v1.0.3). Both these assemblies show low levels of obvious contaminants such as bacterial or fungal sequences, with 78 % of sequences in A classified as within the Viridiplantae. Both plots show the order Sapindales (to which *Aesculus* belongs) as the largest single order – 24 % of sequences in A.

The transcriptome assemblies were searched for *ApCYP1* homologues through blastn, for which two local BLAST databases were made using BLAST+ (v2.9.0). The *ApCYP1* sequence from the OneKP transcriptome was recovered from the new *A. pavia* assembly as *ApDN_968c0g1i5*. *ApCYP1* also had a single, very clear homologue in the *A. hippocastanum* transcriptome (*Ah4815c0g1i1*), with 99 % nucleotide sequence similarity, which could support searching for homologues in both transcriptomes as a means of prioritising candidates. The expression levels of the *ApCYP1* gene in the new transcriptome were investigated, and are shown in table 5.8.3 below. *ApCYP1* shows considerably higher levels of expression in conkers compared with shells and leaves, which is promising for use of *ApCYP1* as bait in co-expression, as this pattern is consistent with the metabolite profiling results.

Sequence	**Expression in Conkers**	**Expression in Shell**	**Expression in Leaves**	**Functional Annotation**	**InterPro Identifier**
DN_968c0g1i5	1940.748	37.462	11.354	Cytochrome P450	IPR001128

Table 5.8.3 – TMM cross-species normalised (Robinson and Oshlack 2010) expression levels across different tissues of *A. pavia* of Ap968c0g1i5 (*ApCYP1*). *ApCYP1* is expressed at significantly higher levels in conkers compared to leaves and shells.

5.9 – Conclusions

To better inform RNAseq and validate a transcriptomic approach, tissues from four *Aesculus* species were assessed for saponin content. Significant differences in saponin accumulation were found. For the most part, the results were in line with previous reports in the literature, aescin-type saponins being found in high concentrations in the inner conker but not in leaves. The presence of small amounts of saponin in *A. pavia* flowers and shells had not previously been reported but was not surprising, previous reports having detected small quantities of aescin in tissues other than the seeds, such as bark and buds (Ćalić-Dragosavac, Zdravković-Korać et al. 2010). The apparent lack of saponins in the outer coating of the conkers (the seed coat) was unexpected, as it had been assumed this tissue was likely to be similar to the rest of the conker. Previous reports had not distinguished between the tissues of the conker, and so this may have been overlooked. The detection of traces of alpha- and beta-amyrin in some samples had also not previously been described, but in the case of beta-amyrin could be rationalised as an intermediate to aescin production. The presence of alpha-amyrin was more unusual, as no ursane-type (alpha-amyrin derived) saponins have been reported in the genus. Overall, this profiling gave confidence to the potential for transcriptomics to aid in candidate identification, as the differences in aescin accumulation between tissues are likely to correspond to differences in biosynthetic enzyme expression levels.

RNAseq was thus carried out, including three tissues from *A. pavia* and two from *A. hippocastanum*. The tissues were selected to include both producing and non-producing tissues, in order to enable identification of candidate genes based on differential expression analysis. Inclusion of a second species will also allow for prioritisation of candidates on the basis of presence of a homologue in the other species. Whilst it would have been desirable to include replicates for each tissue sample for a more robust analysis, it is still hoped that the data generated will be of great use to the project. From the RNAseq output, transcriptomes were assembled that were of good size and completeness, with low levels of apparent contamination from other species. The differential expression levels of *ApCYP1* (the C-21β hydroxylase from the previous chapter) in these new transcriptomes indicate that differential expression and co-expression analysis show promise for identification of new candidate aescin biosynthetic genes, to be explored in the following two chapters.

Chapter 6 – Identification and Cloning of OSC and CYP Candidate genes Using *de novo* Transcriptome data

Synopsis

Previous chapters have set out the identification of the first biosynthetic enzyme from the aescin pathway (a beta-amyrin C-21β hydroxylase) (Chapter 4), and the generation of novel transcriptomic resources for *Aesculus* (Chapter 5). Here these new transcriptome resources are used to search for new candidate aescin pathway genes by co-expression analysis, and a second active CYP – ApCYP2, a likely C-24 hydroxylase – from *A. pavia* is identified. An OSC that makes beta-amyrin, the core scaffold of aescin, was also characterised from *A. hippocastanum*. Since the biosynthesis of beta-amyrin is the first committed step in aescin biosynthesis and so this OSC likely represents the first committed pathway enzyme, subsequent effort was focussed on *A. hippocastanum* rather than *A. pavia*. This OSC (AhOSC1) was mixed amyrin synthase, forming both beta- and alpha- amyrin. Homologues to the two CYPs identified from *A. pavia* (AhCYP1 and AhCYP3) were cloned from *A. hippocastanum* and their functions validated. A third *A. hippocastanum* CYP (AhCYP2) which was found to modify beta-amyrin derivatives at two different positions sequentially was also discovered. Isolation and structural determination of the products of these enzymes following co-expression in *N. benthamiana* identified AhCYP2 as a C-22α, C-28 dihydroxylase and AhCYP3 as a C-24 hydroxylase, allowing formation of 21β,22α,24,28-OH beta-amyrin. Inclusion of a C-16α hydroxylase from the triterpenoid toolkit enabled formation of protoaescigenin through transient plant expression, a significant milestone in the elucidation of aescin biosynthesis. Together, these enzymes provide an excellent foundation for the testing of downstream biosynthetic candidates, such as acyltransferases and glycosyltransferases.

Introduction

6.1 – Introduction

Using the OneKP transcriptome of *A. pavia*, an active aescin biosynthetic enzyme (ApCYP1) was identified and characterised as a C-21β hydroxylase (see chapter 4). Coupled with other enzymes from various species in the triterpenoid toolkit, this enabled generation of a number of beta-amyrin derivatives relevant to the biosynthesis of aescin.

To aid in identification and prioritisation of further candidates, new transcriptomic resources were generated for both *A. pavia* and *A. hippocastanum*, as discussed in chapter 5. Crucially, these transcriptomes are for 3 different tissues, so opening up opportunities to identify candidate genes based on differential expression or co-expression, an approach not possible with the limited OneKP transcriptome data for *Aesculus pavia*.

Differential gene expression analysis is one potential way of searching for candidate pathway genes using the new transcriptome data. By searching for sequences that are expressed at higher levels in tissues producing the metabolite of interest when compared non-producing tissues, it is possible to identify candidate genes based on predicted functions in specialised metabolism and correlation of expression levels with metabolite content in different tissues (Bhambhani, Lakhwani et al. 2017, Kim, Jung et al. 2018). There exist a number of bio-informatic tools for this purpose, such as EdgeR (Robinson, McCarthy et al. 2010) and DEseq2 (Love, Huber et al. 2014). However, most of these tools require the RNAseq experiments to have biological replicates in order to enable robust analysis. This was not possible for the *Aesculus* RNAseq carried out in this work due to cost constraints. However, one method of differential expression is able to cope with data lacking replicates – NOISeq (Tarazona, García-Alcalde et al. 2011, Tarazona, Furió-Tarí et al. 2015). The NOISeq-sim program simulates replicates through a multinomial distribution for each condition, enabling differential expression analysis despite lack of replicates.

A further, more stringent means of candidate gene identification is co-expression analysis. This involves identifying genes whose expression is highly correlated with one or more known biosynthetic pathway genes (Rao and Dixon 2019). This is now possible for the aescin pathway, since an early pathway enzyme (ApCYP1) has now been characterised. As with differential expression analysis, multiple bio-informatic tools for co-expression analysis have been developed (Watson 2006, Hou, Ye et al. 2021), but the majority of these also rely upon biological replicates for the raw data in order to provide meaningful results. An alternative to these programs is co-expression analysis through Self-Organising Maps (SOM) analysis (Wehrens and Buydens 2007, Kohonen 2013), which is more flexible in terms of the inputs required.

A putative biosynthetic pathway for aescin was proposed in chapter 2. This predicted formation of beta-amyrin by an OSC (Thimmappa, Geisler et al. 2014) followed by a sequence of five hydroxylations, likely carried out by CYPs (Hamberger and Bak 2013,

Seki, Tamura et al. 2015). Of these hydroxylations, it was predicted that C-21β, C-22α, and C-28 were likely to come first, followed by C-16α and then C-24. C-21β hydroxylation is now accounted for, which leaves four sequential hydroxylations to elucidate. These are most likely to be carried out by four separate CYPs, although it is possible that fewer are required since CYPs capable of more than one transformation have previously been reported (Geisler, Hughes et al. 2013, Zhou, Ma et al. 2016). Indeed, up to four sequential oxygenations carried out by a single CYP have been reported from sesquiterpenoid biosynthetic pathways in *Fusarium* (Tokai, Koshino et al. 2007).

Of the four remaining hydroxylations, C-22α and C-28 are the priorities. These are the two positions needed to for the predicted minimum scaffold for ACT testing (see figure 6.1.1). They are also the two hydroxylations that could not be achieved using the existing triterpenoid toolkit. The ability to engineer the minimal scaffold 21β,22α,28-OH beta-amyrin will be crucial, as without it comprehensive evaluation of the functions of downstream candidate genes for ACTs and glycosyltransferases will not be possible. If the *Aesculus* CYPs responsible for hydroxylation at C-16α or C-24 are not found, however, it should still be possible to form protoaescigenin using the toolkit, which would still allow for investigation of the pathway downstream. The work presented in this chapter focusses on use of the new transcriptomic resources to identify and characterise the OSC and the remaining four CYPs, with the goal of forming the minimum scaffold and protoaescigenin for testing of downstream enzymes.

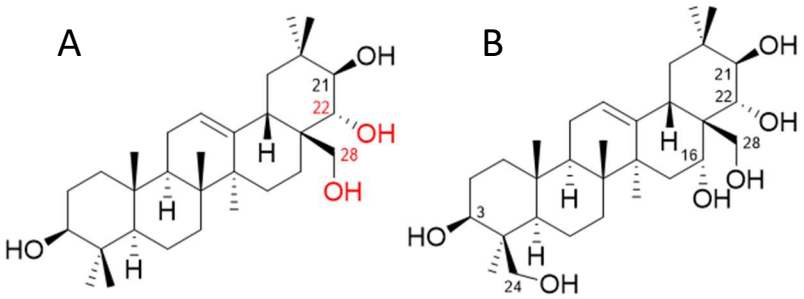


Figure 6.1.1 – The chemical structures of the target molecules for this work. A: 21β,22α,28-OH beta-amyrin, the predicted required substrate for acylation and glycosylation in the aescin pathway. Marked in red are hydroxylations at C-22α and C-28, which cannot be formed using the toolkit and so must employ *Aesculus* CYPs. B: protoaescigenin, the full des-acyl aglycone of aescin.

6.2 – Methods

Differential expression analysis of RNAseq data was carried out in R using NOISeq (v2.38.0) (Tarazona, Furió-Tarí et al. 2015), using the NOISeq-sim program selecting for 5 simulated replicates for each condition, TMM normalised reads, and probability threshold of 0.95. Co-expression analysis was carried out by Anastasia Orme (formerly of the Osbourn Lab, JIC) through generation of self-organising maps (SOMs) (Wehrens and Buydens 2007, Kohonen 2013) in R, using *ApCYP1* as bait. BLAST searches were carried out with BLAST+ (v2.9.0) using the local BLAST databases for *A. pavia* and *A. hippocastanum* generated in chapter 5.

Primers for candidate sequences were obtained from Merck (primers used detailed in 9.1.3), with the exception of *Ah1968c0g1i4*, *Ah3545c0g2i4*, and *Ah15579c0g1i1*, which were obtained as synthetic genes from IDT – sequences for these are provided in Appendix C. RNA was extracted from conker tissue for both *A. pavia* and *A. hippocastanum* as per 9.1.1 and cDNA synthesised according to 9.1.2. Amplification of candidate sequences was carried out using the Q5 protocol (9.1.3) before gateway cloning into expression vectors pEAQ-HT-DEST (Sainsbury, Thuenemann et al. 2009) (9.1.4) and transformation into *A. tumefaciens* LBA 4404 according to 9.1.7. LBA 4404 strains containing expression constructs for GFP, tHMGR, SAD1, CYP716A12, and CYP87D16 were obtained as glycerol stocks from the triterpenoid toolkit (see chapter 3). Protein sequences of cloned candidates are provided in Appendix B.

Heterologous expression in *N. benthamiana* was carried out according to 9.2.1, and the harvested leaf tissue treated as per 9.3.1 or 9.3.2, depending on whether GC-MS or LC-MS analysis of extracts was carried out. Conditions for GC-MS of derivatised extracts were those described in 9.3.4, and conditions for LC-MS were those in 9.3.5 using gradient B. Large-scale vacuum infiltration experiments were carried out using 100 *N. benthamiana* plants for each gene combination according to 9.2.2 and dried leaf material extracted through pressurised solvent extraction with EtOAC as per 9.3.3. Compounds were purified through flash column chromatography as described in 9.3.7, using chlorosulphonic acid and vanillin as visualising agents (9.3.8) and NMR carried out in pyridine D-5 following the general considerations set out in 9.3.9. The NMR referred to here can be found in Appendices D3-D5. Commercial standards of lupeol, lanosterol, alpha-, and beta-amyrin were obtained from Extrasynthese.

6.3 – Differential Expression & Co-Expression Analysis of *A. pavia* Transcriptome

Differential expression analysis was carried out using NOISeq to compare gene expression levels in different *A. pavia* tissues. Figure 6.3.1 shows plots comparing expression levels in conkers to shell and leaf tissue. At a probability threshold of 0.95, 9,352 genes were determined to be expressed at at least ten-fold higher levels in conkers compared to shell tissue, and 23,125 genes in conkers compared to leaf tissue. To complement this, co-expression analysis was carried out by Anastasia Orme, using SOM analysis with *ApCYP1* as bait. Figure 6.3.2 shows the clusters of genes with similar expression patterns generated by the SOM analysis. This analysis identified 3,470 genes clustered with *ApCYP1*, 2,072 of which were clustered with over 95% probability.

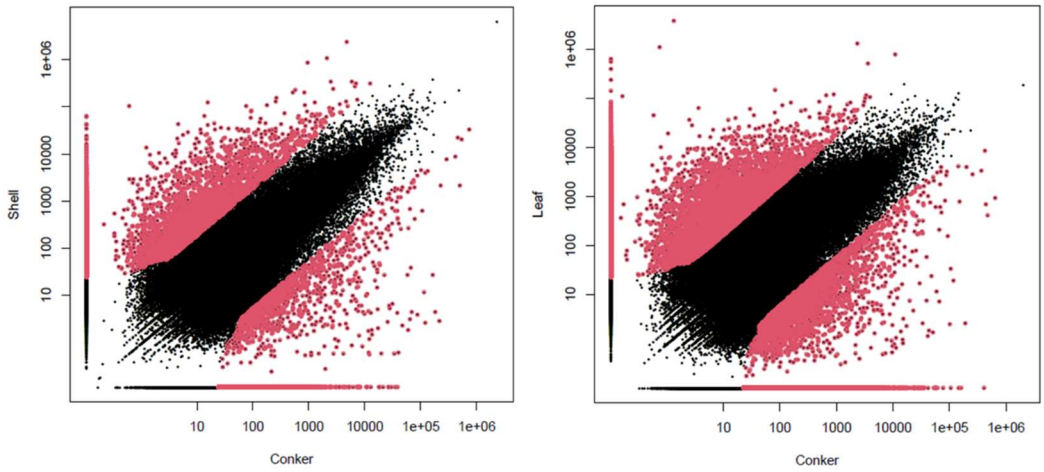


Figure 6.3.1 – Plots showing the distribution of genes by expression levels in different *A. pavia* tissues, generated using NOISeq. Left: expression in conkers against shell tissue. Right: expression in conkers against leaf tissue. In pink in the lower right-hand corner of both plots are those sequences expressed at levels at least ten-fold greater in conkers than in shells or leaves.

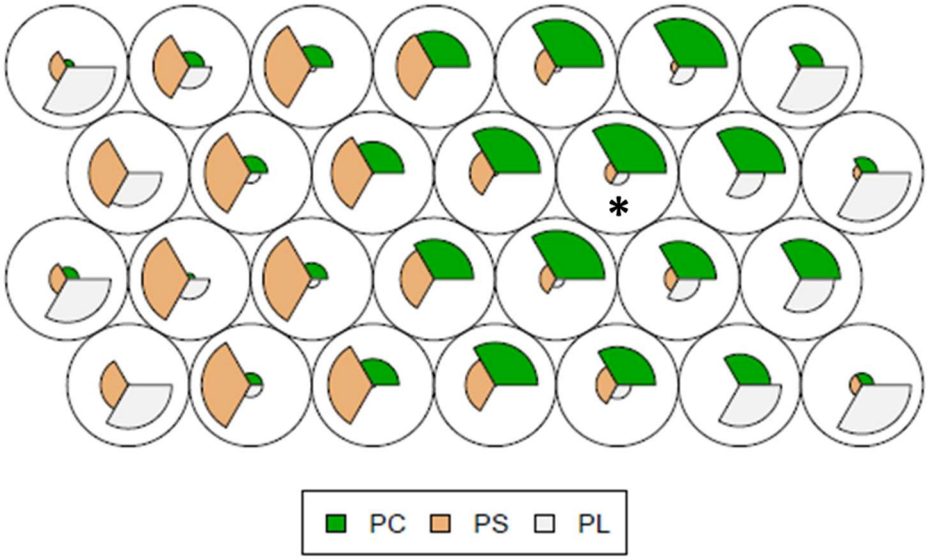


Figure 6.3.2 – A plot of the clusters generated by SOM analysis of the *A. pavia* transcriptome carried out by Anastasia Orme. Genes are sorted into the clusters that most closely match their expression levels across the three tissues (PC – conker, PS – shell, PL – leaf). The cluster with the asterisk * is that containing *ApCYP1*, with high expression levels in the conker and very low expression levels in both shell and leaf tissue. 3,470 other genes were found associated with this cluster.

From the differential expression and co-expression analysis, the functional annotation of sequences (carried out using AHRD (v3.3.3) – see chapter 5) was used to search for predicted OSC and CYP genes. CYP genes were found annotated either as "Cytochrome p450" or as "Beta-amyrin 28-oxidase". In addition to the candidates identified through differential expression and co-expression, tBLASTn searches were also carried out to search for sequences similar to CYP716A12, CYP716A44, and CYP93E1, CYPs that carry out transformations relevant to the biosynthesis of aescin (see chapter 3). Candidates were evaluated to exclude truncated sequences or those missing conserved domains, and priority candidates were identified as listed in table 6.3.1 below.

CANDIDATE GENE	DIFFERENTIALLY EXPRESSED IN CONKERS	% CLUSTERED WITH *APCYP1* (SOM)	CLONED AND TESTED	NOTES
DN5582C0G1I10	✓	100	✓	ApOSC1
DN968C0G1I5	✓	Bait	✓	ApCYP1
DN225C0G1I4	✓	100		Homologue active as AhCYP2
DN2476C0G1I4		-		Sequence similarity to CYP93E1
DN4078C1G1I4		39	✓	Sequence similarity to CYP93E1
DN4414C1G1I4	✓	100	✓	Putative CYP 82
DN8439C0G1I3	✓	100	✓	Putative CYP 92
DN15243C0G1I1 (APCYP2)	✓	100	✓	Potential activity – homologue active as AhCYP3
DN16823C0G1I2	✓	100		Putative CYP 82
DN24067C1G1I2		-		Sequence similarity to CYP716A44
DN40570C0G1I1	✓	100	✓	Putative CYP 71
DN103483C0G1I1		-		Sequence similarity to CYP716A44

Table 6.3.1 – The initial priority OSC and CYP candidates identified from *A. pavia* using both differential and co-expression analysis. An OSC and 10 CYPs (aside ApCYP1) were identified as strong candidates. The OSC and five of the CYPs were cloned and tested, one CYP of which appeared to demonstrate activity (ApCYP2).

As can been seen from Table 6.3.1, only one full-length OSC candidate was identified. This is a homologue to ApOSC1, which was identified from the OneKP transcriptome (see chapter 4). The sequence from this new transcriptome has 98.7 % nucleotide sequence similarity to the sequence from the OneKP transcriptome, apart from having an additional 23 amino acid extension at the C-terminus. Primers were designed for this new sequence, and the coding sequence cloned and transiently expressed in *N. benthamiana*. However, it failed to demonstrate activity.

Of the 10 CYP candidates (not including ApCYP1 which was already characterised), five were successfully cloned and evaluated in combinatorial assays with toolkit biosynthetic enzymes. GC-MS analysis of leaf extracts revealed that one of these candidates, DN15243c0g1i1 (ApCYP2) showed activity when co-expressed with tHMGR, SAD1, ApCYP1 and CYP 716A12, as shown in figure 6.3.3 below. ApCYP2 shares 97% nucleotide sequence similarity to a truncated (685 bp) sequence from the OneKP transcriptome, and so represents a CYP which would have been missed from previous searches.

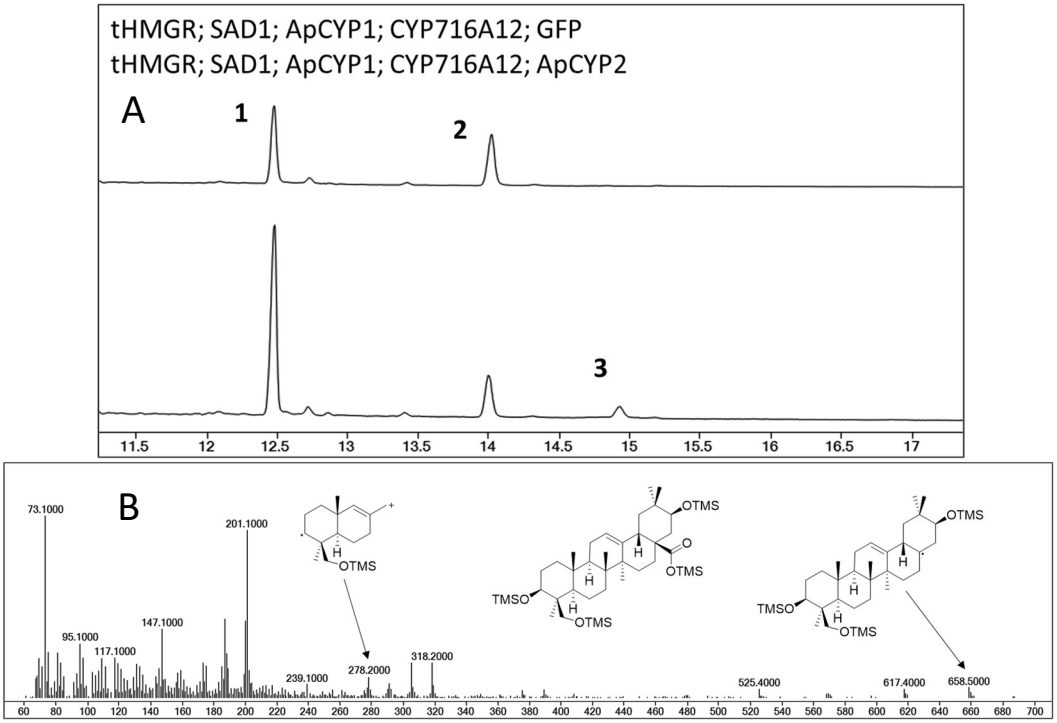


Figure 6.3.3 – ApCYP2 appears to hydroxylate 21β-OH oleanolic acid. A: GC traces (TMS derivatised) showing the activity of ApCYP2 on co-expression with tHMGR, SAD1, ApCYP1, & CYP 716A12. (1) 21β-OH beta-amryin (2) 21β-OH oleanolic acid (3) new product seen with ApCYP2. B: Mass spectrum of (3) – the loss of the characteristic ABC* ring fragment peak at m/z 279 suggests modification on this side of the molecule. Its replacement with a new peak at 278 could indicate addition of a leaving group such as an alcohol. A peak at 658 is consistent with decarboxylation of doubly-hydroxylated oleanolic acid derivative.

ApCYP2 is likely responsible for addition of an alcohol to the ABC* ring fragment side of 21β-OH oleanolic acid. The only position on that side of aescin or its intermediates that is hydroxylated is C-24, a position which is already covered by the triterpenoid toolkit through CYP 93E1 (Shibuya, Hoshino et al. 2006). Interestingly, this CYP only showed activity on 21β-OH oleanolic acid, not on any other oxidised beta-amyrin derivatives that could be formed using the toolkit. If ApCYP2 is indeed a C-24 hydroxylase, this observation fits with the putative biosynthetic pathway predicted in chapter 2, which hypothesised that C-24 hydroxylation was one of the later hydroxylations in the pathway, likely requiring a substrate that is already oxidised. Due to the change in focus to *A. hippocastanum*, this CYP was not investigated further. However its *A. hippocastanum* homologue (AhCYP3) was identified and demonstrated C-24 hydroxylating activity (see 6.7).

6.4 – Identification of an active OSC from *A. hippocastanum*

The lack of activity of the sole candidate OSC from the *A. pavia* transcriptome led to a search of the *A. hippocastanum* transcriptome for potential homologues. A BLASTn search using *ApOSC1* returned two full-length OSC sequences in the *A. hippocastanum* assembly – *Ah331c0g1i3* and *Ah331c0g2i1*; these had 83.6% amino acid sequence similarity to one another. Both of these were successfully cloned from *A. hippocastanum* conker cDNA, transformed into *Agrobacterium tumefaciens* and transiently expressed in *N. benthamiana*. Upon co-expression with tHMGR, Ah331c0g1i3 (named AhOSC1) yielded new products that were detectable by GC-MS (see figure 6.4.1 below).

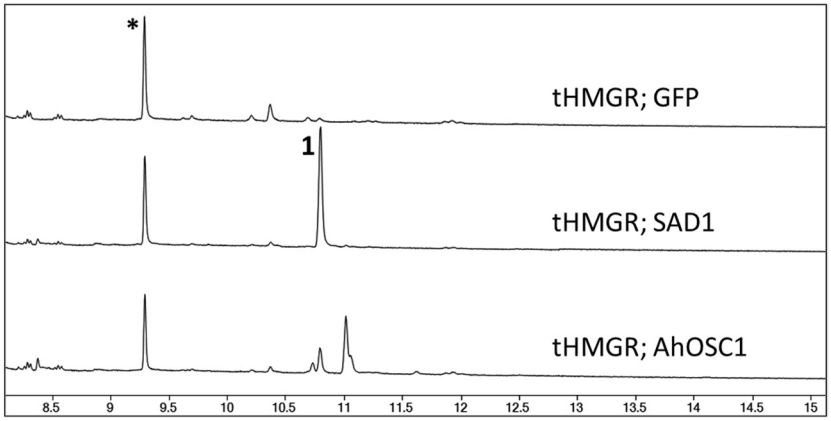


Figure 6.4.1 – GC-MS traces (TMS-derivatised) comparing the activity of AhOSC1 with SAD1 when co-expressed in *N. benthamiana* with tHGMR. * Internal standard (coprostanol 100 ppm) visible at 9.3 mins. Peak at (1) is beta-amyrin. The trace for AhOSC1 shows new peaks in the expected region of OSC products – one of these peaks appears to overlap with (1), and so could be beta-amyrin.

Multiple potential new products are observed in leaf extracts expressing AhOSC1. OSCs forming more than one product have been reported previously (Phillips, Rasbery et al. 2006, Huang, Li et al. 2012, Thimmappa, Geisler et al. 2014, Andre, Legay et al. 2016). To resolve the products of AhOSC1, the GC traces and mass spectra were compared to those of commercial standards of a number of common OSC products: alpha-amyrin, beta-amyrin, lupeol, and lanosterol. Figure 6.4.2 shows traces and spectra matching AhOSC1 products to both alpha- and beta-amyrin.

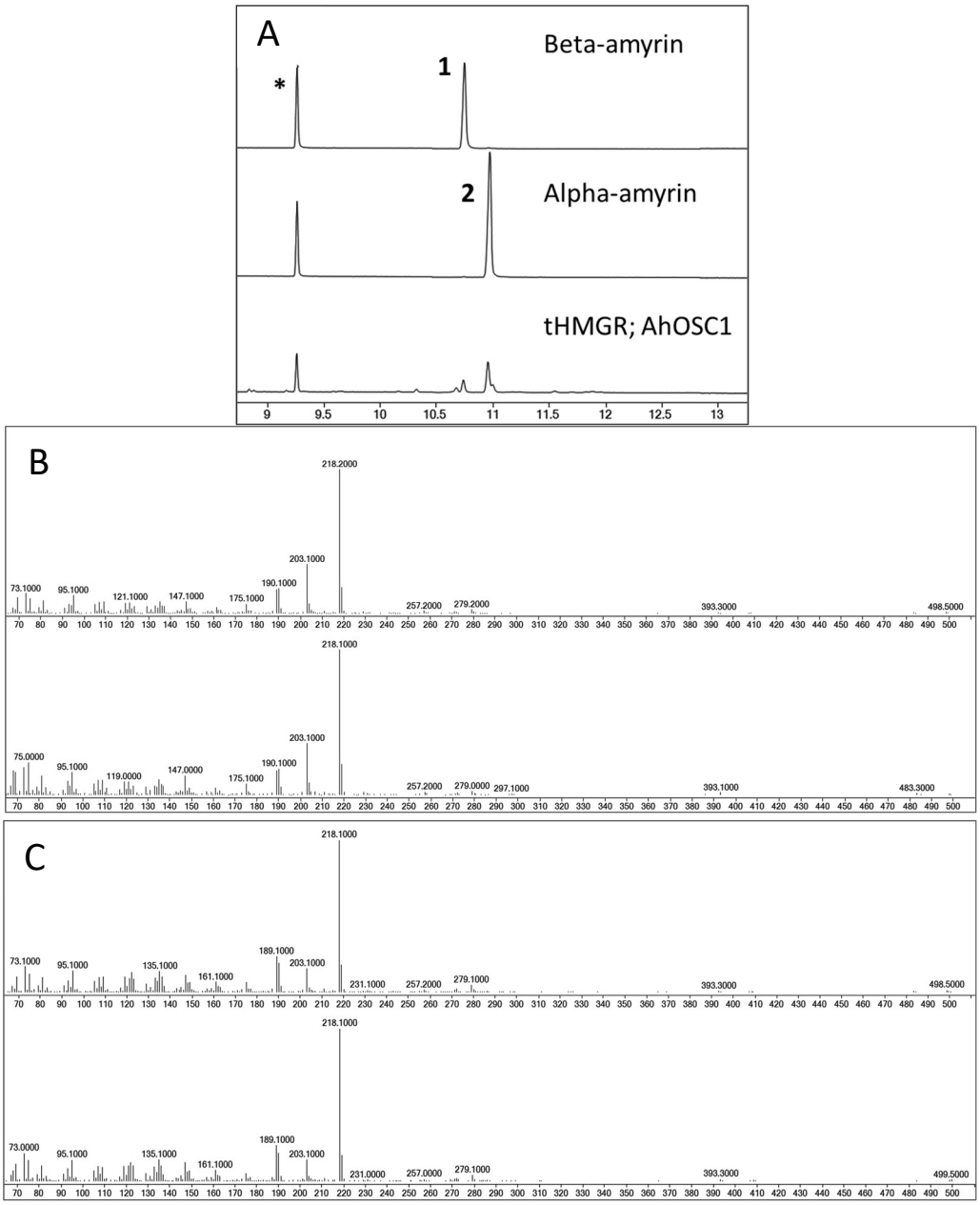


Figure 6.4.2 – The two main products seen in traces for transient expression of AhOSC1 are alpha- and beta-amyrin. A: GC traces (TMS derivatised) comparing the products of AhOSC1 with commercial standards of beta- and alpha-amyrin. (1) beta-amyrin (2) alpha-amyrin. B: Mass spectra comparing peak (1) in the commercial standard (top) and AhOSC1 (below) – these are functionally identical. C: Mass spectra comparing peak (2) in the commercial standard (top) and AhOSC1 (below) – these are also identical.

These experiments confirm that AhOSC1 forms both alpha- and beta-amyrin. Mixed amyrin synthases such as this have previously been reported. Examples include CrAS from *Catharanthus roseus* (Huang, Li et al. 2012), which makes both alpha- and beta-amyrin in a 2.5:1 ratio (which happens to be similar to the ratio of products observed from AhOSC1), and PSM from *Pisum sativum* (Morita, Shibuya et al. 2000). Although alpha-amyrin derivatives are not amongst the major saponins reported from *Aesculus*, traces of alpha-

amyrin were detected in leaf samples from one *A. pavia* tree (chapter 5), which might be explained by a mixed amyrin synthase. With an active OSC from *A. hippocastanum* as the first step of the aescin pathway, subsequent experiments to elucidate the other aescin biosynthetic enzymes focussed on *A. hippocastanum*.

6.5 – Cloning and testing of AhCYP1, an *A. hippocastanum* homologue of ApCYP1

Having decided to focus on pathway elucidation in *A. hippocastanum*, it was important to clone and validate the activity of AhCYP1, the homologue of ApCYP1. ApCYP1 had previously been identified as a beta-amyrin C-21β hydroxylase (see chapter 4), active in combination with toolkit enzymes to form a number of oxidised beta-amyrin derivatives. A BLASTn search (Camacho, Coulouris et al. 2009) of a local blast database of the *A. hippocastanum* transcriptome assembly found the sequence of a homologue with 99 % sequence similarity (*Ah4815c0g1i1*, referred to as AhCYP1). This was cloned from *A. hippocastanum* cDNA, transformed, and transiently expressed in *N. benthamiana* in co-expression experiments to compare the activity of ApCYP1 and AhCYP1. Figure 6.5.1 below shows GC traces comparing the activities of these two enzymes when co-expressed with tHMGR and SAD1. The two enzymes appear to have the same activity as assessed by comparison of the retention times and mass spectra of the product.

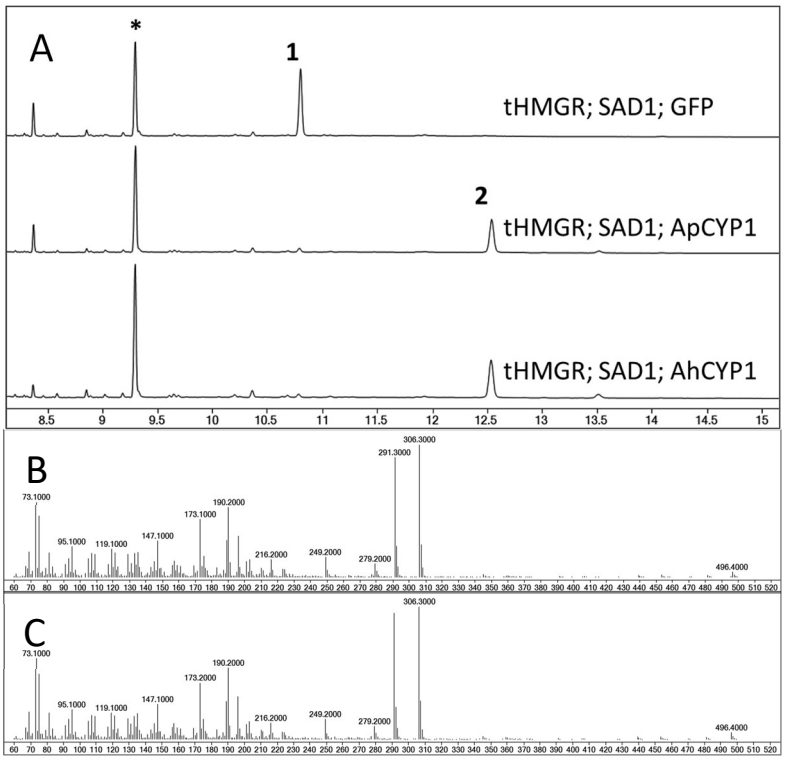


Figure 6.5.1 – GC-MS analysis confirms that ApCYP1 and AhCYP1 have identical activity on beta-amyrin. A: GC traces (TMS-derivatised) comparing the activity of ApCYP1 and AhCYP1 upon co-expression with tHMGR and SAD1. * Internal standard (coprostanol) visible at 9.3 mins. (1) Beta-amyrin (2) 21β-OH beta-amryin. B & C: Mass spectra (positive mode) of peak 2 in the traces for ApCYP1 and AhCYP1 respectively, which appear to be functionally identical.

Interestingly, the expression levels of AhCYP1 in leaves and conkers are the opposite to what might be expected, with significantly greater expression in the leaves than in the conkers. This is at odds with the observed expression levels of ApCYP1 in *A. pavia* tissues, where ApCYP1 is expressed predominantly in the conkers. Similarly, the expression levels of AhOSC1 are also unexpected, with moderate expression in leaves and none in conkers; this is despite AhOSC1 having been successfully amplified and cloned from *A. hippocastanum* conker tissue. Table 6.5.1 compares the expression levels of the two sequences, with TMM cross-species normalised counts. Given this, it may be unwise to rely expression levels in *A. hippocastanum* for identification of subsequent candidates.

GENE	CONKER (TMM)	LEAF (TMM)
AHOSC1	0	13.168
AHCYP1	0.609	15.207

Table 6.5.1 – The expression levels of active *A. hippocastanum* biosynthetic genes is in variance with the expression levels of their homologues in *A. pavia*, with lower or non-existent expression in conkers. Expression levels are TMM-normalised counts-per-million.

6.6 – Identification of AhCYP2 from the *A. hippocastanum* transcriptome

The potential unreliability of the expression levels for *A. hippocastanum* led to use of a different approach for candidate identification, based on sequence similarity to either known active CYPs relevant to aescin biosynthesis, or to promising candidates identified through differential or co-expression in *A. pavia*. Ten initial CYP candidates were chosen, as detailed in table 6.6.1 below.

A. HIPPOCASTANUM CYP	A. PAVIA QUERY SEQUENCE	CLONED & TESTED	NOTES
AH1968C0G1I4	DN2476c0g1i4	✓	Obtained as synthetic gene
AH3545C0G2I4	DN40570c0g1i1	✓	Obtained as synthetic gene
AH3872C1G1I1	DN968c0g1i5		
AH4924C0G1I7	DN4078c1g1i4	✓	
AH8376C0G2I2	DN4414c1g1i4		
AH13762C0G1I2	DN24067c1g1i2	✓	
AH15779C0G1I1	DN16823c0g1i2	✓	Obtained as synthetic gene
AH16676C0G1I1 (AHCYP3)	DN15243c0g1i1	✓	Active – C24 hydroxylase
AH21534C0G2I1 (AHCYP2)	DN225c0g1i4	✓	Active – C22, 28 hydroxylase
AH21534C0G3I1	DN225c0g1i4	✓	

Table 6.6.1 – CYP candidates from *A. hippocastanum* identified through BLASTn searches using *A. pavia* candidates and triterpenoid toolkit enzymes as query sequences. Two of these proved active: AhCYP2 and AhCYP3

One initially very promising candidate from *A. pavia*, *DN225c0g1i4*, that had been annotated as a "Beta-amyrin 28-oxidase" and was strongly co-expressed with *AhCYP1* (but had not been successfully cloned), was found to have two homologues in *A. hippocastanum*. These were both cloned, transformed, and tested, and one shown to have activity in *N. benthamiana* – AhCYP2 (*Ah21534c0g2i1*) GC-MS results for co-expression of AhCYP2 with tHMGR and SAD1 are shown in figure 6.6.1.

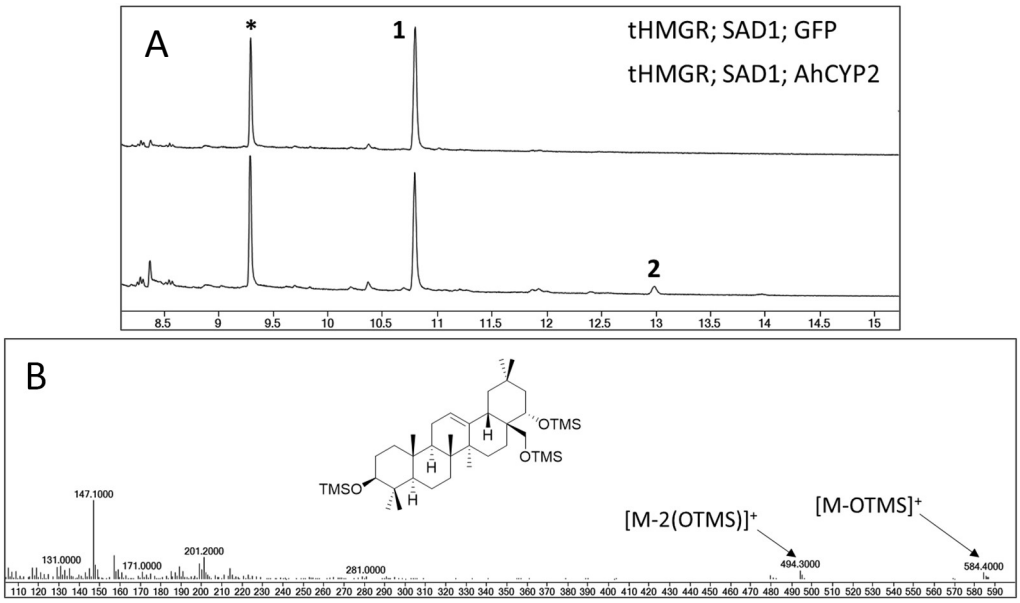


Figure 6.6.1 – GC-MS analysis reveals activity of AhCYP2 on beta-amryin. A: GC traces (TMS-derivatised) comparing co-expression of AhCYP2, tHMGR, and SAD1 with a GFP control. * Internal standard (coprostanol) visible at 9.3 mins. (1) Beta-amyrin (2) new product peak. B: Mass spectrum (positive mode) of peak 2 in A – peaks suggesting successive loss of two derivatised hydroxyl groups could point to a doubly-hydroxylated product such as that depicted.

This experiment showed accumulation of a small quantity of a new product, the mass spectrum of which could indicate a doubly-hydroxylated derivative. To investigate the apparent activity of this CYP further, it was also tested for activity towards 21β-OH beta-amyrin by co-expression with tHMGR, SAD1, and AhCYP1. The GC-MS traces for this experiment are shown in figure 6.6.2.

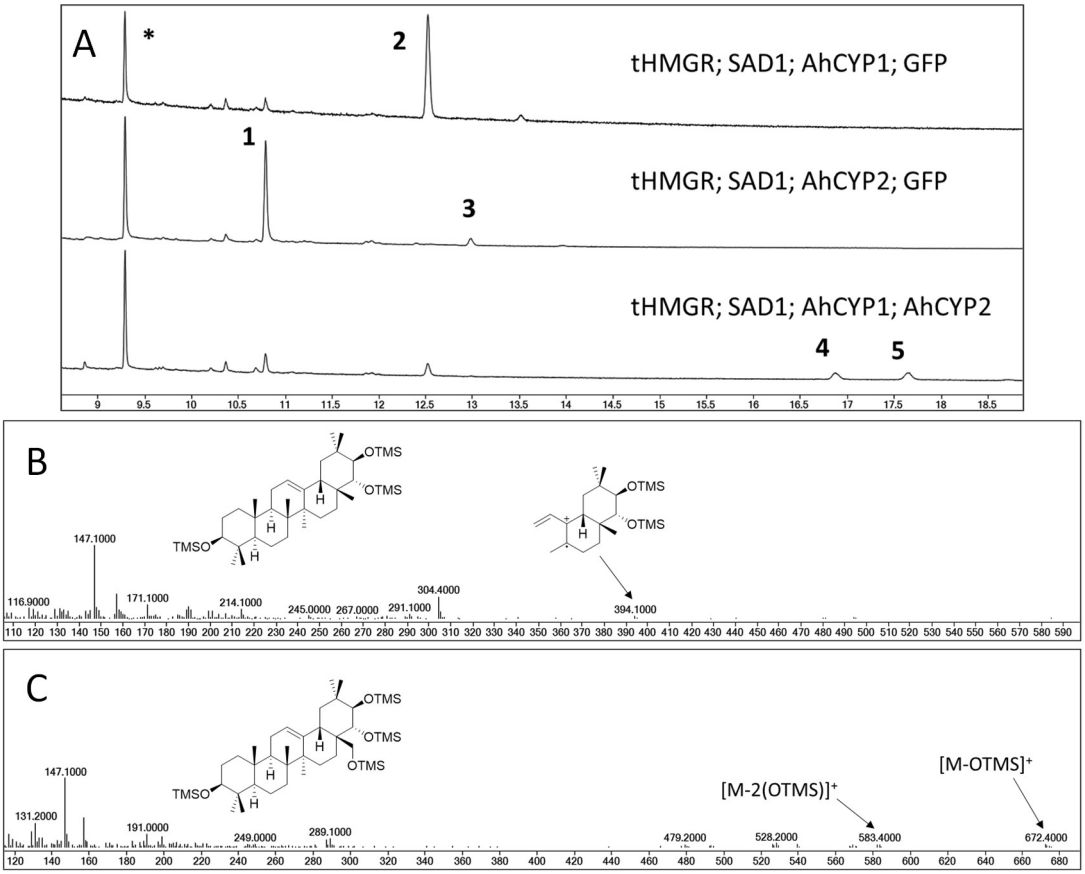


Figure 6.6.2 – AhCYP2 shows apparent double-hydroxylating activity on 21β-OH beta-amyrin. A: GC traces (TMS-derivatised) showing the effects of co-expression of AhCYP1 and AhCYP2 – new product peaks are observed at (4) and (5). * Internal standard (coprostanol) visible at 9.3 mins. (1) Beta-amyrin (2) 21β-OH beta-amyrin (3) AhCYP2 double-hydroxylated beta-amyrin product (4) & (5) new products. B: Mass spectrum (positive mode) of peak 4 in A. A CDE* ring fragment peak at m/z 394 suggest double hydroxylation on the right-hand side of the molecule. C: Mass spectrum (positive mode) of peak 5 in A. Peaks resulting from loss of hydroxyl groups suggest a trebly-hydroxylated beta-amyrin derivative.

AhCYP2 was demonstrated in these experiments to form new products when co-expressed with AhCYP1, the mass spectra suggesting it is capable of adding either one or two hydroxyl groups to 21β-OH beta-amyrin. A double activity like this is not unprecedented – CYPs that oxidise compounds at multiple positions sequentially have been reported previously (Tokai, Koshino et al. 2007, Zhou, Ma et al. 2016). The low levels of accumulation of new products , as detected by GC-MS, warranted further investigation. As these new products are becoming more polar with each additional hydroxyl group, their visibility through GC may be diminished. As such, leaf extracts of these experiments were also analysed by LC-MS, as shown in figure 6.6.3 below.

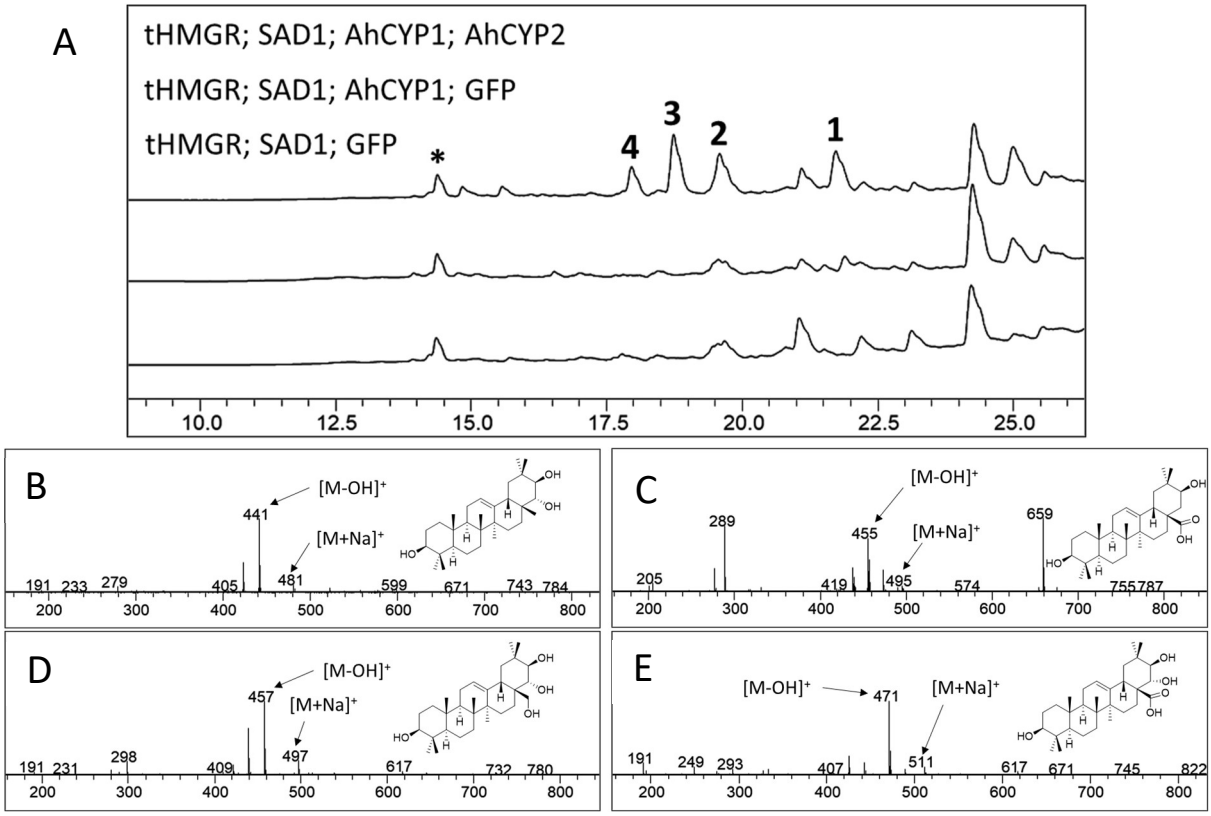


Figure 6.6.3 – LC-MS analysis reveals previously undetected AhCYP2 products. A: LC traces of leaf extracts from AhCYP2 co-expression experiments. * Internal standard (digoxin) seen at 14.4 mins. Whereas GC traces only picked up 2 new products from co-expression of AhCYP2 with AhCYP1 and beta-amyrin, four are detected here. B: Mass spectrum of peak (1) in A. This appears to be a doubly-hydroxylated beta-amyrin derivative. C: Mass spectrum of peak (2) in A. Here, the addition of both an acid and a hydroxyl group to beta-amyrin is indicated. D: Mass spectrum of peak (3) in A. This appears to be a trebly-hydroxylated beta-amyrin derivative, as was seen in figure 6.6.2. E: Mass spectrum of peak (4) in A. Compared to the mass spectrum in D, it would appear that one of the hydroxyl groups is replaced with an acid group.

LC analysis led to detection of new triterpenoid products that were not seen by GC analysis. The mass spectra of these products strongly suggest that AhCYP2 is capable of addition of either two hydroxyl groups or a single -OH and oxidisation up to an acid on 21β-OH beta-amyrin. This is interesting, as it potentially narrows down the positions on which AhCYP2 might be active. Of the positions that are hydroxylated in aescin, oxidation up to an acid is only possible on C-24 and C-28 (both methyl groups). Addition of either two hydroxyl groups, or a hydroxyl group and an acid group to C-21β beta-amyrin was detected, but not two hydroxyl groups and an acid group. This would suggest that the oxidation to an acid occurs instead or subsequently to hydroxylation at the same position. The localisation of hydroxylation to the CDE* ring fragment seen in GC-MS fragmentation, and the lack of activity on oleanolic acid, would favour activity at C-28. Amongst the reported CYPs active on C-28, it is common for a mixture of alcohol, aldehyde, and acid products to be formed

(Fukushima, Seki et al. 2011, Yasumoto, Seki et al. 2017), though C-28 carboxylic acid does not typically form part of the principle saponins isolated from *Aesculus*. Further investigation of this activity through NMR is provided in section 6.8.

6.7 – Identification & Characterisation of AhCYP3

The identification of ApCYP2 as an active enzyme led to investigation of its homologue in *A. hippocastanum*, AhCYP3 (*Ah16676c0g1i1*). AhCYP3 was cloned and co-expressed with tHMGR and SAD1 in *N. benthamiana* as for other CYP candidates, but no activity was detected by either GC or LC analysis. Similar results were observed on co-expression with AhCYP1 with tHMGR and SAD1. However, inclusion of AhCYP2 into these experiments led to detection of activity, as seen in GC-MS results in figure 6.7.1.

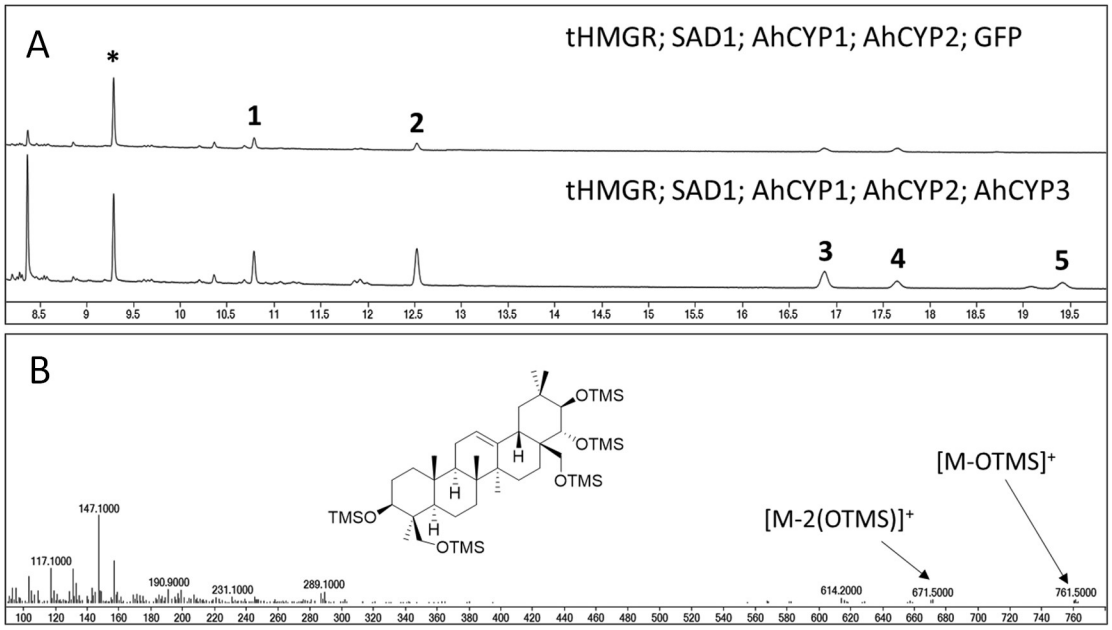


Figure 6.7.1 – AhCYP3 demonstrates activity on co-expression with AhCYP1 & AhCYP2. A: GC traces (TMS-derivatised) showing activity of AhCYP3 on hydroxylated beta-amyrin derivatives. * Internal standard (coprostanol) visible at 9.3 mins. (1) Beta-amyrin (2) 21-OH beta-amryin (3) doubly-hydroxylated beta-amryin (4) trebly-hydroxylated beta-amyrin (5) new product peak. B: Mass spectrum (positive mode) of peak 5 in A – two peaks for loss of successive -OTMS groups suggest a beta-amyrin derivative with four hydroxyl groups.

Detection of a new product from AhCYP3 only upon co-expression with the other two *A. hippocastanum* CYPs suggest the previous modifications may be requirements for its substrate specificity. The mass spectrum of this new product is consistent with a further hydroxylation of the AhCYP1 and AhCYP2 products, for a total of four hydroxylations to beta-amyrin – however, these compounds appear to be towards the limit of detection

through GC-MS, so LC-MS analysis was also carried out. The results of this are shown in figure 6.7.2 below.

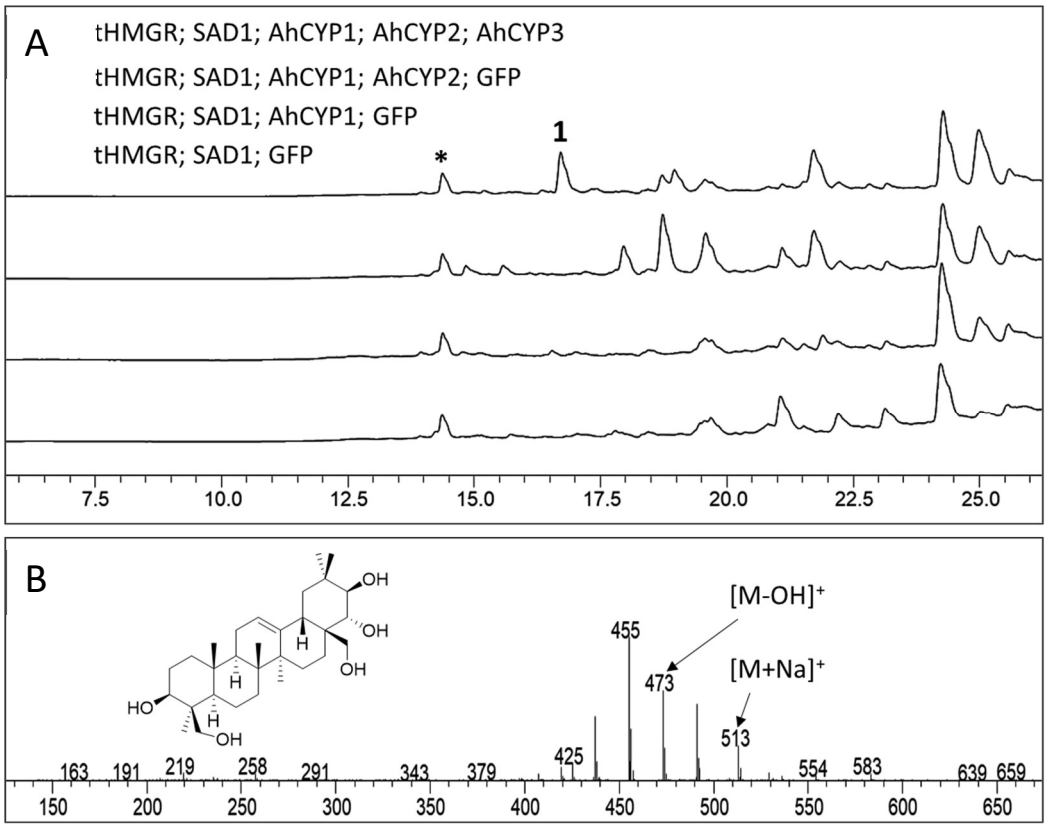


Figure 6.7.2 – On co-expression with tHMGR & SAD1, all three *A. hippocastanum* CYPs make a new product. A: LC traces of leaf extracts from AhCYP3 co-expression experiments. * Internal standard (digoxin) seen at 14.4 mins. Peak at (1) is the product previously observed through GC in figure 6.7.1, though is much more visible here. B: Mass spectrum of peak (1) in A. Peaks for both loss of a hydroxyl group and a sodium adduct are visible, suggestive of a beta-amyrin derivative with four hydroxyl group modifications.

The product previously seen through GC is much more visible through LC, as might be expected given the more polar nature of a beta-amyrin derivative with four additional hydroxyl groups. The mass spectrum of this product clearly supports a beta-amyrin derivative with four hydroxyl groups added. Interestingly, acid derivatives seen on co-expression of AhCYP1 and AhCYP2 do not appear to be modified by AhCYP3, and no acid product of AhCYP3 is detected.

6.8 – Isolation & structural elucidation of aescin intermediates

To determine the structures of AhCYP2 and AhCYP3 products, two large scale vacuum infiltrations were carried out. In one, tHMGR, SAD1, AhCYP1, and AhCYP2 were co-

expressed, and in the other the same enzymes were employed but with the addition of AhCYP3. From the first set of gene combinations was isolated 5.4 mg of dihydroxylated beta-amyrin product (peak 4 in figure 6.6.2). This was investigated through ^{1}H, ^{13}C, and 2-D NMR (400 MHz and 100 MHz respectively, recorded in pyridine D-5, see appendix D3), and the structure was determined to be 21β,22α-OH beta-amyrin. Spectra for this compound have not previously been reported, so a full assignment was carried out. The identification of all eight beta-amyrin -C*H*$_3$ signals indicated that the second hydroxylation was not at C-28. An apparent quartet with integral of 2H at δ 3.72 ppm in the ^{1}H spectrum was resolved through HSQC into a pair of roofed doublets, as these were linked to separate ^{13}C signals (see figure 6.8.1). These doublets were found to be the signals for -C*H*(OH)- of H-21 and H-22, and the coupling constant between these two peaks (*J* = 10.0 Hz) was used to determine the stereochemistry of these positions. A coupling constant of this magnitude would indicate a dihedral angle of 180^{o} between the two protons, ie. an antiperiplanar alignment, placing H-21 and H-22 in an axial position on the E ring. This leads to assignment of the hydroxyl group stereochemistry at these positions to be 21β,22α-OH, which is the same stereochemistry as that seen in aescin and its intermediates.

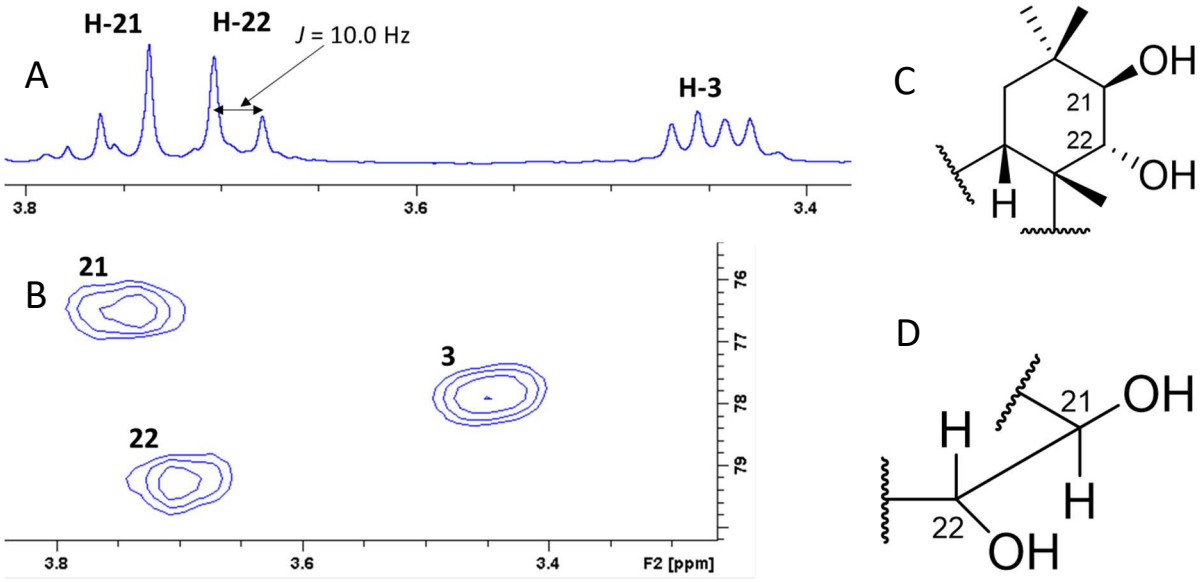


Figure 6.8.1 – The stereochemistry of the dihydroxylated beta-amyrin product of co-expression of AhCYP1 and AhCYP2 is 21β,22α-OH beta-amyrin. A: Detail from ^{1}H spectrum showing apparent quartet signal – this is actually a pair of roofed doublets corresponding to H-21 and H-22. The coupling constant of these doublets is *J* = 10.0 Hz, suggesting an antiperiplanar dihedral angle. B: Detail from HSQC spectrum, showing that the apparent quartet is linked to two separate carbon environments. C: Assigned stereochemistry on the E ring of 21β,22α-OH beta-amyrin. D: Sawhorse projection of the C-22 – C-21 bond, showing protons H-22 and H-21 in axial configuration.

Full assignment of ^{1}H and ^{13}C spectra is provided in table 6.8.1, using the numbering, COSY couplings, and HMBC couplings shown in figure 6.8.2. Though spectra for this compound have not previously been reported, ^{13}C spectra have been reported for many hydroxylated beta-amyrin derivatives (Mahato and Kundu 1994); comparison of these to the ^{13}C reported here gives confidence in these assignments. Previous reports of general trends in ^{13}C spectra of beta-amyrin derivatives also aided distinction between paired methyl groups C-23/C-24 and C-29/C-30 – these are coupled through HMBC to the same signals, so are difficult to distinguish otherwise (Mahato and Kundu 1994).

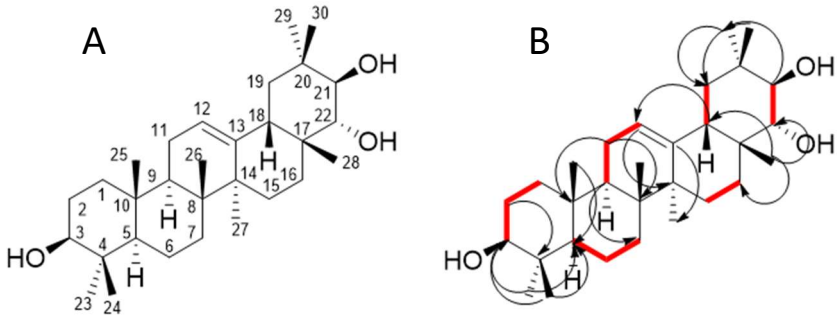


Figure 6.8.2 – Structure of 21β,22α-OH beta-amyrin as assigned. A: with all positions numbered. B: with selected COSY couplings (bold red bonds) and HMBC coupling (arrows) used in the assignment shown.

Position	*δ ^{1}H (ppm) (J in Hz)*	*δ ^{13}C (ppm)*
1	1.08 (m)	38.9
2	1.84 (m)	27.9
3	3.45 (dd) (*J* = 10.8, 5.2)	77.9
4	-	39.2
5	0.86 (m)	55.5
6	1.44 (m)	18.6
7	1.62 (m)	32.8
8	-	40.0
9	1.69 (m)	47.8
10	-	37.0
11	1.95 (dd) (*J* = 8.9, 3.6)	23.7
12	5.33 (app t) (*J* = 3.6)	122.7
13	-	144.0
14	-	42.3
15	1.28 (m)	26.0
16	1.98 (m)	21.2
17	-	39.7
18	2.28 (dd) (*J* = 13.0, 3.8)	47.1
19	2.12 (m)	46.6
20	-	36.2
21	3.75 (d) (*J* = 10.0)	76.6
22	3.69 (d) (*J* = 10.0)	79.3
23	(3H) 1.25 (s)	28.5
24	(3H) 1.06 (s)	16.3
25	(3H) 0.98 (s)	15.6
26	(3H) 1.04 (s)	16.8
27	(3H) 1.27 (s)	26.1
28	(3H) 1.31 (s)	25.7
29	(3H) 1.24 (s)	30.3
30	(3H) 1.20 (s)	18.9

Table 6.8.1 – Full ^{1}H and ^{13}C assignment of 21β,22α-OH beta-amyrin, recorded in pyridine D-5 at 400 MHz and 100 MHz, respectively.

In addition to 21β,22α-OH beta-amyrin, 0.6 mg of a trihydroxylated beta-amyrin product was isolated from the same leaf extracts (peak 5 in figure 6.6.2). This is suspected to be 21β,22α,28-OH beta-amyrin, on the basis of investigation of this product through ^{1}H NMR (400 MHz, pyridine D-5 – see appendix D4). The presence of only seven -C*H*$_3$ signals indicated modification of a methyl group. The -C*H*(OH)- portion of the spectrum contained 5 environments (see figure 6.8.3), compared to the three environments seen for 21β,22α-

OH beta-amyrin – this would further support addition of a hydroxyl group to a methyl, as two environments would be created due to the protons becoming diastereotopic. Two methyl groups are hydroxylated in the biosynthesis of aescin: C-24 and C-28. The localisation of hydroxylation to the CDE* ring fragment through MS fragmentation, and the observation of carboxylic acid products of AhCYP2 (6.6) would support C-28 hydroxylation. This was later corroborated by comparison of the proton environments observed here with those for 21β,22α,24,28-OH beta-amyrin – the signals for H-21, H-22, and H-28 (see table 6.8.2) are very similar to the peaks seen in this ^{1}H spectrum – and by formation of barringtogenol C with CYP 87D16 and protoaescigenin with CYP 87D16 and AhCYP3 (see 6.9).

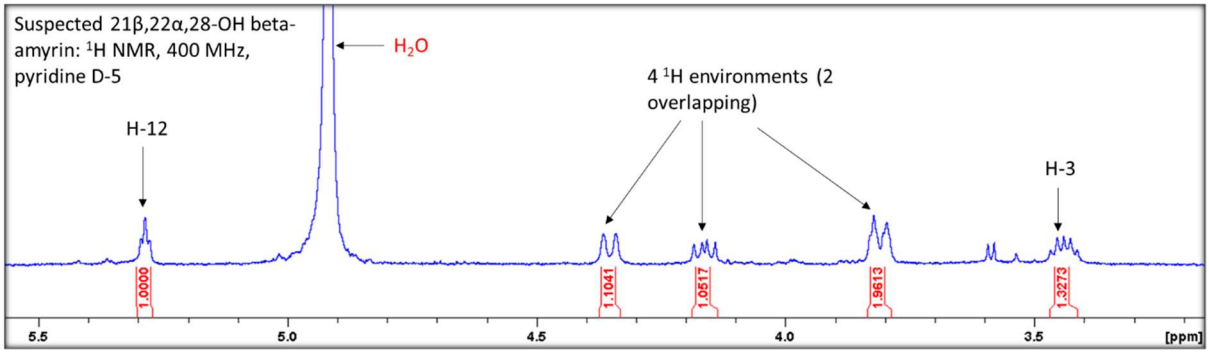


Figure 6.8.3 – detail from the ^{1}H NMR spectrum (400 MHz, pyridine D-5) of the trihydroxylated product of AhCYP1 and AhCYP2, suspected to be 21β,22α,28-OH beta-amyrin. Five proton environments geminal to an -OH group are observed, which, when compared to the assignments for 21β,22α,24,28-OH beta-amyrin, appear to match those for H-21, H-22, H-28 (x2 as diastereotopic) and H-3. The presence of only seven -C*H*$_3$ signals supports addition of -OH to C-28.

From the second large-scale infiltration experiment (tHMGR, SAD1, AhCYP1, AhCYP2, and AhCYP3) 13.5 mg of a new product (peak 5 in figure 6.7.1) was isolated. The structure of this was investigated by NMR in pyridine D-5. A ^{13}C DEPTQ experiment identified 30 carbon environments, as is expected of a triterpene, but with 6 CH_3 groups rather than the expected 8 for unmodified beta-amyrin, suggesting modification of two methyl groups. Furthermore, five ^{13}C signals were detected in the alcohol region: three CH and two CH_2, which would suggest addition of four alcohol groups to beta-amyrin, two of which were added onto methyl groups. The ^{1}H spectrum of this compound indicated seven ^{1}H signals in the region associated with alcohols, which were linked to the five ^{13}C signals through HSQC – this supports hydroxylation at both C-24 and C-28, as these positions give two ^{1}H -C*H*$_2$(OH) signals each due to diastereotopic protons (see figure 6.8.4). Five peaks were noted in the ^{1}H spectrum between δ 5.35 and 6.70 ppm which were not linked to ^{13}C signals through HSQC (see figure 6.8.4), consequentially, these were not assigned, though

as there are five of these peaks and this spectrum was recorded in an aprotic solvent, it is possible that these are -O*H* signals.

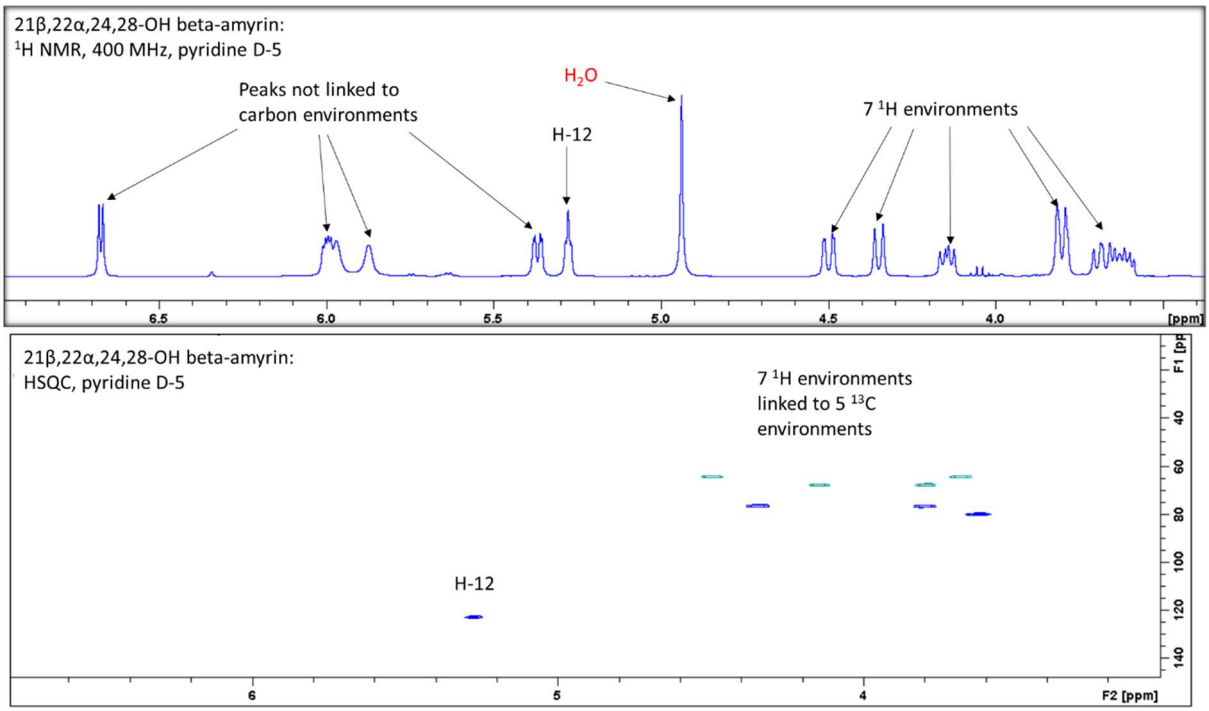


Figure 6.8.4 – detail from the ^{1}H and HSQC NMR spectra (400 MHz, pyridine D-5) of the product of AhCYP1, AhCYP2, and AhCYP3, assigned as 21β,22α,24,28-OH beta-amyrin. Seven proton environments geminal to an -OH group are observed, which are found to be linked to five ^{13}C environments through HSQC. Peaks in the ^{1}H spectrum not linked to carbon environments may represent -O*H* protons.

Full assignment was undertaken, starting from H-12 and working through the molecule using COSY, HSQC, and HMBC (table 6.8.2). The structure assigned was that of 21β,22α,24,28-OH beta-amyrin, which is shown in figure 6.8.5 along with selected COSY and HMBC couplings employed in structural determination. The determination of C-24 hydroxylation over C-23 was based on the ^{13}C signal (δ 65.0 ppm). Comparison with the ^{13}C spectra of other beta-amyrin derivatives provided in Mahato and Kundu (1994) show that the -C*H*$_2$(OH) signal for C-24 is typically around δ 64.0 – 66.0 ppm, whereas hydroxylation at C-23 results in a slightly less shielded ^{13}C environment (typically δ 68.0 – 71.0 ppm). The signal observed here (δ 65.0 ppm) supports 24-OH.

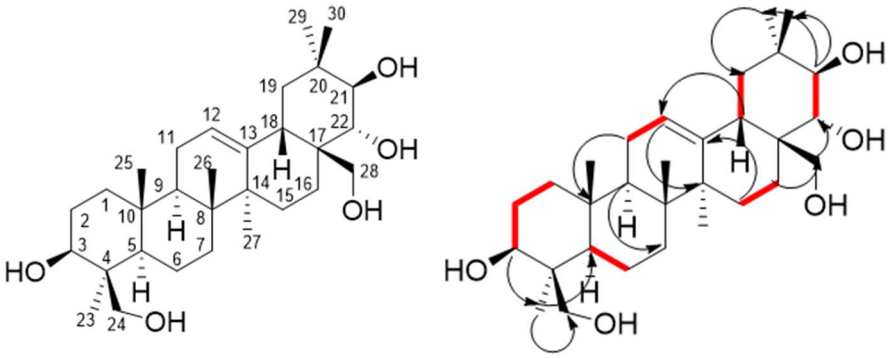


Figure 6.8.5 – Structure of the product of co-expression of AhCYP1, AhCYP2, and AhCYP3 on beta-amyrin – 21β,22α,24,28-OH beta-amyrin. On the right are shown selected COSY and HMBC couplings used to aid determination of this structure.

Position	*δ ^{1}H (ppm) (J in Hz)*	*δ ^{13}C Assignment (ppm)*
1	1.09 (m)	39.4 (CH_2)
2	2.04 (m)	28.9 (CH_2)
3	3.65 (m)	80.6 (CH)
4	-	43.7 (quaternary)
5	0.97 (m)	56.8 (CH)
6	1.52 (m)	19.6 (CH_2)
7	1.22 (m), 1.49 (m)	33.7 (CH_2)
8	-	39.3 (quaternary)
9	1.69 (m)	48.5 (CH)
10	-	37.4 (quaternary)
11	1.87 (m)	24.6 (CH_2)
12	5.31 (app t) (*J* = 3.4)	123.5 (CH)
13	-	144.2 (quaternary)
14	-	42.7 (quaternary)
15	2.36 (m)	18.4 (CH_2)
16	1.92 (m)	26.2 (CH_2)
17	-	43.9 (quaternary)
18	2.55 (dd) (*J* = 13.5, 4.2)	42.4 (CH)
19	1.36 (m)	47.2 (CH_2)
20	-	36.9 (quaternary)
21	3.83 (d) (*J* = 10.3)	77.3 (CH)
22	4.38 (d) (J = 10.3)	77.1 (CH)
23	(3H) 1.58 (s)	24.0 (CH_3)
24	3.71 (m), 4.53 (dd) (*J* = 10.6, 1.8)	65.0 (CH_2)
25	(3H) 0.93 (s)	16.6 (CH_3)
26	(3H) 0.94 (s)	17.2 (CH_3)
27	(3H) 1.32 (s)	26.7 (CH_3)
28	3.82 (m), 4.18 (dd) (*J* = 10.6, 6.7)	68.4 (CH_2)
29	(3H) 1.29 (s)	30.9 (CH_3)
30	(3H) 1.28 (s)	19.6 (CH_3)

Table 6.8.1 – Full ^{1}H and ^{13}C assignments for 21β,22α,24,28-OH beta-amyrin, recorded in pyridine D-5 at 400 MHz and 100 MHz respectively.

NMR spectra for this compound have not been previously reported, though comparison of these spectra with those for protoaescigenin (Agrawal, Thakur et al. 1991) (differing only in hydroxylation at C-16α) and other closely-related triterpenoids (Mahato and Kundu 1994) gives confidence in these assignments. Determination of the

stereochemistry of hydroxyl groups on C-21 and C-22 was through the coupling constants being consistent with an antiperiplanar alignment of protons, as previously.

6.9 – Combinatorial Biosynthesis of Protoaescigenin in *N. benthamiana*

With the three CYPs from the *A. hippocastanum* transcriptome it is now possible to hydroxylate four of the five positions on beta-amyrin necessary for the biosynthesis of protoaescigenin (the des-acyl variant of the aescin aglycone). An *Aesculus* CYP capable of effecting the remaining transformation – C-16α hydroxylation – is yet to be characterised. However, CYPs carrying out this transformation are present in the triterpenoid toolkit (see chapter 3). One CYP capable of C-16α hydroxylation is CYP 87D16, from *Maesa lanceolata* (Moses, Pollier et al. 2015), accordingly, this was included into combinatorial assays with the three *A. hippocastanum* CYPs. Figure 6.9.1 shows LC traces and mass spectra for the co-expression of CYP 87D16 in *N. benthamiana* with tHMGR, SAD1, AhCYP1, and AhCYP2.

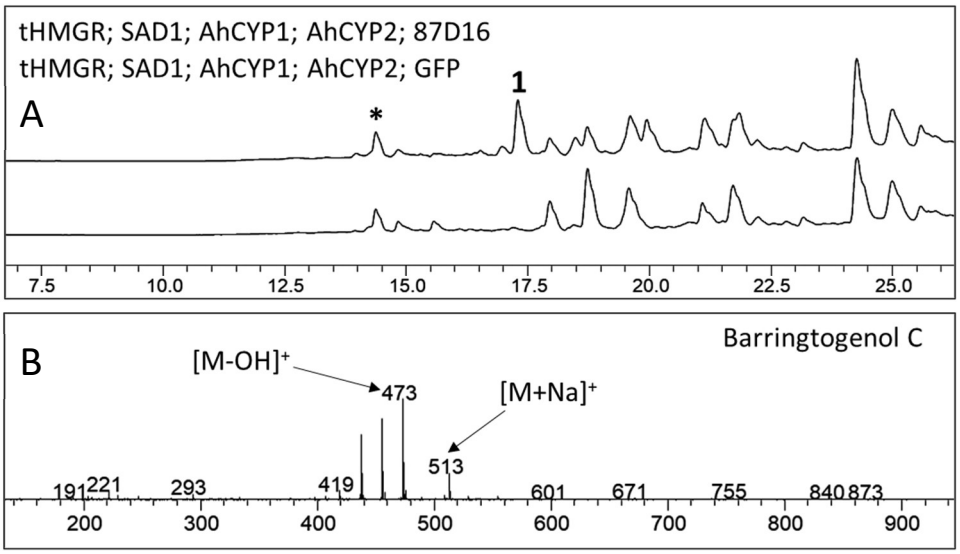


Figure 6.9.1 – Co-expression of CYP 87D16 with AhCYP1 and AhCYP2 gives access to new aescin intermediates. A: **LC** traces showing the activity of CYP 87D16 on 21β,22α,28-OH beta-amyrin. * Internal standard (digoxin 20 ppm) visible at 14.4 mins. A major new product peak is seen at (1). B: Mass spectrum (positive mode) of peak 1 in A, appearing to correspond to the mass spectrum of barringtogenol C, with a peak at m/z 473 from loss of -OH and a sodium adduct peak at 513. This matches the reported mass spectrum of barringtogenol C in the literature (Konoshima and Lee 1986). Other minor peaks in the LC traces appear to be C-28 acid derivatives.

A new product was observed from this experiment, which would appear to be barringtogenol C (16α,21β,22α,28-OH beta-amyrin) on the basis of comparison with reported mass spectra (Konoshima and Lee 1986). This activity matches that predicted, and

leaves only one step remaining to reach protoaescigenin (figure 6.9.2 shows the chemical structures of barringtogenol C and protoaescigenin, which differ only in C-24 hydroxylation). To complete the biosynthesis of this moiety of aescin, CYP 87D16 was co-expressed with tHGMR, SAD1, and all three characterised *A. hippocastanum* CYPs in *N. benthamiana*. The results are shown in figure 6.9.3.

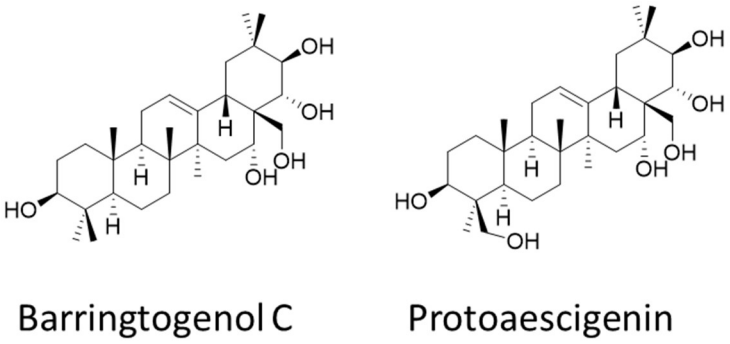


Figure 6.9.2 – The chemical structures of barringtogenol C and protoaescigenin, the two aescin intermediates formed using CYP 87D16 in combination with *A. hippocastanum* CYPs. These represent significant milestones in aescin biosynthesis, and are ideal scaffolds for testing acyltransferase and glycosyltransferase candidates.

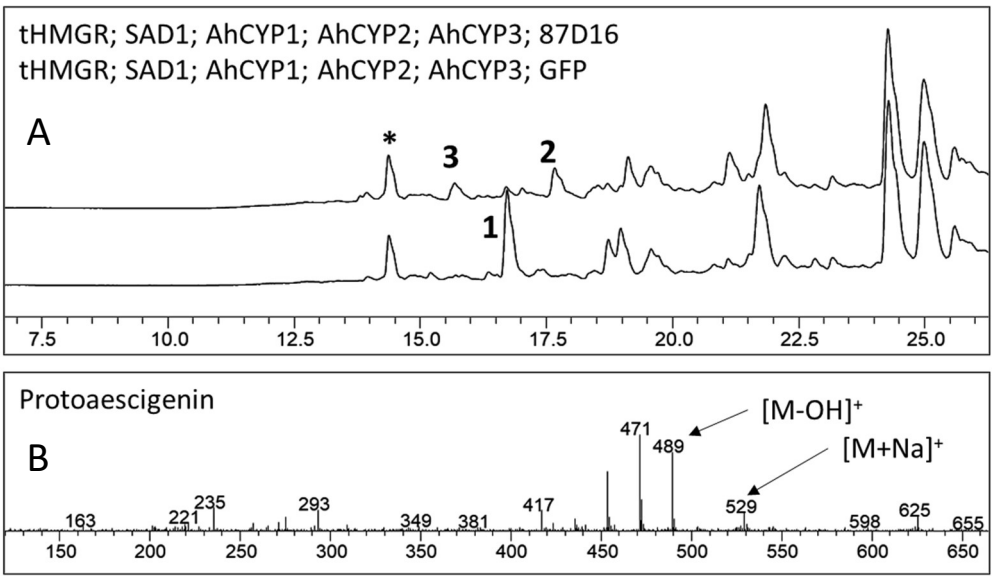


Figure 6.9.1 – Formation of protoaescigenin is possible with CYP 87D16. A: **LC** traces demonstrating the effect of co-expression of CYP 87D16 with tHMGR, SAD1, and all three *A. hippocastanum* CYPs characterised so far. * Internal standard (digoxin 20 ppm) visible at 14.4 mins. Peaks of note: (1) 21β,22α,24,28-OH beta-amyrin (2) barringtogenol C (3) protoaescigenin. B: Mass spectrum (positive mode) of peak (3) in A – this appears to be protoaescigenin. The sodium adduct peak at m/z 529 matches that reported for protoaescigenin (reported m/z 529.3) (Yang, Long et al. 2019).

From these traces can be seen a new peak with a mass spectrum consistent with that of protoaescigenin (Yang, Long et al. 2019). Formation of protoaescigenin is a key achievement for this project, since this compound is a key intermediate in aescin

biosynthesis. It will also form a robust scaffold for testing of downstream enzymes such as acyltransferases, to be tested alongside the other hydroxylated intermediates formed in this chapter. Full structural determination of this product (along with barringtogenol C) would be desirable for confirmation, and would likely form part of future work – full ^{1}H and ^{13}C NMR spectra have previously been reported for comparison (Agrawal, Thakur et al. 1991, Mahato and Kundu 1994). That aside, the use of CYP 87D16 to carry out C-16α hydroxylation is validated as a workable strategy to obtain the full complement of hydroxylated aescin intermediates in the absence of the corresponding *Aesculus* CYP.

6.10 - Conclusions

The new transcriptomic data for *Aesculus* has proven to have great utility in candidate identification. A switch of focus to *A. hippocastanum* over *A. pavia* enabled rapid identification and characterisation of biosynthetic enzymes which, in combination, allow for significant progress towards elucidation of the aescin biosynthetic pathway. The first committed step of aescin biosynthesis – the formation of beta-amyrin – is carried out by the *A. hippocastanum* OSC, AhOSC1. This enzyme is a mixed function amyrin synthase, forming both alpha- and beta-amyrin on transient expression in *N. benthamiana*. Characterisation of the *A. hippocastanum* homologue of ApCYP1, AhCYP1, has revealed that ApCYP1 and AhCYP1 have identical activities. With the aid of co-expression data for *A. pavia*, it has been possible to identify a further two *A. hippocastanum* CYPs based on homology with *A. pavia* CYPs.

The characterisation of AhCYP2 is highly significant to this work. This CYP enables hydroxylation of both the key positions on the scaffold identified earlier as relevant for acyltransferase testing, namely C-22α and C-28. Neither of these transformations could be achieved using the triterpenoid toolkit, and so it was crucial to identify *Aesculus* enzymes that fulfil these functions before other candidate downstream pathway enzymes (ACTs and glycosyltransferases) can be evaluated. The discovery that AhCYP2 is able to oxidise C-28 to either an alcohol or a carboxylic acid may lead to yet further aescin intermediates being identified when downstream enzymes are tested.

With a third *A. hippocastanum* CYP characterised, AhCYP3, four of the five hydroxylations found in aescin biosynthesis can now be carried out. The final hydroxylation at C-16α remains unaccounted for from the *Aesculus* candidates so far tested, although

some CYP have not yet been tested. This hydroxylation can, however, be carried out using enzymes from the triterpenoid toolkit, and their use to form protoaescigenin on co-expression with *A. hippocastanum* CYPs has now been validated. This now provides the full range of hydroxylated beta-amyrin derivatives likely to be aescin intermediates – an excellent platform on which to test downstream modifying enzymes in combinatorial trials. The pathway as it appears so far is summarised in figure 6.10.1, which bears some resemblance to that proposed in chapter 2.

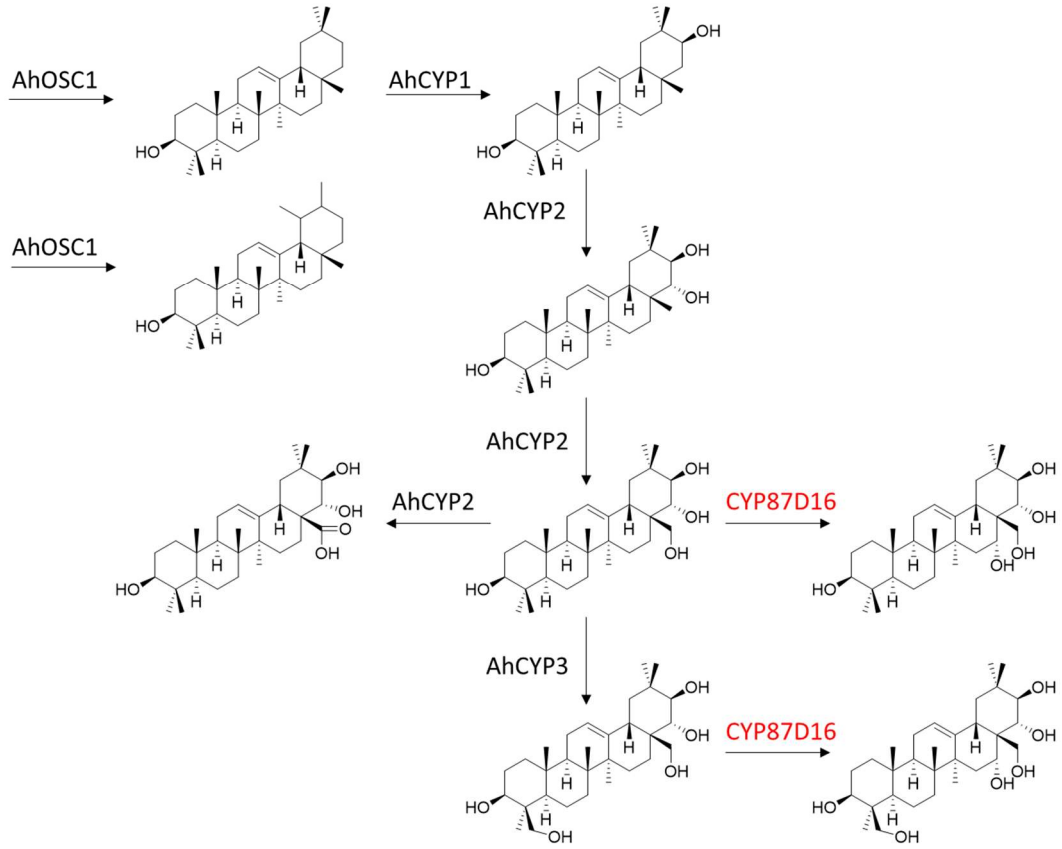


Figure 6.10.1 – The biosynthetic pathway so far. With an OSC and three CYPs from *A. hippocastanum* characterised, it is possible to form beta-amyrin and carry out four of the five hydroxylations required. The final hydroxylation (C-16α) can be effected using the triterpenoid toolkit (in this case CYP 87D16).

Chapter 7 – Acylation and Glycosylation of the Aescin Aglycone

Synopsis

Previous chapters predicted a biosynthetic pathway for aescin, identified BAHD acyltransferases as the most likely to be responsible for acylation of the aglycone, obtained new transcriptomic resources for *A. pavia* and *A. hippocastanum*, and identified four biosynthetic enzymes from *A. hippocastanum* which, when paired with the triterpenoid toolkit, enabled production of protoaescigenin through transient expression in *N. benthamiana*. The new transcriptome data was used here to search for BAHD acyltransferases, which led to identification and characterisation of an active ACT, *AhBAHD1*. This ACT appears to be capable of addition of two different acyl groups to the aglycone. Informed by recent reports implicating Cellulose Synthase-Like (CSL) enzymes in triterpenoid glycosylation, an active CSL (*AhCSL1*) was characterised and found to be responsible for D-glucuronylation of the aescin aglycone. Further progress was made in construction of the glycosyl moiety of aescin through discovery of two UGTs: *AhUGT1*, a D-galactosyl transferase; and *AhUGT2*, which is likely to be responsible for addition of either D-glucose or D-xylose to the sugar chain. These four enzymes bring the total number of aescin biosynthetic enzymes discovered to eight. Combination of these enzymes and the toolkit enables production of a number of aescin pathway intermediates, with only one step in the pathway unaccounted for.

Introduction

7.1 – Introduction to Acylation & Glycosylation of Triterpenoids

The triterpenoid toolkit, as evaluated in chapter 3, enabled hydroxylation at C-16α, C-21β, and C-24 of beta-amyrin, in addition to oxidation at C-28. The discovery of three CYPs from *A. hippocastanum* using the new transcriptome data (chapter 6) has enabled sequential hydroxylation of the β-amyrin scaffold at C-21β, C-22α, C-28, and C-24. Co-expression of these three CYPs with CYP 87D16 (Moses, Pollier et al. 2015), a C-16α hydroxylase from the toolkit, enabled detection of protoaescigenin, the full des-acyl aglycone of aescin (figure 7.1.1). The ability to produce protoaescigenin and its intermediates in *N. benthamiana* now makes it possible to test candidates for downstream biosynthetic enzymes.

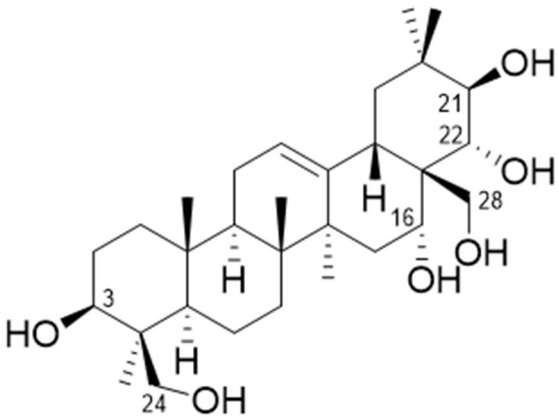


Figure 7.1.1 – The structure of protoaescigenin. Hydroxylations at C-21β, C-22α, C-24, and C-28 can be achieved using native *A. hippocastanum* CYPs AhCYP1, AhCYP2, and AhCYP3. Hydroxylation at C-16α requires use of the toolkit enzyme CYP 87D16.

Aescin is acylated with two acyl groups – an acetyl group, typically at C-22α, and either an angeloyl or tigloyl group at C-21β (these differ only in the stereochemistry about the carbon-carbon double bond). Aescins Ia, IIa, and IIIa are C-21β acylated with a tigloyl group, and aescins Ib, IIb, and IIIb with an angeloyl group (Yoshikawa, Harada et al. 1994). Figure 7.1.2 below shows the structure and position of these two acyl groups. These six principal saponins making up the bulk of aescin in *A. hippocastanum* are only acylated on these two positions, though other saponins are reported with acetylation at C-28 or C-16α rather than C-22α (Price, Johnson et al. 1987, Yang, Long et al. 2019). It is thought that migration of the acetyl group from C-22α to C-28 may occur spontaneously (Sirtori 2001), making it possible that some of these compounds may be artefacts of extraction or isolation. Database mining in chapter 2 identified positions C-21 and C-22 as the two most frequently acylated in reported beta-amyrin derivatives by some margin. Furthermore, the two types of acyl group found here were also identified as the two most frequent acyl groups. However, an acyltransferase capable of adding either group to either position of a beta-amyrin derivative is yet to be reported. The discovery of acyltransferases with this activity could therefore have wider significance for triterpenoid pathway engineering.

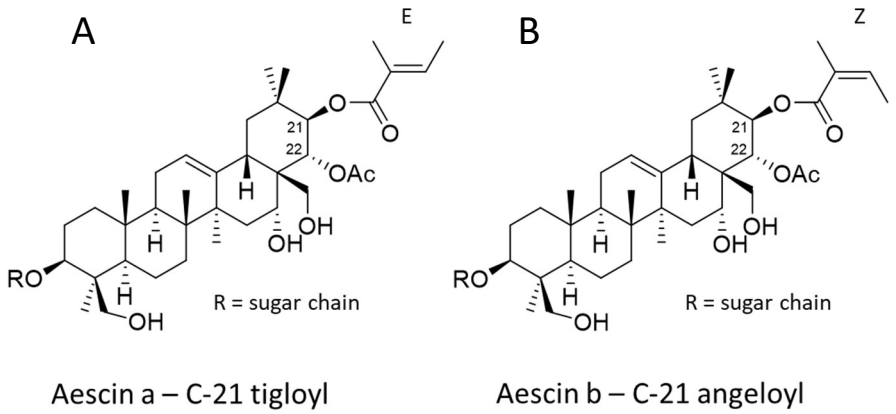


Figure 7.1.2 – The acylation of aescin shows some variation. A: Structure of the aescin a aglycone. The pattern of acylation seen here is 21β-tigloyl, 22α-acetyl. B: Structure of the aescin b aglycone. Here the acylation is 21β-angeloyl, 22α-acetyl. Tigloyl and angeloyl differ only in the geometry about the carbon-carbon double bond: tigloyl is E or trans, and angeloyl is Z or cis.

The acyltransferases active on plant secondary metabolites can be divided into two large families: the BAHDs and the SCPLs (D'Auria 2006, Bontpart, Cheynier et al. 2015). These two families differ in their sub-cellular localisation, substrate specificity, and mode of action. Analysis of reported characterised plant natural product ACTs in chapter 2 found that acetylation and tigloylation were associated with BAHDs rather than SCPLs, and so BAHD enzymes were prioritised as potential candidates for aescin biosynthesis. Unlike SCPLs, BAHDs are cytosolic so do not require transporter enzymes to move their substrates to the vacuole. Also unlike SCPLs, BAHD enzymes use co-enzyme A-activated acyl donors, which, for the three acyl groups found in aescin, are products of primary metabolism in plants (Robinson, Bachhawat et al. 1956, Kochevenko and Fernie 2011, Shi and Tu 2015). As such, BAHDs might be able to acylate protoaescigenin on transient expression in *N. benthamiana* without the addition of any other enzymes. If acetyl, angeloyl, and tigloyl groups are each added by different ACTs, a total of three distinct ACTs will be required to fully acylate the aglycone. However, it is possible that the angeloyl and tigloyl moieties could be added by the same ACT, given their close structural resemblance. A degree of promiscuity is common in BAHDs, substrate specificity being more apparent with respect to acyl acceptor than acyl donor (Bontpart, Cheynier et al. 2015, Rinaldo, Cavallini et al. 2015, Kruse, Weigle et al. 2020, Oshikiri, Watanabe et al. 2020), and so it is possible that fewer than three ACTs may be required. Unlike some other biosynthetic enzyme families, amino acid sequence similarity is not considered a good predictor of function, so a broader approach to candidate selection may be warranted (Beekwilder, Alvarez-Huerta et al. 2004).

The sugar chain of aescin shows some variability, but has two sugars in common: a D-glucuronic acid at C-3 of the aglycone, and a D-glucose at C'-4 of the glucuronic acid. A third sugar at C'-2 of the glucuronic acid varies: in aescin I D-glucose is found, in aescin II this sugar is D-xylose, and in aescin III it is D-galactose (Yoshikawa, Harada et al. 1994). Figure 7.1.3 shows the constituents of the sugar chain of different aescins. If each sugar is added by different glycosyltransferases, then a total of five glycosyltransferase enzymes will be required to form the full range of aescins. This number could be lower, however – substrate promiscuity with respect to the sugar donor has been reported in UGTs, the most common glycosyltransferases in plant secondary metabolism (De Bruyn, Maertens et al. 2015). UGTs, or Uridine Diphosphate-Glycosyltransferases, have been shown to be involved in triterpenoid biosynthesis in many different plant species (Gachon, Langlois-Meurinne et al. 2005, Thimmappa, Geisler et al. 2014, Tiwari, Sangwan et al. 2016, Rahimi, Kim et al. 2019). Examples of UGTs that are able to transfer the sugars found in aescin have previously been characterised, for instance the D-galactosyltransferase GmUGT73P2, the D-glucosyltransferase GmUGT73F2, and the D-xylosyltransferase GmUGT73F4 (all from soybean) (Shibuya, Nishimura et al. 2010, Sayama, Ono et al. 2012), and the D-glucuronyltransferase GuUGT73P12 from liquorice (Nomura, Seki et al. 2019).

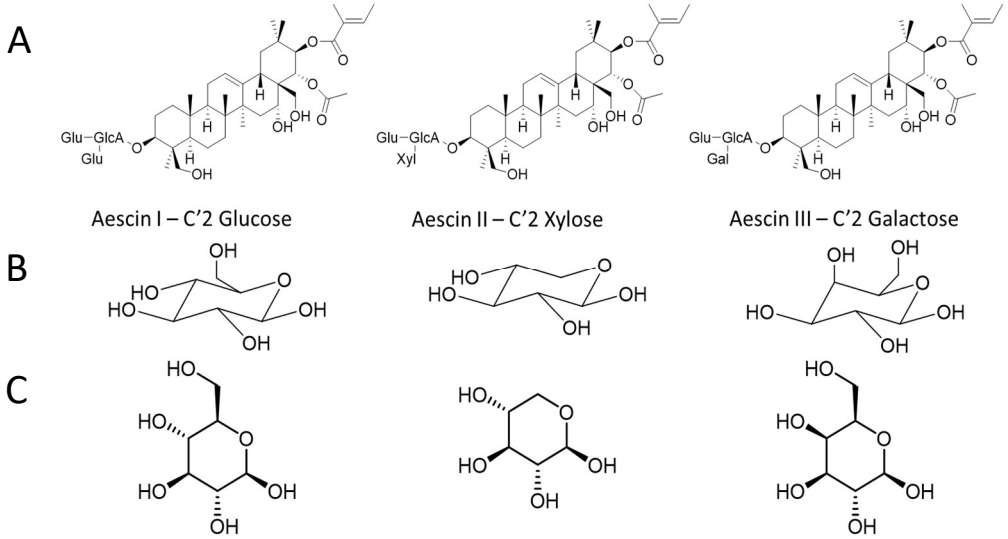


Figure 7.1.3 – The configurations of the three main sugar chains found in aescin. A: Structures of aescins I, II, and III. B: Chair representations of the sugars found at C'-2 in each aescin. C: Planar representations of these sugars, showing the difference in stereochemistry between D-galactose and D-glucose.

There have been recent reports of D-glucuronylation of triterpenoids in plants by enzymes other than UGTs, specifically enzymes from the cellulose synthase-like (CSL) family, which have been shown to glucuronylate triterpenes in spinach (Jozwiak, Sonawane

et al. 2020) and liquorice (Chung, Seki et al. 2020). CSL enzymes had previously only been associated with cell wall biosynthesis (Hazen, Scott-Craig et al. 2002, Burton, Wilson et al. 2006), but have increasingly been implicated in triterpenoid glucuronylation too (Zhang, Hua et al. 2022). A CSL enzyme, *QsCSL* from *Quillaja saponaria* (Osbourn, Reed et al. 2020), is represented in the triterpenoid toolkit, and so could be used to glucuronylate the aescin aglycone to aid in evaluation of UGTs for ability to add the other sugars. Potential CSLs from the *A. hippocastanum* transcriptome are also investigated alongside the UGTs.

Methods

7.2 – Chapter-Specific Methods

BAHD, CSL, and UGT candidates were amplified from *A. hippocastanum* conker cDNA (made according to 9.1.2) according to 9.1.3 using Q5 and the primers listed there, with the exception of *AhUGT1*, which was ordered as a synthetic gene from IDT (sequence provided in appendix C). Amplified sequences were cloned into expression vector pEAQ-HT-DEST1 (Sainsbury, Thuenemann et al. 2009) according to the protocol in 9.1.4 and transformed into LBA4404 as per 9.1.7. LBA4404 strains expressing toolkit enzymes were obtained as glycerol stocks (see chapter 3).

Heterologous expression experiments in *N. benthamiana* were carried out as per 9.2.1, using the candidates cloned and transformed as above and strains expressing toolkit enzymes and *Aesculus* enzymes previously cloned (see chapters 3 and 6). Harvested leaf material was prepared for GC analysis according to 9.3.1 using 100 ppm internal standard, and GC analysis run according to 9.3.4. Samples for LC analysis were prepared as per 9.3.2 and LC performed according to 9.3.5, using gradient A (7.3.2, 7.3.3, 7.4.5), gradient B (7.3.4, 7.4.3, 7.4.4, 7.4.6, 7.5.1) gradient C (7.7.1, 7.7.2, 7.7.3) or gradient D (7.4.1, 7.4.2, 7.6.1, 7.6.2).

Results & Discussion

7.3 – Identification & Characterisation of *AhBAHD1*

Analysis of the activities of published acyltransferases in chapter 2 justified a focus on BAHDs over SCPLs in this work. As such, the *A. hippocastanum* transcriptome was searched for predicted BAHD sequences. In light of the experience of the previous chapter

with use of expression levels in *A. hippocastanum*, a broader approach was taken to candidate selection. The functional annotation of sequences performed by the Earlham Institute using AHRD (v3.3.3), which classifies sequences by putative protein family (see chapter 5), was searched to identify putative BAHD candidates. The *A. hippocastanum* assembly contained potential BAHD sequences annotated as "BAHD acyltransferase DCR-like", "HXXXD-type acyl-transferase family protein, putative", "Vinorine synthase-like" or "Chloramphenicol acetyltransferase-like", vinorine synthase and chloramphenicol acetyltransferase being widely studied ACTs sharing the catalytic HxxxDG conserved domain of BAHDs (Shaw 1992, Bayer, Ma et al. 2004). 2 "BAHD acyltransferase DCR-like", 7 "HXXXD-type acyl-transferase family protein, putative", 16 "Vinorine synthase-like", and 42 "Chloramphenicol acetyltransferase-like" sequences were found in the transcriptome. These sequences were manually evaluated and excluded if they were obviously truncated (<900 bp) or lacking either of the two conserved domains required for BAHDs to be active (HxxxDG & and DFGWG) (Unno, Ichimaida et al. 2007, Molina and Kosma 2015). Thirty-one sequences were retained as candidates. These are listed in table 7.3.1, along with their functional annotation and expression levels in conkers and leaves. Expression levels were not taken into account when selecting candidates, due to their apparent lack of reliability in *A. hippocastanum* as demonstrated in chapter 6, but candidates with greater expression in conkers were given priority in cloning efforts. Thirteen of these candidates were cloned and co-expressed in *N. benthamiana* with biosynthetic enzymes forming a range of hydroxylated scaffolds, including 21β-OH beta-amyrin (AhCYP1), 21β-OH oleanolic acid (AhCYP1; CYP716A12), 21β,22α,28-OH beta-amyrin (AhCYP1; AhCYP2), 21β,22α,24,28-OH beta-amyrin (AhCYP1; AhCYP2; AhCYP3), and protoaescigenin (AhCYP1; AhCYP2; AhCYP3; CYP87D16).

CANDIDATE	ANNOTATION	EXPRESSION - CONKER	EXPRESSION - LEAF	CLONED & TESTED
AH2407C0G1I3	C	0	3.488	
AH3616C0G1I1	C	0	49.220	
AH4415C0G1I6	V	27.080	1.250	✓
AH4755C0G1I1	H	0	0.483	
AH4755C0G1I8	H	0.089	1.303	✓
AH5291C0G2I3	C	0	45.548	✓
AH6318C0G3I1	C	0.049	12.424	
AH6318C0G3I2	C	0.071	7.719	
AH9064C0G1I7	C	0.063	3.277	✓
AH9861C0G1I2	V	0	2.198	
AH10410C0G1I1	B	0.380	36.581	
AH10648C0G1I2	H	0	2.158	✓
AH13680C0G1I1	V	0.008	1.671	
AH16012C0G1I1	C	0.906	1.948	✓
AH16012C0G1I2	C	0.482	0.220	
AH16796C0G3I1	C	0	1.273	
AH17020C0G1I3	V	0.911	0.545	✓
AH22311C0G1I1	C	84.092	0.008	
AH23384C0G1I2	C	0	0.221	
AH23384C0G2I1	C	0	2.000	
AH25465C0G1I1 (*AHBAHD1*)	V	0.371	0.464	✓
AH27763C0G1I1	V	2.037	0.512	✓
AH28436C1G1I1	C	0	0.161	✓
AH30407C0G1I1	C	0	1.552	
AH31072C0G1I1	C	0	0.186	
AH36402C1G1I1	H	0.026	0.217	
AH37835C0G1I1	C	0.035	0.474	
AH43477C0G1I1	C	0	1.335	
AH46030C0G1I1	C	0.146	0.207	✓
AH77698C0G1I1	C	105.763	13.262	✓
AH95885C0G1I1	V	4.213	0.008	✓

Table 7.3.1 – The BAHD candidates selected from the *A. hippocastanum* assembly. Key to annotation: B (BAHD acyltransferase DCR-like) C (Chloramphenicol acetyltransferase-like) H (HXXXD-type acyl-transferase family, putative) V (Vinorine synthase-like). Expression values are TMM cross-species normalised counts per million in the two *A. hippocastanum* tissues. Of the 31 candidates, 13 were successfully cloned and co-expressed with relevant biosynthetic enzymes in *N. benthamiana* – one was found to be active (*Ah25465c0g1i1* or *AhBAHD1*).

From these candidates one BAHD enzyme demonstrated activity on 21β,22α-OH beta-amyrin: Ah25465c0g1i1 (*AhBAHD1*). Figure 7.3.1 below shows GC-MS analysis of leaf extracts on co-expression of AhBAHD1 with tHGMR, SAD1, AhCYP1, and AhCYP2 in *N. benthamiana*.

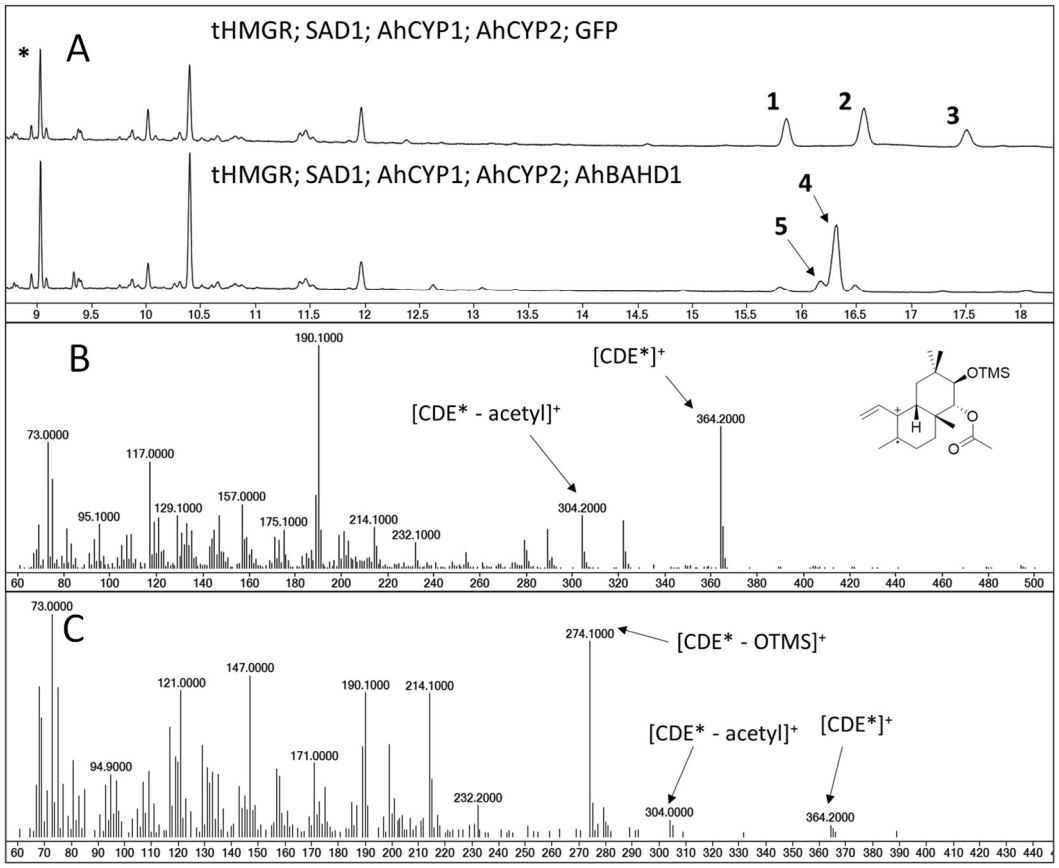


Figure 7.3.1 – AhBAHD1 is active when co-expressed with *A. hippocastanum* CYPs. A: **GC** traces (TMS derivatised) showing the effect of co-expression of AhBAHD1 with tHMGR, SAD1, AhCYP1, and AhCYP2. * Internal standard (coprostanol) visible at 9.1 mins. (1) 21β,22α-OH beta-amyrin (2) 21β,22α,28-OH beta-amyrin (3) 21β,22α-OH oleanolic acid. New peaks seen at (4) and (5). B: Mass spectrum of peak (4). Signals at m/z 364 and 304 could indicate addition of an acetyl group to the CDE* ring fragment of doubly hydroxylated beta-amyrin, as shown. C: Mass spectrum of peak (5). The peaks seen in B are seen here too, alongside another peak at m/z 274, possibly due to loss of -OTMS from the acetylated CDE* ring fragment.

Two new products were observed whose mass spectra consistent with addition of an acetyl group to the CDE* ring fragment. The detection of two products with such similar retention times and mass spectra could suggest isomerism of some kind. Comparable activity was not observed through GC on 21β-OH beta-amyrin (AhCYP1 only), nor were additional products seen with 21β,22α,24,28-OH beta-amyrin (inclusion of AhCYP3). Further investigation of these leaf extracts was carried out through LC-MS, as shown in figure 7.3.2.

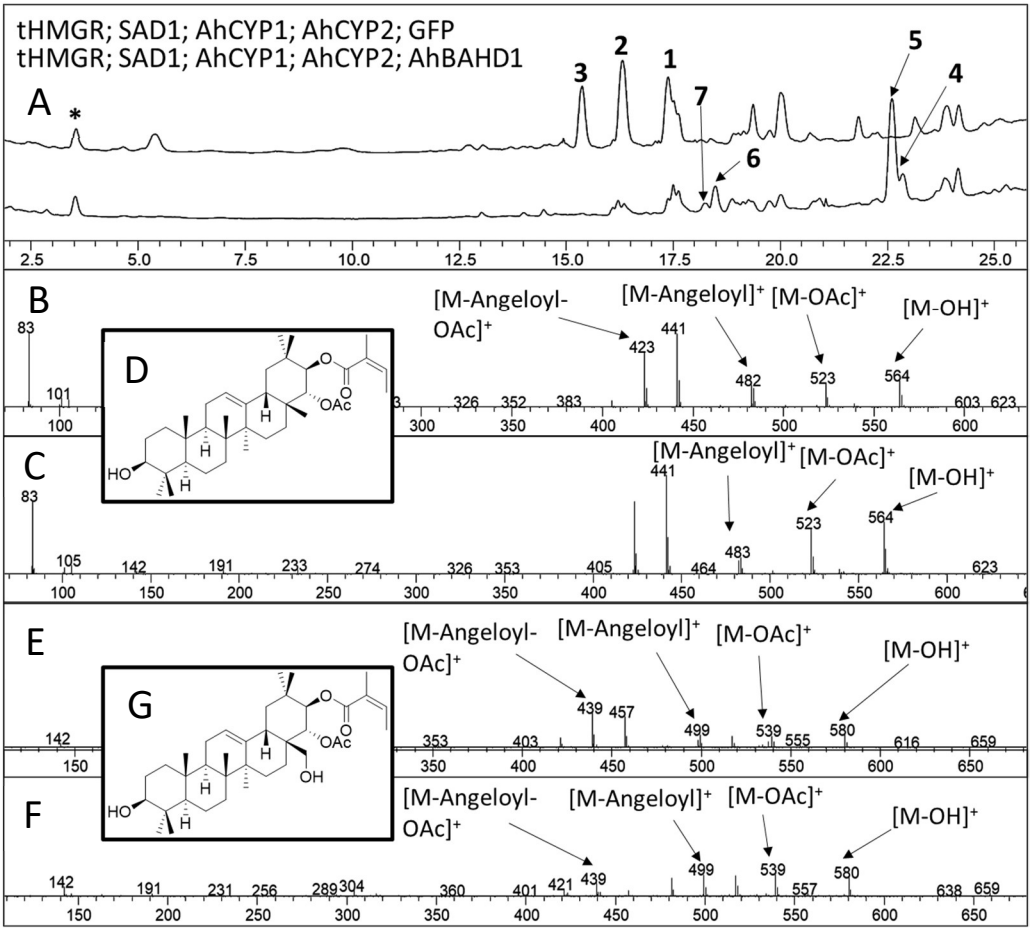


Figure 7.3.2 – New acylated products are detected on LC-MS analysis of leaf extracts expressing AhBAHD1. A: **LC** traces (grad A) showing the activity of AhBAHD1 on 21β,22α,28 beta-amyrin. * Internal standard visible at 3.6 mins. (1) 21β,22α-OH beta-amyrin (2) 21β,22α,28-OH beta-amyrin (3) 21β,22α-OH oleanolic acid. New acylated products seen at (4)-(7). B: Mass spectrum of peak (4). C: Mass spectrum of peak (5). D, inset: Possible structure for B or C – fragmentation patterns suggest addition of two acyl groups, acetyl and angeloyl or tigloyl, to 21,22-OH beta-amyrin. E: Mass spectrum of peak (6). F: Mass spectrum of peak (7). G, inset: Possible structure for E or F. The same fragmentation pattern is present, but on a 21,22,28-OH beta-amyrin

LC analysis revealed multiple new product peaks resulting from activity of AhBAHD1 on 21β,22α-OH beta-amyrin and 21β,22α,28-OH beta-amyrin. The mass spectra of these peaks suggested compounds with additional acylation – fragmentation suggested an acyl group with a mass of 59 (acetyl) and an acyl group with a mass of 99 (consistent with addition of angeloyl or tigloyl moieties). These products were observed as pairs with very similar retention times and mass spectra, possibly indicative of isomerism. This could take the form of regioisomerism (the position of the acetyl group can vary in the minor aescins) or stereoisomerism (the geometry about the double bond in angeloyl / tigloyl). Stereoisomerism about the double bond would be consistent with the main pattern of acylation observed in aescin – this would involve AhBAHD1 being capable of transferring three different types of acyl group: acetyl, angeloyl, and tigloyl.

A single BAHD being responsible for multiple acylation events is not without precedent. There have been reports of sequential addition at multiple sites on the same substrate by BAHDs, for instance triacylation of glycerol by OsHCT4, diacylation of polyamines by AtSDT (Eudes, Mouille et al. 2016), or triacylation of spermidine by AtSHT (Wang, Li et al. 2021). A BAHD from chicory, CiSHT2, has even been found to be capable of sequential addition of four acyl groups to polyamines, and shows some promiscuity in the acyl groups employed (Delporte, Bernard et al. 2018).

The apparent ability of AhBAHD1 to transfer different types of acyl group is very common amongst the BAHDs, which often display acyl donor promiscuity (D'Auria 2006). Promiscuity with respect to the acyl group transferred by BAHDs appears to depend on the nature of the acyl group – BAHDs transferring aromatic acyl groups are generally inactive on aliphatic groups, and vice versa (Nakayama, Suzuki et al. 2003), though there are some possible exceptions. For instance, CcAT2 from montbretia is capable of transferring malonyl, acetyl, coumaroyl, feruloyl and caffeoyl; acyl groups that vary significantly in their structure (Irmisch, Jo et al. 2018). Similarly, VvMYBA from grapevine is able to mediate acylation with both aromatic and aliphatic groups (Rinaldo, Cavallini et al. 2015). The angeloyl, tigloyl, and acetyl groups are all aliphatic, so the ability of AhBAHD1 to add all three would be consistent with the general rule. Interestingly, a tigloyltransferase from *Datura stramonium* has previously been reported as being able to use both tigloyl-CoA and acetyl-CoA as acyl donors, similar to the activity seen here (Rabot 1995).

To test the activity of AhBAHD1 further, experiments were carried out including AhCYP3 (C-24 hydroxylase). Figure 7.3.3 shows the results of co-expression of AhBAHD1 with all three *A. hippocastanum* CYPs, tHMGR, and SAD1.

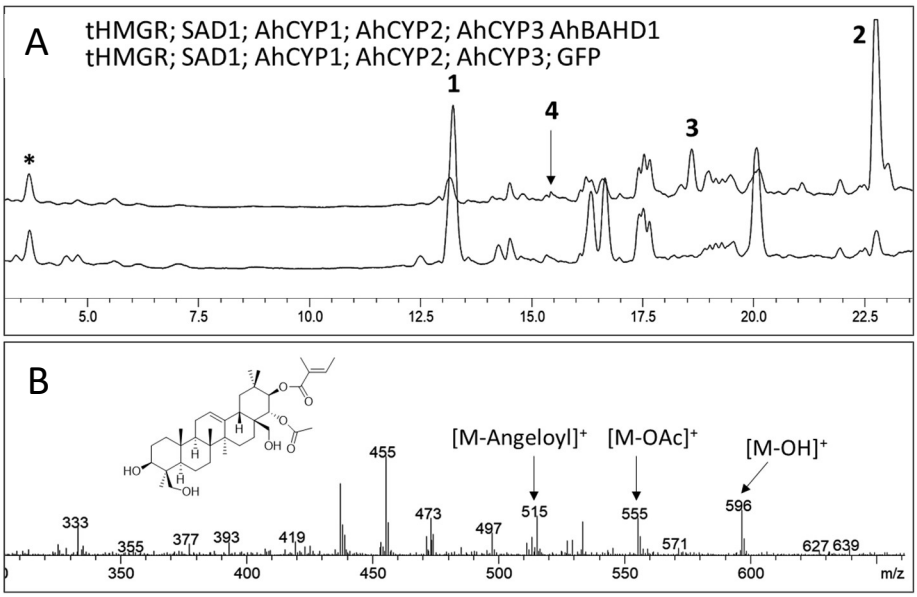


Figure 7.3.3 – Low accumulation of a new acylated product is detected from action of AhBAHD1 on 21β,22α,24,28-OH beta-amyrin. A: **LC** traces (grad A) showing the activity of AhBAHD1 on 21β,22α,24,28 beta-amyrin. * Internal standard visible at 3.6 mins. (1) 21β,22α,24,28-OH beta-amyrin (2) suspected diacylated 21β,22α-OH beta-amyrin derivatives (3) suspected diacylated 21β,22α,28-OH beta-amyrin derivatives. Small peak of new acylated product seen at (4). B: Mass spectrum of peak (4). This is consistent with addition of the same pattern of acylation as previously to 21β,22α,24,28-OH beta-amyrin, namely an acetyl and an angeloyl or a tigloyl, possible structure shown.

The expected product of this experiment, diacylated 21β,22α,24,28-OH beta-amyrin (peak 4 in figure 7.3.3), could be detected through MS, but appeared to only accumulate at low levels. The acylated derivatives of earlier hydroxylated intermediates (21β,22α-OH beta-amyrin and 21β,22α,28-OH beta-amyrin) accumulated at higher levels. Nonetheless, a further experiment was carried out to test the activity of AhBAHD1 on the full protoaescigenin scaffold, formed by co-expression of the three *A. hippocastanum* CYPs with CYP 87D16 (Moses, Pollier et al. 2015). Figure 7.3.4 shows LC-MS data for this experiment.

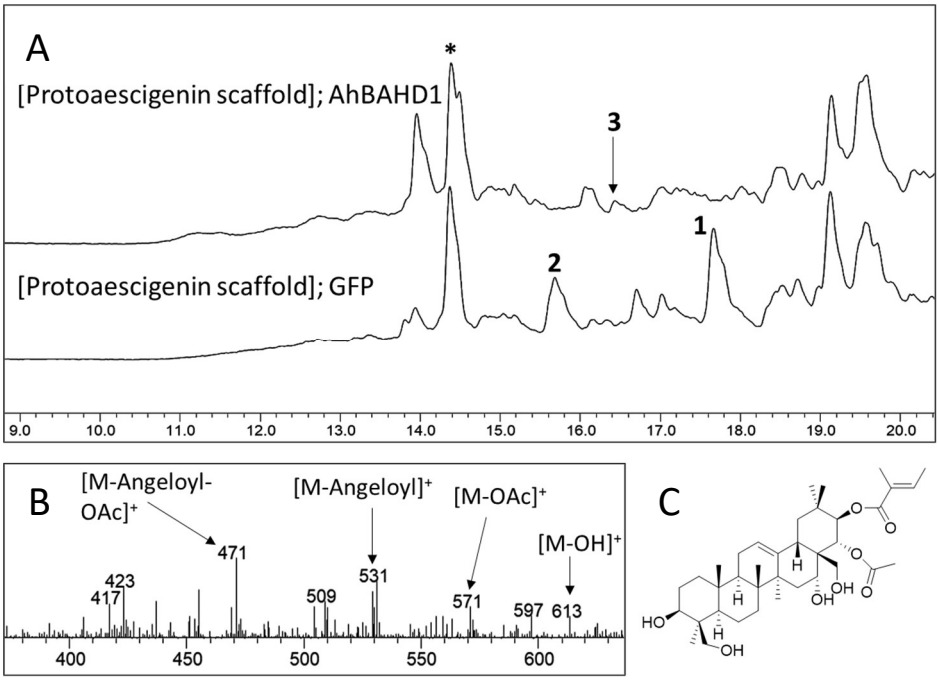


Figure 7.3.4 – Diacylated protoaescigenin is detected on co-expression of AhBAHD1 with the enzymes required to form protoaescigenin. A: Detail from **LC** traces (grad B) showing the activity of AhBAHD1 on protoaescigenin. * Internal standard visible at 14.4 mins. [Protoaescigenin scaffold] consists of tHGMR, SAD1, AhCYP1, AhCYP2, AhCYP3, and CYP 87D16. (1) barringtogenol C (2) protoaescigenin. A small new peak is seen at (3). B: Mass spectrum of peak (3). The same pattern of acylation is seen as in previous AhBAHD1 products, with masses consistent with addition of both acyl groups to the protoaescigenin scaffold. C: Possible structure consistent with B.

As with the experiments on 21β,22α,24,28-OH beta-amyrin, a peak consistent with diacylated protoaescigenin (peak 3 in figure 7.3.4) was detected, albeit only in very small quantities. This represents an important milestone in the project – the possible synthesis of the full aglycone. With the full range of hydroxylated scaffolds using the *A. hippocastanum* and toolkit CYPs, and now the full range of acylated scaffolds using AhBAHD1, testing of glycosyltransferases on a broad range of substrates will be possible.

7.4 – Identification and Characterisation of *AhCSL1*

Addition of the first sugar to the aglycone (a D-glucuronic acid) had previously been assumed to be carried out by a UGT. However, recent reports of the ability of cellulose synthase-like (CSL) enzymes to glucuronate triterpenoids (Chung, Seki et al. 2020, Jozwiak, Sonawane et al. 2020) led to a search for candidate CSL sequences in the *Aesculus* transcriptome. One of these – *DN14871c0g113* – is co-expressed with *ApCYP1* in the *A. pavia* assembly. Its A*. hippocastanum* homologue Ah6798c0g2i1 (*AhCSL1*) was cloned from

conker cDNA into expression vectors and transformed into *Agrobacterium tumefaciens*. To validate potential glucuronylating activity, it was compared against QsCSL, an active CSL from *Quillaja saponaria* in use in the triterpenoid toolkit (Osbourn, Reed et al. 2020). Co-expression of AhCSL1 and QsCSL in *N. benthamiana* together with tHMGR, SAD1, AhCYP1, and CYP 716A12 (the latter four enzymes collectively forming 21β-OH oleanolic acid) is demonstrated through LC-MS data in figure 7.4.1 below.

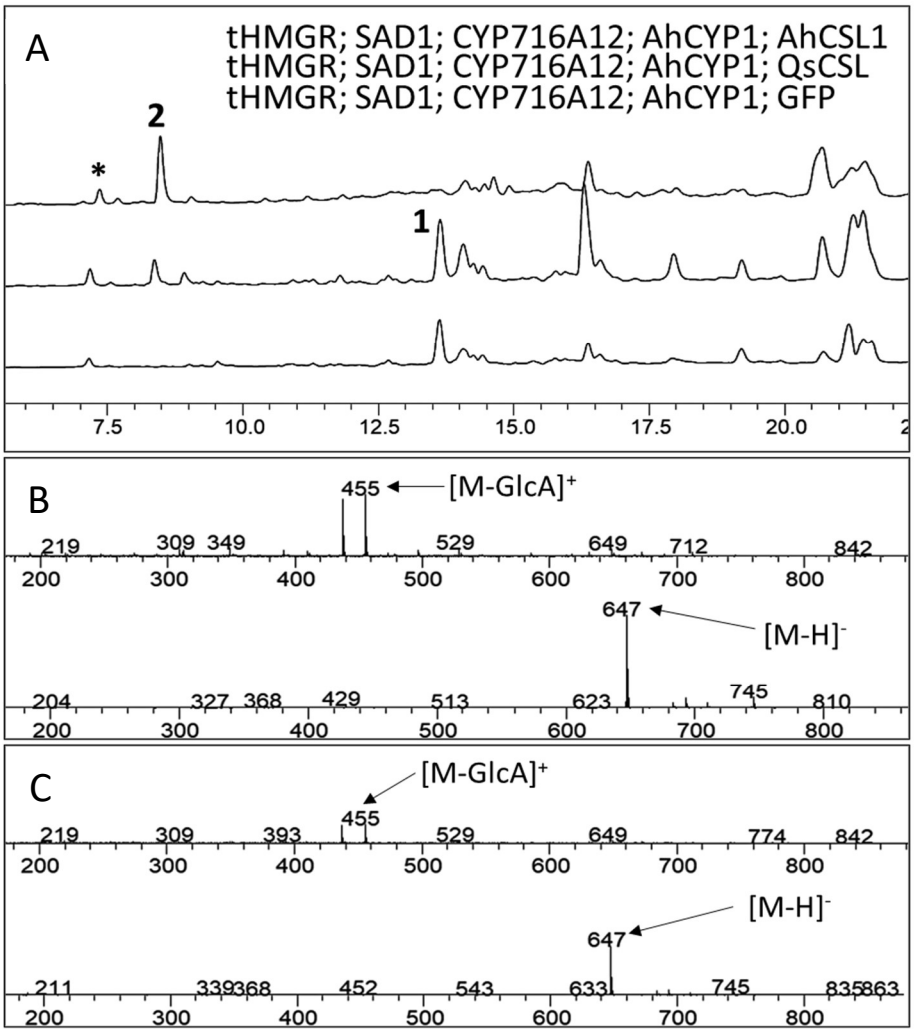


Figure 7.4.1 – The activity of AhCSL1 on 21β-OH oleanolic acid is identical to that of QsCSL. A: **LC** traces (grad D) comparing the activity of AhCSL1 on 21β-OH oleanolic acid with that of QsCSL. * Internal standard (digoxin 20 ppm) visible at 7.2 mins. (1) 21β-OH oleanolic acid. New peak seen at (2) with both AhCSL1 and QsCSL. B: Mass spectra (positive and negative) of (2) with AhCSL1. A negative mode peak caused by loss of a proton is highly characteristic of D-glucuronic acid derivatives. A peak for the aglycone is seen in the positive mode. C: Mass spectra (positive and negative) of (2) with QsCSL. These appear functionally identical to B.

AhCSL1, like QsCSL, was found to be capable of carrying out D-glucuronylation of 21β-OH oleanolic (peak 2 in figure 7.4.1). Interestingly, AhCSL1 appeared to give greater conversion of substrate to product than QsCSL, with the 21β-OH oleanolic acid almost entirely consumed. To further investigate this, the activities of the two CSLs were compared on hydroxylated beta-amyrin scaffolds produced by AhCYP1 and AhCYP2 together, as shown in figure 7.4.2.

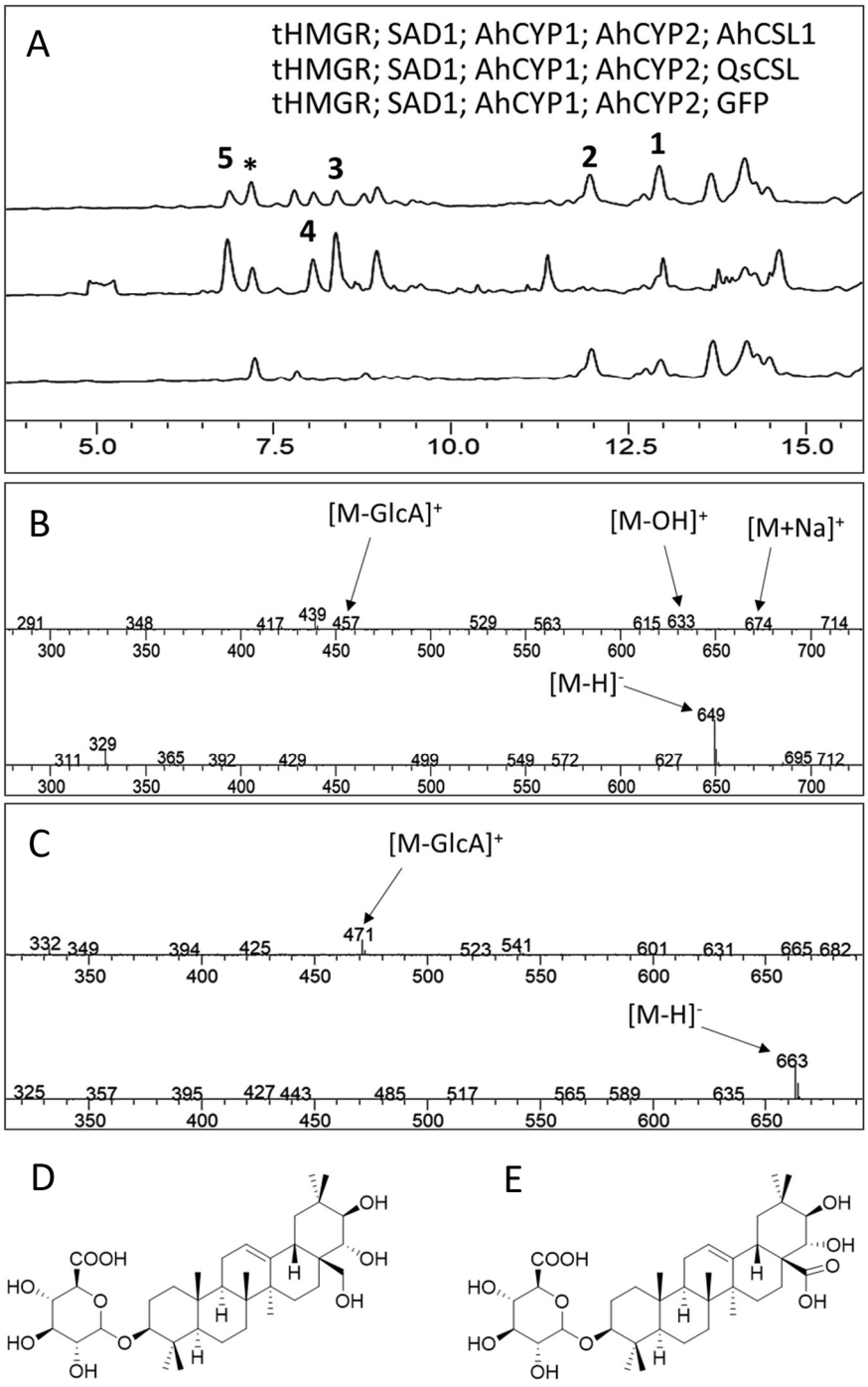


Figure 7.4.2 – AhCSL1 shows identical activity to QsCSL on a range of hydroxylated scaffolds. A: **LC** traces (grad D) comparing the activity of the two CSLs. * Internal standard (digoxin) visible at 7.2 mins. (1) 21β,22α,28-OH beta-amyrin (2) 21β,22α-OH oleanolic acid (3) 21β-OH oleanolic acid D-glucuronate. New peaks seen in both CSL traces at (4) and (5). B: Mass spectra (positive and negative) of peak (4) in trace for AhCSL1. Peaks suggest a trebly-hydroxylated beta-amyrin glucuronate. Mass spectra for QsCSL identical. C: Mass spectra (positive and negative) of peak (5) in trace for AhCSL1, a doubly-hydroxylated oleanolic acid glucuronate is suggested here. Spectra for QsCSL identical. D & E: Inferred structures for B and C, respectively.

These experiments show that AhCSL1 is able to glucuronylate the early aescin intermediates formed by combination of AhCYP1 and AhCYP2. To test the activity of AhCSL1 towards downstream pathway intermediates, further CYPs were included in combinatorial trials. Figure 7.4.3 shows the activity of AhCSL1 towards 21β,22α,24,28-OH beta-amyrin (addition of AhCYP3 to combinations).

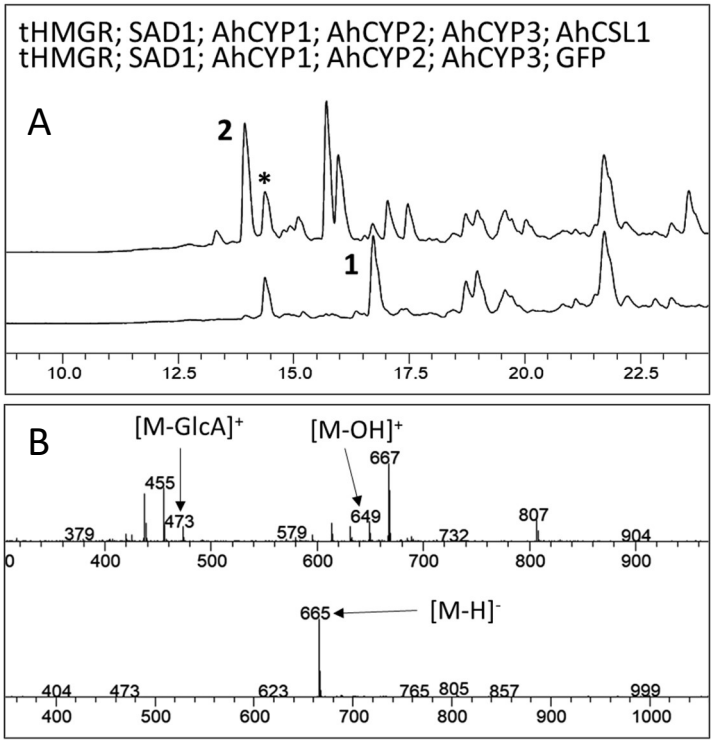


Figure 7.4.3 – AhCSL1 is able to glucuronylate 21β,22α,24,28-OH beta-amyrin. A: **LC** traces (grad B) demonstrating activity of AhCSL1 on 21β,22α,24,28-OH beta-amyrin. * Internal standard (digoxin) visible at 14.4 mins. (1) 21β,22α,24,28-OH beta-amyrin. New peak seen at (2). B: Mass spectra (positive and negative) of peak (2) in A. Addition of D-glucuronic acid to 21β,22α,24,28-OH beta-amyrin is indicated by characteristic negative mode $[M-H]^-$ peak. Aglycone peak in positive mode confirms 21β,22α,24,28-OH beta-amyrin scaffold.

This experiment demonstrates good apparent conversion of 21β,22α,24,28-OH beta-amyrin to the corresponding D-glucuronylated derivative. The resulting product would be only one hydroxylation event removed from protoaescigenin D-glucuronate, which could be an important intermediate for the testing of downstream glycosyltransferases. This final hydroxylation can be achieved through use of CYP 87D16 (Moses, Pollier et al. 2015) from the triterpenoid toolkit (see section 6.9). Figure 7.4.4 below shows the activity of AhCSL1 on protoaescigenin.

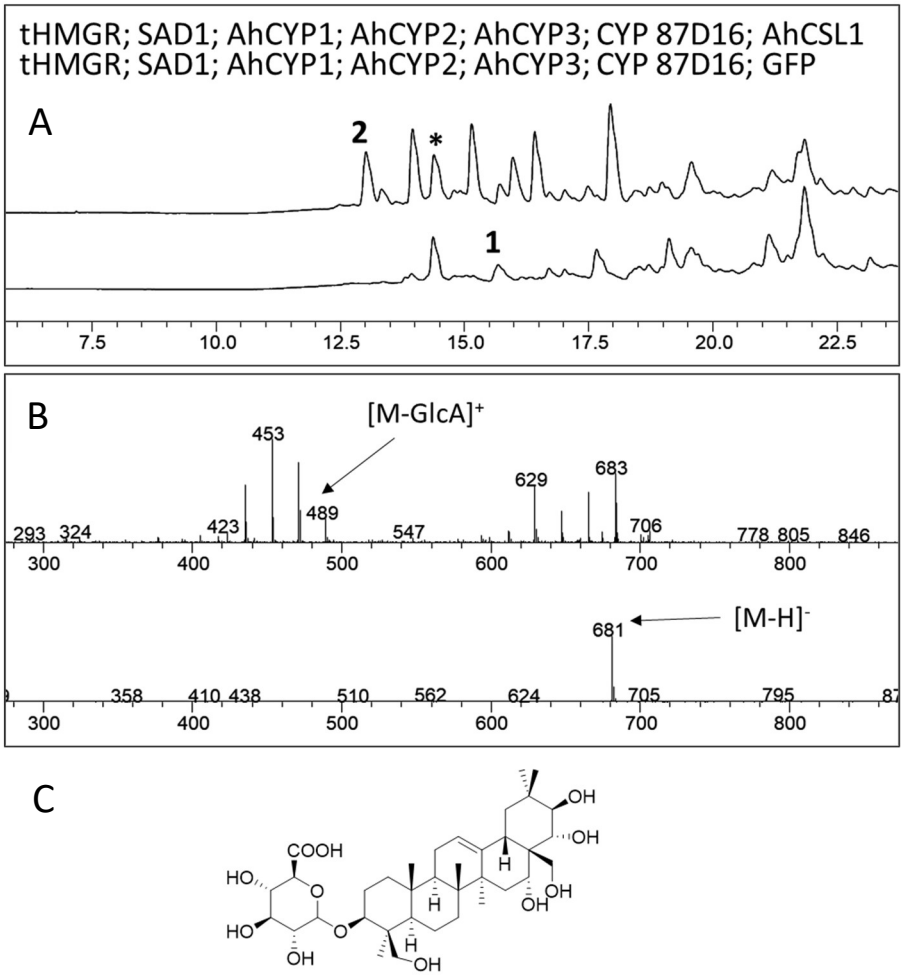


Figure 7.4.4 – Combination of AhCSL1 with *A. hippocastanum* and toolkit enzymes can form protoaescigenin glucuronylate. A: **LC** traces (grad B) demonstrating activity of AhCSL1 on protoaescigenin. * Internal standard (digoxin) visible at 14.4 mins. (1) Protoaescigenin. New peak seen at (2). B: Mass spectra (positive and negative) of peak (2) in A. Characteristic negative mode $[M-H]^-$ peak suggests successful D-glucuronylation of protoaescigenin. C: Inferred structure of B.

These results suggest that AhCSL1 is able to add D-glucuronic acid to the whole range of hydroxylated beta-amyrin derivatives relevant to aescin biosynthesis. This is highly significant, as it allows testing of downstream candidate UGTs on as wide a range of scaffolds as is possible. Experiments were also carried out to establish whether AhBAHD1 and AhCSL1 are active in combination with one another. Figure 7.4.5 shows co-expression of AhCSL1 with tHGMR, SAD1, AhCYP1, AhCYP2, and AhBAHD1.

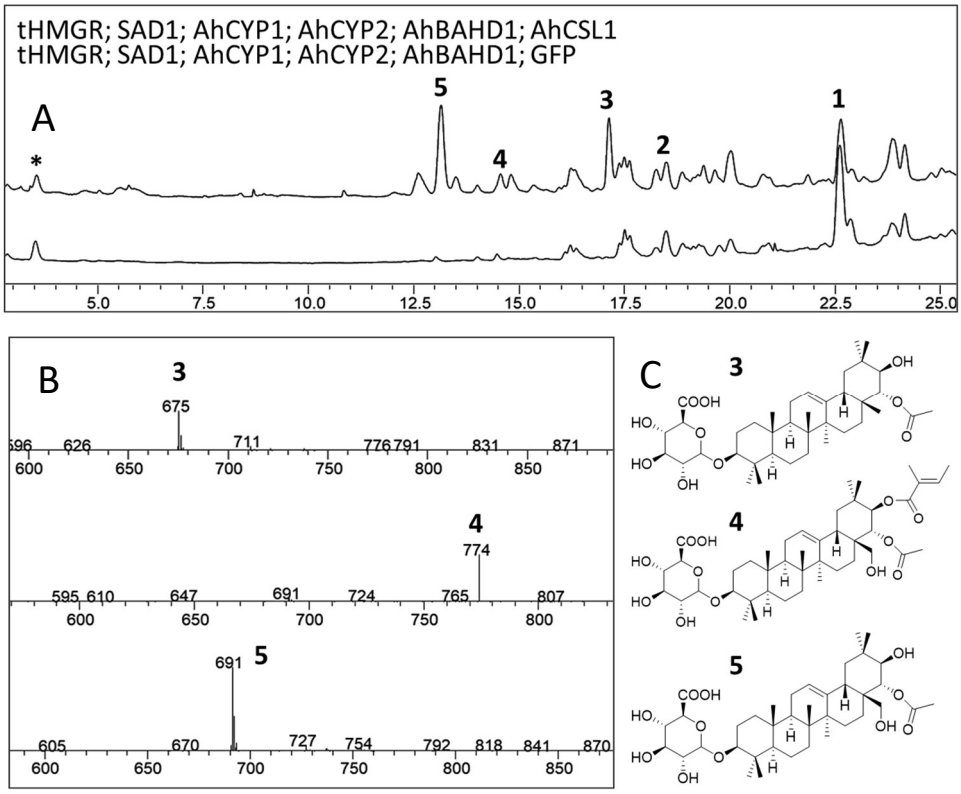


Figure 7.4.5 – AhCSL1 is able to glucuronylate suspected acylated intermediates. A: **LC** traces (grad A) showing activity of AhCSL1 on acylated 21β,22α,28-OH beta-amyrin derivatives. * Internal standard seen at 3.6 mins. (1) suspected diacylated 21β,22α-OH beta-amyrin (2) suspected diacylated 21β,22α,28-OH beta-amyrin. New product peaks seen at (3), (4), and (5). B: Mass spectra (negative mode) of (3) – (5), showing characteristic $[M-H]^-$ peaks. C: Possible structures consistent with mass spectra in B. (3) and (5) appear to be singly acylated with an acetyl group, whereas (4) appears doubly acylated with acetyl and angeloyl / tigloyl.

These traces suggest that AhCSL1 is capable of glucuronylating the presumed acylated products formed by AhBAHD1. Interestingly, when co-expressed with AhCSL1, the suspected singly acylated products of AhBAHD1 are more visible than the doubly-acylated ones, although the latter are still detected. Following the demonstration that these two enzymes are able to work together, combinatorial trials were then carried out incorporating all these enzymes together: the three *A. hippocastanum* CYPs, CYP87D16 from the toolkit, AhCSL1 and AhBAHD1. Figure 7.4.6 shows LC-MS data for these trials.

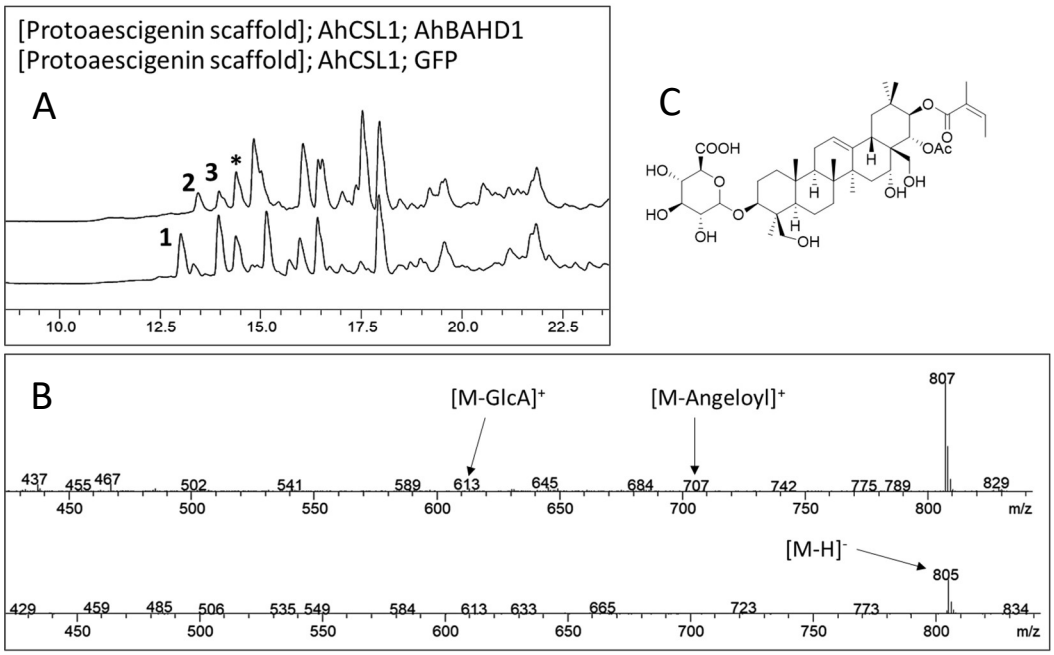


Figure 7.4.6 – AhBAHD1 and AhCSL1 are active in combination with each other on protoaescigenin. A: **LC** traces (grad B) showing the activity of AhBAHD1 on protoaescigenin glucuronate. [Protoaescigenin scaffold] consists of tHMGR, SAD1, AhCYP1, AhCYP2, AhCYP3, and CYP 87D16 * Internal standard (digoxin) visible at 14.4 mins. (1) presumed protoaescigenin glucuronate. New peaks seen at (2) and (3). Peak (2) has mass spectrum consistent with addition of acetyl to (1). B: Mass spectra (positive and negative) of peak (3) in A. The negative mode $[M-H]^-$ peak is consistent with addition of both acetyl and angeloyl or tigloyl to (1). Peaks in the positive mode support D-glucuronylation with an aglycone peak consistent with addition of the two acyl groups. C: Inferred structure of B.

The two enzymes together, in combination with the three *A. hippocastanum* CYPs and toolkit enzymes, appear capable of forming a doubly-acylated protoaescigenin D-glucuronate, which is only two steps removed from aescin. This would provide the full range of aescin intermediates for downstream glycosyltransferase testing. The formation of D-glucuronylated derivatives for so many of the tested intermediates suggests that AhCSL1 is substrate promiscuous, and that transfer of the first sugar onto aescin is not restricted to a strictly linear biosynthetic pathway, but rather can occur at multiple points along the pathway.

7.5 – Characterisation of *ApUGT2*, a D-Galactosyltransferase

The next steps in aescin biosynthesis are glycosylation of the D-glucuronic acid at C'-2 and C'-4. The sugar at C'-4 is always D-glucose, but the sugar at C'-2 can be either D-glucose, D-xylose, or D-galactose. Two relevant D-galactosyltransferases are represented within the triterpene toolkit: GmUGT73P2 (Shibuya, Nishimura et al. 2010), and QsUGT11

(Osbourn, Reed et al. 2020). These enzymes transfer D-galactose onto C'-2 of glucuronic acids in the soyasaponin and QS-21 pathways, respectively. In addition to these two enzymes, ApUGT2 (an *A. pavia* enzyme identified as an aescin biosynthetic UGT candidate by Thomas Louveau from the OneKP (Leebens-Mack, Barker et al. 2019) transcriptome (see chapter 4) and cloned during previous work) was included in combinatorial trials. This was co-expressed with tHMGR, SAD1, AhCYP1, AhCYP2, and AhCSL1 alongside the two toolkit UGTs, and leaf extracts analysed through LC-MS (9.3.5 gradient B) as shown in figure 7.5.1.

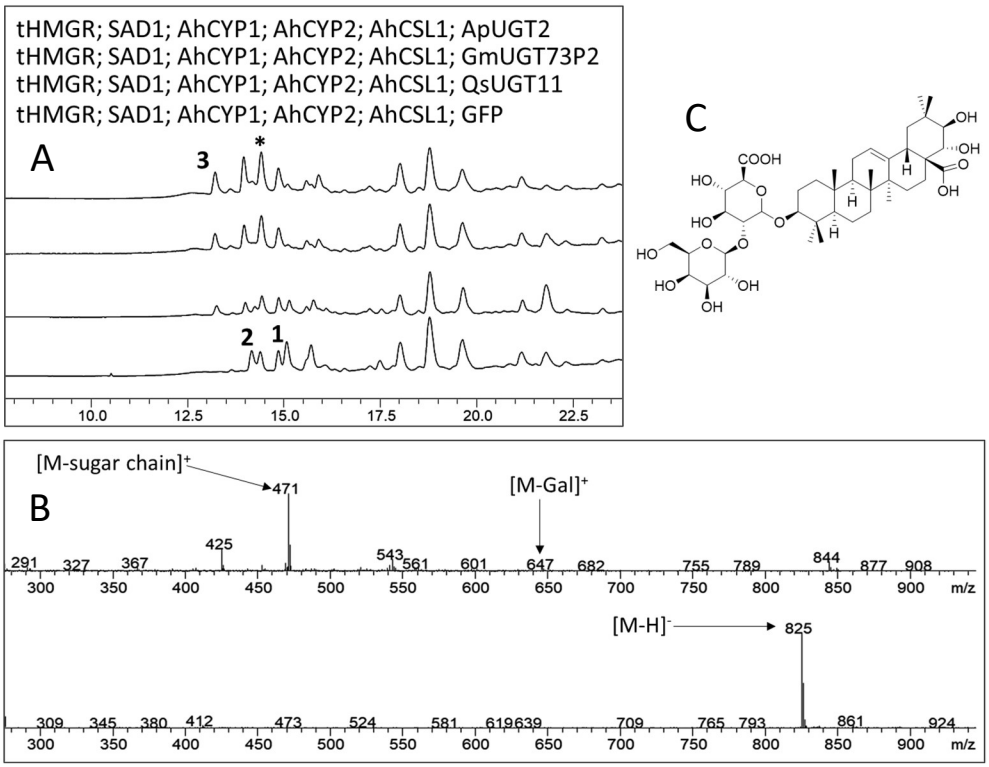


Figure 7.5.1 – ApUGT2 displays identical activity to toolkit D-galactosyltransferases. A: **LC** traces (grad B) comparing the activity of ApUGT2 with toolkit D-galactosyltransferases on glucuronated scaffolds. * Internal standard (digoxin) visible at 14.4 mins. (1) 21β,22α,28-OH beta-amyrin glucuronate (2) 21β,22α-OH oleanolic acid glucuronate. New peak seen at (3). B: Mass spectra (positive and negative) of peak (3) in trace for ApUGT2. The characteristic $[M-H]^-$ peak supports addition of a hexose, as does a peak in the positive spectrum at m/z 647. The spectra for ApUGT2 are identical to those for GmUGT73P2 and QsUGT11. C: Inferred structure of B.

The mass spectra of ApUGT2 products are consistent with addition of a hexose to 21β,22α-OH oleanolic acid glucuronate. Their retention times and mass spectra are identical to the products formed with GmUGT73P2 and QsUGT11, previously characterised triterpene D-galactosyltransferases, which would suggest that ApUGT2 is responsible for addition of the aescin D-galactosyl group. Further investigation into D-galactosylation focussed on the *A. hippocastanum* homologue of ApUGT2 described in the next section.

7.6 – Identification & Characterisation of *AhUGT1*

The discovery that a UGT previously identified from the OneKP transcriptome of *A. pavia* was active as a D-galactosyltransferase led to a search for its homologue in *A. hippocastanum*. This homologue, Ah14918c0g5i1 (*AhUGT1*), was identified using the local blast databases made previously (BLAST+ v2.9.0 – see chapter 5) and obtained as a synthetic gene. *AhUGT1* was cloned and transformed into *Agrobacterium tumefaciens* before co-expression with *Aesculus* CYPs and comparison of its activity with that of ApUGT2. Comparison of the activities of the two enzymes towards 21β,22α,24,28-OH beta-amyrin glucuronate is shown in figure 7.6.1.

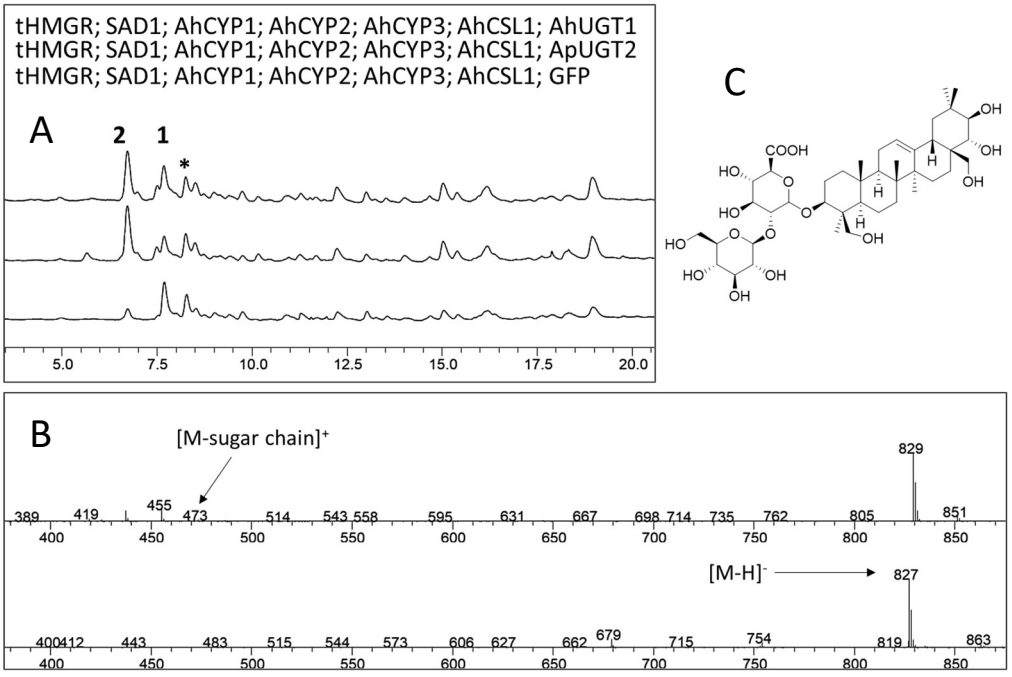


Figure 7.6.1 – AhUGT1 displays identical activity to its homologue, ApUGT2. A: **LC** traces (grad D) comparing the activity of AhUGT1 and ApUGT2 on 21β,22α,24,28-OH beta-amyrin glucuronate. * Internal standard (digoxin) visible at 8.3 mins. (1) 21β,22α,24,28-OH beta-amyrin glucuronate. New peak seen at (2). B: Mass spectra (positive and negative) of peak (2) in trace for AhUGT1. Addition of a hexose is supported by a $[M-H]^-$ peak at m/z 827 in the negative mode spectrum. The spectra for AhUGT1 are identical to those for ApUGT2. C: Inferred structure of B.

The profiles for AhUGT1 and ApUGT2 are very similar, suggesting that AhUGT1 is, like its homologue, active as a C'-2 D-galactosyltransferase. To test the compatibility of AhUGT1 with other aescin biosynthetic enzymes, combinatorial trials were carried out

incorporating AhBAHD1 and CYP 87D16 into co-expression experiments. Figure 7.6.2 shows LC-MS data of these trials.

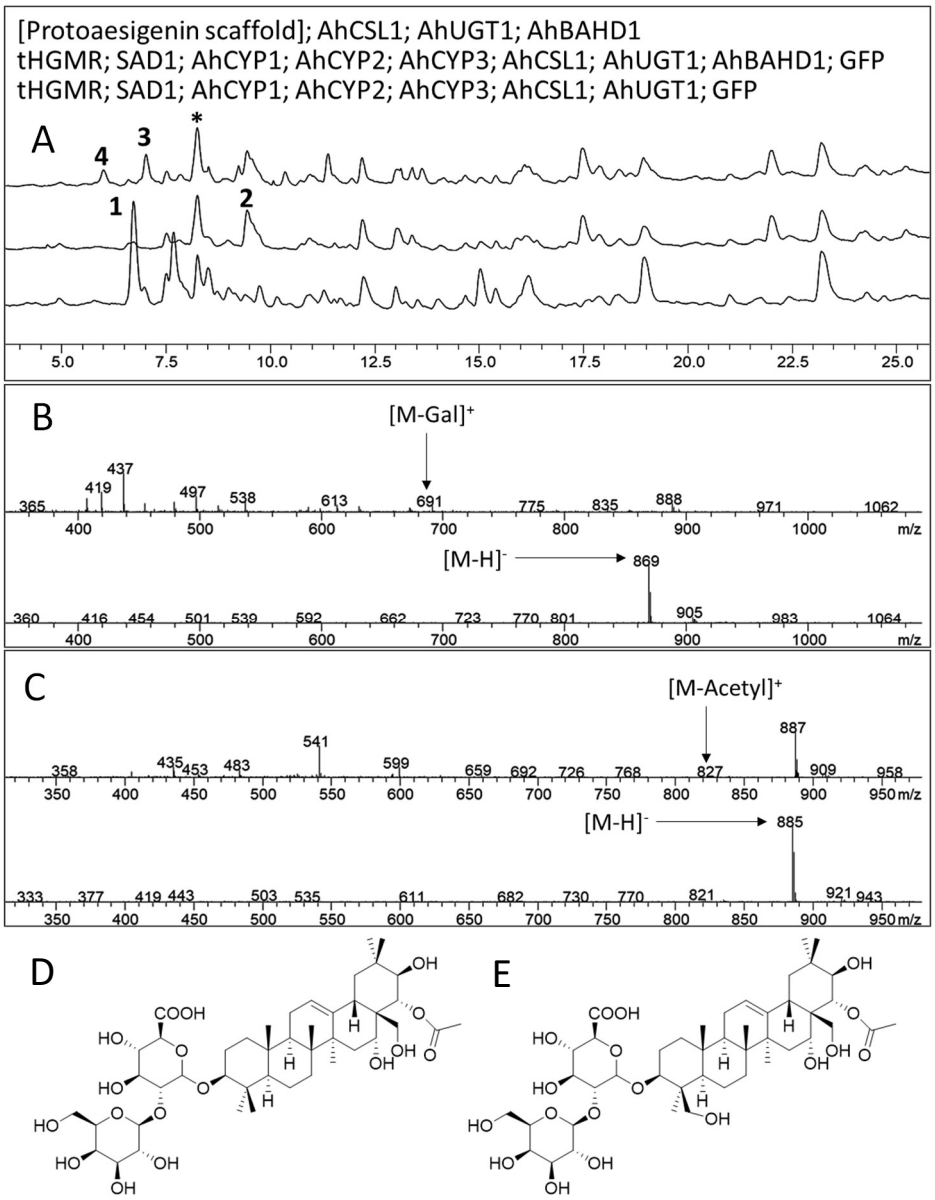


Figure 7.6.2 – AhUGT1 is active in combination with the full range of aescin biosynthetic enzymes. A: **LC** traces (grad D) showing formation of D-galactosyl derivatives using AhUGT1 in combination with aescin biosynthetic genes. * Internal standard (digoxin) visible at 8.3 mins. [Protoaescigenin scaffold] consists of tHGMR, SAD1, AhCYP1, AhCYP2, AhCYP3, CYP87D16. (1) suspected 21β,22α,24,28-OH beta-amyrin glucuronate galactosyl (2) suspected acetyl 21β,22α,24,28-OH beta-amyrin glucuronate galactosyl. Addition of CYP87D16 into the mix lead to formation of new peaks at (3) and (4) B: Mass spectra (positive and negative) of peak (3) in A. Consistent with addition of a hexose to acetyl barrintogenol glucuronate. C: Mass spectra (positive and negative) of peak (4) in A. Consistent with addition of a hexose to acetyl protoaescigenin glucuronate. D: Possible structure for B. E: Possible structure consistent with C.

These combinatorial trials were successful, with AhUGT1 found to be capable of adding D-galactose to presumed acylated protoaescigenin glucuronate (though only the singly acetylated compound could be detected). This would leave only addition of a single sugar (D-glucose to C'-4) remaining to form aescin III.

7.7 – Identification and Characterisation of *AhUGT2*

Enzymes capable of addition of three sugars were required for formation of the full range of aescins – D-glucose to C'-4, D-glucose to C'-2, and D-xylose to C'2. Enzymes capable of these transformations are not represented in the triterpene toolkit, and so must be found from the *A. hippocastanum* transcriptome. The UGT candidates selected by Thomas Louveau from the OneKP transcriptome for *A. pavia* (chosen on the basis of sequence similarity to known active UGTs and conserved domains) were used to search for homologues in the *A. hippocastanum* transcriptome (using local BLAST databases made using BLAST+ v2.9.0). From the 19 candidates identified by Thomas Louveau (minus *ApUGT2*, already cloned and found to be active), fifteen homologues were found in *A. hippocastanum*. Primers were designed for these, and four were cloned and tested: *Ah3338c1g3i1, Ah7803c0g2i1, Ah19168c0g1i1*, and *Ah5240c0g1i1* (*AhUGT2*). *AhUGT2* (homologue to *ApUGT3*) was found to be active when tested on 21β,22α,28-OH beta-amyrin glucuronate, as shown in figure 7.7.1.

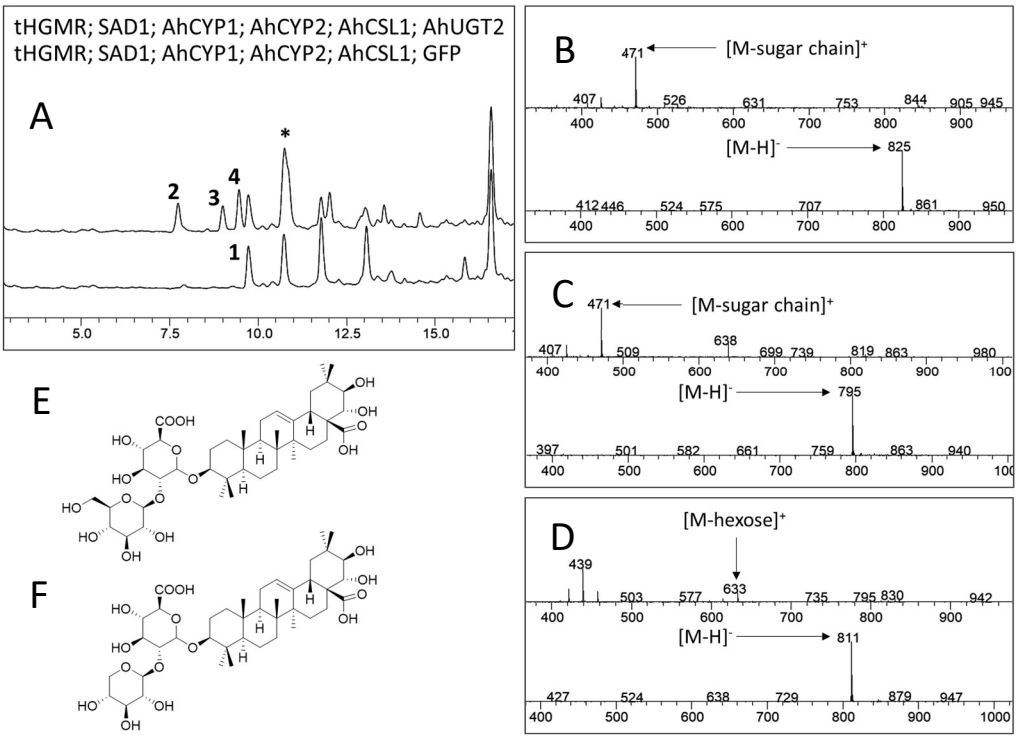


Figure 7.7.1 – AhUGT2 is responsible for formation of new glycosylated products. A: **LC** traces (grad C) showing activity of AhUGT2 on hydroxylated beta-amyrin glucuronates. * Internal standard (digoxin) visible at 10.7 mins. (1) 21β,22α-OH oleanolic acid glucuronate. New products seen at (2), (3), and (4) B: Mass spectra (positive and negative) of peak (2) in A. Consistent with addition of a hexose to 21β,22α-OH oleanolic acid glucuronate. C: Mass spectra (positive and negative) of peak (3) in A. Consistent with addition of a pentose to 21β,22α-OH oleanolic acid glucuronate. D: Mass spectra (positive and negative) of peak (4) in A. Consistent with addition of a hexose to 21β,22α,28-OH beta-amyrin glucuronate. E: Possible structure consistent with B – addition of glucose to C'-2. F: Possible structure consistent with C – addition of xylose to C'-2.

AhUGT2 was found to form multiple new products, which the characteristic $[M-H]^-$ peaks in the negative mode spectra suggested included addition of either a hexose or a pentose to hydroxylated beta-amyrin glucuronates. The only pentose involved in the biosynthesis of aescin is D-xylose, with D-glucose and D-galactose both being hexoses. D-Xylose is found at the C'-2 position of the glucuronic acid in aescin II, as with D-galactose in aescin III. The hexose could be D-glucose at C'-2, at C'-4, or D-galactose at C'-2. To clarify this AhUGT2 was co-expressed with AhUGT1 (C'-2 galactosyltransferase) – if the hexose is D-galactose, the product peaks will be the same as those for AhUGT1, if D-glucose at C'-4, new products might be formed in combination with AhUGT1, and if D-glucose at C'-2, no new products will be formed. Figure 7.7.2 shows LC traces comparing the two UGTs separately and together.

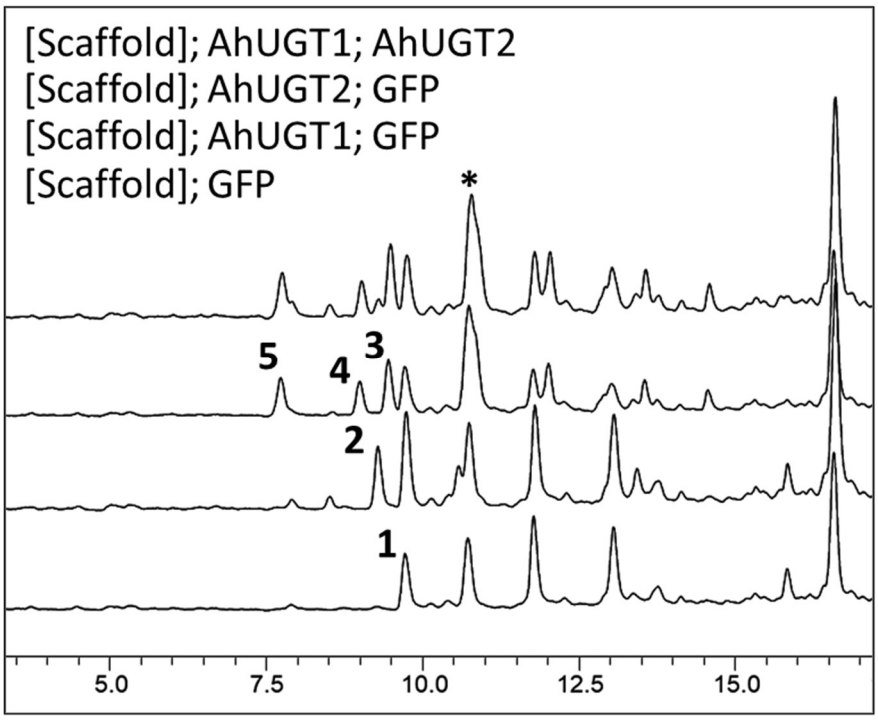


Figure 7.7.2 – AhUGT1 and AhUGT2 form different products which appear to be in competition with each other for substrate. Detail from **LC** traces (grad C) comparing the two UGTs. * Internal standard (digoxin) visible at 10.7 mins. [Scaffold] consists of tHMGR, SAD1, AhCYP1, AhCYP2, and AhCSL1. (1) 21β,22α-OH oleanolic acid glucuronate. (2) 21β,22α-OH oleanolic acid glucuronate galactosyl (3) suspected 21β,22α,28-OH beta-amyrin glucuronate glucosyl (4) suspected 21β,22α-OH oleanolic acid glucuronate xylosyl (5) suspected 21β,22α-OH oleanolic acid glucuronate glucosyl. The products of AhUGT1 and AhUGT2 are distinct from one another, and no new products are formed on combination with each other. AhUGT2 is therefore likely to be responsible for C'-2 D-glucosylation and C'-2 D-xylosylation.

From these traces it can be seen that no new products are formed, and that the hexose product of AhUGT2 is distinct from the D-galactose product of AhUGT1. This would fit addition of D-glucose to C'-2 by AhUGT2, in addition to addition of D-xylose to the same site. Interestingly, the two enzymes appear to be in competition with each other for substrate – on co-expression with AhUGT2, the accumulation of AhUGT1 products is significantly reduced. This further supports addition to C'-2 by AhUGT2, and matches the observed ratio of C'-2 sugars in aescin reported (D-glucose most abundant, then D-xylose, then small quantities of D-galactose) (Price, Johnson et al. 1987). The addition of either D-glucose or D-xylose to C'-2 of glucuronic acid by AhUGT2 is consistent with the known activities of UGTs; though UGTs typically show strong sugar specificity, promiscuity with respect to sugar donor has been reported in some cases, especially between similar sugars (Osmani, Bak et al. 2009). In this case, though D-glucose is a hexose and D-xylose a pentose, the two sugars share the same stereochemistry (see figure 7.1.3), and so the

transfer of either sugar by the AhUGT2 can be readily rationalised. Indeed, there are reports of glucosyltransferases using UDP-xylose as a substrate, but being inactive on UDP-mannose and UDP-galactose (which both involve a change in stereochemistry at one position) (Kubo, Arai et al. 2004).

To test the activity of AhUGT2 on acylated compounds, it was co-expressed with AhBAHD1 in addition to AhCYP1, AhCYP2, AhCSL1, tHGMR, and SAD1. Figure 7.7.3 shows these results through LC-MS data.

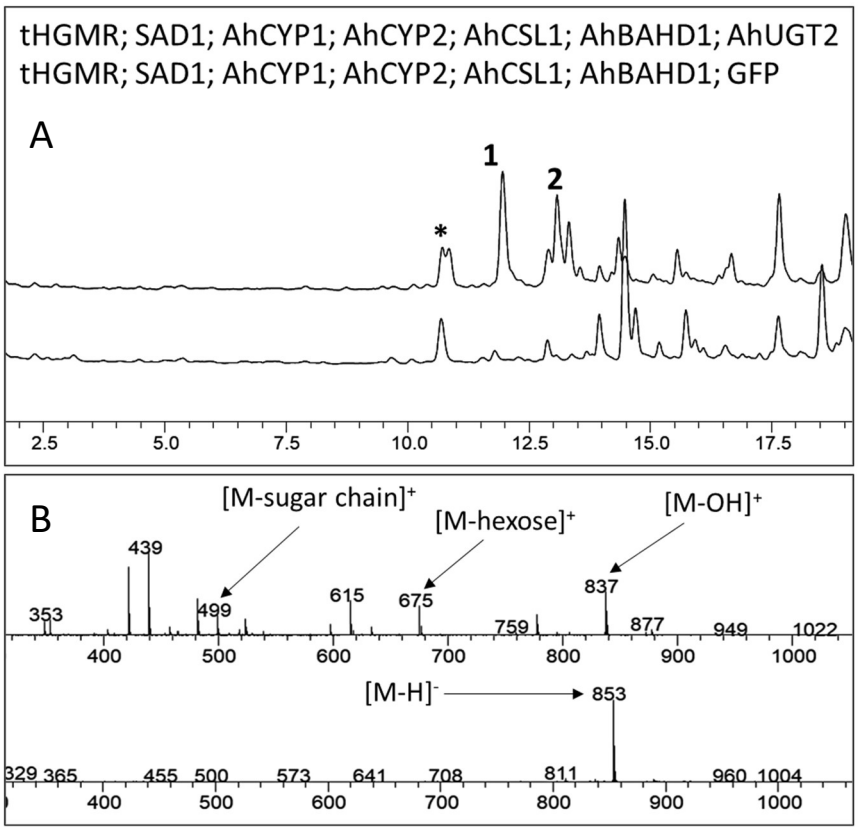


Figure 7.7.3 – AhUGT2 is active on presumed acylated beta-amyrin glucuronates. A: **LC** traces (grad C) showing activity of AhUGT2 on suspected acylated 21β,22α,28-OH beta-amyrin glucuronate. * Internal standard (digoxin) visible at 10.7 mins. New products seen at (1) and (2). (2) is consistent with addition of a pentose. B: Mass spectra (positive and negative) of peak (1) in A. Consistent with addition of a hexose to acetyl 21β,22α,28-OH beta-amyrin glucuronate.

AhUGT2 is active on suspected acylated aescin intermediates, forming peaks with mass spectra consistent with addition of either a hexose (D-glucose) or a pentose (D-xylose) to acetylated 21β,22α,28-OH beta-amyrin glucuronate. This is significant, as with AhUGT1 it suggests formation of all three C'-2 sugars on a range of hydroxylated and acylated beta-amyrin glucuronates, which paves the way for synthesis of the full range of aescins.

7.8 – Conclusion

Significant progress along the aescin biosynthetic pathway has been achieved with the discovery of four new biosynthetic enzymes – AhBAHD1, AhCSL1, AhUGT1, and AhUGT2. AhBAHD1 appears to be capable of addition of two acyl groups simultaneously to the scaffold, and is active upon the whole range of hydroxylated beta-amyrin derivatives relevant to aescin biosynthesis. This enabled detection of the presumed complete aglycone, a major milestone in aescin biosynthesis, and represents novel multifunctionality in triterpenoid acylation. Future work on this project would likely seek to clarify this activity further through full structural determination.

The characterisation of AhCSL1 as an active cellulose synthase-like enzyme responsible for D-glucuronylation of aescin aglycone intermediates adds further to the growing body of CSL enzymes implicated in triterpenoid biosynthesis. AhCSL1 was found to glycosylate a wide range of aescin intermediates, achieving good accumulation of presumed acylated protoaescigenin glucuronates. This first step in the synthesis of the glycosyl moiety of aescin enabled testing of UGT candidates and the subsequent discovery of AhUGT1 (found to be a C'-2 D-galactosyltransferase) and AhUGT2 (suspected to be a C'-2 D-glucosyl- and D-xylosyltransferase). These two UGT candidates were also active in combination with the full range of aescin biosynthetic enzymes, which led to detection of advanced aescin intermediates only a few steps removed from aescin. The multifunctional activity of AhUGT2 makes it the fourth multifunctional enzyme from the aescin pathway (alongside AhOSC1, AhCYP2, and AhBAHD1) – it is possible that these are partially responsible for the apparent diversity of aescins and related saponins isolated from *Aesculus*. Combination of the four newly discovered enzymes from this chapter with the four previously characterised (chapter 6) and toolkit enzymes (chapter 3) enables further re-construction of the aescin biosynthetic pathway, as depicted in figure 7.8.1. Two transformations remain unaccounted for from *A. hippocastanum* enzymes (C-16α hydroxylation and C'-4 D-glucosylation), which, through use of the toolkit, is reduced to a single transformation remaining – the D-glucosylation at C'-4. Future work should seek to identify the enzyme responsible for this D-glucosylation, with which it should be possible to form the full range of aescins.

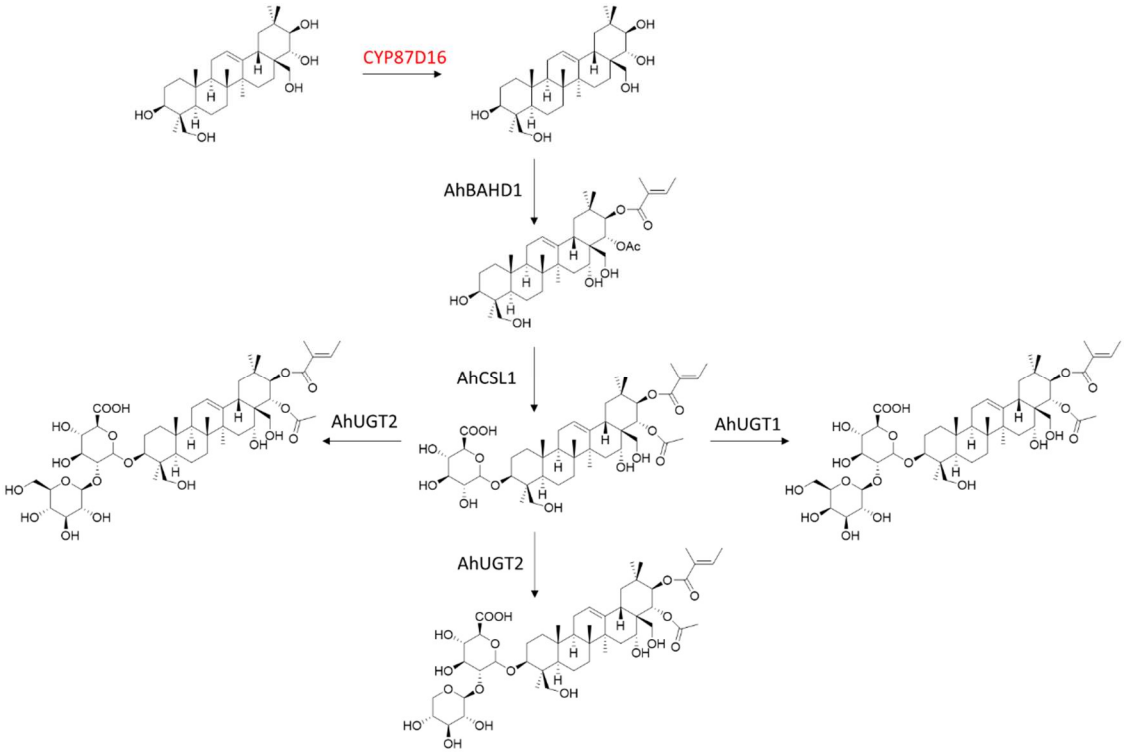


Figure 7.8.1 – Proposed 2nd part of the aescin biosynthetic pathway, using the four new *A. hippocastanum* enzymes discovered here. For the first part of the pathway, see figure 6.10.1. 21β,22α,24,28-OH beta-amyrin can be formed using the *Aesculus hippocastanum* CYPs previously identified, and inclusion of the toolkit enzyme CYP 87D16 achieves protoaescigenin. AhBAHD1 is then responsible for acylation – the exact nature of this is not clear, but it is suspected to add both acyl groups simultaneously. AhCSL1, a cellulose synthase-like enzyme, is responsible for glucuronylation of the aglycone at C-3. These two steps do not seem to follow a strict pathway; the enzymes show activity on a range of intermediates. Addition of sugars to C'-2 is then effected by AhUGT1 and AhUGT2. AhUGT1 adds a D-galactose to this position, as would be required in aescin III, whereas AhUGT2 is thought to add a mixture of D-glucose and D-xylose, as seen in aescins I and II respectively. The only transformation required for aescin synthesis yet unaccounted for is D-glucosylation at C'-4.

Chapter 8 – General Discussion & Conclusions

8.1 – Aescin biosynthetic pathway elucidation using publicly available bioinformatic resources

This work represents the first reported attempt to elucidate the biosynthetic pathway of aescin, a commercially important mixture of triterpenoid saponins from trees of the genus *Aesculus*. Initially, this was informed by database mining of reported triterpenoids isolated from *Aesculus* to infer a predicted biosynthetic pathway (chapter 2). Database mining was also used to focus on the acylation of beta-amyrin derivatives, as this aspect of triterpenoid biosynthesis was one of the least well understood. The results of this work demonstrated that aescin was an excellent model for the study of triterpenoid acylation, having the most common acyl groups on the most frequent positions of the scaffold. Investigation of the activities of reported acyltransferases led to identification of BAHDs as the class of enzymes most likely to be responsible for acylation in aescin biosynthesis. The identification of 21β,22α,28-OH beta-amyrin as the minimum scaffold required for thorough testing of downstream candidate biosynthetic enzymes was highly significant to later work.

The triterpenoid toolkit was subsequently used to reconstitute steps of the aescin biosynthetic pathway (chapter 3), with a particular goal to make the minimum scaffold previously identified. This validated the use of heterologous expression in *N. benthamiana* as a means to investigate triterpenoid biosynthesis, and identified a number of enzymes from the toolkit able to carry out transformations relevant to aescin biosynthesis. The use of these enzymes in combination with one another enabled generation of a range of scaffolds on which to test the activity of candidate biosynthetic enzymes.

The subsequent use of publicly available transcriptomic resources for *A. pavia* provided by the OneKP project (Leebens-Mack, Barker et al. 2019) enabled identification of the first active aescin biosynthetic enzyme: ApCYP1 (chapter 4). This was found to be responsible for 21β- hydroxylation of beta-amyrin and its derivatives. Co-expression of ApCYP1 with toolkit enzymes enabled detection of a new product which could not be achieved through use of the toolkit alone: 21β,24-OH beta-amyrin. Isolation of this product confirmed its identity through NMR and provides the first reported ^{1}H NMR spectrum for this molecule. The OneKP transcriptome also furnished a candidate UGT identified by Thomas Louveau (ApUGT2) that was later shown (chapter 7) to be a D-galactosyltransferase implicated in biosynthesis of aescin III.

Though the OneKP transcriptome was successfully used to identify and characterise the first biosynthetic enzyme of the aescin pathway, this work exposed the limitations of relying on this particular assembly. Without expression data across different tissues, candidate identification from this data proved challenging. The need for improved resources identified, RNAseq experiments were designed and carried out for both *A. pavia* and *A. hippocastanum*, these experiments having been informed by metabolic profiling work detailed in chapter five. These experiments, though limited in their scope by cost constraints, were able to generate data for aescin-type saponin producing and non-producing tissues across the two species. The data generated has been shown by subsequent work in chapters 6 and 7 to be a valuable resource in aescin biosynthetic gene discovery, enabling the discovery of a further *A. pavia* CYP and all eight *A. hippocastanum* genes identified in chapters six and seven. The decision to include both species in transcriptomic efforts has been justified by the switch in focus to *A. hippocastanum* over *A. pavia*, the original focus of this project.

The new transcriptome assemblies have also enabled identification of candidates through multiple means – sequence similarity to known active enzymes, homology to active enzymes in the other *Aesculus* species, functional annotation, differential expression in aescin producing tissues, and co-expression with other biosynthetic enzymes from the pathway. Many of these means were not previously possible using the OneKP resources. Having multiple means of candidate selection has led to rapid identification of active enzymes from both species, which, paired with enzymes from other species represented in the triterpenoid toolkit, allows for production of a range of aescin intermediates in *N. benthamiana*.

8.2 – Characterisation of aescin biosynthetic enzymes from *A. hippocastanum*

The new transcriptomic resources have enabled identification of eight active biosynthetic enzymes from *A. hippocastanum*, namely an OSC (AhOSC1), three CYPs (AhCYP1, AhCYP2, AhCYP3), a BAHD (AhBAHD1), a CSL (AhCSL1) and two UGTs (AhUGT1 and AhUGT2), see chapters 6 and 7; investigation of these enzymes using the transient plant expression system has reconstituted much of the aescin biosynthetic pathway. The discovery of AhOSC1, a mixed amyrin synthase responsible for the first committed step of the aescin pathway (cyclisation of 2,3-oxidosqualene to beta-amyrin), justified a change in focus to *A. hippocastanum* over *A. pavia*. This led to identification of three CYPs, and subsequent validation of their activities through scale-up, purification, and NMR. AhCYP1,

like its homologue ApCYP1, is active as a beta-amyrin C-21β hydroxylase. AhCYP2 is multifunctional, first hydroxylating C-22α and then C-28 (where it can either add a hydroxyl group or oxidise the position up to a carboxylic acid). AhCYP3 is subsequently able to hydroxylate C-24. Co-expression of all three of these CYPs together has led to purification of 21β,22α,24,28-OH beta-amyrin, a putative aescin intermediate which had not previously been reported.

Characterisation of AhCSL1 as responsible for the addition of D-glucuronic acid to the aescin aglycone was achieved through comparison of its activity to that of a known active CSL from *Quillaja saponaria* (Osbourn, Reed et al. 2020). This adds to the growing body of CSL enzymes implicated in D-glucuronylation of triterpenoids, and provides the basis for synthesis of the aescin glycosyl moiety. By comparison with the activity of known C-2' D-galactosyltransferases, the activities of ApUGT2 and AhUGT1 were validated – this transformation is necessary for the synthesis of aescin III. A further UGT was identified but remains to be fully characterised: AhUGT2, which is believed on the basis of LC-MS data to be responsible for the transfer of either D-glucose or D-xylose to C-2'. This is highly significant, as it would imply that AhUGT1 and AhUGT2 are able to carry out the full range of C-2' glycosylations required for formation of the principal aescins I, II, and III.

The active ACT AhBAHD1 has been shown to acylate a range of glycosylated and non-glycosylated aescin intermediates, however, the nature of this acylation is yet to be determined through NMR. LC traces indicate isomeric products, for which MS fragmentation data support addition of both acetyl and angeloyl or tigloyl to the aescin aglycone. Such activity would enable completion of the aescin aglycone. Co-expression of AhBAHD1 with the *A. hippocastanum* CYPs and a C-16α hydroxylase appeared to show accumulation of a product consistent with the full aescin aglycone – structural elucidation of this product will be a key aim of future work on this pathway.

Four of the aescin biosynthetic enzymes identified (AhOSC1, AhCYP2, AhBAHD1, and AhUGT2) demonstrated a degree of multifunctionality – having the ability to produce more than one product from the same substrate, which has led to fewer enzymes being required to complete aspects of the pathway than were initially expected. For example, the sequential addition of two hydroxyl groups by AhCYP2 has meant that the pathway as a whole appears to only require four CYPs, rather than the five that had been predicted in chapter 2. Another feature of some of the biosynthetic enzymes characterised that has become evident is promiscuity: activity with a wide range of putative aescin intermediates

as substrates. Amongst the enzymes displaying comparatively little substrate specificity are AhCSL1 and the two UGTs, which are capable of glycosylation of a wide range of aglycone substrates (see chapter seven). This has implications for the potential use of these enzymes in further bioengineering efforts, making them promising candidates for wider use, for example as part of the triterpenoid toolkit.

Of the enzymes characterised here, three are of note for carrying out activity novel to the toolkit: AhCYP2, which is responsible for hydroxylation at the orphaned positions of C-22β and C-28 (previously only oxidised to an acid in *N. benthamiana*); AhBAHD1, whose activity remains to be fully elucidated, but appears capable of addition of sequential addition of two different acyl groups to C-21α and C-22β; and AHUGT2, which is suspected of addition of either glucose or xylose to C-2'. The activity of AhBAHD1 is particularly relevant to work on other triterpenoid biosynthetic pathways – given how few ACTs have so far been identified as active on triterpenoids, the discovery of a multifunctional ACT active on the most frequently acylated positions of beta-amyrin derivatives (see chapter 2) will likely serve to inform ACT candidate selection in other species.

A genome for *A. hippocastanum* (or any other *Aesculus sp.*) is not presently available, however, were that to change it would be interesting to investigate the locations of the genes identified here. Genes from several previously characterised triterpenoid biosynthetic pathways are clustered into BGCs (biosynthetic gene clusters), including the avenacin pathway in oat (Qi, Bakht et al. 2004, Mugford, Louveau et al. 2013) and the yossoside pathway in spinach (Jozwiak, Sonawane et al. 2020) – it is possible that these genes could be clustered in *Aesculus* too. Such an approach could also form part of a strategy to identify the remaining genes from the pathway, which are likely to be a fourth CYP and a third UGT.

8.3 – Progress along the aescin biosynthetic pathway

Using the *A. hippocastanum* enzymes thus far discovered, it is possible to get to within two steps of the main aescin saponins. When enzymes from the triterpenoid toolkit are included, in particular a CYP capable of C-16α hydroxylation such as CYP 87D16 (Moses, Pollier et al. 2015), only one step is unaccounted for (C-4' D-glucosylation). Figure 8.3.1 shows the biosynthetic pathway so far.

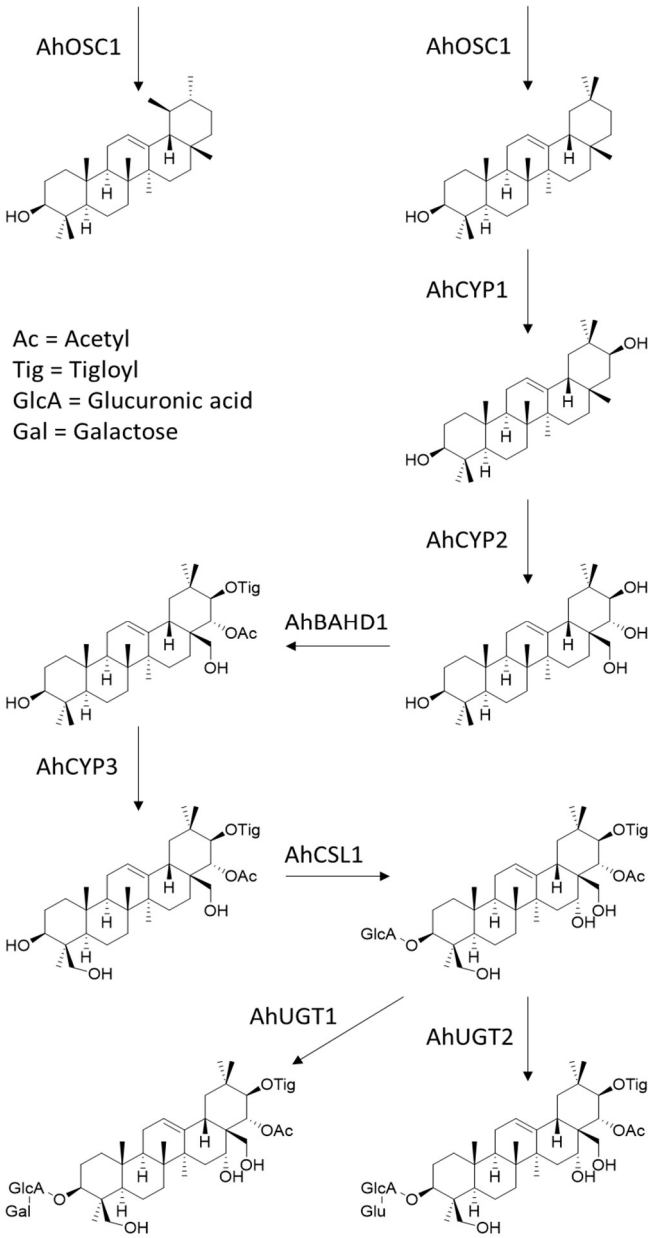


Figure 8.3.1 –Characterised aescin biosynthetic pathway steps. Products of some multifunctional enzymes, eg. AhCYP2, omitted for clarity. The products formed by AhBAHD1 and AhUGT2 are yet to be clarified through NMR. The order of the first few steps is clear, but after AhCYP2 enzymes appear to be flexible in their substrate specificity, such that the order of transformations cannot be deduced. In addition to those shown, AhCYP2 is capable of forming C-28 acid products; AhBAHD1 is likely to form C-21β angleloyl products, and AhUGT2 appears to form C-2' D-xylose products.

The pathway thus far is overall consistent with that predicted in chapter 2. Hydroxylations at C-21β, C-22α, and possibly C-28 appear to be necessary for subsequent transformations, as was anticipated based on knowledge of compounds reported from *Aesculus.* AhBAHD1, AhCSL1, and AhCYP3 all appear to require modifications carried out by AhCYP1 and AhCYP2 in order to show activity. The apparent promiscuity of AhCSL1 and

subsequent UGTs with respect to the aglycone is consistent with the challenges encountered in attempting to predict at which stage these transformations might occur. Overall, later parts of the pathway do not appear likely to follow a strict order.

The C-24 hydroxylating activity of AhCYP3 on 21β,22α,28-OH beta-amyrin had not been predicted – the putative pathway postulated in chapter 5 had predicted that C-24 hydroxylation was likely to be dependent on C-16α hydroxylation. No compounds isolated from *Aesculus* were reported in the Reaxys database as having C-24 hydroxylation but lacking 16α-OH, yet 21β,22α,24,28-OH beta-amyrin was successfully isolated from large-scale transient expression of *A. hippocastanum* CYPs in *N. benthamiana*. It may be the case that the C-16α hydroxylase from the aescin pathway achieves greater rates of conversion to product, so intermediates lacking 16α-OH are only poorly accumulated. Further investigation of this will require identification of the aescin pathway biosynthetic enzyme responsible for this, which is likely to be a CYP.

8.4 – Relevance of findings to bioengineering of saponins

Aescins are of considerable interest to the pharmaceutical (Sirtori 2001, Idris, Mishra et al. 2020) and materials science (Geisler, Dargel et al. 2019, Tucker, Burley et al. 2021) sectors, owing to their vasoprotective, detergent, and emulsifying properties, and as such a practical route to large scale production is highly desirable. In this study the biosynthetic pathway for the major aescins has been reconstituted through use of *Aesculus* and heterologous biosynthetic enzymes, with the exception of the final transformation (C-4' D-glucosylation). Therefore, this work will provide a strong platform for future endeavours to bioengineer aescins and analogues thereof for potential commercial applications. Bio-production of aescins at scale could be achieved through transient expression in *N. benthamiana* – production of triterpenoids at gram scale in *N. benthamiana* has been validated (Reed, Stephenson et al. 2017). Transient expression in *N. benthamiana* or *N. tabacum* is already in use for scale production of high-value pharmaceuticals such as vaccines (Ortega-Berlanga and Pniewski 2022) (a covid vaccine produced in this manner by Medicago has recently been granted approval in Canada). Alternatively, yeast offers another possible heterologous host (Tatsis and O'Connor 2016): yields in excess of 600 mg/l of simple triterpenoids such as oleanolic acid have been reported (Zhao and Li 2018). Furthermore, the identification of the *A. hippocastanum* enzymes responsible for aescin biosynthesis may lead to production of a range of aescin derivatives for use in further pharmacological research.

The characterised enzymes make ideal candidates for incorporation into the triterpenoid toolkit. Some of these enzymes provide for transformations which have for some time been desired additions to the toolkit, for instance hydroxylation at C-28. As part of the toolkit, these enzymes could enable further progress on production of other triterpenoids. This might be especially true for those saponins structurally similar to the aescins, many of which have also seen industrial interest. The saponins from tea, for instance, share many structural features with aescin and are also desirable to the pharmaceutical industry (Murakami, Nakamura et al. 1999, Toshio Morikawa, Ning Li et al. 2006, Yoshikawa, Morikawa et al. 2007). The enzymes characterised from this work could potentially find use in reconstructing the pathways to saponins such as these.

8.5 – Prospects for future work

To fully leverage the progress made thus far on this project, some work remains in elucidating the aescin biosynthetic pathway. Future work should seek to clarify the nature of the action of AhBAHD1 and AhUGT2 through NMR. Another priority is identification of the remaining two enzymes – most likely a CYP responsible for hydroxylation at C-16α and a UGT capable of C-4' D-glucosylation. There remain candidates for both of these enzymes that this work has identified but that have not yet been tested; this should form the starting point of future work. Once these two transformations can be achieved, testing can be carried out to determine whether all these enzymes, upon co-expression, are capable of forming appreciable quantities of aescins in *N. benthamiana*. Investigation of alternative routes to aescin production using these enzymes ought also to be investigated, for example through stable germ-line *N. benthamiana* transformants, or through heterologous expression in yeast.

Concluding Remarks

This work has pioneered the elucidation of the aescin biosynthetic pathway in *Aesculus*. This has been greatly aided by generation of fresh transcriptomic resources for *A. pavia* and *A. hippocastanum*, and use of the triterpenoid toolkit, a collection of active triterpenoid biosynthetic enzymes held by the Osbourn group. Selection of candidates using these transcriptomic resources and co-expression in *N. benthamiana* alongside toolkit enzymes enabled identification of aescin biosynthetic enzymes. Altogether, eight active enzymes have been identified from *A. hippocastanum*: an OSC, three CYPs, a BAHD, a CSL, and two UGTs. A number of these show multifunctional activity, which has led to production of a range of advanced intermediates from the aescin pathway. It is hoped that this work will lead to identification of the remaining enzymes, and thus the elucidation of the entire aescin pathway. This work brings the end goal of a bio-engineering approach to production of aescin significantly closer.

Chapter 9 – General Materials and Methods

9.1 – Molecular Biology Methods

9.1.1 – RNA extraction – Sigma spectrum

Plant material was taken from -70 ^{o}C storage and kept on dry ice to keep frozen. Material was ground to a fine powder in a pre-cooled, sterile mortar and pestle with liquid nitrogen to avoid thawing. 100 mg of powdered plant material was measured out into DNase/RNase-free Eppendorf tubes. RNA was extracted from these samples using the Spectrum Plant Total RNA (Sigma-Aldrich/Merck) following the manufacturer's instructions for protocol B, but with repetition of binding column wash steps with wash solutions 1 and 2 to improve RNA purity. Before elution of RNA, binding columns were treated with 10 µl of RQ1 RNase-free DNase (Promega) made up with buffer according to manufacturer's instructions, left for 20 minutes and treated with stop solution before continuing with elution. The purity and yield of RNA obtained was assessed through optical measurement with a NanoDrop spectrophotometer (Thermo). RNA for RNAseq experiments was subsequently assessed by the Earlham Institute using a Perkin-Elmer GX II to generate RNA quality scores. RNA was stored at -70 ^{o}C.

9.1.2 – Synthesis of cDNA

RNA was removed from cold storage and thawed on ice. cDNA synthesis was carried out with SuperScript IV Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions, using DEPC-treated water and RNasin Recombinant Ribonuclease Inhibitor (Promega). Resultant cDNA was treated with DNase-free RNase ONE Ribonuclease (Promega) according to manufacturer's instructions and stored at -20 ^{o}C for later use.

9.1.3 – Cloning PCR conditions (iProof and Q5) and primers

Cloning PCR reactions were carried out on a 50 µl scale using a Mastercycler Pro VapoProtect PCR machine (Eppendorf). Cloning PCR reactions were either carried out using iProof High-Fidelity DNA Polymerase (Bio-Rad) or Q5 High-Fidelity DNA Polymerase (New England BioLabs), both using nuclease-free water according to manufacturer's instructions, for 30 cycles per reaction. Exact conditions (eg. annealing temperature, cDNA concentration, $MgCl_2$ concentration) varied according to the melting point of the primers used. Amplification of products with the correct size was verified through gel

electrophoresis. PCR products were purified using the QIAquick PCR Purification kit (Qiagen), using the lower volumes of wash buffer and elution buffer as recommended by the manufacturer's protocol. The primers used in each chapter are listed in the table below; forward primers include an AttB1 site and reverse primers an AttB2 site for subsequent Gateway cloning. Primers for CYPs and BAHDs cloned in chapter 4 were designed for 2-step PCR – an initial 25-cycle 10 µl reaction with the specific primer was used as a template for a subsequent 15-cycle 50 µl reaction using AttB1 and AttB2 primers. Subsequent primers were designed for 1-step PCR with the full AttB1 and AttB2 site included. All primers were purchased from Sigma-Aldrich.

Name	**Sequence**
Gateway site primers	
AttB1	GGGGACAAGTTTGTACAAAAAAGCAGGCT
AttB2	GGGGACCACTTTGTACAAGAAAGCTGGGT
AttL1	CAAATAATGATTTTATTTTGACTGATAGT
AttL2	CTATCAGTCAAAATAAAATCATTATTTG
Tomato CYP cloning (chapter 3)	
CYP716A44 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGTTGTTGTATGTCTGTCTTG
CYP716A44 R	GGGGACCACTTTGTACAAGAAAGCTGGGTTATCACGGGTTGTGGTGGTGAGGGTAGAGTC
CYP716A46 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAATTGTTCTATGCCTCTCTTGT
CYP716A46 R	GGGGACCACTTTGTACAAGAAAGCTGGGTTATTAAGTTTTAACGTGATGAGGATAGAGTCT
A. pavia CYP & OSC cloning from OneKP transcriptome (chapter 4) (* indicates 2-step PCR)	
2033231 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTGGAAGCTGAAGACAGC
2033231 R	GGGGACCACTTTGTACAAGAAAGCTGGGTTACTAGTTCCTATAGGCTGCATAA
2020895 F *	AAAAAGCAGGCTTAATGGAGTTCTTCTTTGTTTGTGG
2020895 R *	AGAAAGCTGGGTATCACTTGTGAGGGATAAGACG
2029750 F *	AAAAAGCAGGCTTAATGGAGTTGTTCTTGGTGATC
2029750 R *	AGAAAGCTGGGTATCAATGAGGAATAAGACGAACG
2008407 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGACTCCCTCTTGGGCTC
2008407 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTACATCTCATAAAGATGAGCAGG
2015618 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGTTCTCGCCTTTTACC
2015618 R	GGGGACCACTTTGTACAAGAAAGCTGGGTTATCACAGATGTTGGAAGATGATTG
A. pavia BAHD cloning from OneKP transcriptome (chapter 4) (* indicates 2-step PCR)	
2005581 F *	AAAAAGCAGGCTTAATGGCTATCCACTCCAC
2005581 R *	AGAAAGCTGGGTATCAAATTTCAATTACTGTAGCAGC
2017756 F *	AAAAAGCAGGCTTAATGGAGGTTTCAGTTGTTTC
2017756 R *	AGAAAGCTGGGTATCATATTCTTTGACATTTTCTAATATCTGG

2027681 F *	AAAAAGCAGGCTTAATGGCTGAAAACACTTTACAAG
2027681 R *	AGAAAGCTGGGTATCAACGTCTACTCTCCGCC
2263010 F *	AAAAAGCAGGCTTAATGGCGTCAATGGAAGTG
2263010 R *	AGAAAGCTGGGTATCAATGCTCCATGAGAAACTC
2000297 F *	AAAAAGCAGGCTTAATGGAAATCCATATCATTTCTAAAG
2000297 R *	AGAAAGCTGGGTACTAAGAAGAAGAAGCATAAGC
2010294 F *	AAAAAGCAGGCTTAATGAAAATGCATGTAAAGAAATC
2010294 R *	AGAAAGCTGGGTACTAAAAGTCATACAAACAGTTTTTG
2010319 F *	AAAAAGCAGGCTTAATGTCTGATGTTCTTAAACAGC
2010319 R *	AGAAAGCTGGGTACTAAGAAGCATAAGCAACGAC
2016716 F *	AAAAAGCAGGCTTAATGGACGTTCAAATTCTTTCC
2016716 R *	AGAAAGCTGGGTACTATATATCTTTATCATTTTGGAAGAAAGC
2029913 F *	AAAAAGCAGGCTTAATGGAAGTTAAAATCCTGGAAAC
2029913 R *	AGAAAGCTGGGTACTATTTGAGCTCTGGTTTGAG
2032487 F *	AAAAAGCAGGCTTAATGAAGCTTCACTACTTGATAGG
2032487 R *	AGAAAGCTGGGTACTATTTCTTCTTGGCTCGC
2038261 F *	AAAAAGCAGGCTTAATGGCAGCCGGGTGGAT
2038261 R *	AGAAAGCTGGGTACTAAGGAATTGAGACAATAATGTTGAGTCC
2040175 F *	AAAAAGCAGGCTTAATGTTGTGGTTGAATGTGCC
2040175 R *	AGAAAGCTGGGTACTAGAGCCCATCTACAAAAAGGG
2042989 F *	AAAAAGCAGGCTTAATGCTCTCCTGTCATTACATTC
2042989 R *	AGAAAGCTGGGTACTACGATACGTACTGCATGAAC
2043977 F *	AAAAAGCAGGCTTAATGGAGATCCGTATCATTTCTAG
2043977 R *	AGAAAGCTGGGTACTAGGCCGGAGCAAAAC
2261566 F *	AAAAAGCAGGCTTAATGTGCGACGCCACAGG
2261566 R *	AGAAAGCTGGGTACTACACCGGCACCACAATG
A. pavia OSC & CYP cloning from new transcriptome (chapter 6)	
5582c0g1i10 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTGGAAACTGAAGACAGC
5582c0g1i10 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAAGCAGCAGTAATCTTGGATG
225c0g1i4 F *	AAAAAGCAGGCTTAATGGCACCCCTTTTCCTTG
225c0g1i4 R *	AGAAAGCTGGGTATCATATGTATGTGATAACATTACATTTAATAAAAAGTC
2476c0g1i4 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGCTCATCAGTACTCC
2476c0g1i4 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTACATCTCATAAAGATGAGCAGG
4078c1g1i4 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCAATCACAGCTCTC
4078c1g1i4 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAGGCAGCAAATGGG
4414c1g1i4 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGACTTTCTTCTTTCTCCAC
4414c1g1i4 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTACAGGGCACTGTGATAGAC
8439c0g1i3 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGTTTGAAGCTTCTTTGTG

8439c0g1i3 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCACCATCCCCTTTCTTCAAC
15243c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGGGATTCGAGTTTG
15243c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTATAACTTTCTCACCAAAAGATCG
16823c0g1i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAATTTCCTTCTTCAATCTC
16823c0g1i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTATTGGTAGAGATGAAGGG
24067c1g1i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGTTATTCTTCTTCTCTGG
24067c1g1i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGCTTGGTGAGGTATAAGAC
40570c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGACTCTCTACAGTCTACCC
40570c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTACTGGGTCTGGAGGTAC
103483c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGTTCTTCGTTAATGGC
103483c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAAAGCTTGTGAGGAATAAGACG
A. hippocastanum OSC & CYP cloning (chapter 6)	
331c0g1i3 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTGGAGGCTGAAGATTGC
331c0g1i3 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAACCATTTTTATAGGTAGGCAAAGG
331c0g2i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTGGAAGCTGAAGACAGC
331c0g2i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAAGCAGCAGTAGTCTTGGAAG
4815c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGTTCTTCTTTGTTTG
4815c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAAGCCTTGTGAGGGA
3872c1g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGTTGTTCTTGGTTATAC
3872c1g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAGTGATGAGGAATAAGACG
4924c0g1i7 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCAATCACAGCTCTCATG
4924c0g1i7 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAGGCAGCAAATGGGTTG
8376c0g2i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGGTGTCCATCGAGC
8376c0g2i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTACAGGGCACGGTGATAC
13762c0g1i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGTTGTTCATTTTCTG
13762c0g1i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAACTAGATCTTTCACCGG
16676c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGGGTTTTGGGTTTG
16676c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTACAACTTTCTCACCAGAAGATC
21534c0g2i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCACCCCTTTTCCT
21534c0g2i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAGGTACATTGTTCATGATGG
21534c0g3i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTCCCATTCTCCTTG
21534c0g3i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAGTACTTTGTTGGTGGG
A. hippocastanum BAHD cloning (chapter 7)	
2407c0g1i3 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGTTGCCAAAAAAGGAG
2407c0g1i3 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAGTTCTTTTCTGAGGTAATTAAG
3616c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCATCAAAATCGAGTCAAG

3616c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAAGCTTGCATAACATGTTG
4415c0g1i6 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGGTTTCAGTTGTTTCC
4415c0g1i6 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCATATTCTTTGACATTTTCTAATATCTGG
4755c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTGGTATTCGGGTG
4755c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAGTAGTAATGGTGACAGCA
4755c0g1i8 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTGGTATTCGGGTG
4755c0g1i8 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGTGGACAGTACAGTATCCAT
5291c0g2i3 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCACCACCCCCC
5291c0g2i3 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTACAAAGCAGAAGTGATAAAACTGG
6318c0g3i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGATAATCAACGTGAAGCAG
6318c0g3i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAATGTCATATATGAACTTCTCAAAC
6318c0g3i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGATAATCAATGTGAAGCAGTC
6318c0g3i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAATGTCATATATGAACTTCTGAAAC
9064c0g1i7 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTGAAAACACTATACA
9064c0g1i7 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAATAAATTTGAGGCCTTGA
9861c0g1i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGGTTCGAATTATCTCC
9861c0g1i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAGATTTTGGATTGGAGAATCG
10410c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCGTCAATGGAAGTG
10410c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATGCTCCATGAGAAACTCC
10648c0g1i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGATTCCAAAACTGTCAAAC
10648c0g1i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTACAAAGAGGCGGCAG
13680c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGATCCGTATCATTTCTAGG
13680c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTATGGAGAAAGAGAATATTCCGG
16012c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTTCAAATCTGAAGAGCT
16012c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTATTCACTATCCAAAAACTCTTTC
16012c0g1i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTTAAAATCTCAAGAGCTCC
16012c0g1i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTATTCACTATCCAAAAACTCTTTC
16796c0g3i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCATCAAACCATAACT
16796c0g3i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCATCTATCTCCTTCCGG
17020c0g1i3 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAAAAAGCAGCATCATATC
17020c0g1i3 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCATGTGGTTGGATTAAAGGAAG
22311c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCAGCACAGCCAAG
22311c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAGCTAGTGCTCAAAAGGG
23384c0g1i2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCGGCATTTCAGC
23384c0g1i2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATGTACATCCAAACCTTTCTG
23384c0g2i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTTGGCTTTTCTGT
23384c0g2i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATGTACATCCAAACCTTTCTG
25465c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGATGTCAGTATCATTTCC

25465c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAAGAAGCATAAGCAATAATTTCAG
27763c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAGTGCTGTGCAAATCATATG
27763c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTATATGGTTGGGTTGGATGAAG
28436c1g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCGAGCTCGTCTGT
28436c1g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAGCATACACTCCTGATAAGATGC
30407c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCAAACCACTGCAA
30407c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAGCACAATTATTATTATTATCAGTAT
31072c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTGTTTATTTATTTACGGATTTG
31072c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAAAGTCGTAAAAACAGTT
36402c1g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAACTGGAGGCTCA
36402c1g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCACAGTCGAAGATGAGC
37835c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCATCTTCCTCTACTAC
37835c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAAGACGGCATTGCA
43477c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTCTCGCAGTAATC
43477c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAATTGTTTCCTTCTTCATGA
46030c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGTTTTGGGGTTTGAAG
46030c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTAGATTTGCACAAGTTCTTG
77698c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAGTTAAAATCCTGGAA
77698c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAAATTCTTTGTTACTGAACTT
95885c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAATTCTTCGTGCCTCC
95885c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTACTATGTTCGAGAAGTGATGGC
A. hippocastanum CSL & UGT cloning (chapter 7)	
6798c0g2i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGGTGATCTCCTTCATTCC
6798c0g2i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAATTTGTTTTGCTCTTGCTTGA
3338c1g3i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCTACTTCCAGTTCCAG
3338c1g3i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAAGTGTTGGTCTTCTTCC
5240c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAAGCAAGCAAGAG
5240c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATTACAGTACATGAGTTTTCTTTG
7803c0g2i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGATTCAGGGTCGAGC
7803c0g2i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAATGAGAATTCAGCTTCAATTCT
19168c0g1i1 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAAGTATAGTTATATATCCAGC
19168c0g1i1 R	GGGGACCACTTTGTACAAGAAAGCTGGGTATCAGTTCCGCATCCATAAC

9.1.4 – Gateway cloning into expression vectors

Gateway cloning was used to clone amplified sequences into expression vectors, as described in Esposito, Garvey et al. (2009). Purified PCR products were used in a BP cloning reaction using Gateway BP Clonase II (Invitrogen) to transfer the amplified sequence into

donor vector pDNR207, according to the manufacturer's instructions. This was transformed into MAX Efficiency DH5α *E. coli* competent cells (Invitrogen) by heat shock for 30 seconds at 42^{o}C. These were incubated for 1 hour at 37^{o}C in SOC medium (see 9.1.8), before being plated onto LB-agar with 20 µg ml^{-1} gentamycine (Melford) and incubated with agitation at 37^{o}C overnight.

Colonies on the plate were subjected to colony PCR (see 9.1.5) to determine colonies with a correct size insert. These were incubated in 10 ml LB (see 9.1.8) with 20 µg ml^{-1} gentamycine at 37^{o}C overnight. Plasmid extraction from liquid cultures was carried out using the QIAprep Spin Miniprep kit (Qiagen) according to manufacturer's instructions, and the yield of plasmids quantified using a NanoDrop spectrophotometer (Thermo). The primers AttL1 and AttL2 (forward and reverse, respectively, see table) were used to verify the sequence of the plasmid inserts (sequencing carried out by GATC and Eurofins).

Plasmids with correct inserts were used for LR cloning reactions using Gateway LR Clonase II (Invitrogen) to transfer the insert into the expression vector pEAQ-HT-DEST1 (Sainsbury, Thuenemann et al. 2009). This was then transformed through heat shock as for BP plasmids into DH5α cells before plating onto LB-agar (see 9.1.8) with 50 µg ml^{-1} kanamycine (Formedium) and incubation at 37 ^{o}C overnight. Colonies were used to set up liquid cultures in 10 ml LB with 50 µg ml^{-1} kanamycine, incubated with agitation overnight at 37 ^{o}C and used for plasmid extraction as previously. Plasmids were stored at -20 ^{o}C.

Candidates obtained as synthetic genes (Ah1968c0g1i4, Ah3545c0g2i4, Ah15779c0g1i1, and Ah14918c0g5i1 (AhUGT1)) had sequences designed with the AttB1 and AttB2 sites added onto the ends, as with the primers. The sequences of these are listed in Appendix C1. These were resuspended and used in the Gateway BP reaction as with purified PCR products.

9.1.5 – Colony PCR

Colonies resultant from transformation of BP products were sampled for colony PCR. At least 4 colonies per construct were sampled using a sterile pipette tip and dipped into 50 µl of sterile deionised water. This was heated to 95 ^{o}C for 15 mins, before 2 µl was used as material for PCR. These PCR reactions were carried out using GoTaq G2 Flexi DNA polymerase (Promega) according to the manufacturer's instructions, for 30 cycles (annealing temperature of 55 ^{o}C) in a 20 µl reaction. Amplification was verified through gel electrophoresis.

9.1.6 – Preparation of competent cell aliquots of *Agrobacterium tumefaciens*

Agrobacterium tumefaciens strain LBA 4404 was plated onto LB-agar with 50 µg ml^{-1} rifampicin (Melford) and 100 µg ml^{-1} streptomycin (Sigma-Aldrich) and incubated at 28 ^{o}C for three days. From this was started a 60 ml liquid culture (LB, 50 µg ml^{-1} rifampicin and 100 µg ml^{-1} streptomycin), which was incubated overnight at 28 ^{o}C with agitation at 200 rpm. This culture was cooled on ice and pelleted through centrifugation. The cells were re-suspended in 1 ml of a pre-cooled aqueous solution of 20 mM $CaCl_2$ and aliquots of 50 µl made. These aliquots were flash-frozen in liquid nitrogen and stored at -70 ^{o}C for later use.

9.1.7 – Transformation of expression vectors into *A. tumefaciens*

Aliquots of competent LBA 4404 cells were thawed on ice over two hours before addition of 100 ng of pEAQ-HT-DEST1 plasmids with relevant insert (9.1.4) were added and mixed gently. Transformation was carried out through cold shock in liquid nitrogen for five minutes. Aliquots were then thawed at room temperature, before addition to 1 ml of LB and incubation at 28 ^{o}C for 4 hours. Cells were then plated onto LB-agar with 50 µg ml^{-1} rifampicin, 50 µg ml^{-1} kanamycine, and 100 µg ml^{-1} streptomycin and incubated at 28 ^{o}C for three days. Colonies were used to start 10 ml liquid cultures (LB-rif-kan-strep) and incubated overnight at 28 ^{o}C with agitation. Glycerol stocks of these cultures were formed by mixing 1200 µl of culture with 300 µl of glycerol for a 20 % glycerol solution; these were vortexed very thoroughly for a homogenous solution before storage at -70 ^{o}C.

9.1.8 – Recipes for media used

Ingredients for LB broth	Concentration (g l^{-1})
Typtone	10
Yeast extract	5
NaCl	10

Ingredients for LB-Agar	Concentration (g l^{-1})
Tryptone	10
Yeast extract	5
NaCl	10
Agar (Formedium)	11

Ingredients for SOC broth	Concentration (g l^{-1})
Tryptone	20
Yeast extract	5
NaCl	0.58
KCl	0.186
$MgCl_2$	2.03
$MgSO_4$	2.46
Glucose	3.6

9.2 – Transient expression in *Nicotiana benthamiana* methods

9.2.1 – Small scale infiltration of *N. benthamiana*

Experiments were carried out as described by Reed, Stephenson et al. (2017). Glycerol stocks of desired strains of *Agrobacterium tumefaciens* were struck onto LB-agar (with 50 µg ml^{-1} rifampicin, 50 µg ml^{-1} kanamycine, and 100 µg ml^{-1} streptomycin) plates and incubated for 3 days at 28 ^{o}C. From these plates LB-rif-kan-strep liquid cultures were inoculated and incubated overnight at 28 ^{o}C with agitation at 200 rpm. Centrifugation of liquid cultures at 4000 rpm was used to pellet the cells, before re-suspension in a minimal quantity of MMA buffer (an aqueous solution of 10 mM MES hydrate (Sigma-Aldrich, buffered to pH 5.5 with NaOH), 10 mM $MgCl_2$, and 150 µM acetosyringone (Sigma-Aldrich)) and incubation in the dark at room temperature for 2 hours. Using a spectrophotometer (7315 – Jenway) to measure the optical density of cultures at 600 nm (OD_{600}), cultures were diluted with further MMA buffer such that each strain would have an OD_{600} of 0.2 in each combination for infiltration. The combinations were thus mixed, with a strain transformed

with GFP (control) used to maintain equal concentrations of *Agrobacterium tumefaciens* in each combination.

N. benthamiana plants kept in containment glasshouses were used at approximately 5 weeks old. Young leaves from near the top of the plant were punctured with a sterile pipette tip to aid infiltration. The combination of strains of *Agrobacterium tumefaciens* suspended in MMA was infiltrated using a syringe, with three leaves per combination infiltrated. These were harvested with scissors between five and seven days after infiltration, frozen at -70 ^{o}C, and freeze-dried for analysis (see 9.3).

9.2.2 – Large-scale infiltration for triterpenoid purification

Large-scale experiments were carried out using 100 plants, as described in Stephenson, Reed et al. (2018). This involved preparation of liquid cultures as in 9.2.1, but on a 1 litre scale. As with small scale experiments, these cultures were pelleted through centrifugation and re-suspended in MMA, before dilution and mixing strains together to the correct concentration (measured by OD_{600}). This solution was used to fill a basin part-way, into which 4 inverted *N. benthamiana* plants at a time were placed, only the aerial parts of the plants touching the solution. The basin with the plants was placed into a vacuum chamber and subjected to reduced air pressure (~100 mbar), before a rapid return to atmospheric pressure. Plants are then removed from the solution and placed upright again. As with small scale experiments, harvest of the leaves occurs after 5-7 days. The leaves are frozen and freeze-dried for subsequent extraction.

9.3 – Extraction & Analysis of Infiltrated Leaves

9.3.1 – EtOAC extraction of leaf material for GC-MS analysis

Dried leaf material was crushed and 20 mg measured. This was powdered through agitation with tungsten carbide beads. To this was added 300 µl EtOAc with coprostanol (Sigma-Aldrich) as an internal standard (either 100 ppm or 200 ppm). Extraction was carried out using sonication at 55 ^{o}C for 30 mins. After brief centrifugation to settle leaf material, 50 µl supernatant was transferred to a glass vial. This was evaporated under nitrogen, and the dry extract re-suspended in 50 µl TMS-imidazole (Sigma-Aldrich) as a derivatising agent. The re-suspended extracts were heated to 70 ^{o}C for 30 mins and transferred to a glass vial insert to be analysed undiluted.

9.3.2 – MeOH extraction of leaf material for LC-MS analysis

Dried leaf material was crushed and 10 mg measured. This was powdered through agitation with tungsten carbide beads. To this was added 550 µl aqueous 80% MeOH with 20 µg ml^{-1} digoxin (Sigma-Aldrich) as an internal standard. Extraction was carried out at room temperature through agitation at 1400 rpm for 20 mins. After brief centrifugation to settle leaf material, 400 µl supernatant was transferred to a fresh vial. This was twice washed with 400 µl ice-cold hexane to remove fats. The methanolic extract was then dried using an EZ-2 Elite evaporator (Genevac) and re-suspended in 75 µl pure MeOH. The re-suspended extracts were filtered through 0.22 µm nylon filter columns (Norgen Biotek). 50 µl of filtrate was transferred to a glass vial insert and diluted with 50 µl of 50% aqueous MeOH, ready for injection into LC.

9.3.3 – Extraction of large-scale experiments using pressurised solvent extraction

Dry leaf material was broken up and mixed 4:1 with clean quartz sand. This was used to fill 120 ml cells for the SpeedExtractor E-914 (Büchi), as described in the manufacturer's instructions. Extraction was carried out at 100 bar and 90 ^{o}C in four cycles, the first with hexane and the subsequent three with EtOAc. The hold time for each cycle was 5 mins, and the discharge time 3 mins. Hexane extracts and EtOAc extracts were collected separately. The EtOAc extracts were evapourated under reduced pressure before re-suspension in EtOH. This re-suspended extract was treated with Ambersep 900 Hydroxide (Sigma-Aldrich) ion-exchange beads, following the protocol set out in Stephenson, Reed et al. (2018). The ion-exchange beads were removed by filtration through a layer of celite (Sigma-Aldrich). The ethanolic extract is then adsorbed onto a small volume of silica gel and dried under reduced pressure, before loading onto a column for separation through flash column chromatography (hexane-EtOAc gradient).

9.3.4 – GC-MS analysis of derivatised leaf extracts

GC-MS was carried out using a 7890 B GC (Agilent) fitted with a ZB5-HT Inferno (30 m length, 0.25 mm diameter) column (Zebron) connected to a 5977 B MSD (Agilent). 1 µl of sample prepared as described in 9.3.1 was injected to the inlet (inlet temperature 250 ^{o}C, inlet flow 24 ml min^{-1}) in pulsed split mode (pulse pressure 30 psi, splitting ratio 20:1). A column temperature gradient was programmed as follows: hold 2 min at 170 ^{o}C, from 2-8.5 min ramping 170-300^{o}C, from 8.5-20 min hold 300 ^{o}C. MS detection (EI) was started after 8

min solvent delay, scanning 7.2 scans sec^{-1} from m/z 60-800. Data was analysed on MassHunter software (Agilent).

9.3.5 – LC-MS analysis of methanolic leaf extracts

LC-MS was carried out on a Nexera X2 Prominence LC linked to a LC-2020 mass spectrometer (Shimadzu). The LC was fitted with a 50x2.1 mm 100 Å 2.6 µm Kinetex XB-C18 column (Phenomex) using water (0.1% formic acid) and MeCN (0.1% formic acid) gradients in a column oven set to 30 ^{o}C. A total flow rate of 0.6 ml min^{-1} was split between the MS and a CAD detector (Thermo Scientific). The MS was fitted for DUIS using both electrospray ionisation and atmospheric pressure chemical ionisation using a corona needle. The heat block temperature was 400 ^{o}C and the desorbation line temperature was 250 ^{o}C. Nebulising gas was set to 1.5 l min^{-1}, and drying gas to 15 l min^{-1}. Mass spectra were collected from m/z 50-1500 in both positive and negative modes.

Gradient A (chapter 5, chapter 7): 25% MeCN from 0-10 min. From 10-26 min gradient from 25-70%, from 26.5-28 min 70-95%. Held at 95% to flush column from 28-29 min. From 29-30 min return from 95-25% MeCN.

Gradient B (chapter 6, chapter 7): 15% MeCN from 0-10 min. From 10-26 min gradient from 15-70%, from 26.5-28 min 70-95%. Held at 95% to flush column from 28-29 min. From 29-30 min return from 95-15% MeCN.

Gradient C (chapter 7): 15-25% MeCN from 0-10 min, followed by 25-60% from 10-26 min, then 60-95% from 26.5-28 min. Held at 95% to flush column from 28-29 min. From 29-30 min return from 95-15% MeCN.

Gradient D (Chapter 7): 15-30% MeCN from 0-10 min, followed by 30-70% from 10-26 min, then 70-95 from 26.5-28 min. Held at 95% to flush column from 28-29 min. From 29-30 min return from 95-15% MeCN.

9.3.6 – Quantification of the accumulation of 21β-OH beta-amryin (chapter 4)

10 *N. benthamiana* leaves of a similar size were infiltrated for each combination. After five days, leaves were harvested and dried. Samples were prepared for analysis as in 9.3.1, with the exact mass of leaf material used for each sample recorded. GC analysis was carried out as in 9.3.4. The area of the peaks in each sample for the internal standard (200 ppm), beta-amyrin, and 21β-OH beta-amyrin was integrated using MassHunter analysis software. The peak areas for beta-amyrin and 21β-OH beta-amyrin were divided by the peak area for the

internal standard and multiplied by 200 to give the estimated concentration in the extract in ppm. This was converted into mg present in the extract, and then divided by the mass of leaf material of the sample to give the estimated accumulation in mg g^{-1} dry leaf tissue.

9.3.7 – Separation of compounds through flash column chromatography

Crude extracts prepared according to 9.3.3 were separated using an Isolera One flash column chromatography machine (Biotage) with appropriately sized Sfär Silica HC or SNAP Ultra columns (Biotage). Initial separation was with a hexane-EtOAc gradient, subsequent separation of fractions was achieved with isocratic DCM-EtOAc elution. Investigation of the contents of fractions was carried out through TLC (visualised using the reagents described in 9.3.8) or through GC or LC (as 9.3.4 and 9.3.5).

9.3.8 – TLC visualising agents used

Vanillin was prepared with 15 g vanillin (Sigma-Aldrich), 2.5 ml conc. H_2SO_4, and 250 ml EtOH. This was used as a TLC plate dip before heating to visualise spots.

p-Anisaldehyde was prepared with 3.5 ml *p*-anisaldehyde (Sigma-Aldrich), 15 ml AcOH, 350 ml EtOH, and 50 ml conc. H_2SO_4. This was sprayed onto TLC plates, heating to visualise spots.

Chlorosulphonic acid was prepared using 5 ml chlorosulphonic acid (Sigma-Aldrich) and 10 ml acetic acid. This was sprayed onto TLC plates and heated. Spots visualised under UV light.

9.3.9 – Triterpenoid NMR (general considerations)

Products purified through 9.3.7 were dried thoroughly under reduced pressure using an EZ-2 Elite evaporator (Genevac) before being dissolved in NMR solvent. This was either $CDCl_3$ (Sigma-Aldrich) for 21β-OH beta-amyrin or pyridine D-5 (Sigma-Aldrich) for other compounds. NMR were measured on a AVIII machine (Brucker) at 400 MHz for ^{1}H NMR and 100 MHz for ^{13}C NMR. Chemical shifts were measured in ppm referenced to residual solvent peaks. Data was analysed using TopSpin software (Brucker).

Chapter 10 - Bibliography

Abe, Y., M. Tera, N. Sasaki, M. Okamura, N. Umemoto, M. Momose, N. Kawahara, H. Kamakura, Y. Goda, K. Nagasawa and Y. Ozeki (2008). "Detection of 1-O-malylglucose: Pelargonidin 3-O-glucose-6''-O-malyltransferase activity in carnation (Dianthus caryophyllus)." Biochemical and Biophysical Research Communications **373**(4): 473-477.

Afendi, F. M., T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, M. Altaf-Ul-Amin, L. K. Darusman, K. Saito and S. Kanaya (2012). "KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research." Plant and Cell Physiology **53**(2): e1-e1.

Agrawal, P. K., R. S. Thakur and J. N. Shoolery (1991). "Application of 2D Nmr Spectroscopy to the Structural Establishment of the Major Hydrolysis Product of Aescin." Journal of Natural Products **54**(5): 1394-1396.

Aharoni, A., L. Gaidukov, O. Khersonsky, S. M. Gould, C. Roodveldt and D. S. Tawfik (2005). "The 'evolvability' of promiscuous protein functions." Nature Genetics **37**(1): 73-76.

Akagi, T., A. Ikegami, Y. Suzuki, J. Yoshida, M. Yamada, A. Sato and K. Yonemori (2009). "Expression balances of structural genes in shikimate and flavonoid biosynthesis cause a difference in proanthocyanidin accumulation in persimmon (Diospyros kaki Thunb.) fruit." Planta **230**(5): 899-915.

Aktas, R. N. and I. Tontul (2021). "Usability of soapwort and horse chestnut saponin extracts as foaming agents in foam mat drying of pomegranate juice." Journal of the Science of Food and Agriculture **101**(2): 786-793.

Alabdul Magid, A., N. Lalun, C. Long, N. Borie, H. Bobichon, C. Moretti and C. Lavaud (2012). "Triterpene saponins from Antonia ovata leaves." Phytochemistry **77**: 268-274.

Alqathama, A., L. Shao, A. Bader, P. Khondkar, S. Gibbons and J. M Prieto (2020). "Differential Anti-Proliferative and Anti-Migratory Activities of Ursolic Acid, 3-O-Acetylursolic Acid and Their Combination Treatments with Quercetin on Melanoma Cells." Biomolecules **10**(6): 894.

Andre, C. M., S. Legay, A. Deleruelle, N. Nieuwenhuizen, M. Punter, C. Brendolise, J. M. Cooney, M. Lateur, J. F. Hausman, Y. Larondelle and W. A. Laing (2016). "Multifunctional oxidosqualene cyclases and cytochrome P450 involved in the biosynthesis of apple fruit triterpenic acids." New Phytologist **211**(4): 1279-1294.

Asres, K., A. Seyoum, C. Veeresham, F. Bucar and S. Gibbons (2005). "Naturally derived anti-HIV agents." Phytotherapy Research **19**(7): 557-581.

Avtzis, N. D., D. N. Avtzis, S. G. Vergos and S. Diamandis (2007). "A contribution to the natural distribution of Aesculus hippocastanum (Hippocastanaceae) in Greece." Phytologia Balcanica **13**(2): 183-187.

Bai, Y., P. Fernández-Calvo, A. Ritter, A. C. Huang, S. Morales-Herrera, K. U. Bicalho, M. Karady, L. Pauwels, D. Buyst, M. Njo, K. Ljung, J. C. Martins, S. Vanneste, T. Beeckman, A. Osbourn, A. Goossens and J. Pollier (2021). "Modulation of Arabidopsis root growth by specialized triterpenes." New Phytologist **230**(1): 228-243.

Bak, S., F. Beisson, G. Bishop, B. Hamberger, R. Höfer, S. Paquette and D. Werck-Reichhart (2011). "Cytochromes P450." The Arabidopsis Book **9**: e0144.

Barr, I. G., A. Sjölander and J. C. Cox (1998). "ISCOMs and other saponin based adjuvants." Advanced Drug Delivery Reviews **32**(3): 247-271.

Bartley, L. E., M. L. Peck, S.-R. Kim, B. Ebert, C. Manisseri, D. M. Chiniquy, R. Sykes, L. Gao, C. Rautengarten, M. E. Vega-Sánchez, P. I. Benke, P. E. Canlas, P. Cao, S. Brewer, F. Lin, W. L. Smith, X. Zhang, J. D. Keasling, R. E. Jentoff, S. B. Foster, J. Zhou, A. Ziebell, G. An, H. V. Scheller and P. C. Ronald (2013). "Overexpression of a BAHD Acyltransferase, OsAt10,

Alters Rice Cell Wall Hydroxycinnamic Acid Content and Saccharification " Plant Physiology **161**(4): 1615-1633.
Bayer, A., X. Ma and J. Stöckigt (2004). "Acetyltransfer in natural product biosynthesis—functional cloning and molecular analysis of vinorine synthase." Bioorganic & Medicinal Chemistry **12**(10): 2787-2795.
Beekwilder, J., M. Alvarez-Huerta, E. Neef, F. W. A. Verstappen, H. J. Bouwmeester and A. Aharoni (2004). "Functional Characterization of Enzymes Forming Volatile Esters from Strawberry and Banana." Plant Physiology **135**(4): 1865-1878.
Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2012). "GenBank." Nucleic Acids Research **41**(D1): D36-D42.
Beyraghdar Kashkooli, A., A. R. Van Der Krol, P. Rabe, J. S. Dickschat and H. Bouwmeester (2019). "Substrate promiscuity of enzymes from the sesquiterpene biosynthetic pathways from Artemisia annua and Tanacetum parthenium allows for novel combinatorial sesquiterpene production." Metabolic Engineering **54**: 12-23.
Bezerra, K. G. O., R. D. Rufino, J. M. Luna and L. A. Sarubbo (2018). "Saponins and microbial biosurfactants: Potential raw materials for the formulation of cosmetics." Biotechnology Progress **34**(6): 1482-1493.
Bhambhani, S., D. Lakhwani, P. Gupta, A. Pandey, Y. V. Dhar, S. Kumar Bag, M. H. Asif and P. Kumar Trivedi (2017). "Transcriptome and metabolite analyses in Azadirachta indica: identification of genes involved in biosynthesis of bioactive triterpenoids." Scientific Reports **7**(1): 5043.
Boatright, J., F. Negre, X. Chen, C. M. Kish, B. Wood, G. Peel, I. Orlova, D. Gang, D. Rhodes and N. Dudareva (2004). "Understanding in Vivo Benzenoid Metabolism in Petunia Petal Tissue." Plant Physiology **135**(4): 1993-2011.
Bontpart, T., V. Cheynier, A. Ageorges and N. Terrier (2015). "BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds." New Phytologist **208**(3): 695-707.
Böttger, S. and M. F. Melzig (2013). "The influence of saponins on cell membrane cholesterol." Bioorganic & Medicinal Chemistry **21**(22): 7118-7124.
Boycheva, S., L. Daviet, J.-L. Wolfender and T. B. Fitzpatrick (2014). "The rise of operon-like gene clusters in plants." Trends in Plant Science **19**(7): 447-459.
Bradshaw, R. H. W. (2004). "Past anthropogenic influence on European forests and some possible genetic consequences." Forest Ecology and Management **197**(1): 203-212.
Burhenne, K., B. K. Kristensen and S. K. Rasmussen (2003). "A New Class of N-Hydroxycinnamoyltransferases." Journal of Biological Chemistry **278**(16): 13919-13927.
Burton, R. A., S. M. Wilson, M. Hrmova, A. J. Harvey, N. J. Shirley, A. Medhurst, B. A. Stone, E. J. Newbigin, A. Bacic and G. B. Fincher (2006). "Cellulose Synthase-Like CslF Genes Mediate the Synthesis of Cell Wall (1,3;1,4)-ß-d-Glucans." Science **311**(5769): 1940-1942.
Calgaro-Kozina, A., K. M. Vuu, J. D. Keasling, D. Loqué, E. S. Sattely and P. M. Shih (2020). "Engineering Plant Synthetic Pathways for the Biosynthesis of Novel Antifungals." ACS Central Science **6**(8): 1394-1400.
Ćalić-Dragosavac, D., S. Zdravković-Korać, K. Šavikin-Fodulović, L. Radojević and B. Vinterhalter (2010). "Determination of escin content in androgenic embryos and hairy root culture ofAesculus hippocastanum." Pharmaceutical Biology **48**(5): 563-567.
Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden (2009). "BLAST+: architecture and applications." BMC Bioinformatics **10**(1): 421.
Cannon, J. G., R. A. Burton, S. G. Wood and N. L. Owen (2004). "Naturally Occurring Fish Poisons from Plants." Journal of Chemical Education **81**(10): 1457.
Cárdenas, P. D., A. Almeida and S. Bak (2019). "Evolution of Structural Diversity of Triterpenoids." Frontiers in Plant Science **10**: 1523-1523.

Carelli, M., E. Biazzi, F. Panara, A. Tava, L. Scaramelli, A. Porceddu, N. Graham, M. Odoardi, E. Piano, S. Arcioni, S. May, C. Scotti and O. Calderini (2011). "Medicago truncatula CYP716A12 Is a Multifunctional Oxidase Involved in the Biosynthesis of Hemolytic Saponins." The Plant Cell **23**(8): 3070-3081.
Carpenter, E. J., N. Matasci, S. Ayyampalayam, S. Wu, J. Sun, J. Yu, R. Fabio, C. Bowler, R. G. Dorrell, M. A. Gitzendanner, L. Li, W. Du, Kristian, N. J. Wickett, T. J. Barkmann, M. S. Barker, J. H. Leebens-Mack and G. K.-S. Wong (2019). "Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP)." GigaScience **8**(10).
Chan, P. K. (2007). "Acylation with diangeloyl groups at C21–22 positions in triterpenoid saponins is essential for cytotoxcity towards tumor cells." Biochemical Pharmacology **73**(3): 341-350.
Chedgy, R. J., T. G. Köllner and C. P. Constabel (2015). "Functional characterization of two acyltransferases from Populus trichocarpa capable of synthesizing benzyl benzoate and salicyl benzoate, potential intermediates in salicinoid phenolic glycoside biosynthesis." Phytochemistry **113**: 149-159.
Chen, Y.-F., C.-H. Yang, M.-S. Chang, Y.-P. Ciou and Y.-C. Huang (2010). "Foam Properties and Detergent Abilities of the Saponins from Camellia oleifera." International Journal of Molecular Sciences **11**(11): 4417-4425.
Cheng, A.-X., J.-Y. Gou, X.-H. Yu, H. Yang, X. Fang, X.-Y. Chen and C.-J. Liu (2013). "Characterization and Ectopic Expression of a Populus Hydroxyacid Hydroxycinnamoyltransferase." Molecular Plant **6**(6): 1889-1903.
Cheong, D. H. J., F. Arfuso, G. Sethi, L. Wang, K. M. Hui, A. P. Kumar and T. Tran (2018). "Molecular targets and anti-cancer potential of escin." Cancer Letters **422**: 1-8.
Christianson, D. W. (2006). "Structural Biology and Chemistry of the Terpenoid Cyclases." Chemical Reviews **106**(8): 3412-3442.
Chung, S. Y., H. Seki, Y. Fujisawa, Y. Shimoda, S. Hiraga, Y. Nomura, K. Saito, M. Ishimoto and T. Muranaka (2020). "A cellulose synthase-derived enzyme catalyses 3-O-glucuronosylation in saponin biosynthesis." Nature Communications **11**(1).
Ciarkowska, A., M. Ostrowski and A. Jakubowska (2018). "A serine carboxypeptidase-like acyltransferase catalyzes synthesis of indole-3-acetic (IAA) ester conjugate in rice (Oryza sativa)." Plant Physiology and Biochemistry **125**: 126-135.
Colson, E., C. Decroo, D. Cooper-Shepherd, G. Caulier, C. Henoumont, S. Laurent, J. De Winter, P. Flammang, M. Palmer, J. Claereboudt and P. Gerbaux (2019). "Discrimination of Regioisomeric and Stereoisomeric Saponins from Aesculus hippocastanum Seeds by Ion Mobility Mass Spectrometry." Journal of The American Society for Mass Spectrometry **30**(11): 2228-2237.
Corey, E. J. and J. Lee (1993). "Enantioselective total synthesis of oleanolic acid, erythrodiol, .beta.-amyrin, and other pentacyclic triterpenes from a common intermediate." Journal of the American Chemical Society **115**(19): 8873-8874.
Cravotto, G., L. Beltramo, S. Sapino, A. Binello and M. E. Carlotti (2011). "A new cyclodextrin-grafted viscose loaded with aescin formulations for a cosmeto-textile approach to chronic venous insufficiency." Journal of Materials Science: Materials in Medicine **22**(10): 2387-2395.
D'Auria, J. C. (2006). "Acyltransferases in plants: a good time to be BAHD." Current Opinion in Plant Biology **9**(3): 331-340.
D'Auria, J. C., F. Chen and E. Pichersky (2002). "Characterization of an Acyltransferase Capable of Synthesizing Benzylbenzoate and Other Volatile Esters in Flowers and Damaged Leaves of Clarkia breweri " Plant Physiology **130**(1): 466-476.
De Bruyn, F., J. Maertens, J. Beauprez, W. Soetaert and M. De Mey (2015). "Biotechnological advances in UDP-sugar based glycosylation of small molecules." Biotechnology Advances **33**(2): 288-302.

De Geyter, E., L. Swevers, T. Soin, D. Geelen and G. Smagghe (2012). "Saponins do not affect the ecdysteroid receptor complex but cause membrane permeation in insect culture cell lines." Journal of Insect Physiology **58**(1): 18-23.
Deli, J., Z. Matus and G. Tóth (2000). "Comparative study on the carotenoid composition in the buds and flowers of differentAesculus species." Chromatographia **51**(1): S179-S182.
Delporte, C., M. Rodríguez-Díaz and B. K. Cassels (2021). Quillaja saponaria Molina, Springer International Publishing**:** 461-473.
Delporte, M., G. Bernard, G. Legrand, B. Hielscher, A. Lanoue, R. Molinié, C. Rambaud, D. Mathiron, S. Besseau, N. Linka, J.-L. Hilbert and D. Gagneul (2018). "A BAHD neofunctionalization promotes tetrahydroxycinnamoyl spermine accumulation in the pollen coats of the Asteraceae family." Journal of Experimental Botany.
Dhaubhadel, S., M. Farhangkhoee and R. Chapman (2008). "Identification and characterization of isoflavonoid specific glycosyltransferase and malonyltransferase from soybean seeds." Journal of Experimental Botany **59**(4): 981-994.
Dinda, B., S. Debnath, B. C. Mohanta and Y. Harigaya (2010). "Naturally Occurring Triterpenoid Saponins." Chemistry & Biodiversity **7**(10): 2327-2580.
Dizes, C., F. Gerald, C. Lavaud, R. Elias, R. Faure, G. Massiot and G. Balansard (1998). "Harpuloside a triterpenoid saponin from Harpullia ramiflora." Phytochemistry **48**(7): 1229-1232.
Do, R., R. Kiss, D. Gaudet and J. Engert (2009). "Squalene synthase: a critical enzyme in the cholesterol biosynthesis pathway." Clinical Genetics **75**(1): 19-29.
Domanski, D., O. Zegrocka-Stendel, A. Perzanowska, M. Dutkiewicz, M. Kowalewska, I. Grabowska, D. Maciejko, A. Fogtman, M. Dadlez and K. Koziak (2016). "Molecular Mechanism for Cellular Response to β-Escin and Its Therapeutic Implications." PLOS ONE **11**(10): e0164365-e0164365.
Dudareva, N., J. C. D'Auria, K. H. Nam, R. A. Raguso and E. Pichersky (1998). "Acetyl-CoA:benzylalcohol acetyltransferase – an enzyme involved in floral scent production in Clarkia breweri." The Plant Journal **14**(3): 297-304.
Dudek-Makuch, M. and I. Matławska (2011). "Flavonoids from the flowers of Aesculus hippocastanum." Acta Pol Pharm **68**(3): 403-408.
Dudek-Makuch, M. and E. Studzińska-Sroka (2015). "Horse chestnut–efficacy and safety in chronic venous insufficiency: an overview." Revista Brasileira de Farmacognosia **25**: 533-541.
Duhamel du Monceau, H.-L. (1819). Traite des arbres et arbustes que l'on cultive en France en pleine terre. Paris.
Edem, E., B. Kahyaoğlu and M. A. Çakar (2016). "Acute Effusive Pericarditis due to Horse Chestnut Consumption." American Journal of Case Reports **17**: 305-308.
Eljounaidi, K. and B. R. Lichman (2020). "Nature's Chemists: The Discovery and Engineering of Phytochemical Biosynthesis." Frontiers in Chemistry **8**(1041).
Englert, J., B. Weniger, A. Lobstein, R. Anton, E. Krempp, D. Guillaume and Y. Leroy (1995). "Triterpenoid Saponins from Mimosa pigra." Journal of Natural Products **58**(8): 1265-1269.
Escamilla-Treviño, L. L., H. Shen, T. Hernandez, Y. Yin, Y. Xu and R. A. Dixon (2014). "Early lignin pathway enzymes and routes to chlorogenic acid in switchgrass (Panicum virgatum L.)." Plant Molecular Biology **84**(4): 565-576.
Esposito, D., L. A. Garvey and C. S. Chakiath (2009). Gateway Cloning for Protein Expression. High Throughput Protein Expression and Purification: Methods and Protocols. S. A. Doyle. Totowa, NJ, Humana Press**:** 31-54.
Eudes, A., M. Mouille, D. S. Robinson, V. T. Benites, G. Wang, L. Roux, Y.-L. Tsai, E. E. K. Baidoo, T.-Y. Chiu, J. L. Heazlewood, H. V. Scheller, A. Mukhopadhyay, J. D. Keasling, S. Deutsch and D. Loqué (2016). "Exploiting members of the BAHD acyltransferase family to

synthesize multiple hydroxycinnamate and benzoate conjugates in yeast." Microbial Cell Factories **15**(1).
Evelyn, J. (1776). Sylva, or a discourse of forest trees, with notes by A. Hunter (first published 1664). London.
Fang, Y., L. Zhao, F. Yan, X. Xia, D. Xu and X. Cui (2010). "Escin improves sperm quality in male patients with varicocele-associated infertility." Phytomedicine **17**(3): 192-196.
Ferracini, C., P. Curir, M. Dolci, V. Lanzotti and A. Alma (2010). "Aesculus pavia foliar saponins: Defensive role against the leafminer Cameraria ohridella." Pest Management Science **66**(7): 767-772.
Fink, R. and S. Filip (2022). "Surface-active natural saponins. Properties, safety, and efficacy." International Journal of Environmental Health Research: 1-10.
Fink, R., A. Potočnik and M. Oder (2020). "Plant-based natural saponins for Escherichia coli surface hygiene management." LWT **122**: 109018.
Fraser, C. M., M. G. Thompson, A. M. Shirley, J. Ralph, J. A. Schoenherr, T. Sinlapadech, M. C. Hall and C. Chapple (2007). "Related Arabidopsis Serine Carboxypeptidase-Like Sinapoylglucose Acyltransferases Display Distinct But Overlapping Substrate Specificities." Plant Physiology **144**(4): 1986-1999.
Freeman, E. G., Z. Wisotsky and A. Dahanukar (2014). "Detection of sweet tastants by a conserved group of insect gustatory receptors." Proceedings of the National Academy of Sciences **111**(4): 1598-1603.
Friedman, M. and G. M. McDonald (1999). Steroidal Glycoalkaloids. Naturally Occurring Glycosides. R. Ikan. Chichester, John Wiley & Sons.
Fujiwara, H., Y. Tanaka, K. Yonekura-Sakakibara, M. Fukuchi-Mizutani, M. Nakao, Y. Fukui, M. Yamaguchi, T. Ashikari and T. Kusumi (1998). "cDNA cloning, gene expression and subcellular localization of anthocyanin 5-aromatic acyltransferase from Gentiana triflora." The Plant Journal **16**(4): 421-431.
Fukushima, E. O., H. Seki, K. Ohyama, E. Ono, N. Umemoto, M. Mizutani, K. Saito and T. Muranaka (2011). "CYP716A Subfamily Members are Multifunctional Oxidases in Triterpenoid Biosynthesis." Plant and Cell Physiology **52**(12): 2050-2061.
Fukushima, E. O., H. Seki, S. Sawai, M. Suzuki, K. Ohyama, K. Saito and T. Muranaka (2013). "Combinatorial Biosynthesis of Legume Natural and Rare Triterpenoids in Engineered Yeast." Plant and Cell Physiology **54**(5): 740-749.
Gachon, C. M. M., M. Langlois-Meurinne and P. Saindrenan (2005). "Plant secondary metabolism glycosyltransferases: the emerging functional analysis." Trends in Plant Science **10**(11): 542-549.
Gallelli, L. (2019). "Escin: a review of its anti-edematous, anti-inflammatory, and venotonic properties." Drug Design, Development and Therapy **Volume 13**: 3425-3437.
Gallelli, L., L. Zhang, T. Wang and F. Fu (2020). "Severe Acute Lung Injury Related to COVID-19 Infection: A Review and the Possible Role for Escin." The Journal of Clinical Pharmacology **60**(7): 815-825.
Geisler, K., R. K. Hughes, F. Sainsbury, G. P. Lomonossoff, M. Rejzek, S. Fairhurst, C. E. Olsen, M. S. Motawia, R. E. Melton, A. M. Hemmings, S. Bak and A. Osbourn (2013). "Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants." Proceedings of the National Academy of Sciences of the United States of America **110**(35).
Geisler, R., C. Dargel and T. Hellweg (2019). "The Biosurfactant β-Aescin: A Review on the Physico-Chemical Properties and Its Interaction with Lipid Model Membranes and Langmuir Monolayers." Molecules 2020, Vol. 25, Page 117 **25**(1): 117-117.
Geisler, R., S. Prévost, R. Dattani and T. Hellweg (2020). "Effect of Cholesterol and Ibuprofen on DMPC-β-Aescin Bicelles: A Temperature-Dependent Wide-Angle X-ray Scattering Study." Crystals 2020, Vol. 10, Page 401 **10**(5): 401-401.

Geissman, T. A. and D. H. G. Grout (1969). Organic chemistry of secondary plant metabolism. San Francisco, Freeman, Cooper, & Co.
Ghosh, S. (2017). "Triterpene Structural Diversification by Plant Cytochrome P450 Enzymes." Frontiers in Plant Science **8**(1886).
Glensk, M., M. Wlodarczyk, C. Bassarello and C. Pizza (2011). "Aescin determination in seeds and capsules of three pure and hybrid Aesculus species." Chemistry of Natural Compounds **47**(1): 142-144.
Gou, J.-Y., X.-H. Yu and C.-J. Liu (2009). "A hydroxycinnamoyltransferase responsible for synthesizing suberin aromatics in Arabidopsis." Proceedings of the National Academy of Sciences **106**(44): 18855-18860.
Grienenberger, E., S. Besseau, P. Geoffroy, D. Debayle, D. Heintz, C. Lapierre, B. Pollet, T. Heitz and M. Legrand (2009). "A BAHD acyltransferase is expressed in the tapetum of Arabidopsis anthers and is involved in the synthesis of hydroxycinnamoyl spermidines." The Plant Journal **58**(2): 246-259.
Grothe, T., R. Lenz and T. M. Kutchan (2001). "Molecular Characterization of the Salutaridinol 7-O-Acetyltransferase Involved in Morphine Biosynthesis in Opium Poppy Papaver somniferum." Journal of Biological Chemistry **276**(33): 30717-30723.
Guo, C. F., X. C. Xiong, H. Dong and X. Q. Qi (2021). "Genome-wide investigation and transcriptional profiling of the oxidosqualene cyclase (OSC) genes in wheat (Triticum aestivum)." Journal of Systematics and Evolution.
Guo, Q., X. Ma, S. Wei, D. Qiu, I. W. Wilson, P. Wu, Q. Tang, L. Liu, S. Dong and W. Zu (2014). "De novo transcriptome sequencing and digital gene expression analysis predict biosynthetic pathway of rhynchophylline and isorhynchophylline from Uncaria rhynchophylla, a non-model plant with potent anti-alzheimer's properties." BMC Genomics **15**(1): 676.
Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman and A. Regev (2013). "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." Nature Protocols **8**(8): 1494-1512.
Hamberger, B. and S. Bak (2013). "Plant P450s as versatile drivers for evolution of species-specific chemical diversity." Philosophical Transactions of the Royal Society B: Biological Sciences **368**(1612): 20120426.
Haralampidis, K., G. Bryan, X. Qi, K. Papadopoulou, S. Bakht, R. Melton and A. Osbourn (2001). "A new class of oxidosqualene cyclases directs synthesis of antimicrobial phytoprotectants in monocots." Proceedings of the National Academy of Sciences **98**(23): 13431-13436.
Harris, A., Q.-Y. Xiang and D. T. Thomas (2009). "Phylogeny, origin, and biogeographic history ofAesculusL. (Sapindales) - an update from combined analysis of DNA sequences, morphology, and fossils." TAXON **58**(1): 108-126.
Hasegawa, S. (1999). Limonoid Glycosides. Naturally Occurring Glycosides. R. Ikan. Chichester, John Wiley & Sons.
Hazen, S. P., J. S. Scott-Craig and J. D. Walton (2002). "Cellulose Synthase-Like Genes of Rice." Plant Physiology **128**(2): 336-340.
Hodgson, H., R. De La Peña, M. J. Stephenson, R. Thimmappa, J. L. Vincent, E. S. Sattely and A. Osbourn (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): 17096-17104.
Hoffmann, L., S. Maury, F. Martz, P. Geoffroy and M. Legrand (2003). "Purification, Cloning, and Properties of an Acyltransferase Controlling Shikimate and Quinate Ester Intermediates in Phenylpropanoid Metabolism." Journal of Biological Chemistry **278**(1): 95-103.

Holstein, S. A. and R. J. Hohl (2004). "Isoprenoids: Remarkable diversity of form and function." Lipids **39**(4): 293-309.
Hong, K.-J., Y.-K. Choi, S. Tokunaga, Y. Ishigami and T. Kajiuchi (1998). "Removal of cadmium and lead from soil using aescin as a biosurfactant." Journal of Surfactants and Detergents **1**(2): 247-250.
Hostettmann, K. and A. Marston (1995). Saponins, Cambridge University Press.
Hou, J., X. Ye, C. Li and Y. Wang (2021). "K-Module Algorithm: An Additional Step to Improve the Clustering Results of WGCNA Co-Expression Networks." Genes **12**(1): 87.
Howard, A. L. (1945). "The Horse-Chestnut Tree (Aesculus Hippocastanum)." Nature **155**(3939): 521-522.
Huang, A. C., T. Jiang, Y.-X. Liu, Y.-C. Bai, J. Reed, B. Qu, A. Goossens, H.-W. Nützmann, Y. Bai and A. Osbourn (2019). "A specialized metabolic network selectively modulates Arabidopsis root microbiota." Science **364**(6440): eaau6389.
Huang, A. C. and A. Osbourn (2019). "Plant terpenes that mediate below-ground interactions: prospects for bioengineering terpenoids for plant protection." Pest Management Science.
Huang, H.-C., M.-D. Wu, W.-J. Tsai, S.-C. Liao, C.-C. Liaw, L.-C. Hsu, Y.-C. Wu and Y.-H. Kuo (2008). "Triterpenoid saponins from the fruits and galls of Sapindus mukorossi." Phytochemistry **69**(7): 1609-1616.
Huang, L., J. Li, H. Ye, C. Li, H. Wang, B. Liu and Y. Zhang (2012). "Molecular characterization of the pentacyclic triterpenoid biosynthetic pathway in Catharanthus roseus." Planta **236**(5): 1571-1581.
Idris, S., A. Mishra and M. Khushtar (2020). "Phytochemical, ethanomedicinal and pharmacological applications of escin from Aesculus hippocastanum L. towards future medicine." Journal of Basic and Clinical Physiology and Pharmacology **31**(5).
Ikegami, A., S. Eguchi, A. Kitajima, K. Inoue and K. Yonemori (2007). "Identification of genes involved in proanthocyanidin biosynthesis of persimmon (Diospyros kaki) fruit." Plant Science **172**(5): 1037-1047.
Ikram, N. K. B. K., X. Zhan, X.-W. Pan, B. C. King and H. T. Simonsen (2015). "Stable heterologous expression of biologically active terpenoids in green plant cells." Frontiers in Plant Science **6**(129).
Irmisch, S., S. Jo, C. R. Roach, S. Jancsik, M. Man Saint Yuen, L. L. Madilao, M. O'Neil-Johnson, R. Williams, S. G. Withers and J. Bohlmann (2018). "Discovery of UDP-Glycosyltransferases and BAHD-Acyltransferases Involved in the Biosynthesis of the Antidiabetic Plant Metabolite Montbretin A." The Plant Cell **30**(8): 1864-1886.
Itkin, M., R. Davidovich-Rikanati, S. Cohen, V. Portnoy, A. Doron-Faigenboim, E. Oren, S. Freilich, G. Tzuri, N. Baranes, S. Shen, M. Petreikov, R. Sertchook, S. Ben-Dor, H. Gottlieb, A. Hernandez, D. R. Nelson, H. S. Paris, Y. Tadmor, Y. Burger, E. Lewinsohn, N. Katzir and A. Schaffer (2016). "The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from Siraitia grosvenorii." Proceedings of the National Academy of Sciences **113**(47): E7619-E7628.
Jæger, D., C. P. Ndi, C. Crocoll, B. S. Simpson, B. Khakimov, R. M. Guzman-Genuino, J. D. Hayball, X. Xing, V. Bulone, P. Weinstein, B. L. Møller and S. J. Semple (2017). "Isolation and Structural Characterization of Echinocystic Acid Triterpenoid Saponins from the Australian Medicinal and Food PlantAcacia ligulata." Journal of Natural Products **80**(10): 2692-2698.
Jain, M., P. L. Srivastava, M. Verma, R. Ghangal and R. Garg (2016). "De novo transcriptome assembly and comprehensive expression profiling in Crocus sativus to gain insights into apocarotenoid biosynthesis." Scientific Reports **6**(1): 22456.
Jatczak, K. and G. Grynkiewicz (2014). "Triterpene sapogenins with oleanene skeleton: chemotypes and biological activities." Acta Biochimica Polonica **61**(2).

Jiang, N., W. Xin, T. Wang, L. Zhang, H. Fan, Y. Du, C. Li and F. Fu (2011). "Protective effect of aescin from the seeds of Aesculus hippocastanum on liver injury induced by endotoxin in mice." Phytomedicine **18**(14): 1276-1284.
Jones, D. T. and D. Woods (1986). "Acetone-butanol fermentation revisited." Microbiological reviews **50**(4): 484-524.
Jozwiak, A., P. D. Sonawane, S. Panda, C. Garagounis, K. K. Papadopoulou, B. Abebie, H. Massalha, E. Almekias-Siegl, T. Scherf and A. Aharoni (2020). "Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery." Nature Chemical Biology 2020 16:7 **16**(7): 740-748.
Jurado Gonzalez, P. and P. M. Sörensen (2020). "Characterization of saponin foam from Saponaria officinalis for food applications." Food Hydrocolloids **101**: 105541.
Kahn, R. A., T. Fahrendorf, B. A. Halkier and B. L. Møller (1999). "Substrate Specificity of the Cytochrome P450 Enzymes CYP79A1 and CYP71E1 Involved in the Biosynthesis of the Cyanogenic Glucoside Dhurrin inSorghum bicolor(L.) Moench." Archives of Biochemistry and Biophysics **363**(1): 9-18.
Kasai, R., K. Yamasaki and O. Tanaka (1999). Saponins. Naturally Occurring Glycosides. R. Ikan. Chichester, John Wiley & Sons.
Kautsar, S. A., G. Hernando, K. Blin, A. Osbourn and M. H. Medema (2017). "plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters." Nucleic Acids Research **45**(W1): W55-W63.
Khakimov, B., V. Kuzina, P. Ø. Erthmann, E. O. Fukushima, J. M. Augustin, C. E. Olsen, J. Scholtalbers, H. Volpin, S. B. Andersen, T. P. Hauser, T. Muranaka and S. Bak (2015). "Identification and genome organization of saponin pathway genes from a wild crucifer, and their use for transient production of saponins inNicotiana benthamiana." The Plant Journal **84**(3): 478-490.
Kilari, E. K. and S. Putta (2016). "Biological and Phytopharmacological Descriptions of Litchi Chinensis." Pharmacognosy reviews **10**(19): 60-65.
Kim, B.-G., I.-A. Kim and J.-H. Ahn (2011). "Characterization of hydroxycinnamoyl-coenzyme a Shikimate hydroxycinnamoyltransferase from Populus euramericana." Journal of the Korean Society for Applied Biological Chemistry **54**(5): 817-821.
Kim, D., M. Jung, I. J. Ha, M. Y. Lee, S.-G. Lee, Y. Shin, S. Subramaniyam and J. Oh (2018). "Transcriptional Profiles of Secondary Metabolite Biosynthesis Genes and Cytochromes in the Leaves of Four Papaver Species." Data **3**(4): 55.
Kim, D., L. Song, F. P. Breitwieser and S. L. Salzberg (2016). "Centrifuge: rapid and sensitive classification of metagenomic sequences." Genome Research **26**(12): 1721-1729.
Kim, I. A., B.-G. Kim, M. Kim and J.-H. Ahn (2012). "Characterization of hydroxycinnamoyltransferase from rice and its application for biological synthesis of hydroxycinnamoyl glycerols." Phytochemistry **76**: 25-31.
Kim, S., P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant (2015). "PubChem Substance and Compound databases." Nucleic Acids Research **44**(D1): D1202-D1213.
Kinjo, J., I. Miyamoto, K. Murakami, K. Kida, T. Tomimatsu, M. Yamasaki and T. Nohara (1985). "Oleanene-sapogenols from puerariae radix." Chemical and Pharmaceutical Bulletin **33**(3): 1293-1296.
Kochevenko, A. and A. R. Fernie (2011). "The genetic architecture of branched-chain amino acid accumulation in tomato fruits." Journal of Experimental Botany **62**(11): 3895-3906.
Kohonen, T. (2013). "Essentials of the self-organizing map." Neural Networks **37**: 52-65.
Konoshima, T. and K.-H. Lee (1986). "Antitumor Agents, 82. Cytotoxic Sapogenols from Aesculus hippocastanum." Journal of Natural Products **49**(4): 650-656.

Kosma, D. K., I. Molina, J. B. Ohlrogge and M. Pollard (2012). "Identification of an Arabidopsis Fatty Alcohol:Caffeoyl-Coenzyme A Acyltransferase Required for the Synthesis of Alkyl Hydroxycinnamates in Root Waxes1 " Plant Physiology **160**(1): 237-248.
Koul, O. (2008). "Phytochemicals and Insect Control: An Antifeedant Approach." Critical Reviews in Plant Sciences **27**(1): 1-24.
Kruse, L. H., A. T. Weigle, J. Martínez-Gómez, J. D. Chobirko, J. E. Schaffer, A. A. Bennett, C. D. Specht, J. M. Jez, D. Shukla and G. D. Moghe (2020). "Ancestral class-promiscuity as a driver of functional diversity in the BAHD acyltransferase family in plants." bioRxiv: 2020.2011.2018.385815-382020.385811.385818.385815.
Kubo, A., Y. Arai, S. Nagashima and T. Yoshikawa (2004). "Alteration of sugar donor specificities of plant glycosyltransferases by a single point mutation." Archives of Biochemistry and Biophysics **429**(2): 198-203.
Kunatsa, Y. and D. R. Katerere (2021). "Checklist of African Soapy Saponin—Rich Plants for Possible Use in Communities' Response to Global Pandemics." Plants **10**(5): 842.
Kushiro, T., M. Shibuya, K. Masuda and Y. Ebizuka (2000). "Mutational Studies on Triterpene Synthases: Engineering Lupeol Synthase into β-Amyrin Synthase." Journal of the American Chemical Society **122**(29): 6816-6824.
Landmann, C., S. Hücherig, B. Fink, T. Hoffmann, D. Dittlein, H. A. Coiner and W. Schwab (2011). "Substrate promiscuity of a rosmarinic acid synthase from lavender (Lavandula angustifolia L.)." Planta **234**(2): 305-320.
Lanzotti, V., P. Termolino, M. Dolci and P. Curir (2012). "Paviosides A–H, eight new oleane type saponins from Aesculus pavia with cytotoxic activity." Bioorganic & Medicinal Chemistry **20**(10): 3280-3286.
Lee, S., Y. Kaminaga, B. Cooper, E. Pichersky, N. Dudareva and C. Chapple (2012). "Benzoylation and sinapoylation of glucosinolate R-groups in Arabidopsis." The Plant Journal **72**(3): 411-422.
Leebens-Mack, J. H., M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, S. W. Graham, I. Grosse, Z. Li, M. Melkonian, S. Mirarab, M. Porsch, M. Quint, S. A. Rensing, D. E. Soltis, P. S. Soltis, D. W. Stevenson, K. K. Ullrich, N. J. Wickett, L. DeGironimo, P. P. Edger, I. E. Jordon-Thaden, S. Joya, T. Liu, B. Melkonian, N. W. Miles, L. Pokorny, C. Quigley, P. Thomas, J. C. Villarreal, M. M. Augustin, M. D. Barrett, R. S. Baucom, D. J. Beerling, R. M. Benstein, E. Biffin, S. F. Brockington, D. O. Burge, J. N. Burris, K. P. Burris, V. Burtet-Sarramegna, A. L. Caicedo, S. B. Cannon, Z. Çebi, Y. Chang, C. Chater, J. M. Cheeseman, T. Chen, N. D. Clarke, H. Clayton, S. Covshoff, B. J. Crandall-Stotler, H. Cross, C. W. dePamphilis, J. P. Der, R. Determann, R. C. Dickson, V. S. Di Stilio, S. Ellis, E. Fast, N. Feja, K. J. Field, D. A. Filatov, P. M. Finnegan, S. K. Floyd, B. Fogliani, N. García, G. Gâteblé, G. T. Godden, F. Goh, S. Greiner, A. Harkess, J. M. Heaney, K. E. Helliwell, K. Heyduk, J. M. Hibberd, R. G. J. Hodel, P. M. Hollingsworth, M. T. J. Johnson, R. Jost, B. Joyce, M. V. Kapralov, E. Kazamia, E. A. Kellogg, M. A. Koch, M. Von Konrat, K. Könyves, T. M. Kutchan, V. Lam, A. Larsson, A. R. Leitch, R. Lentz, F.-W. Li, A. J. Lowe, M. Ludwig, P. S. Manos, E. Mavrodiev, M. K. McCormick, M. McKain, T. McLellan, J. R. McNeal, R. E. Miller, M. N. Nelson, Y. Peng, P. Ralph, D. Real, C. W. Riggins, M. Ruhsam, R. F. Sage, A. K. Sakai, M. Scascitella, E. E. Schilling, E.-M. Schlösser, H. Sederoff, S. Servick, E. B. Sessa, A. J. Shaw, S. W. Shaw, E. M. Sigel, C. Skema, A. G. Smith, A. Smithson, C. N. Stewart, J. R. Stinchcombe, P. Szövényi, J. A. Tate, H. Tiebel, D. Trapnell, M. Villegente, C.-N. Wang, S. G. Weller, M. Wenzel, S. Weststrand, J. H. Westwood, D. F. Whigham, S. Wu, A. S. Wulff, Y. Yang, D. Zhu, C. Zhuang, J. Zuidof, M. W. Chase, J. C. Pires, C. J. Rothfels, J. Yu, C. Chen, L. Chen, S. Cheng, J. Li, R. Li, X. Li, H. Lu, Y. Ou, X. Sun, X. Tan, J. Tang, Z. Tian, F. Wang, J. Wang, X. Wei, X. Xu, Z. Yan, F. Yang, X. Zhong, F. Zhou, Y. Zhu, Y. Zhang, S. Ayyampalayam, T. J. Barkman, N.-p. Nguyen, N. Matasci, D. R. Nelson, E. Sayyari, E. K. Wafula, R. L. Walls, T. Warnow, H. An, N. Arrigo, A. E. Baniaga, S. Galuska, S. A. Jorgensen, T. I. Kidder, H. Kong, P. Lu-Irving, H. E.

Marx, X. Qi, C. R. Reardon, B. L. Sutherland, G. P. Tiley, S. R. Welles, R. Yu, S. Zhan, L. Gramzow, G. Theißen, G. K.-S. Wong and I. One Thousand Plant Transcriptomes (2019). "One thousand plant transcriptomes and the phylogenomics of green plants." Nature **574**(7780): 679-685.
Lepelley, M., G. Cheminade, N. Tremillon, A. Simkin, V. Caillet and J. McCarthy (2007). "Chlorogenic acid synthesis in coffee: An analysis of CGA content and real-time RT-PCR expression of HCT, HQT, C3H1, and CCoAOMT1 genes during grain development in C. canephora." Plant Science **172**(5): 978-996.
Leveau, A., J. Reed, X. Qiao, M. J. Stephenson, S. T. Mugford, R. E. Melton, J. C. Rant, R. Vickerstaff, T. Langdon and A. Osbourn (2019). "Towards take-all control: a C-21β oxidase required for acylation of triterpene defence compounds in oat." New Phytologist **221**(3): 1544-1555.
Li, A. X. and J. C. Steffens (2000). "An acyltransferase catalyzing the formation of diacylglucose is a serine carboxypeptidase-like protein." Proceedings of the National Academy of Sciences **97**(12): 6902-6907.
Li, Y., A. Leveau, Q. Zhao, Q. Feng, H. Lu, J. Miao, Z. Xue, A. C. Martin, E. Wegel, J. Wang, A. Orme, M.-D. Rey, M. Karafiátová, J. Vrána, B. Steuernagel, R. Joynson, C. Owen, J. Reed, T. Louveau, M. J. Stephenson, L. Zhang, X. Huang, T. Huang, D. Fan, C. Zhou, Q. Tian, W. Li, Y. Lu, J. Chen, Y. Zhao, Y. Lu, C. Zhu, Z. Liu, G. Polturak, R. Casson, L. Hill, G. Moore, R. Melton, N. Hall, B. B. H. Wulff, J. Doležel, T. Langdon, B. Han and A. Osbourn (2021). "Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals." Nature Communications **12**(1).
Liao, Y., Z. Li, Q. Zhou, M. Sheng, Q. Qu, Y. Shi, J. Yang, L. Lv, X. Dai and X. Shi (2021). "Saponin surfactants used in drug delivery systems: A new application for natural medicine components." International Journal of Pharmaceutics **603**: 120709.
Lichota, A., L. Gwozdzinski and K. Gwozdzinski (2019). "Therapeutic potential of natural compounds in inflammation and chronic venous insufficiency." European Journal of Medicinal Chemistry **176**: 68-91.
Liu, M.-H., B.-R. Yang, W.-F. Cheung, K. Y. Yang, H.-F. Zhou, J. S.-L. Kwok, G.-C. Liu, X.-F. Li, S. Zhong, S. M.-Y. Lee and S. K.-W. Tsui (2015). "Transcriptome analysis of leaves, roots and flowers of Panax notoginseng identifies genes involved in ginsenoside and alkaloid biosynthesis." BMC Genomics **16**(1).
Liu, Y., L. Gao, L. Liu, Q. Yang, Z. Lu, Z. Nie, Y. Wang and T. Xia (2012). "Purification and Characterization of a Novel Galloyltransferase Involved in Catechin Galloylation in the Tea Plant (Camellia sinensis)." Journal of Biological Chemistry **287**(53): 44406-44417.
Liu, Z., J. Cheema, M. Vigouroux, L. Hill, J. Reed, P. Paajanen, L. Yant and A. Osbourn (2020). "Formation and diversification of a paradigm biosynthetic gene cluster in plants." Nature Communications **11**(1).
Liu, Z., H. G. Suarez Duran, Y. Harnvanichvech, M. J. Stephenson, M. E. Schranz, D. Nelson, M. H. Medema and A. Osbourn (2020). "Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae." New Phytologist **227**(4): 1109-1123.
Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology **15**(12).
Luo, J., C. Fuell, A. Parr, L. Hill, P. Bailey, K. Elliott, S. A. Fairhurst, C. Martin and A. J. Michael (2009). "A Novel Polyamine Acyltransferase Responsible for the Accumulation of Spermidine Conjugates in Arabidopsis Seed." The Plant Cell **21**(1): 318-333.
Maeda, H. A. (2019). "Evolutionary Diversification of Primary Metabolism and Its Contribution to Plant Chemical Diversity." Frontiers in Plant Science **10**.
Mahato, S. B. and A. P. Kundu (1994). "13C NMR Spectra of pentacyclic triterpenoids—a compilation and some salient features." Phytochemistry **37**(6): 1517-1575.

Majinda, R. R. T. (2012). Extraction and Isolation of Saponins, Humana Press**:** 415-426.
Mann, J., R. S. Davidson, J. B. Hobbs, D. V. Banthorpe and J. B. Harborne (1994). Natural products: their chemistry and biological significance. Harlow, Essex, Longman.
Matasci, N., L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, S. Ayyampalayam, M. Barker, J. G. Burleigh, M. A. Gitzendanner, E. Wafula, J. P. Der, C. W. Depamphilis, B. Roure, H. Philippe, B. R. Ruhfel, N. W. Miles, S. W. Graham, S. Mathews, B. Surek, M. Melkonian, D. E. Soltis, P. S. Soltis, C. Rothfels, L. Pokorny, J. A. Shaw, L. Degironimo, D. W. Stevenson, J. C. Villarreal, T. Chen, T. M. Kutchan, M. Rolf, R. S. Baucom, M. K. Deyholos, R. Samudrala, Z. Tian, X. Wu, X. Sun, Y. Zhang, J. Wang, J. Leebens-Mack and G. K.-S. Wong (2014). "Data access for the 1,000 Plants (1KP) project." GigaScience **3**(1): 17.
Matsuda, H., Y. Li, T. Murakami, K. Ninomiya, J. Yamahara and M. Yoshikawa (1997). "Effects of Escins Ia, Ib, IIa, and IIb from Horse Chestnut, the Seeds of Aesculus hippocastanum L., on Acute Inflammation in Animals." Biological and Pharmaceutical Bulletin **20**(10): 1092-1095.
Matsuura, H. N., S. Malik, F. De Costa, M. Yousefzadi, M. H. Mirjalili, R. Arroo, A. S. Bhambra, M. Strnad, M. Bonfill and A. G. Fett-Neto (2018). "Specialized Plant Metabolism Characteristics and Impact on Target Molecule Biotechnological Production." Molecular Biotechnology **60**(2): 169-183.
Menon, A., N. B. Krdzavac and M. Kraft (2019). "From database to knowledge graph — using data in chemistry." Current Opinion in Chemical Engineering **26**: 33-37.
Miettinen, K., S. Iñigo, L. Kreft, J. Pollier, Christof, A. Botzki, F. Coppens, S. Bak and A. Goossens (2018). "The TriForC database: a comprehensive up-to-date resource of plant triterpene biosynthesis." Nucleic Acids Research **46**(D1): D586-D594.
Miettinen, K., J. Pollier, D. Buyst, P. Arendt, R. Csuk, S. Sommerwerk, T. Moses, J. Mertens, P. D. Sonawane, L. Pauwels, A. Aharoni, J. Martins, D. R. Nelson and A. Goossens (2017). "The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis." Nature Communications **8**(1): 14153.
Milkowski, C., A. Baumert, D. Schmidt, L. Nehlin and D. Strack (2004). "Molecular regulation of sinapate ester metabolism in Brassica napus: expression of genes, properties of the encoded proteins and correlation of enzyme activities with metabolite accumulation." The Plant Journal **38**(1): 80-92.
Milkowski, C. and D. Strack (2004). "Serine carboxypeptidase-like acyltransferases." Phytochemistry **65**(5): 517-524.
Miziorko, H. M. (2011). "Enzymes of the mevalonate pathway of isoprenoid biosynthesis." Archives of Biochemistry and Biophysics **505**(2): 131-143.
Moglia, A., S. Lanteri, C. Comino, L. Hill, D. Knevitt, C. Cagliero, P. Rubiolo, S. Bornemann and C. Martin (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): 1777-1787.
Molina, I. and D. Kosma (2015). "Role of HXXXD-motif/BAHD acyltransferases in the biosynthesis of extracellular lipids." Plant Cell Reports **34**(4): 587-601.
Morant, M., S. Bak, B. L. Møller and D. Werck-Reichhart (2003). "Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation." Current Opinion in Biotechnology **14**(2): 151-162.
Morita, M., M. Shibuya, T. Kushiro, K. Masuda and Y. Ebizuka (2000). "Molecular cloning and functional expression of triterpene synthases from pea (Pisum sativum)." European Journal of Biochemistry **267**(12): 3453-3460.
Moses, T., J. Pollier, L. Almagro, D. Buyst, M. Van Montagu, M. A. Pedreño, J. C. Martins, J. M. Thevelein and A. Goossens (2014). "Combinatorial biosynthesis of sapogenins and saponins in Saccharomyces cerevisiae using a C-16α hydroxylase from Bupleurum

falcatum." Proceedings of the National Academy of Sciences of the United States of America **111**(4): 1634-1639.
Moses, T., J. Pollier, A. Faizal, S. Apers, L. Pieters, Johan, D. Geelen and A. Goossens (2015). "Unraveling the Triterpenoid Saponin Biosynthesis of the African Shrub Maesa lanceolata." Molecular Plant **8**(1): 122-135.
Mugford, S. T., T. Louveau, R. Melton, X. Qi, S. Bakht, L. Hill, T. Tsurushima, S. Honkanen, S. J. Rosser, G. P. Lomonossoff and A. Osbourn (2013). "Modularity of plant metabolic gene clusters: A trio of linked genes that are collectively required for acylation of triterpenes in oat." Plant Cell **25**(3): 1078-1092.
Mugford, S. T. and A. Osbourn (2010). "Evolution of serine carboxypeptidase-like acyltransferases in the monocots." Plant Signaling & Behavior **5**(2): 193-195.
Mugford, S. T., X. Qi, S. Bakht, L. Hill, E. Wegel, R. K. Hughes, K. Papadopoulou, R. Melton, M. Philo, F. Sainsbury, G. P. Lomonossoff, A. D. Roy, R. J. M. Goss and A. Osbourn (2009). "A serine carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats." Plant Cell **21**(8): 2473-2484.
Murakami, T., J. Nakamura, T. Kageura, H. Matsuda and M. Yoshikawa (2000). "Bioactive Saponins and Glycosides. XVII. Inhibitory Effect on Gastric Emptying and Accelerating Effect on Gastrointestinal Transit of Tea Saponins : Structures of Assamsaponins F, G, H, I, and J from the Seeds and Leaves of the Tea Plant." Chemical and Pharmaceutical Bulletin **48**(11): 1720-1725.
Murakami, T., J. Nakamura, H. Matsuda and M. Yoshikawa (1999). "Bioactive Saponins and Glycosides. XV. Saponin Constituents with Gastroprotective Effect from the Seeds of Tea Plant, Camellia sinensis L. var. assamica PIERRE, Cultivated in Sri Lanka : Structures of Assamsaponins A, B, C, D, and E." Chemical and Pharmaceutical Bulletin **47**(12): 1759-1764.
Muroi, A., A. Ishihara, C. Tanaka, A. Ishizuka, J. Takabayashi, H. Miyoshi and T. Nishioka (2009). "Accumulation of hydroxycinnamic acid amides induced by pathogen infection and identification of agmatine coumaroyltransferase in Arabidopsis thaliana." Planta **230**(3): 517.
Mutyam, V., M. Du, X. Xue, K. M. Keeling, E. L. White, J. R. Bostwick, L. Rasmussen, B. Liu, M. Mazur, J. S. Hong, E. Falk Libby, F. Liang, H. Shang, M. Mense, M. J. Suto, D. M. Bedwell and S. M. Rowe (2016). "Discovery of Clinically Approved Agents That Promote Suppression of Cystic Fibrosis Transmembrane Conductance Regulator Nonsense Mutations." American Journal of Respiratory and Critical Care Medicine **194**(9): 1092-1103.
Nakayama, T., H. Suzuki and T. Nishino (2003). "Anthocyanin acyltransferases: specificities, mechanism, phylogenetics, and applications." Journal of Molecular Catalysis B: Enzymatic **23**(2): 117-132.
Navarro Del Hierro, J., T. Herrera, T. Fornari, G. Reglero and D. Martin (2018). "The gastrointestinal behavior of saponins and its significance for their bioavailability and bioactivities." Journal of Functional Foods **40**: 484-497.
Nelson, D. and D. Werck-Reichhart (2011). "A P450-centric view of plant evolution." The Plant Journal **66**(1): 194-211.
Nicholas, H. J. (1963). Biogenesis of Terpenes in Plants. Biogenesis of Natural Compounds. P. Bernfeld. Oxford, Pergammon Press.
Niggeweg, R., A. J. Michael and C. Martin (2004). "Engineering plants with increased levels of the antioxidant chlorogenic acid." Nature Biotechnology **22**(6): 746-754.
Nishizaki, Y., M. Yasunaga, E. Okamoto, M. Okamoto, Y. Hirose, M. Yamaguchi, Y. Ozeki and N. Sasaki (2013). "p-Hydroxybenzoyl-Glucose Is a Zwitter Donor for the Biosynthesis of 7-Polyacylated Anthocyanin in Delphinium." The Plant Cell **25**(10): 4150-4165.
Nomura, Y., H. Seki, T. Suzuki, K. Ohyama, M. Mizutani, T. Kaku, K. Tamura, E. Ono, M. Horikawa, H. Sudo, H. Hayashi, K. Saito and T. Muranaka (2019). "Functional specialization

of UDP -glycosyltransferase 73P12 in licorice to produce a sweet triterpenoid saponin, glycyrrhizin." The Plant Journal **99**(6): 1127-1143.
Nützmann, H.-W. and A. Osbourn (2014). "Gene clustering in plant specialized metabolism." Current Opinion in Biotechnology **26**: 91-99.
Ober, D. (2010). "Gene duplications and the time thereafter - examples from plant secondary metabolism." Plant Biology.
Ohtani, K., K. Ogawa, R. Kasai, C.-R. Yang, K. Yamasaki, J. Zhou and O. Tanaka (1992). "Oleanane glycosides from Glycyrrhiza yunnanensis roots." Phytochemistry **31**(5): 1747-1752.
Okada, T., M. Y. Hirai, H. Suzuki, M. Yamazaki and K. Saito (2005). "Molecular Characterization of a Novel Quinolizidine Alkaloid O-Tigloyltransferase: cDNA Cloning, Catalytic Activity of Recombinant Protein and Expression Analysis in Lupinus Plants." Plant and Cell Physiology **46**(1): 233-244.
Omar, H. R., I. Komarova, M. El-Ghonemi, A. Fathy, R. Rashad, H. D. Abdelmalak, M. R. Yerramadha, Y. Ali, E. Helal and E. M. Camporesi (2012). "Licorice abuse: time to send a warning message." Therapeutic Advances in Endocrinology and Metabolism **3**(4): 125-138.
Onkokesung, N., E. Gaquerel, H. Kotkar, H. Kaur, I. T. Baldwin and I. Galis (2012). "MYB8 Controls Inducible Phenolamide Levels by Activating Three Novel Hydroxycinnamoyl-Coenzyme A:Polyamine Transferases in Nicotiana attenuata." Plant Physiology **158**(1): 389-407.
Ortega-Berlanga, B. and T. Pniewski (2022). "Plant-Based Vaccines in Combat against Coronavirus Diseases." Vaccines **10**(2): 138.
Osbourn, A. E. and J. Reed (2019). Metabolic Engineering. WIPO. **WO 2019/122259 A1**.
Osbourn, A. E., J. Reed, A. Orme and T. Louveau (2020). Transferase Enzymes. WIPO. **WO 2020/260475 A1**.
Oshikiri, H., B. Watanabe, H. Yamamoto, K. Yazaki and K. Takanashi (2020). "Two BAHD Acyltransferases Catalyze the Last Step in the Shikonin/Alkannin Biosynthetic Pathway." Plant Physiology **184**(2): 753-761.
Osmani, S. A., S. Bak and B. L. Møller (2009). "Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling." Phytochemistry **70**(3): 325-347.
Owen, C., N. J. Patron, A. Huang and A. Osbourn (2017). "Harnessing plant metabolic diversity." Current Opinion in Chemical Biology **40**: 24-30.
Patro, R., G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford (2017). "Salmon provides fast and bias-aware quantification of transcript expression." Nature Methods **14**(4): 417-419.
Patterson, G. W. and W. D. Nes (1991). Physiology and biochemistry of sterols. Illinois, The American Oil Chemists Society.
Pekdemir, T., Y. Ishigami and H. Uchiyama (1999). "Characterization of aescin as a biosurfactant for environmental remediation." Journal of Surfactants and Detergents **2**(3): 337-341.
Pence, H. E. and A. Williams (2010). "ChemSpider: An Online Chemical Information Resource." Journal of Chemical Education **87**(11): 1123-1124.
Peng, M., Y. Gao, W. Chen, W. Wang, S. Shen, J. Shi, C. Wang, Y. Zhang, L. Zou, S. Wang, W. Jian, X. Liu, L. Gong and J. Luo (2016). "Evolutionarily Distinct BAHD N-acyltransferases are Responsible for Natural Variation of Aromatic Amine Conjugates in Rice." The Plant Cell: tpc.00265.02016.
Percival, G. C., I. Barrow, K. Noviss, I. Keary and P. Pennington (2011). "The impact of horse chestnut leaf miner (Cameraria ohridella Deschka and Dimic; HCLM) on vitality, growth and reproduction of Aesculus hippocastanum L." Urban Forestry & Urban Greening **10**(1): 11-17.

Petrik, D. L., S. D. Karlen, C. L. Cass, D. Padmakshan, F. Lu, S. Liu, P. Le Bris, S. Antelme, N. Santoro, C. G. Wilkerson, R. Sibout, C. Lapierre, J. Ralph and J. C. Sedbrook (2014). "p-Coumaroyl-CoA:monolignol transferase (PMT) acts specifically in the lignin biosynthetic pathway in Brachypodium distachyon." The Plant Journal **77**(5): 713-726.
Phillips, D. R., J. M. Rasbery, B. Bartel and S. P. T. Matsuda (2006). "Biosynthetic diversity in plant triterpene cyclization." Current Opinion in Plant Biology **9**(3): 305-314.
Pichersky, E. and D. R. Gang (2000). "Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective." Trends in Plant Science **5**(10): 439-445.
Pichersky, E. and E. Lewinsohn (2011). "Convergent Evolution in Plant Specialized Metabolism." Annual Review of Plant Biology **62**(1): 549-566.
Pittler, M. H. and E. Ernst (1998). "Horse-Chestnut Seed Extract for Chronic Venous Insufficiency." Archives of Dermatology **134**(11).
Podolak, I., A. Galanty and D. Sobolewska (2010). "Saponins as cytotoxic agents: a review." Phytochemistry Reviews **9**(3): 425-474.
Polakowski, T., U. Stahl and C. Lang (1998). "Overexpression of a cytosolic hydroxymethylglutaryl-CoA reductase leads to squalene accumulation in yeast." Applied Microbiology and Biotechnology **49**(1): 66-71.
Polturak, G. and A. Osbourn (2021). "The emerging role of biosynthetic gene clusters in plant defense and plant interactions." PLOS Pathogens **17**(7): e1009698.
Price, K. R., I. T. Johnson, G. R. Fenwick and M. R. Malinow (1987). "The chemistry and biological significance of saponins in foods and feedingstuffs." C R C Critical Reviews in Food Science and Nutrition **26**(1): 27-135.
Qi, X., S. Bakht, M. Leggett, C. Maxwell, R. Melton and A. Osbourn (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): 8233-8238.
Rabot, S. (1995). "Tigloyl-CoA:pseudotropine acyl transferase—An enzyme of tropane alkaloid biosynthesis." Phytochemistry **39**(2): 315-322.
Rahimi, S., J. Kim, I. Mijakovic, K.-H. Jung, G. Choi, S.-C. Kim and Y.-J. Kim (2019). "Triterpenoid-biosynthetic UDP-glycosyltransferases from plants." Biotechnology Advances **37**(7): 107394.
Rao, X. and R. A. Dixon (2019). "Co-expression networks for plant biology: why and how." Acta Biochimica et Biophysica Sinica **51**(10): 981-988.
Rautengarten, C., B. Ebert, M. Ouellet, M. Nafisi, E. E. K. Baidoo, P. Benke, M. Stranne, A. Mukhopadhyay, J. D. Keasling, Y. Sakuragi and H. V. Scheller (2011). "Arabidopsis Deficient in Cutin Ferulate Encodes a Transferase Required for Feruloylation of ω-Hydroxy Fatty Acids in Cutin Polyester " Plant Physiology **158**(2): 654-665.
Reed, J. and A. Osbourn (2018). "Engineering terpenoid production through transient expression in Nicotiana benthamiana." Plant Cell Reports **37**(10): 1431-1441.
Reed, J., M. J. Stephenson, K. Miettinen, B. Brouwer, A. Leveau, P. Brett, R. J. M. Goss, A. Goossens, M. A. O'Connell and A. Osbourn (2017). "A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules." Metabolic Engineering **42**: 185-193.
Rinaldo, A., E. Cavallini, Y. Jia, S. M. A. Moss, D. A. J. McDavid, L. C. Hooper, S. P. Robinson, G. B. Tornielli, S. Zenoni, C. M. Ford, P. K. Boss and A. R. Walker (2015). "A grapevine anthocyanin acyltransferase, transcriptionally regulated by VvMYBA, can produce most acylated anthocyanins present in grape skins." Plant Physiology: pp.01255.02015.
Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.
Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." Genome Biology **11**(3): R25.

Robinson, W. G., B. K. Bachhawat and M. J. Coon (1956). "TIGLYL COENZYME A AND α-METHYLACETOACETYL COENZYME A, INTERMEDIATES IN THE ENZYMATIC DEGRADATION OF ISOLEUCINE." Journal of Biological Chemistry **218**(1): 391-400.
Ru, M., K. Wang, Z. Bai, L. Peng, S. He, T. Pei, Y. Jia, H. Li and Z. Liang (2017). "Molecular cloning and characterisation of two enzymes involved in the rosmarinic acid biosynthesis pathway of Prunella vulgaris L." Plant Cell, Tissue and Organ Culture (PCTOC) **128**(2): 381-390.
Sainsbury, F., E. C. Thuenemann and G. P. Lomonossoff (2009). "pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants." Plant Biotechnology Journal **7**(7): 682-693.
Sander, M. and M. Petersen (2011). "Distinct substrate specificities and unusual substrate flexibilities of two hydroxycinnamoyltransferases, rosmarinic acid synthase and hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl-transferase, from Coleus blumei Benth." Planta **233**(6): 1157-1171.
Sarker, L. S. and S. S. Mahmoud (2015). "Cloning and functional characterization of two monoterpene acetyltransferases from glandular trichomes of L. x intermedia." Planta **242**(3): 709-719.
Sayama, T., E. Ono, K. Takagi, Y. Takada, M. Horikawa, Y. Nakamoto, A. Hirose, H. Sasama, M. Ohashi, H. Hasegawa, T. Terakawa, A. Kikuchi, S. Kato, N. Tatsuzaki, C. Tsukamoto and M. Ishimoto (2012). "The Sg-1 Glycosyltransferase Locus Regulates Structural Diversity of Triterpenoid Saponins of Soybean." The Plant Cell **24**(5): 2123-2138.
Schilmiller, A. L., A. L. Charbonneau and R. L. Last (2012). "Identification of a BAHD acetyltransferase that produces protective acyl sugars in tomato trichomes." Proceedings of the National Academy of Sciences **109**(40): 16377-16382.
Schläpfer, P., P. Zhang, C. Wang, T. Kim, M. Banf, L. Chae, K. Dreher, A. K. Chavali, R. Nilo-Poyanco, T. Bernard, D. Kahn and S. Y. Rhee (2017). "Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants." Plant Physiology **173**(4): 2041-2059.
Schlösser, E. and G. Wulff (1969). "Über die Strukturspezifität der Saponinhämolyse / Structural Specifity of Saponin Hemolysis." Zeitschrift für Naturforschung B **24**(10): 1284-1290.
Seki, H., K. Tamura and T. Muranaka (2015). "P450s and UGTs: Key Players in the Structural Diversity of Triterpenoid Saponins." Plant and Cell Physiology **56**(8): 1463-1471.
Serra, O., C. Hohn, R. Franke, S. Prat, M. Molinas and M. Figueras (2010). "A feruloyl transferase involved in the biosynthesis of suberin and suberin-associated wax is required for maturation and sealing properties of potato periderm." The Plant Journal **62**(2): 277-290.
Shalit, M., I. Guterman, H. Volpin, E. Bar, T. Tamari, N. Menda, Z. Adam, D. Zamir, A. Vainstein, D. Weiss, E. Pichersky and E. Lewinsohn (2003). "Volatile Ester Formation in Roses. Identification of an Acetyl-Coenzyme A. Geraniol/Citronellol Acetyltransferase in Developing Rose Petals." Plant Physiology **131**(4): 1868-1876.
Sharma, P. K., N. S. Sangwan, S. K. Bose and R. S. Sangwan (2013). "Biochemical characteristics of a novel vegetative tissue geraniol acetyltransferase from a monoterpene oil grass (Palmarosa, Cymbopogon martinii var. Motia) leaf." Plant Science **203-204**: 63-73.
Shaw, W. V. (1992). "Chemical anatomy of antibiotic resistance: chloramphenicol acetyltransferase." Science Progress (1933-) **76**(3/4 (301/302)): 565-580.
Shi, L. and B. P. Tu (2015). "Acetyl-CoA and the regulation of metabolism: mechanisms and consequences." Current Opinion in Cell Biology **33**: 125-131.
Shibuya, M., M. Hoshino, Y. Katsube, H. Hayashi, T. Kushiro and Y. Ebizuka (2006). "Identification of beta-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay." FEBS Journal **273**(5): 948-959.

Shibuya, M., K. Nishimura, N. Yasuyama and Y. Ebizuka (2010). "Identification and characterization of glycosyltransferases involved in the biosynthesis of soyasaponin I in Glycine max." FEBS Letters **584**(11): 2258-2264.
Shimura, K., A. Okada, K. Okada, Y. Jikumaru, K.-W. Ko, T. Toyomasu, T. Sassa, M. Hasegawa, O. Kodama, N. Shibuya, J. Koga, H. Nojiri and H. Yamane (2007). "Identification of a Biosynthetic Gene Cluster in Rice for Momilactones." Journal of Biological Chemistry **282**(47): 34013-34018.
Singh, H., S. Sidhu, K. Chopra and M. U. Khan (2017). "The novel role of β-aescin in attenuating CCl4-induced hepatotoxicity in rats." Pharmaceutical Biology **55**(1): 749-757.
Sirtori, C. R. (2001). "Aescin: pharmacology, pharmacokinetics and therapeutic profile." Pharmacological Research **44**(3): 183-193.
Sonawane, P. D., J. Pollier, S. Panda, J. Szymanski, H. Massalha, M. Yona, T. Unger, S. Malitsky, P. Arendt, L. Pauwels, E. Almekias-Siegl, I. Rogachev, S. Meir, P. D. Cárdenas, A. Masri, M. Petrikov, H. Schaller, A. A. Schaffer, A. Kamble, A. P. Giri, A. Goossens and A. Aharoni (2017). "Plant cholesterol biosynthetic pathway overlaps with phytosterol metabolism." Nature Plants **3**(1): 16205.
Sonnante, G., R. D'Amore, E. Blanco, C. L. Pierri, M. De Palma, J. Luo, M. Tucci and C. Martin (2010). "Novel Hydroxycinnamoyl-Coenzyme A Quinate Transferase Genes from Artichoke Are Involved in the Synthesis of Chlorogenic Acid " Plant Physiology **153**(3): 1224-1238.
Sparg, S. G., M. E. Light and J. van Staden (2004). "Biological activities and distribution of plant saponins." Journal of Ethnopharmacology **94**(2): 219-243.
Springob, K. and T. M. Kutchan (2009). Introduction to the different classes of natural products. Plant-derived natural products. A. E. Osbourn and V. Lanzotti. Heidelberg, Springer**:** 3-50.
Sreij, R., C. Dargel, R. Schweins, S. Prévost, R. Dattani and T. Hellweg (2019). "Aescin-Cholesterol Complexes in DMPC Model Membranes: A DSC and Temperature-Dependent Scattering Study." Scientific Reports **9**(1).
Srivastava, P. L., A. Shukla and R. M. Kalunke (2018). "Comprehensive metabolic and transcriptomic profiling of various tissues provide insights for saponin biosynthesis in the medicinally important Asparagus racemosus." Scientific Reports **8**(1): 9098.
St-Pierre, B., P. Laflamme, A.-M. Alarco, V. D and E. Luca (1998). "The terminal O-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer." The Plant Journal **14**(6): 703-713.
Stephenson, M. J., J. Reed, B. Brouwer and A. Osbourn (2018). "Transient Expression in Nicotiana Benthamiana Leaves for Triterpene Production at a Preparative Scale." Journal of Visualized Experiments(138).
Sullivan, M. (2009). "A Novel Red Clover Hydroxycinnamoyl Transferase Has Enzymatic Activities Consistent with a Role in Phaselic Acid Biosynthesis." Plant Physiology **150**(4): 1866-1879.
Suzuki, H., T. Nakayama, S. Nagae, M.-A. Yamaguchi, T. Iwashita, Y. Fukui and T. Nishino (2004). "cDNA cloning and functional characterization of flavonol 3-O-glucoside-6″-O-malonyltransferases from flowers of Verbena hybrida and Lamium purpureum." Journal of Molecular Catalysis B: Enzymatic **28**(2): 87-93.
Suzuki, H., T. Nakayama, M.-a. Yamaguchi and T. Nishino (2004). "cDNA cloning and characterization of two Dendranthema×morifolium anthocyanin malonyltransferases with different functional activities." Plant Science **166**(1): 89-96.
Suzuki, H., T. Nakayama, K. Yonekura-Sakakibara, Y. Fukui, N. Nakamura, M. Nakao, Y. Tanaka, M.-A. Yamaguchi, T. Kusumi and T. Nishino (2001). "Malonyl-CoA:Anthocyanin 5-O-Glucoside-6‴-O-Malonyltransferase from Scarlet Sage (Salvia splendens) Flowers." Journal of Biological Chemistry **276**(52): 49013-49019.

Suzuki, H., T. Nakayama, K. Yonekura-Sakakibara, Y. Fukui, N. Nakamura, M.-a. Yamaguchi, Y. Tanaka, T. Kusumi and T. Nishino (2002). "cDNA Cloning, Heterologous Expressions, and Functional Characterization of Malonyl-Coenzyme A:Anthocyanidin 3-O-Glucoside-6"-O-Malonyltransferase from Dahlia Flowers." Plant Physiology **130**(4): 2142-2151.
Suzuki, H., T. Nishino and T. Nakayama (2007). "cDNA cloning of a BAHD acyltransferase from soybean (Glycine max): Isoflavone 7-O-glucoside-6”-O-malonyltransferase." Phytochemistry **68**(15): 2035-2042.
Suzuki, H., S.-Y. Sawada, K. Yonekura-Sakakibara, T. Nakayama, M.-A. Yamaguchi and T. Nishino (2003). "Identification of a cDNA Encoding Malonyl-Coenzyme A: Anthocyanidin 3-O-Glucoside 6"-O-Malonyltransferase from Cineraria(Senecio cruentus) Flowers." Plant Biotechnology **20**(3): 229-234.
Suzuki, H., S. y. Sawada, K. Watanabe, S. Nagae, M.-a. Yamaguchi, T. Nakayama and T. Nishino (2004). "Identification and characterization of a novel anthocyanin malonyltransferase from scarlet sage (Salvia splendens) flowers: an enzyme that is phylogenetically separated from other anthocyanin acyltransferases." The Plant Journal **38**(6): 994-1003.
Szakiel, A., C. Pączkowski and M. Henry (2011). "Influence of environmental abiotic factors on the content of saponins in plants." Phytochemistry Reviews **10**(4): 471-491.
Taguchi, G., Y. Shitchi, S. Shirasawa, H. Yamamoto and N. Hayashida (2005). "Molecular cloning, characterization, and downregulation of an acyltransferase that catalyzes the malonylation of flavonoid and naphthol glucosides in tobacco cells." The Plant Journal **42**(4): 481-491.
Taguchi, G., T. Ubukata, H. Nozue, Y. Kobayashi, M. Takahi, H. Yamamoto and N. Hayashida (2010). "Malonylation is a key reaction in the metabolism of xenobiotic phenolic glucosides in Arabidopsis and tobacco." The Plant Journal **63**(6): 1031-1041.
Tamura, K., H. Seki, H. Suzuki, M. Kojoma, K. Saito and T. Muranaka (2017). "CYP716A179 functions as a triterpene C-28 oxidase in tissue-cultured stolons of Glycyrrhiza uralensis." Plant Cell Reports **36**(3): 437-445.
Tarazona, S., P. Furió-Tarí, D. Turrà, A. D. Pietro, M. J. Nueda, A. Ferrer and A. Conesa (2015). "Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package." Nucleic Acids Research: gkv711.
Tarazona, S., F. García-Alcalde, J. Dopazo, A. Ferrer and A. Conesa (2011). "Differential expression in RNA-seq: A matter of depth." Genome Research **21**(12): 2213-2223.
Tatsis, E. C. and S. E. O’Connor (2016). "New developments in engineering plant metabolic pathways." Current Opinion in Biotechnology **42**: 126-132.
Thimmappa, R., K. Geisler, T. Louveau, P. O'Maille and A. Osbourn (2014). "Triterpene Biosynthesis in Plants." Annual Review of Plant Biology **65**(1): 225-257.
Thomas, P. A., O. Alhamd, G. Iszkuło, M. Dering and T. A. Mukassabi (2019). "Biological Flora of the British Isles:Aesculus hippocastanum." Journal of Ecology **107**(2): 992-1030.
Tiwari, P., R. S. Sangwan and N. S. Sangwan (2016). "Plant secondary metabolism linked glycosyltransferases: An update on expanding knowledge and scopes." Biotechnology Advances **34**(5): 714-739.
Tokai, T., H. Koshino, N. Takahashi-Ando, M. Sato, M. Fujimura and M. Kimura (2007). "Fusarium Tri4 encodes a key multifunctional cytochrome P450 monooxygenase for four consecutive oxygenation steps in trichothecene biosynthesis." Biochemical and Biophysical Research Communications **353**(2): 412-417.
Töpfer, N., L.-M. Fuchs and A. Aharoni (2017). "The PhytoClust tool for metabolic gene clusters discovery in plant genomes." Nucleic Acids Research **45**(12): 7049-7063.
Toshio Morikawa, Ning Li, Akifumi Nagatomo, Hisashi Matsuda, a. Xian Li and Masayuki Yoshikawa* (2006). "Triterpene Saponins with Gastroprotective Effects from Tea Seed (the Seeds of Camellia sinensis)1." Journal of Natural Products **69**(2): 185-190.

Trdá, L., M. Janda, D. Macková, R. Pospíchalová, P. I. Dobrev, L. Burketová and P. Matušinsky (2019). "Dual Mode of the Saponin Aescin in Plant Protection: Antifungal Agent and Plant Defense Elicitor." Frontiers in Plant Science **10**.
Tucker, I. M., A. Burley, R. E. Petkova, S. L. Hosking, J. R. P. Webster, P. X. Li, K. Ma, J. Doutch, J. Penfold and R. K. Thomas (2021). "Self-assembly in saponin mixtures: Escin/tea, tea/glycyrrhizic acid, and escin/glycyrrhizic acid mixtures." Colloids and Surfaces A: Physicochemical and Engineering Aspects **629**: 127420-127420.
Tucker, I. M., A. Burley, R. E. Petkova, S. L. Hosking, J. R. P. Webster, P. X. Li, K. Ma, J. Doutch, J. Penfold and R. K. Thomas (2021). "Self-assembly in saponin/surfactant mixtures: Escin and sodium dodecylsulfate." Colloids and Surfaces A: Physicochemical and Engineering Aspects **626**: 127019-127019.
Underland, V., I. Sæterdal and E. S. Nilsen (2012). "Cochrane Summary of Findings: Horse Chestnut seed Extract for Chronic Venous Insufficiency." Global Advances in Health and Medicine **1**(1): 122-123.
Unno, H., F. Ichimaida, H. Suzuki, S. Takahashi, Y. Tanaka, A. Saito, T. Nishino, M. Kusunoki and T. Nakayama (2007). "Structural and Mutational Studies of Anthocyanin Malonyltransferases Establish the Features of BAHD Enzyme Catalysis." Journal of Biological Chemistry **282**(21): 15812-15822.
Varbanova, M., K. Porter, F. Lu, J. Ralph, R. Hammerschmidt, A. D. Jones and B. Day (2011). "Molecular and Biochemical Basis for Stress-Induced Accumulation of Free and Bound p-Coumaraldehyde in Cucumber " Plant Physiology **157**(3): 1056-1066.
Voutquenne-Nazabadioko, L. (2010). "Etude chimiotaxonomique de la famille des Sapindaceae." Ethnopharmacologia **45**: 49-52.
Voutquenne, L., P. Guinot, C. Froissard, O. Thoison, M. Litaudon and C. Lavaud (2005). "Haemolytic acylated triterpenoid saponins from Harpullia austro-caledonica." Phytochemistry **66**(7): 825-835.
Voutquenne, L., C. Lavaud, G. Massiot and C. Delaude (1998). "Saponins from Harpullia cupanioides." Phytochemistry **49**(7): 2081-2085.
Voutquenne, L., C. Lavaud, G. Massiot and L. L. Men-Olivier (2002). "Structure-Activity Relationships of Haemolytic Saponins." Pharmaceutical Biology **40**(4): 253-262.
Wagner, A., J. Ralph, T. Akiyama, H. Flint, L. Phillips, K. Torr, B. Nanayakkara and L. T. Kiri (2007). "Exploring lignification in conifers by silencing hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyltransferase in Pinus radiata." Proceedings of the National Academy of Sciences **104**(28): 11856-11861.
Walker, A. M., R. P. Hayes, B. Youn, W. Vermerris, S. E. Sattler and C. Kang (2013). "Elucidation of the Structure and Reaction Mechanism of Sorghum Hydroxycinnamoyltransferase and Its Structural Relationship to Other Coenzyme A-Dependent Transferases and Synthases " Plant Physiology **162**(2): 640-651.
Walker, K. and R. Croteau (2000). "Molecular cloning of a 10-deacetylbaccatin III-10-O-acetyl transferase cDNA from Taxus and functional expression in Escherichia coli." Proceedings of the National Academy of Sciences **97**(2): 583-587.
Walker, K. and R. Croteau (2000). "Taxol biosynthesis: Molecular cloning of a benzoyl-CoA:taxane 2α-O-benzoyltransferase cDNA from Taxus and functional expression in Escherichia coli." Proceedings of the National Academy of Sciences **97**(25): 13591-13596.
Walker, K., S. Fujisaki, R. Long and R. Croteau (2002). "Molecular cloning and heterologous expression of the C-13 phenylpropanoid side chain-CoA acyltransferase that functions in Taxol biosynthesis." Proceedings of the National Academy of Sciences **99**(20): 12715-12720.
Walker, K., R. Long and R. Croteau (2002). "The final acylation step in Taxol biosynthesis: Cloning of the taxoid C13-side-chain N-benzoyltransferase from Taxus." Proceedings of the National Academy of Sciences **99**(14): 9166-9171.

Walker, K., A. Schoendorf and R. Croteau (2000). "Molecular Cloning of a Taxa-4(20),11(12)-dien-5α-ol-O-Acetyl Transferase cDNA from Taxus and Functional Expression in Escherichia coli." Archives of Biochemistry and Biophysics **374**(2): 371-380.
Wang, C., J. Li, M. Ma, Z. Lin, W. Hu, W. Lin and P. Zhang (2021). "Structural and Biochemical Insights Into Two BAHD Acyltransferases (AtSHT and AtSDT) Involved in Phenolamide Biosynthesis." Frontiers in Plant Science **11**.
Wang, J. and V. D. Luca (2005). "The biosynthesis and regulation of biosynthesis of Concord grape fruit esters, including 'foxy' methylanthranilate." The Plant Journal **44**(4): 606-619.
Wang, J. P., P. P. Naik, H.-C. Chen, R. Shi, C.-Y. Lin, J. Liu, C. M. Shuford, Q. Li, Y.-H. Sun, S. Tunlaya-Anukit, C. M. Williams, D. C. Muddiman, J. J. Ducoste, R. R. Sederoff and V. L. Chiang (2014). "Complete Proteomic-Based Enzyme Reaction and Inhibition Kinetics Reveal How Monolignol Biosynthetic Enzyme Families Affect Metabolic Flux and Lignin in Populus trichocarpa " The Plant Cell **26**(3): 894-914.
Wang, T., F. Fu, L. Zhang, B. Han, M. Zhu and X. Zhang (2009). "Effects of escin on acute inflammation and the immune system in mice." Pharmacological Reports **61**(4): 697-704.
Watson, M. (2006). "CoXpress: differential co-expression in gene expression data." BMC Bioinformatics **7**(1).
Wehrens, R. and L. M. C. Buydens (2007). "Self- and Super-organizing Maps in R: The kohonen Package." Journal of Statistical Software **21**(5): 1 - 19.
Wei, F., L.-Y. Ma, W.-T. Jin, S.-C. Ma, G.-Z. Han, I. A. Khan and R.-C. Lin (2004). "Antiinflammatory Triterpenoid Saponins from the Seeds of Aesculus chinensis." Chemical and Pharmaceutical Bulletin **52**(10): 1246-1248.
Weier, D., J. Mittasch, D. Strack and C. Milkowski (2008). "The genes BnSCT1 and BnSCT2 from Brassica napus encoding the final enzyme of sinapine biosynthesis: molecular characterization and suppression." Planta **227**(2): 375-385.
Wendt, K. U., G. E. Schulz, E. J. Corey and D. R. Liu (2000). "Enzyme Mechanisms for Polycyclic Triterpene Formation." Angewandte Chemie International Edition **39**(16): 2812-2833.
Weng, J.-K., R. N. Philippe and J. P. Noel (2012). "The Rise of Chemodiversity in Plants." Science **336**(6089): 1667-1670.
Winzer, T., V. Gazda, Z. He, F. Kaminski, M. Kern, T. R. Larson, Y. Li, F. Meade, R. Teodor, F. E. Vaistij, C. Walker, T. A. Bowser and I. A. Graham (2012). "A Papaver somniferum 10-Gene Cluster for Synthesis of the Anticancer Alkaloid Noscapine." Science **336**(6089): 1704-1708.
Withers, S., F. Lu, H. Kim, Y. Zhu, J. Ralph and C. G. Wilkerson (2012). "Identification of Grass-specific Enzyme That Acylates Monolignols with p-Coumarate." Journal of Biological Chemistry **287**(11): 8347-8355.
Wulff, G. and R. Tschesche (1969). "Über triterpene—XXVI: Über die struktur der rosskastaniensaponnie (Aescin) und die aglykone verwandter glykoside." Tetrahedron **25**(2): 415-436.
Xu, J., X.-Y. Wang and W.-Z. Guo (2015). "The cytochrome P450 superfamily: Key players in plant development and defense." Journal of Integrative Agriculture **14**(9): 1673-1686.
Xue, Z., L. Duan, D. Liu, J. Guo, S. Ge, J. Dicks, P. Ómáille, A. Osbourn and X. Qi (2012). "Divergent evolution of oxidosqualene cyclases in plants." New Phytologist **193**(4): 1022-1038.
Xue, Z., Z. Tan, A. Huang, Y. Zhou, J. Sun, X. Wang, R. B. Thimmappa, M. J. Stephenson, A. Osbourn and X. Qi (2018). "Identification of key amino acid residues determining product specificity of 2,3-oxidosqualene cyclase in Oryza species." New Phytologist **218**(3): 1076-1088.
Yang, Q., K. Reinhard, E. Schiltz and U. Matern (1997). "Characterization and heterologous expression of hydroxycinnamoyl/benzoyl-CoA:anthranilate N-

hydroxycinnamoyl/benzoyltransferase from elicited cell cultures of carnation, Dianthus caryophyllus L." Plant Molecular Biology **35**(6): 777-789.
Yang, Q., H. Xuan Trinh, S. Imai, A. Ishihara, L. Zhang, H. Nakayashiki, Y. Tosa and S. Mayama (2004). "Analysis of the Involvement of Hydroxyanthranilate Hydroxycinnamoyltransferase and Caffeoyl-CoA 3-O-Methyltransferase in Phytoalexin Biosynthesis in Oat." Molecular Plant-Microbe Interactions® **17**(1): 81-89.
Yang, X.-W., J. Zhao, J.-R. Cui and W. Guo (2004). "Studies on the biotransformation of escin Ia by human intestinal bacteria and the anti-tumor activities of desacylescin I." Journal of Peking University Health Sciences **36**(1): 31-35.
Yang, Y., L. Long, X. Zhang, K. Song, D. Wang, X. Xiong, H. Gao and L. Sha (2019). "16-Tigloyl linked barrigenol-like triterpenoid from Semen Aesculi and its anti-tumor activity in vivo and in vitro." RSC Advances **9**(54): 31758-31772.
Yano, R., K. Takagi, Y. Takada, K. Mukaiyama, C. Tsukamoto, T. Sayama, A. Kaga, T. Anai, S. Sawai, K. Ohyama, K. Saito and M. Ishimoto (2017). "Metabolic switching of astringent and beneficial triterpenoid saponins in soybean is achieved by a loss-of-function mutation in cytochrome P450 72A69." The Plant Journal **89**(3): 527-539.
Yao, J., Y. Weng, A. Dickey and K. Wang (2015). "Plants as Factories for Human Pharmaceuticals: Applications and Challenges." International Journal of Molecular Sciences **16**(12): 28549-28565.
Yasumoto, S., H. Seki, Y. Shimizu, E. O. Fukushima and T. Muranaka (2017). "Functional Characterization of CYP716 Family P450 Enzymes in Triterpenoid Biosynthesis in Tomato." Frontiers in Plant Science **8**(21).
Yokosuka, A., S. Kawakami, M. Haraguchi and Y. Mimaki (2008). "Stryphnosides A–F, six new triterpene glycosides from the pericarps of Stryphnodendron fissuratum." Tetrahedron **64**(7): 1474-1481.
Yokosuka, A., G. Okabe, S. Tatsuno and Y. Mimaki (2016). "Stryphnosides G–P, 10 new triterpene glycosides from the pericarps of Stryphnodendron fissuratum." Carbohydrate Research **434**: 18-26.
Yonekura-Sakakibara, K., Y. Tanaka, M. Fukuchi-Mizutani, H. Fujiwara, Y. Fukui, T. Ashikari, Y. Murakami, M. Yamaguchi and T. Kusumi (2000). "Molecular and Biochemical Characterization of a Novel Hydroxycinnamoyl-CoA: Anthocyanin 3-O-Glucoside-6”-O-Acyltransferase from Perilla frutescens." Plant and Cell Physiology **41**(4): 495-502.
Yoshikawa, M., E. Harada, T. Murakami, H. Matsuda, N. Wariishi, J. Yamahara, N. Murakami and I. Kitagawa (1994). "Escins-Ia, Ib, IIa, IIb, and IIIa, Bioactive Triterpene Oligoglycosides from the Seeds of Aesculus hippocastanum L.: Their Inhibitory Effects on Ethanol Absorption and Hypoglycemic Activity on Glucose Tolerance Test." Chemical and Pharmaceutical Bulletin **42**(6): 1357-1359.
Yoshikawa, M., T. Morikawa, S. Nakamura, N. Li, X. Li and H. Matsuda (2007). "Bioactive Saponins and Glycosides. XXV. Acylated Oleanane-Type Triterpene Saponins from the Seeds of Tea Plant (Camellia sinensis)." Chemical and Pharmaceutical Bulletin **55**(1): 57-63.
Yu, Z. and P. Su (2013). "Effect of beta-aescin extract from Chinese Buckeye Seed on chronic venous insufficiency." Die Pharmazie - An International Journal of Pharmaceutical Sciences **68**(6): 428-430.
Yuan, W., P. Wang, Z. Su, V. S. Wang and S. Li (2013). "Cytotoxic triterpenoid saponins from husks of Aesculus californica (Spach) Nutt." Phytochemistry **90**: 95-105.
Zdravković-Korać, S., J. Milojević, M. Belić and D. Ćalić (2022). "Tissue Culture Response of Ornamental and Medicinal Aesculus Species—A Review." Plants **11**(3): 277.
Zhang, H., X. Hua, D. Zheng, H. Wu, C. Li, P. Rao, M. Wen, Y.-E. Choi, Z. Xue, Y. Wang and Y. Li (2022). "De Novo Biosynthesis of Oleanane-Type Ginsenosides in Saccharomyces cerevisiae Using Two Types of Glycosyltransferases from Panax ginseng." Journal of Agricultural and Food Chemistry **70**(7): 2231-2240.

Zhang, X., S. Zhang, Y. Yang, D. Wang and H. Gao (2019). "Natural barrigenol–like triterpenoids: A comprehensive review of their contributions to medicinal chemistry." Phytochemistry **161**: 41-74.
Zhang, Z. and S. Li (2007). "Cytotoxic triterpenoid saponins from the fruits of Aesculus pavia L." Phytochemistry **68**(15): 2075-2086.
Zhang, Z., S. Li and X.-Y. Lian (2010). "An Overview of Genus Aesulus L.: Ethnobotany, Phytochemistry, and Pharmacological Activities." Pharmaceutical Crops **1**: 24-51.
Zhang, Z., S. Li, S. Zhang and D. Gorenstein (2006). "Triterpenoid saponins from the fruits of Aesculus pavia." Phytochemistry **67**(8): 784-794.
Zhao, J., D. Huhman, G. Shadle, X.-Z. He, L. W. Sumner, Y. Tang and R. A. Dixon (2011). "MATE2 Mediates Vacuolar Sequestration of Flavonoid Glycosides and Glycoside Malonates in Medicago truncatula " The Plant Cell **23**(4): 1536-1555.
Zhao, Y.-J. and C. Li (2018). "Biosynthesis of Plant Triterpenoid Saponins in Microbial Cell Factories." Journal of Agricultural and Food Chemistry **66**(46): 12155-12165.
Zhou, Q.-M., H. Zhu, R. Feng, C. Peng, H. Yang, F. Liu, Y.-N. Wang and L. Xiong (2019). "New triterpenoids from Leonurus japonicus (Lamiaceae)." Biochemical Systematics and Ecology **82**: 27-30.
Zhou, W., Q. Huang, X. Wu, Z. Zhou, M. Ding, M. Shi, F. Huang, S. Li, Y. Wang and G. Kai (2017). "Comprehensive transcriptome profiling of Salvia miltiorrhiza for discovery of genes associated with the biosynthesis of tanshinones and phenolic acids." Scientific Reports **7**(1).
Zhou, Y., Y. Ma, J. Zeng, L. Duan, X. Xue, H. Wang, T. Lin, Z. Liu, K. Zeng, Y. Zhong, S. Zhang, Q. Hu, M. Liu, H. Zhang, J. Reed, T. Moses, X. Liu, P. Huang, Z. Qing, X. Liu, P. Tu, H. Kuang, Z. Zhang, A. Osbourn, D.-K. Ro, Y. Shang and S. Huang (2016). "Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae." Nature Plants **2**(12): 16183.

Appendices

Appendix A1 – Characterised BAHD acyltransferases

BAHD	SPECIES	MAIN ACYL GROUP	SUBSTRATE	REFERENCE
ACT	*Hordeum vulgare*	Hydroxycinnamoyl	Agmatine	(Burhenne, Kristensen et al. 2003)
AMAT	*Vitis labrusca*	Anthranilate	Methanol	(Wang and Luca 2005)
ASHHT	*Avena sativa*	Hydroxycinnamoyl	Alkaloid	(Yang, Xuan Trinh et al. 2004)
AT1	*Nicotiana attenuata*	Caffeoyl	Putrescine	(Onkokesung, Gaquerel et al. 2012)
ATACT	*Arabidopsis thaliana*	Coumaroyl	Agmatine	(Muroi, Ishihara et al. 2009)
ATHCT	*Arabidopsis thaliana*	Hydroxycinnamoyl	Chlorogenic acids	(D'Auria 2006)
ATHHT	*Arabidopsis thaliana*	Ferrulate	Suberin	(Gou, Yu et al. 2009)
ATPMAT1	*Arabidopsis thaliana*	Malonyl	Flavonoids	(Taguchi, Ubukata et al. 2010)
ATPMAT2	*Arabidopsis thaliana*	Malonyl	Flavonoids	(Taguchi, Ubukata et al. 2010)
ATSCT	*Arabidopsis thaliana*	Coumaroyl	Spermidine	(Luo, Fuell et al. 2009)
ATSDT	*Arabidopsis thaliana*	Sinapoyl	Spermidine	(Luo, Fuell et al. 2009)
BANAAT	*Musa sapientum*	Acetyl	Cinnamic acid	(Beekwilder, Alvarez-Huerta et al. 2004)
BAPT	*Taxus cuspidata*	Phenylalanoyl	Baccatin	(Walker, Fujisaki et al. 2002)
BDPMT	*Brachypodium distachyon*	Coumaroyl	Lignin	(Petrik, Karlen et al. 2014)
BPBT	*Petunia x hybridia*	Benzoyl	Benzyl alcohol	(Boatright, Negre et al. 2004)
BRACT	*Brassica rapa*	Acetyl	Triterpenoid	(Liu, Suarez Duran et al. 2020)
CBBEAT	*Clarkia breweri*	Acetyl	Benzyl alcohol	(Dudareva, D'Auria et al. 1998)
CBBEBT	*Clarkia breweri*	Benzoyl	Benzyl alcohol	(D'Auria, Chen et al. 2002)
CBHST	*Coleus blumei*	Hydroxycinnamoyl	Shikimates	(Sander and Petersen 2011)
CBRAS	*Coleus blumei*	Caffeate; coumaroyl	Phenyllactates	(Sander and Petersen 2011)
CCACT1	*Crocosmia x crocosmiiflora*	Caffeate & others	Flavonoids	(Irmisch, Jo et al. 2018)
CCACT2	*Crocosmia x crocosmiiflora*	Caffeate & others	Flavonoids	(Irmisch, Jo et al. 2018)
CCHQT	*Coffea canephora*	Hydroxycinnamoyl	Quinic acid	(Lepelley, Cheminade et al. 2007)
CHAT	*Arabidopsis thalana*	Acetyl	Hexenol	(D'Auria, Chen et al. 2002)
CRACT	*Capsella rubella*	Acetyl	Triterpenoid	(Liu, Suarez Duran et al. 2020)
CLACT	*Citrullus lanatus*	Acetyl	Triterpenoid	(Zhou, Ma et al. 2016)
CSHCT	*Cucumis sativus*	Coumaroyl	Shikimates	(Varbanova, Porter et al. 2011)
DAT	*Catharanthus roseus*	Acetyl	Alkaloids	(St-Pierre, Laflamme et al. 1998)

DBAT	*Taxus cuspidata*	Acetyl	Diterpenoids	(Walker and Croteau 2000)
DBBT	*Taxus cuspidata*	Benzoyl	Diterpenoids	(Walker and Croteau 2000)
DBNTBT	*Taxus canadensis*	Benzoyl	Diterpenoids	(Walker, Long et al. 2002)
DCF	*Arabidopsis thaliana*	Ferrulate; Sinapoyl	Fatty acids	(Rautengarten, Ebert et al. 2011)
DH29	*Nicotiana attenuata*	Caffeate	Spermidine	(Onkokesung, Gaquerel et al. 2012)
DM3MAT1	*Chrysanthemum x*	Malonyl	Anthocyanins	(Suzuki, Nakayama et al. 2004)
DM3MAT2	*Chrysanthemum x*	Malonyl	Anthocyanins	(Suzuki, Nakayama et al. 2004)
DV3MAT	*Dahlia variabilis*	Malonyl	Anthocyanins	(Suzuki, Nakayama et al. 2002)
FACT	*Arabidopsis thaliana*	Caffeate	Fatty acids	(Kosma, Molina et al. 2012)
FHT	*Solanum tuberosum*	Ferulate	Suberin	(Serra, Hohn et al. 2010)
GAAT	*Cymbopogon martinii*	Acetyl	Monoterpenoids	(Sharma, Sangwan et al. 2013)
GMIF7MAT	*Glycine max*	Malonyl	Flavonoids	(Suzuki, Nishino et al. 2007)
GMMT7	*Glycine max*	Malonyl	Flavonoids	(Dhaubhadel, Farhangkhoee et al. 2008)
GT5AT	*Gentiana triflora*	Hydroxycinnamoyl	Anthocyanins	(Fujiwara, Tanaka et al. 1998)
HCBT	*Dianthus carophyllus*	Hydroxycinnamoyl	Dianthramides	(Yang, Reinhard et al. 1997)
HMT/HLT	*Lupinus albus*	Tigloyl	Alkaloids	(Okada, Hirai et al. 2005)
HQT	*Nicotiana tabacum*	Hydroxylcinnamoyl	Quinic acids	(Niggeweg, Michael et al. 2004)
HQT1	*Cynara cardunculus*	Caffeate; coumaroyl	Quinic acid	(Sonnante, D'Amore et al. 2010)
HQT2	*Cynara cardunculus*	Caffeate; coumaroyl	Quinic acid	(Sonnante, D'Amore et al. 2010)
LAAT1	*Lavandula angustifolia*	Hydroxycinnamoyl	Shikimates	(Landmann, Hücherig et al. 2011)
LIAAT3	*Lavandula intermedia*	Acetyl	Monoterpenoids	(Sarker and Mahmoud 2015)
LIAAT4	*Lavandula intermedia*	Acetyl	Monoterpenoids	(Sarker and Mahmoud 2015)
LP3MAT1	*Lamium purpureum*	Malonyl	Flavonoids	(Suzuki, Nakayama et al. 2004)
MAT4	*Medicago trunculata*	Malonyl	Flavonoids	(Zhao, Huhman et al. 2011)
MAT5	*Medicago trunculata*	Malonyl	Anthocyanins	(Zhao, Huhman et al. 2011)
MAT6	*Medicago trunculata*	Malonyl	Anthocyanins	(Zhao, Huhman et al. 2011)
NTBEBT	*Nicotiana tabacum*	Benzoyl	Benzyl alcohol	(D'Auria, Chen et al. 2002)
NTHCT	*Nicotiana tabacum*	Hydroxycinnamoyl	Quinic acid	(Hoffmann, Maury et al. 2003)
NTMAT1	*Nicotiana tabacum*	Malonyl	Flavonoids	(Taguchi, Shitchi et al. 2005)
OSAHT1	*Oryza sativa*	Multiple aromatics	Agmatine	(Peng, Gao et al. 2016)
OSAT10	*Oryza sativa*	Coumaroyl	Xylan	(Bartley, Peck et al. 2013)
OSHCT4	*Oryza sativa*	Hydroxycinnamoyl	Shikimates; glycerol	(Kim, Kim et al. 2012)

OSPHT1	*Oryza sativa*	Coumaroyl	Putrescine	(Peng, Gao et al. 2016)
OSPHT2	*Oryza sativa*	Coumaroyl	Putrescine	(Peng, Gao et al. 2016)
OSPHT3	*Oryza sativa*	Multiple aromatics	Putrescine	(Peng, Gao et al. 2016)
OSPMT	*Oryza sativa*	Coumaroyl	Lignol	(Withers, Lu et al. 2012)
OSTBT1	*Oryza sativa*	Multiple aromatics	Tryptamine	(Peng, Gao et al. 2016)
OSTBT2	*Oryza sativa*	Cinnamoyl; benzoyl	Tryptamine	(Peng, Gao et al. 2016)
OSTHT1	*Oryza sativa*	Multiple aromatics	Tryptamine	(Peng, Gao et al. 2016)
OSTHT2	*Oryza sativa*	Multiple aromatics	Tryptamine	(Peng, Gao et al. 2016)
PEHCT	*Populus euramericana*	Coumaroyl	Shikimates	(Kim, Kim et al. 2011)
PF3AT	*Perilla frutescens*	Hydroxylcinnamoyl	Anthocyanins	(Yonekura-Sakakibara, Tanaka et al. 2000)
PF5MAT	*Perilla frutescens*	Malonyl	Anthocyanins	(Suzuki, Nakayama et al. 2001)
PRHCT	*Pinus radiata*	Hydroxycinnamoyl	Shikimates	(Wagner, Ralph et al. 2007)
PTBEBT	*Populus trichocarpa*	Benzoyl	Benzyl alcohol	(Chedgy, Köllner et al. 2015)
PTFHT1	*Populus trichocarpa*	Ferulate; coumaroyl	Fatty acids	(Cheng, Gou et al. 2013)
PTHCT1	*Populus trichocarpa*	Caffeate; coumaroyl	Shikimates	(Wang, Naik et al. 2014)
PTHCT6	*Populus trichocarpa*	Caffeate; coumaroyl	Shikimates	(Wang, Naik et al. 2014)
PTSABT	*Populus trichocarpa*	Benzoyl	Salicilic acid	(Chedgy, Köllner et al. 2015)
PVHCT1A	*Panicum virgatum*	Caffeate; coumaroyl	Shikimates	(Escamilla-Treviño, Shen et al. 2014)
PVHCT2A	*Panicum virgatum*	Caffeate; coumaroyl	Shikimates	(Escamilla-Treviño, Shen et al. 2014)
PVHCT-LIKE1	*Panicum virgatum*	Caffeate; coumaroyl	Quinic acid	(Escamilla-Treviño, Shen et al. 2014)
PVRAS	*Prunella vulgaris*	Coumaroyl	Phenyllactates	(Ru, Wang et al. 2017)
RHAAT1	*Rosa hybrid*	Acetyl	Monoterpenes	(Shalit, Guterman et al. 2003)
SALAT	*Papaver somniferum*	Acetyl	Alkaloid	(Grothe, Lenz et al. 2001)
SBHCT	*Sorghum bicolor*	Coumaroyl	Shikimates	(Walker, Hayes et al. 2013)
SC3MAT	*Senecio cruentus*	Malonyl	Anthocyanins	(Suzuki, Sawada et al. 2003)
SHT	*Arabidopsis thaliana*	Hydroxycinnamoyl	Spermidine	(Grienenberger, Besseau et al. 2009)
SLAT2	*Solanum lycopersicon*	Acetyl	Acyl sugars	(Schilmiller, Charbonneau et al. 2012)
SLHQT	*Solanum lycopersicon*	Caffeate; coumaroyl	Chlorogenic acid	(Moglia, Lanteri et al. 2014)
SOAP10	*Spinacia oleracea*	Acetyl	Triterpenoids	(Jozwiak, Sonawane et al. 2020)
SS5MAT1	*Salvia splendens*	Malonyl	Anthocyanins	(Suzuki, Nakayama et al. 2001)
SS5MAT2	*Salvia splendens*	Malonyl	Anthocyanins	(Suzuki, Sawada et al. 2004)
TAT	*Taxus cuspidata*	Acetyl	Taxol	(Walker, Schoendorf et al. 2000)
THAA1	*Arabidopsis thaliana*	Acetyl	Triterpenoids	(Huang, Jiang et al. 2019)
THAA2	*Arabidopsis thaliana*	Acetyl	Triterpenoids	(Huang, Jiang et al. 2019)
THAA3	*Arabidopsis thaliana*	Palmitoyl	Triterpenoids	(Huang, Jiang et al. 2019)
TPHCT1A	*Trifolium pratense*	Caffeate; coumaroyl	Shikimates	(Sullivan 2009)

TPHCT1B	*Trifolium pratense*	Caffeate; coumaroyl	Shikimates	(Sullivan 2009)
TPHCT2	*Trifolium pratense*	Caffeate; coumaroyl	Malic acid	(Sullivan 2009)
VAAT	*Fragaria vesca*	Acetyl	Aliphatics	(Beekwilder, Alvarez-Huerta et al. 2004)
VH3MAT1	*Verbena hybrida*	Malonyl	Flavonoids	(Suzuki, Nakayama et al. 2004)
VINORINE SYNTHASE	*Rauvolfia serpentina*	Acetyl	Alkaloids	(Bayer, Ma et al. 2004)

Appendix A2 – Characterised SCPL acyltransferases

SCPL	SPECIES	MAIN ACYL GROUP	SUBSTRATE	REFERENCE
AA7G-AT	*Delphinium grandiflorum*	Benzoyl	Anthocyanins	(Nishizaki, Yasunaga et al. 2013)
AA7GBG-AT	*Delphinium grandiflorum*	Benzoyl	Anthocyanins	(Nishizaki, Yasunaga et al. 2013)
AMALT	*Dianthus caryophyllus*	Malonyl	Anthocyanins	(Abe, Tera et al. 2008)
BNSCT1	*Brassica napus*	Sinapoyl	Choline	(Milkowski, Baumert et al. 2004)
BNSCT2	*Brassica napus*	Sinapoyl	Choline	(Weier, Mittasch et al. 2008)
CTAT1	*Clitoria ternatea*	Coumaroyl	Anthocyanins	(Bontpart, Cheynier et al. 2015)
DKSCPL1	*Diosypros kaki*	Galloyl	Flavonoids	(Ikegami, Eguchi et al. 2007)
DKSCPL2	*Diospyros kaki*	Galloyl	Flavonoids	(Akagi, Ikegami et al. 2009)
ECGT	*Camellia sinensis*	Galloyl	Flavonoids	(Liu, Gao et al. 2012)
GAC	*Solanum pennelli*	Iso-Butyrate	Acyl sugars	(Li and Steffens 2000)
IAINOS SYNTHASE	*Oryza sativa*	Indole-acetyl	Myoinositol	(Ciarkowska, Ostrowski et al. 2018)
SAD7	*Avena strigosa*	N-methylanthranilate; Benzoyl	Triterpenoids	(Mugford, Qi et al. 2009)
SAT	*Arabidopsis thaliana*	Sinapoyl	Anthocyanins	(Fraser, Thompson et al. 2007)
SCPL17	*Arabidopsis thaliana*	Sinapoyl; Benzoyl	Glucosinolates	(Lee, Kaminaga et al. 2012)
SCT	*Arabidopsis thaliana*	Sinapoyl	Choline	(Fraser, Thompson et al. 2007)
SMT	*Arabidopsis thaliana*	Sinapoyl	Malic acid	(Fraser, Thompson et al. 2007)
SST	*Arabidopsis thaliana*	Sinapoyl	Acyl sugars	(Fraser, Thompson et al. 2007)
VVGAT1	*Vitis vinifera*	Galloyl	Flavonoids	(Bontpart, Cheynier et al. 2015)
VVGAT2	*Vitis vinifera*	Galloyl	Flavonoids	(Bontpart, Cheynier et al. 2015)
UNNAMED SCPL	*Iponema batatas*	Hydroyxcinnamoyl	Quinic acid	(Bontpart, Cheynier et al. 2015)
UNNAMED SCPL	*Quercus sp.*	Galloyl	Tannins	(Bontpart, Cheynier et al. 2015)
UNNAMED SCPL	*Zea mays*	Indole-acetyl	Myoinostinol	(Bontpart, Cheynier et al. 2015)
UNNAMED SCPL	*Daucus carota*	Hydroxylcinnamoyl	Anthocyanins	(Bontpart, Cheynier et al. 2015)

Appendix B – Protein sequences of cloned *Aesculus* candidates

Chapter 4:

>2033231

MWKLKTAEGANGNPYLYSTNNFVGRQTWEFDPNAGTLEERAEVEAARQKFYENRFQVKPSGDLLWR
MQFLKEKNFNQTIAPVKVEDGEEITYETASTALKRAVHFFSALQASDGHWPAENAGPLFFLPPLVMCVY
ITGHLNTVFPAEHRKEILRYIYYHQNKDGGWGLHIEGHSTMFCTVLSYVCMRILGEGPDGGQDNACTR
ARKWILDHGGATHIPSWGKTWLSIFGVCEWAASNPMPPEFWILPSFLPMHPAKMWCYCRMVYMP
MSYLYGKRFVAPITPLTLQLREELYNEPYDQINWSKARHLCAPEDIYYPHPWIQDMMWDSLYILTEPLLT
RWPFNKLVRDHALKVTMKHIHYEDENSRYITIGCVEKVLCMLACWVEDPNGAYFKKHLARIPDYLWVA
EDGMKMQSFGSQEWDTGFAIQALLASNFTDEILPVLKRGHEFIKASQVKDNPSGDFKSMHRHISKGS
WTFSDQDHGWQVSDCTAEGLKCCLLFSMMPPELVGEKMEAERLYDSVNVLLSLQSKNGGLAAWEPA
GAQEWLELLNPTEFFADIVIEHEYVECTSSAIHTLVLFKKLHPGHRKKEIEIFIANAVRYLENVQMPDGSW
YGNWGVCFTYGTWFALGGLAAVGKTYNNCPTVRKAVEFLLNSQRDNGGWGESYRSCPEKKYVPLEG
NKSNLVHTAWAMMGLIHSGQAERDPTPLHRAAKLIINSQFEDGDFPQEEITGVFMKNCMLHYAAYRN
*

>2020895

MEFFFVCGMLLFTLFVSVSLAYLFFHKPKNLPPGKMGLPYIGESLEFLRTARNGHPDKFIFDRLAKYSSKV
FKTSILTKPTAVFCGPVANKFLFSNENKIVHVWFPESVTKIFPSTGSVKEESLKMRKMLPSFLKPEALMKY
VGVMDHIAKKHFETDWEGKQQVNVFPLAKRYTFWVACKLFLSIEDPKHIAKFADPFKSLIAGIYTIPLNF
PGTPYYNSIKAAKEIRKELIGIIKQRKIDLQENRASPTQDLLSYVLQETDENGQPMSELFIADKIIGLLIGGH
DTASSTISFIVKYLAELPHVYNEVLKEQMEIAKSKKEGEFLNWDDLQKMKYSWNVACEVLRLTPPLQGA
FREAINDFIFNGFSIPKGWKIYWSTYSTHKDAEYFPEPEKFDPTRFEGKGPAPYTYVPFGGGPRMCPGKE
YARLEILVFMHNVIKRFNWKQLLADEKTVAEPIPTPAKGLPVRLIPHK*

>2008407

MHLKFGSFPVVVGSSVEMAKTFLKTHDVVFASRPKIAAGKYTTFNYSNITWSPYGPYWRQARKMCLM
ELFSAKRLESYEYIRVEELKLLLRGLYKSSGRAVLLKDHLSDVSLNVISRMVLGKKYTDDQHDQNDVVTPQ
EFKEMLDELFLLNGVLDIGDSIPWLGFLDLQGYIKRMKVLAKKFDRFLEHVLDEHNERRKGVENYQAKD
MVDVLLQLADDPTLEVKLQRHGVKAFTQDLIAGGTESSAVTVEWAISELLKNPKIFHKAAEELDRVIGRE
RWVEEKDIVNLPYMDAIVKETMRLHPVAPLLVPRMTREDCKVAGYDIPKNTRALVSVWTIGRDPELWE
NPEEFSPERFLGKSIDVKGHDFELLPFGAGRRMCPGYSLGLKVIQAFTWKLPENMTLEDLNMEEIFGLST
PRKYPLEVVLQPRLPAHLYEM*

>2015618

MVLAFYHHWRKIYGATFLIWFGPTVRLAVSDPDLIREVFTSKAHPLVRQLEGDGLLSLKGEKWAHHRKIL
TPTFHMENLKLLVPVVAKSVTDMLDKWVAMSKSGKVEVEVSEWFQIFSEDVITRTAFGSSYEDGKAIFR
LQARQMELAAEAFQTVQIPGYRFTPTKRNIRSWKLDRDIKKSLMKLIRKRKENCGGNAAPEKGPKDLLG
LMIQASDSSSNITDYDIVEECKSFFTTSNLLTWTTVLLAMHPQWQVLARDEVLRVCGARDVPTKDDVG
KLKTLSMIVNESLRLYPPIIATIRRAKADTQLGGYKIPSGTEILIPILAVHHDQAIWGSDANEFNPSRFSEGV
ARAAKHPVGFIPFGLGVRTCIGQNLAILQTKLALAIILQRFSFRLAPTYQHAPTVLMLLYPQHGAPIIFQHL
*

>2027681

MAENTLQVRDVIRVSPQSNSVGPTTESILPLTLFDTFWLKFPPVERLFFYHVTDLTHHLFDSVILPNLKRSL
SLALLHHLPLAGKIKWPPEAPKPAVFYSPNDGVSVTVAESNADFDLLSSDDIHEALDLRPLTPELITSDDFA

DMISFQITLFPNKGFSIGVSTHHAIFDGKSSTTFIKSWAYLCKSQSDEEQPSLPPELTPSFDRTVIQDPTRA
DVSFVEKWFGLGGQYSSNPTSRSLKVFPSFGSTQNSVRATFELTRDDFKKLRDKIHEYEVKESKKLHLSTF
VLTCAYVFSCMVKARGGDAGREVIIGFTADYRARLDPQIPSNYFGNCVGTRVSFAKAGDFMEENGVAI
GAEKLSEMIKGLDNKGVVEGLEEIFGLYKNIKPGIQGLGTAGSTRFGVYGTDFGWGWPKKVEIVSIDRTG
SVSLAESR*

>2000297

MDQFSLNSSFSYVFFYLGACKNSDNLKKSLSETLTHYYPFAGRVKDGFSVDCDDYGATFVEASVAGDMS
ELIKQPKIELLEQLRPYNEEELLSANVNLAVQVNYFGCGGVAICVSFRHVIADASAAAHFVKNWSRVALG
GNGGIDDVIFDCSSIFPSQDVSGISNVKSQGMRHLLQEHVTRWFVFDNSKIAALQAKIGGRTTRFEALV
ALMWGAVITAEEEEKKDESTKTQFTALIPVNLRRKMNPPLPEQCIGNAITIGMTGWPTKEKADYNKIVF
RVRELMRMVDDYINRGFPNGWLHKRAAALDRTCKMKVFTFSSICRLPFYEADFGFGKPVWASIAEMS
DKNVFIPLDTSDGKGIQVMVWLSKEDMAKFEQDPAILAYASSS*

>2027528

MESQNRPLRAHFIPFMTTSHLMPQVDMAMIFAAHGVDVTIITTYYNSLLFENSIDKAIASGHRINVHKL
RFPFKEVGLPEGLENFSDATSGEIRMKMGNGGNMLQKPIEQFVRDNRPDFIVADMFHFWASDASDLG
IPRLTSHGMGHLALCCQDCVHRYAPHKTVQSDKEPFLLPGLPDRIQMTKSEITEWIKVPNQLTGLMHMI
DESIRKSYGVLVNSFYELEKDYVQHYRDVVGIRSWPIGSVRLYNSRNDSGNDDKHGCLSWLDSQNPNS
VVYVSFGSLCSFSSAQLSEIAKALDSLKYPFIWVVRKDGKENEWLPREFEEKITKSNKGLIIKGWAPQISILE
HPAVGGFLTHCGWNSVLETIAAGVPMATWPLGADQFYNEKLIQILRVGVRVGSETWTQLLEESRSYVE
KEKIEKAVNELMSGGEEAETRRKKVKELSAIAKAAVKEGGSSYSNVTSLLDELRQLRQ*

Chapter 6:

>DN5582c0g1i10

MWKLKTAEGANGNPYLYSTNNFVGRQTWEFDPNAGTPEERAKVEAARQKFYENRFQVKPSGDLLWR
MQFLKEKNFKQTIAPVKVEDGEEITYETASTALKRAVHFFSALQASDGHWPAENAGPLFFLPPLVMCVY
ITGHLNTVFPAEHRKEILRYIYYHQNKDGGWGLHIEGHSTMFCTVLSYVCMRILGEGPDGGQDNACTR
ARKWILDHGGATHIPSWGKTWLSIFGVCEWAASNPMPPEFWILPSFLPMHPAKMWCYCRMVYMP
MSYLYGKRFVAPITPLTLQLREELYNEPYDQINWSKARHLCAPEDIYYPHPWIQDMMWDSLYILTEPLLT
RWPFNKLVRDYALKVTMKHIHYEDENSRYITIGCVEKVLCMLACWVEDPNGAYFKKHLARIPDYLWVA
EDGMKMQSFGSQEWDTGFAIQALLASNFTDEILPVLKRGHEFIKASQVKDNPSGDFKSMHRHISKGS
WTFSDQDHGWQVSDCTAEGLKCCLLFSMMPPELVGEKMETERLYDSVNVLLSLQSKNGGLAAWEPA
GAQEWLELLNPTEFFADIVIEHEYVECTSSAIHTLVLFKKLHPGHRKKEIEIFIANAVRYLENIQMPDGSWY
GNWGVCFTYGTWFALGGLAASGKTYNNCPTVRKAVEFLLNSQRDNGGWGESYRSCPEKKYIPLEGNK
SNLVHTAWAMMGLIHSGQAERDPTPLHRAAKLIINSQLEDGDFPQEEITGVFMKNCMLHYAAYRNIYP
LWALAEYRKRVPLPSKITAA*

>DN4078c1g1i4

MAITALMDQYYLVFFFFFLLICLISSKLYQQIFGKPSSHLNLPPSPPALPIIGHLHLLTPALYNSFSNLSSKYG
PLLYLRLGATRCLLVSSASMASEVFKTHDLAFSARPMFDFADKLLYGTSGFVTAPYGDYWRFMKKLCVT
ELLGGRQLERSRTIRHQELRRFLAKVVEKASRNEVLDVGGELMRLTNNVTCRMVMSTRCSGDEGEAE
MCRKMVKESFELAAKLCFGDVLGPLKKLGFWIYGKQALNLSKRYDELLERVLKEHEERAERDGDREDED
LMDILLKVYKDDKKSEVKFTRTHIKAFFLDLFIAGTDTTAEAMQWTIAELLNHPDAFKKVREEAESVVGR

SRLVEESDIPSLPYLQAVAKEALRMYPPAPVITRECRENCVIKGFDIPEKTAVAINLYAIMRDPDALESPDE
FRPERFLIPSGDQEQKLFNFVPFGGGWRGCPGSMLASAVMNTVIAAMVQCLDWDVNGDNSDKVDM
KPGSGMSMTMVKPLLCQPVLHFNPFAA*

>DN4414c1g1i4

MDFLLSPQMLVVVALSAFVIFLSSFLFKSDNVLTRKGCKKRTAPEAGGAWPVIGHLHLLQGPEPAHRVF
ARMADKCGPIFTIKMGVHRALVVCNWEMAKECLSTNDKIFANRPKSLANEILTYNFSMFGFSSYGSYW
RHSRKLAVLELLSNYRLEKLRHVRESEMGASIKELYELCVKRNQSSSSSNSTEVSVEMKRWFGDVTLNMI
LRIIIGKRWSSPEGDEGDGWKEEMKSFFQWTGKFLVSDGLPFLRLLDIGGVQRSMKKTAKELDIALQKW
LEEHKKKIEYSGEVKGGEEQDFMDVMLSILPDFAKQFPDHDSDTIIKATCLNLILAATDTTSVTLTWALSL
LLNHRNVLKKAQDELDINVGSQRQVNESDLKNLVYLQAILKEAMRLYPAGPLSVPHESSEDCAIGGYHV
PAGTRLLVNLWKIHRDPSVWLEPGEFQPERFLTTHKDLDVRGKNFEFMPFSSGRRMCPGVSFALQVA
QLTLASLLHGFDLETPLGEPVDMREGMGLSSTKVSPLHVLLAPRLSVSVYHSAL*

>DN8439c0g1i3

MEFEASLWIQSMSTAFCFLFFSFTALFSLFSVFIFVLRLKPWCNCDVCQSYLKSSWTKDFDSLCDWYTHL
LEKSPTGTIHLHVLGNIVTANAENVEYMLKTRFDNYPKGKPFSVLLSDLLGQGIFNVDGDSWKFQRKMA
SLELGSVSIRSYAFEIVSAEIRSRLIPLLSSVAGHDQKPCDLQDVFRRFSFDNICKFSFGLDPGCLKLNQPISE
FAVNFDLASKLTAERALAPSQMIWKIKRLLNLGSEKKLKEAIRLVNELAQNLIHHRRKMGFSDKSDLLSRF
MGSIDDEKYLRDIVVSFLLAGRDTVASGLTSFFWLISQHPEVQSAIREESDRIMGPDHQDLVSFEQMRN
MNYLNAAIHESMRLYPPVQFDSKFARDDDILPDGTFVRKGTRVTYHPYAMGRMERIWGPDCLEFKPE
RWLKTGKFAPENPFKYPVFQAGLRVCLGKDMALVSMKSVALAVIRRFNIRVAEPNKAPRFVPGLTATVR
GGLPVMVEERGW*

>DN15243c1ig1i1

MEGFEFEYVNILVTVAVAGSLWLLISLYKKLVAKPERVRQMLRKQGVDGPPPTSWLLGNIMDMKKAA
QQSRQYQPTTITTHNVADAIFPFNEQCRKLYGDVFTFSLGNTQILHVSKPDVVREIVTCTSLDLGKPLYLQ
NELGPFLGRGILTTNGAVWAHQNKTMAPELYMSKVKGMINLVTESAMMIVKSWKSKFEKEGGVVDIK
IDEDMRSFSGDVISRACFGSNYSKGEEIFLKLRALQEIMANIAVLLGIPGMRYLPTKSNRRAWKLQKETR
DLILEVVKERKEKAAYEKDLLQMIIEGAESNSKFTQEESEQFIVDNCKNIYVAAFEAIGLATTWCLMLLAS
NQEWQDRVRTEILEICGGRTPDYEMLKKMKQLTMVIQETLRLYPPAAEISRETFKDMKFQDINIPKGVTI
WTMVLTLHTNPEIWGPDAYKFNPERFANGISSSCKLPHVYMPFGMGPRVCIGQNLAMVELKILVALIV
SNFSLSISPKYTHSPIRRLTIEPEYGLDLLVRKL*

>DN40570c0g1i1

MWKLKTAEGANGNPYLYSTNNFVGRQTWEFDPNAGTPEERAKVEAARQKFYENRFQVKPSGDLLWR
MQFLKEKNFKQTIAPVKVEDGEEITYETASTALKRAVHFFSALQASDGHWPAENAGPLFFLPPLVMCVY
ITGHLNTVFPAEHRKEILRYIYYHQNKDGGWGLHIEGHSTMFCTVLSYVCMRILGEGPDGGQDNACTR
ARKWILDHGGATHIPSWGKTWLSIFGVCEWAASNPMPPEFWILPSFLPMHPAKMWCYCRMVYMP
MSYLYGKRFVAPITPLTLQLREELYNEPYDQINWSKARHLCAPEDIYYPHPWIQDMMWDSLYILTEPLLT
RWPFNKLVRDYALKVTMKHIHYEDENSRYITIGCVEKVLCMLACWVEDPNGAYFKKHLARIPDYLWVA
EDGMKMQSFGSQEWDTGFAIQALLASNFTDEILPVLKRGHEFIKASQVKDNPSGDFKSMHRHISKGS
WTFSDQDHGWQVSDCTAEGLKCCLLFSMMPPELVGEKMETERLYDSVNVLLSLQSKNGGLAAWEPA
GAQEWLELLNPTEFFADIVIEHEYVECTSSAIHTLVLFKKLHPGHRKKEIEIFIANAVRYLENIQMPDGSWY
GNWGVCFTYGTWFALGGLAASGKTYNNCPTVRKAVEFLLNSQRDNGGWGESYRSCPEKKYIPLEGNK
SNLVHTAWAMMGLIHSGQAERDPTPLHRAAKLIINSQLEDGDFPQEEITGVFMKNCMLHYAAYRNIYP
LWALAEYRKRVPLPSKITAA*

>Ah4815c0g1i1

MEFFFVCGMLLFTLFVSVSLAYLFFQKPKNLPPGKMGLPYIGESLEFLRTARNGHPDKFIFDRLAKYSSKV
FKTSILTKPTAVFCGPVANKFLFSNENKIVHVWFPESVTKIFPSTGSVKEESLKMRKMLPSFLKPEALMKY
VGVMDHIAKKHFETDWEGKQQVNVFPLAKRYTFWVACKLFLSIEDPKHIAKFADPFKSLIAGIYTIPLNF
PGTPYYNSIKAAKEIRKELIGIVKQRKIDLQENRASPTQDLLSYVLQETDENGQPMSELFIADKIIGLLIGGH
DTASSTISFIIKYLAELPHVYNEVLKEQMEIAKSKKEGEFLNWDDLQKMKYSWNVACEVLRLTPPLQGAF
REAMNDFIFNGFSIPKGWKIYWSTYSTHKDAEYFPEPEKFDPTRFEGKGPAPYTYVPFGGGPRMCPGKE
YARLEILVFMHNVIKRFNWKQLLVDEKTVAEPIPTPAKGLPVRLIPHKA*

>Ah331c0g1i3

MWRLKIAEGGNGPYIYSTNNYVGRQIWEYDPEAGSPEEVAEVEEAQQKFYKNRHQVKPSSDLLWRFQ
FLREKNFKQSIPQVKVKEEEEITYETATAALRRAAHFFSALQANDGHWPAENSGCLYFLPPFLMSLYITG
HLNSVITSEHRKEILRYLYNHQHEDGGWGQHIEDHSSMFGTVFSYVAMRLLGLGPDDGQNNACARAR
KWILDHGGVTYIPSWGKNWLSILGVFEWSGVNPMPPEFWMLPSYLPFHPAKTFCYFRLVYIPLSYFYGK
RFVAPITPLIQQLREELHTQPYNQINWRKVRHLCAKEDLYYPHPLIQDVIWDSLYIATEPLLTRWPFNKLI
REKALRATMDCIHYEDENSRYITIGNVEKPLCMLACWVEDPNGVAFKKHLARIPDYFWLGEDGMKVQS
FGSQLWDTALGLQALLASNLTEEIGPTLAKGHDYLKKSQIKDNPSGDFERMFRHISKGAWGFSDQDHG
WQLSDCTAESLNCCLQFSMLPPEIVGEKMEPEGLYEAVNVIFSLQSENGGLSVWEPACTSPIFEWLNPV
EFLEGVVVDHTYVECTASAIQAFVLFKKLHPNHRNKEIEDFIRKAVKYIEDEQTGDGSWYGNWGICFLYG
TTFALAGLAAAGKTFKNCLAVRKAVEFLLNSQTKDGGWGESFISCPNKIYTPLEGKRSNLVQTAFGLIGLI
HSGQAERDPTPLHRAAKLLINSQLDNGDFPQQEIIGVFMKNSMLHYAQYRNIFPLWALAEYRSKVPLPT
YKNG*

>Ah331c0g2i1

MWKLKTAEGANGNPYLYSTNNFVGRQTWEFDPNAGTPEERAEVEAARQKFYENRFQVKPSGDLLWR
MQFLKEKNFKQTIAPVKVEDGEEITYETASTALKRAVHFFSALQASDGHWPAENAGPLFFLPPLVMCVY
ITGHLNTVFPAEHRKEILRYIYYHQNKDGGWGLHIEGHSTMFCTVLSYVCMRILGEGPDGGQDNACTR
ARKWILDHGGATHIPSWGKTWLSIFGVCEWAASNPMPPEFWILPSFLPMHPAKMWCYCRMVYMP
MSYLYGKRFVAPITPLTLQLREELYNEPYDQINWSKARHLCAPEDIYYPHPWIQDMMWDSLYILTEPLLT
RWPFNNLVRDYALKVTMKHIHYEDENSRYITIGCVEKVLCMLACWVEDPNGAYFKKHLARIPDYLWVA
EDGMKMQSFGSQEWDTGFAIQALLASNFTDEILPVLKRGHEFIKASQVKDNPSGDFKSMHRHISKGS
WTFSDQDHGWQVSDCTAEGLKCCLLFSMMPPELVGEKMETERLYDSVNVLLSLQSKNGGLAAWEPA
GAQEWLELLNPTEFFADIVIEHEYVECTSSAIHTLVLFKKLYPGHRKKEIEIFIANAVRYLQNIQMPDGSWY
GNWGVCFTYGTWFALGGLAAAGKTYENCPTVRRAVEFLLNSQRDNGGWGESYRSCPEKKYIPLEGNK
SNLVHTAWAMMGLIHSGQAERDPTALHRAAKLIINSQFEDGDFPQEEITGVFMKNCMLHYAAYRNIY
PLWALAEYRKRVPLPSKTTAA*

>Ah4924c0g1i7

MAITALMDQYYLVVFFFFLLICLISSKLYQQIFGKPSSHLNLPPSPPALPIIGHLHLLTPALYNSFSNLSSKYG
PLLYLRLGAARCLLVSSASMASEVFKTHDLAFSSRPMFDFADKLLYGTSGFITAPYGDYWRFMKKLCITEL
LGGRQLERSRTIRHQELRRFLAKVVEKASRNEVLDVGGELMRLTNNVTCRMVMSTRCSGDDGEAERCR
EMVKESFELAAKMCFGDVLGPLKKLGFWIYGKQALNLSKRYDELLERVLKEHEERAERDGDREDEDLM
DILLKVYKDDKKSEVKFTRTHIKAFFLDLFIAGTDTTAEAMQWTIAELLNHPDAFKKVREEVESVVGRRSR
LVEESDIPSLPYLQAVAKEALRLYPPAPVTTRECRENCVIKGFDIPEKTAVAINLYAIMRDPDAWENPDEF
RPERFLIPSGDQEQKLFNFVPFGGGRRGCPGSMLASAVMNTVIAAMVQCFDWDVNGDKVDMKPGS
GMSMTMVKPLLCQPVLHFNPFAA*

>Ah13762c0g1i2

MELFIFCALILFTLLICVSLAYLLRSNLPESCLKNLPPGKLGLPYIGETLEYLKTARRGHPEKFIFDRTTKYSSQL
FKTSILAAPTVFVCGAVANKFLFSNDNKLVDIWFPDSLKKIFPATQESSPDRESKKQKMMLSNFLKTESLI
RYISIMDSIAQKHFESDWEGKQQVTVFPLAKKFTFWVACKVFLSIEDQHHIEKLSYTFSALAAGLFSIPINF
PGTTFNSAIKAAKEIRKELIAIIKQRKIDLEENKASPTQDLLSHLLLETDENGHHMTESYIAERILGLLVAGHE
TLTSAITSIVKYLSDLPNIYNQVLEEQLKIVKSKKSGESLNWDDIKRMKYSWNVACEAMRVSPPVLGAYR
EAINDFIFLGFSIPKGWKLYWSMYSTHLNPECFPEPEKFDPTRFEGNGPAPYTFVPFGGGARMCPGQEY
ARLEILVFMHNVIKRFKWEKLLAQEKTIMEPGPIPAKGLPVRLIPHQNQITGERSS*

>Ah16676c0g1i1

MEGFGFEYVNILVTVALAGSLWLFISLYKKLVAKPERVRQMLRKQGVDGPPPTSWLLGNIMDMKKAA
QQTCRYQPPTITTHNVADAIFPFNEQCRKLYGDVFTFSLGSAQILHVSKPDVVREIVTCTSLDLGKPLYLQ
NELGPFLGRGILTTNGAVWAHQNKTMAPELYMSKVKGMINLITESAMTIVKSWKSKFEKEGGVADIKI
DEDMGSFSGDVISRACFGSNYSKGEEIFLKLRALQEIMANIAVLLGIPGMRYLPTKSNRKAWKLQKEIRD
LILEVVKERKEKAACEKDLLQMIIEGAESNSKFTQEESEQFIVDNCKNIYVAAFEAIGLATTWCLMLLASN
QEWQDRVRTEVLEICRGRTPDYEMLKKMKQLTMVIQETLRLYPPAAEISRETFKDMKFQDINIPKGVTI
WTMVLTLHTNPEIWGPDAYKFNPERFANGISSSCKLPHVYMPFGMGPRVCIGQNLAMVELKILVALIV
SNFSLSISPKYTHSPIRRLTIEPEYGLDLLVRKL*

>Ah21534c0g2i1

MAPLFLVLVAFVVIFVSHTLYSLFFNEKSHVDGKVPPGGKGWPIIGETLEYLSLGKKNDIEKFFYDRKNKY
NARHFFKTSYIGENMLFACTTEANKFVFVTANKFVRSWWPKSMEKVLPSTEQTSTVEESSRSRRIAVVF
NEPIVLQKYVGMFDEVAKAHIRKNWKNQQVVKVYPLAKKFSFALACRTILNVSDTKLVEEFEKLFICLSA
GFFSLPIDLPGTKLNRAIKAAKELRRRFELIIRESKRRNKSEDKDSKSEEEARDLLDFLLQEKYANGEYMND
SDVANMTLGILIAGYEAASTTIVSIMKFMAELPDVYQRVRQEQMEILKSKPEGDEFLNMSDIRNMKYS
WNVASEALRLLPPVAGTFREAIRDLNYDGFMIPRGWKMHWNPHATHKNPEYFPDPYKFDPSRFEGN
GPPLFSYVPFGGGPHICPGREFAKIQMLVFMHNVVTNFKWEKVFPEEQIIVDPVLVPTKGLPIRLQSHHE
QCT*

>Ah21534c0g3i1

MAPILLVLLAFVVIFVSHALYSLFFNEKSHDDGQVPPGGRGWPFIGETLEYLSLGKKNNVEKFIYDRKNKY
NARHFFKTSYIGENMLFACTTEANKFVFVTANKFVRSWWPKSMEKVLPSTDKTSTVEESSRSRRIAVVF
NEPIVLQKYVGMFDEVAKAHIQKNWNGQQVVKVYPLAKKFSFALACRTILNVSDTQLVEEFEKLFICLSA
GFFSLPINLPWTKLNRAIKAAKELRRRFEFIIRESKRRNNIEGLDSKSDEARDLLNFLLHEKYSNGDYMNDS
DVANMTLGILIAGYEAASTTIVSIMKFLAELPDVYERVRQEQMEILNSKSAGELLNMNDIRNMKYSWNV
ASEVLRLLPPVAGTFREAIRDLSYDGFTIPKGWKMHWNPHATHKNPEYFPDPYKFDPSRFDGNGPPLFS
YVPFGGGPHICPGREFAKIQMLVFMHNVVTNFKWEKVFPEEQIIVDPVLVPTKGLPVRLQPHQQST*

Chapter 7:

>Ah4415c0g1i6

MEVSVVSQKLIKPSVPTPPHLRNMCISFMDHRLLPRTVPFIFYYTGVHGNYHRLEKSLSEILTLYYPLAGRY
ENGNLSINCNDEGVEYAEALVNGELTRILRGQFEVEELNRFVPDDRRMLEESATSPLLAVQINSFNCGGP
AIGLRISHRICDGFTLSLFTNGWAKASKAGNINNDVTVPNFGSGVLFPPRYTAAMANRAASVATAASGI
VTRRFVFEHGISTAVVIALIWRAQINAERRARYGFLRTSSLLAVAINLRGKTFRKIPENSSGNIFTWGTAKF
EADETKMGLSHFVDKVKDAIRGSVAEFGTPKEEEEDGFFAKLIARPAMEMAEQVGKGEGDVRRFSSSR

NFPFYDVDFGWGKPAWVSLAKRPYKGVILMATKGGGDGVEAWVTMDKEEMAYFQDDPDIRKCQRI
*

>Ah4755c0g1i8

MAGIRVISTSTVCAASHKEPNLRIDLTPWDLQLLLLDTIQKGLVFHKPKPDPNVTLIHHLQAALSRTLDFF
YPLAGRLATIENGDETSFFLDCNNAGAEFVHAVYDGVSVSDVIEHTCVPDHIVYSFFPLNGIRNHEGTSK
PLLAVQVTELVDGIFIGCTINHTALDGTSFWHFFNSWSEISRGLDSLSKPPVLERRFLHDSDYPIRIPFFKN
EQIHNSFVPPPLQQRVFHFTAESIAKLKAKANAEICTDKISSLQALLSHFWRSAVRNKRLDPDEETSFLILI
GVRQRLQPQLPQQYLGNAVQADFIKMKAGELLEQGLGYVALQMNTMIAMHTEEESRKFLESWRENP
KIYKRGAFVSNGLGASSSPRHNVYGTDFGWGRPIAVRSGPGNKSDGKITIFPGAKQGSMDIEACLSPDT
LQAMANDAEFMDTVLST*

>Ah5291c0g2i3

MAPPPTSSLVFTVRRCDPELIVPAKPTPRELKPLSDIDDQEGLRFQIPVIQFYKHDPNSMKGKDPVKVIRQ
GIAETLVFYYPFAGRLREGADRKLMVDCTGEGVMFIEADADVTLDQFGDALQPPFPCWEELLFDVPDS
AGVLNCPLLLIQVTRLKCGGFIFALRLNHTMSDAAGLVQFMAAVGEMSRGARAPSIPPVWERQVLNAR
DPPRVTCTHHEYDEVADTKGTIIPLDDMAHQSFFFGPTEIAAIRRFVPLHLKRCSTFEILTACLWRCRTIAL
QPDPTEEMRVICIVNARAKFNPPLPTGYYGNEFAFPVALTTAEKLSKNPLGYALELVRKAKAEVTEEYMK
SVAAMMVLKGRPHFTVVRTYLVSDVTHAGFGDVDFGWGKAVYGGPAKGGVGAIPGVASFYIPFKNN
KGEDGIVVPVCLPAPAMVRFVKELDAMLKDEMVGESKSSFITSAL*

>Ah9064c0g1i7

MAENTIQVRDVIRVSPQSNSVGSTTESILPLTLFDTFWLKFPPVERLFFYQLTDLTHHLFDSVILPNLKRSLS
LALLHHLPLAGKIKWPPAAPKPAVFYSPNDDVSVTVAESNADFDRLSADDIHEALDLRPLTPELITSDEFA
DMISFQITLFPNKGFSIGVSTHHAIFDGKSSTSFMKSWAYLCKSQSDEEQPSLPPELTQSFDRTVIKDPTG
ADVSFVEKWFGLGGQFSSNPNSRSLKVFPSFGTTRNSVRATFELTRDDFKKLRDKIHEYEVKESKKLHLST
FVLTCAYVFSCMVKARGGDADREVIIGFTADYRARLDPPIPSNYFGNCVGTRISFAKAGDFMEENGVAI
GAEKLSEMIKGLDNKGVVEGLEEIFGLYKNIKPGTQGLSTAGSTRFGVYGTDFGWGGPKKVDMVSIDRT
GAVSLAESRRGDDGVEVGLVLEKHEMEAFASFFNQGLKFI*

>Ah10648c0g1i2

MDSKTVKHISEFFVKPQYVSAEMKQPLYLTPWDLAMLSVHHIQKGLLFIKPEVDDEEDFINTLLDRLKQS
LSLTLVEFYPLAGRFKTVYSENPHSYVVFVDLVDSPGARLIRASLDMTVSEIVSPIYVPLVVQSLFDHDRAI
NHDGHTRPLLSIQVTELVDGVFIGCSMNHALGDGSSYWNFFQAYSETFQAQLGSSTSNSISRPPPIVKR
WFPDGHDQIINLPFTHQDEFLSRHEAPHLLERYFHFTAESIAKLKARANSESKTSKISSFQSLSALIWRSITR
ARRFPNDQTTSCRLATNNRLRLEPPLPQEYFGNSIQTVRGVATAGELVEHDLGWAAWKLHEAVVNHN
NKTVREWVDSWRQSPFIYQLGQFFDPKSVMIGSSPRFNKYGIEFGLGKAVTLRSGYANKSDGAVYCYP
GHEGGGSIDLEVCLPPESMNALESDEEFMSAASL*

>Ah16012c0g1i1

MFKSEELPDCCYHPNEPSLISPNAPTPNHSLYLSNLDDQKFLRFSIKYLYLFKKSVSQDFLKYSLSKLLVDYY
PLAGRLRTSKDDDQKLEVDCNGEGAVFASAFMDISAEEFLELSMKPNRSWRKLLYKVEARSFLDVPPLV
VQVTNLRCGGMIVCTAINHCLCDGIGTSQFLQAWAHMTAKPNLDLPIAPFHTRHVLKPRNPPQVNYT
HHVYTTPAPTTKTRNSSKDSAAAHVVVDINQYLQSQQLVPASLIFTPSHILHLKRQCSPSLKCTTFEALAS
HTWRSWVRSLDLPPSLNVKLLFSVNVRKNLNPEIPQGYYGNGFVLGCAETSARDLLSADNLYSGIKLVQ
QAKSSLTDDYVRSMIDLLEDKTVKTDLSASLVISQWTKLGLDDLDFGEGKALHMGPVTSDIYCLFLPVVG
SFDSVRVLVSLPESVVERFEYYMMKEFLDSE*

>Ah17020c0g1i3

MEKSSIISREIIKPSSPTPTHLRIHNLSRLDQINLDIVYPVILFYSNAPKNSDRLKKSLSEILTHYYSFAGRIKDH
FSVECDDYGVSFVEAHVVSEALRNPECYGLEELMPYKSDELLMSRFNLAVQVSYFGCGGVAICVCFRHA
VADATAVANFLKRWSAVAACDRQDGSLKDVVLDFTSIFPPEEWPGISTKLLKSSTDSTIRKRFVFEASKM
AALREEIGNQPTRFEAVFALLWKAVTSAKGEGEVFVASIPVSLRRKMNPRIPDECIGNFFTAVMANWKL
EETIIYKDIVEKLRETLSLVNADYVKEAFPNGWSLLRMENDVGRNKRRVFHISSCCGLPFYEADFGWGKP
SWVDTRPRNNAGDNIFVTLFDTCEGKGVEAWVAMSKEEMARFEQDSHVIAYASFNPTT*

>Ah25465c0g1i1

MDVSIISTETIKPSSSTPEDLRTYKLNRLDQTIINGHIPIIFFYSAAPKSCDYLKNSLSQLLTHYYPLAGRVKDF
LSIDCNDSGVSFVEARVAANMSDVLKQPEIDPQEQLLPFKTREMMPTRPILAIQVSYFDCGSVAICLLFR
HVVVDATAAANFIKNWAATARGGGDGGKDISGIKDVIFDSTSIFPPKDQSVIPESTGTQGVFYSLNENV
TRRFVLEGSKISSLQEKIGNRPTRFEAVFALMWGALTAAKREGDGFIATIPVNLRKRMNPPLPEQCIGNI
NTIIQANWPPEETARYGDLTAKVHELVGMVNSEYVKKASPNGWLMYRVRRDLIEKRKRLFMISSWCTL
PLYGTDFGWGKPTVVLPIGRLNSGDNIVVILLDTGIGKGIEAWVSMPKEEMDKFERDSEIIAYAS*

>Ah27763c0g1i1

MDVSIISTETIKPSSSTPEDLRTYKLNRLDQTIINGHIPIIFFYSAAPKSCDYLKNSLSQLLTHYYPLAGRVKDF
LSIDCNDSGVSFVEARVAANMSDVLKQPEIDPQEQLLPFKTREMMPTRPILAIQVSYFDCGSVAICLLFR
HVVVDATAAANFIKNWAATARGGGDGGKDISGIKDVIFDSTSIFPPKDQSVIPESTGTQGVFYSLNENV
TRRFVLEGSKISSLQEKIGNRPTRFEAVFALMWGALTAAKREGDGFIATIPVNLRKRMNPPLPEQCIGNI
NTIIQANWPPEETARYGDLTAKVHELVGMVNSEYVKKASPNGWLMYRVRRDLIEKRKRLFMISSWCTL
PLYGTDFGWGKPTVVLPIGRLNSGDNIVVILLDTGIGKGIEAWVSMPKEEMDKFERDSEIIAYAS*

>Ah28436c1g1i1

MASSSVHVKEAVVIVPAEPTPSRVLALSVLDSQLFLRFTIEYLLVYRPRPGSDPCATAARIKNALARALVPY
YPLAGRVRARPEGSGGGCLEVVCRAQGALFIEAVSDHTTTISEFERAPRQVTQWRKLLSFHVEDVLMGA
PPLVVQLTWLRDGSATLGIGFNHCVCDGIGSAEFLNLFAQLSTRQLDEFKPNPVWDRHLLNPSLVKRPRI
NSVSHHPEFTRVPDLCGFVSRFNNERLAPTSVTFNKRCLNELKKLALSTSRPSESTYTSFEVLAAHVWSS
WARSMNLPSNQILKLLFSINVRARVKPSLPRWFYGNAFVLGCAQTTVKELTEKGLGYATMLVKRAKDR
VDGEFVRSVIESVSESLACPDSVGVLILTQWSRLGLERVDFGMGRPVHVGPICCDRYCLFLPVFNQTDAV
KVMVAVPTSAVDKYEHLIRSVC*

>Ah46030c0g1i1

MVLGFEASAAAAAATMKSCNGNVFQLSVKQLGEPTLVAPAEETEKGLYFLSNLDQNIAVIVRTIYCFKSG
EKGNEKSAEVIKDALKKVLVHYYPLAGRLTISAEGKLIVNCNGEGAVFAEAEANCPLEEIGDITKPDPETLG
KLVYDIPGAKNILEIPPLVAQVTKFKCGGFVLGVCMNHCMFDGIGAMEFVNSWGETARGLPLNVPPFL
DRSILKARNPPKAEYPHQEFAEIEDKSGTSDLYKDEMLYRSFCFDPEKIESIKSKAMEDGALEKCTTFEALS
AFVWRARTKALKFLPDQPTKLLFAVDGRPKFKPPLPKGYFGNGIVLTNSICQAGELMEKPLSFAVGQVQ
EAIKMVDDDYMKSAIDYFEVTRLRPSLASTLLITTWSRLSFHTTDFGWGEPVLSGPVALPEKEVILFLSHG
KERRSINVLLGLPASSMKIFQELVQI*

>Ah77698c0g1i1

MEVKILETKLVAPSPPPFNGDHVLSLSHLDNDPNLRIIFRYLRVYVGTATDKRPFDVIAAALSSALVPYYPL
AGTLCPDEDRRFSLSCKEGQGVPLIHAAVDCTLDSLNYLDEPDTVFVERLVPDPDPKEILANPCVLQVTA
FKCGGFTLGAALHNSLCDGMGATQFFNLMAEVARGGTHQSAQPVWNRAGLLGPRDPPRVDGAIKEF
LSLEKGYEPYVQGFGPVERESFDVKDECLDQLKKVLLENCGLNFTTFEALGAFLWRAKVRSSGISGGENV

KFAYLTNIRKAVKPPLPFGYWGNGCVPLYAQINAEELVEQPLWKTAELIKKSKVNANDEYVRSYIDFQEL
HHGDGITAGKEVSGFTDWRHIGHSGVDFGWGGPVTVFPVTRQILGSFEPCFFLPYSAASGGKKDGFKI
QVTLRKTAMPAFKEEMKKFSNKEF*

>Ah95885c0g1i1

MNSSCLHSSGVLYTQVPFQHHFLVQILTKKLIKPATPTPHNLRNLKISSLDQLNHSWHLPFLLFYPLEIGCE
DQRDKNTERSKLLQKSLSQVLTVFYPLAGRYTGDDIFVRCNDHGVEYVEARVCGQLAPFLRRDVDSKQL
LPLVPNQIESPASPLVAVQFSTFECGGAAIGLSISHKIADGFVMARFIDAWAVACSRGGWSDRVMMNP
VSFDLAMVLPTRETIPELKPVVKLIPGMKTRSRTFLFDSLVISNLREVLVKELQWRRRRPSRVQLVTALIW
KALIRMNQTKHGKLVSFVISNTMNLRGRTALPISENSCGNLYRPVIVRFIPENKETTAEFPALVTLLGDAIT
TAVDDCAKPQNGDDLFSMVTDTMREAGEEFNREEVVAYGFTSMCRMPFYGDFGWGKPNRVTHVQS
AAELIMLFDSRDGDGIEAWISLDENDMQLFLQDPDIVAITSRT*

>Ah6798c0g2i1

MGDLLHSRTVHRYSAAVNRLYFLFHFTAILCLFYYRFTNLLFHNVTMIPWSLITVAELVLTFIWLTTQSFR
WLPVTRSIFLENIPADDATLPGVDVFVCTADPNKEPTLEVMNTVISALALDYPTDKMAVYLSDDGGSKIT
FYATKEASKFSKYWIPFCTKYGIKTRNPGAYFTPFLAKYETLPTHDQFKADEAEIKSKYENFKSNVEKFGEK
DENNEVAVERPPCVEIIHENKNDEKEGKMPLLVYVSRERRGSYPHRFKAGALNALLRVSGIMSNGPYILV
LDCDMYCNDPTSARQAMCFHLDPLISPTLAFVQFPQMFYNVNDNDIYEGGGRSIYKSIFRGMDGLRGP
LLSGTGYYLKRIALFTTPHQQDKHLLHPEKNFGNSRKFIDSLKSKNDNDFAIRKEEYSSDSILEEAKQLASC
AYELNTNWGKEMGYLYESMLESTFTGYMLHCKGWISVYLYPKRSCFLGCTNIDMKDSMTQIMKWCSG
LIDVGLSRFSPVIYGFSRMSILQSMCYGYFTFIPLFSVALLLYGTVPQLCLLTGTPLYPKVSNIWFSVFAIVFL
SSLLNQLYDILSTGETLRTLWNEQRFWMIQSVSGNLFGCIDMVMKRIGAKKGKFTISNKAVVKEKLDKY
EKGKYDFEGANMYMVPMTIMVILNIICFCGGIIRLMLEKNMEDMFGQLFLSSYILVLSYPIIEGLISSKSKT
N*

>Ah3338c1g3i1

MSTSSSSGLDPQSQPHVALIPSAGMGHLIPFLRLSASLVHHHCRVTLVTTHPTVSLAESRFISSFLSAFPQ
VTEKQFHLLPLDPTLSNNDPFYLQWEVIRRSAHLLSPLLSSVSPLSAVVSDVTLGSSVTPITANLGLPNYILF
TASARMCSLFGSFPDFVAPKTSSGSVQLGGDLEIPGLPPIPISSIPPALLNSNSLFATIFLQNAQEVVKSEG
VLINTFEALEGDTLHALNEKRVIDGFPPLYAIGPFLPCEFEKRDGGAPLKWLDDQPEGSVVYVSFGSRTAL
SREQMKELGNGLVSSGCRFLWVVKDKKVDKDDEESLEKLLGDELTEKIQDRGLVIKNWVDQDEILSHKA
VGGFVAHGGWNSLAEATWHGVRLLVWPQFGDQRMNAGAVERSGVALWVKSWGWGEQAVVKGD
EIGETIREFMGNDLLKSQAARIREEARKSVGAGGSSERTLKRLIHEWKKTNT*

>Ah7803c0g2i1

MDSGSSQVHMLFFPYMAQGHMIPTVDMARLFARHGAKATVVVTPCNVPLFSSTIRRERELGFDISLRT
VEFPCAEAGLPEACENVNSITSRVMEMNFYKAIGLLQQPFEQILEELRPNCLVADVMIAWATEAAAKFEI
PRLCFHGTGFFALCVYHSLKEHQPHKAVASDSEPFVVPGLPDQVKLTRQQLATSMKAEPDKELVKLVAQ
GIQAEMTSYGVLVNSFYDLEPGYVEHYRKVLGRKAWHIGPVSLCRDRSSIDEHECLRWLESKQPSSVIYV
CFGSMVHFSAADQLIEIAMGLEASGQNFIWVIKKEDEEWLPEGFEKRVEGRGLIIRGWAPQVVILEQEA
IGGFLTHCGWNSVLESVSAGVPMATWPLSAEQFYNEKLVTDVIRIGVGVGAQEWCHWMEDKRGSVV
RREDVRNAVSELMAGEDAEEMRNRAMALREMAKRANEEGGSSYSDINALLEELKLNSH*

>Ah19168c0g1i1

MESIVIYPAPDRGHLISTVELAKLLLSQNPSLTIHIIVIKPPIKLASTAPYIATVAAAIPSIKFHHLPPVSMDPN
SYPNYFAIMFDHYRLVNHHVRVLLRSLSRDFTVRCLIIDFFCFPALEIASELNLPAYYFFTSGACCLALLFQL

PTLYETTSQNFNDLDTTFLKLPDFPPIPLSHMPEPALFRNNPECLYFMDCLARLKKSAGIIVNTFESLEPKA
VKLLSSSIDGFCIPDSPVPPVYCIGPLTTATDQTGDNVHGCMAWLDSQPSRSVVFLCFGSMGLLSKQQL
REIAYGLERSEQRFLWVVRNPPSDNQSKDPMNQPDPDLDSILPAGFLDRTRDCGLVVKSWAPQVAVL
GHVSVGGFVTHCGWNSVLESVNAGVPMVAWPLYAEQKINKAFLVEELKLALPMNSESETGGLVSAAE
VEKRVRQLMDSEEGNSVRERVAAKRDEAIAASSHGGSSRVAVAKLAELWMRN*

>Ah5240c0g1i1

MESKQEKAKSIVMLPYLAHGHVSPFYELAKKLSTRGFHIYLCSTPICLQPIRNITHFDDDHHHHNIHLIDLH
LPSSFHPELPPHYHSTKNLPPHLTNTLLAAFDASKPAFSDLLRTLRPDLVIYDLFQPWAPEAAGEQNIEAV
LFLISAASQNLTFEIPAVTESNGQITDYLVKNVVSGMTYTQRFYRSVTSSSKMILVKTLREIEDRCRNDQR
NPYNFVPVGPLVQKRSQDADNGDADEKMIMDWLSKREPRSVVFVSFGSENFLSEQEIQEMGHGLELS
GVDFIWVVRFHGGNEVKTIHEALPSGFLERTKERGLIVGWAPQVKILHHESTGAFVSHCGWSSTLEAM
VYGVPIIGLPFKLDQPLNAKLAAEMGVGVEVPRRENKITRQEVGKVIKTVVVEEEGKGVRLRTKELSERLK
KKEDEEIDVAVEKLVQIVEESKKTHVL*

Appendix C – Nucleotide sequences of synthetic genes ordered from IDT

>Ah1968c0g1i4

GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAGCTCATCAATACTCCCTATTGGCCTTCATA
CTGGGCCGCATGGCTAGCAACCCTGGCCCTACTCCTCCTCGCCCGCCGCCTCCGCCGCCCCAAACTC
AACTTACCACCGGGTCCAAAACCGTGGCCCATCATCGGAAACCTCAACCTCATCGGCAAGCTTCCCC
ACCGGTCCATCCACGCCCTATCCCAAAAATACGGGCCTATCATGCAACTCAAATTCGGGTCCTTCCC
CGTGGTGGTGGGCTCCTCCGTCGAAATGGCCAAGACCATCCTCAAAACCAATGACGTCATCTTCGC
CACCCGGCCCAAAACTGCCGCCGGGAAGTACACAACCTACAACTACTCCGACATAACCTGGTCGCC
CTACGGCCCCTACTGGCGGCAGGCCCGAAAAATGTGCTTGATGGAGCTGTTCAGCGCCAAACGCCT
CGAATCATACGAGTACATAAGAGTGGAAGAGCTGAAACTCCTCCTCCGCGGCCTGTACAGGTCATC
GGAAAAGCCTGTGATGTTGAAAGACCATCTCTCCGACGTGAGCTTGAACGTGATAAGCCGGATGG
TTCTGGGGAAGAAATACACGGAAGAAGAAGACGGCGGAGGCGGCGAAGATCGAAACGAGGTGG
TGACTGCTCAGGAGTTCAAGGAGATGCTTGATGAGTTGTTTCTGCTAAATGGGGTTTTGGACATTG
GAGACTCGATTCCATGGCTAGGGTTTCTGGATTTGCAGGGCTACATAAAGAGGATGAAGGTTTTGG
CGAAGAAGTTCGACAGGTTTCTGGAACATGTTCTGGATGAACACAATGAGAGGAGAAGAGGGGTT
GAGAACTATCAGGCCAAAGACATGGTGGATGTGCTTTTGCAGTTAGCTGATGATCCTACTCTTGAG
GTCAAGCTTCAAAGACATGGTGTCAAGGCTTTTACACAGGACTTGATAGCTGGAGGGACAGAGAG
CTCCGCCGTGACAGTGGAATGGGCCATTTCCGAGCTCCTGAAAAAGCCGAAAATCTTCCACAAGGC
AGCAGAAGAACTGGACAGAGTGATCAGCAGGGAGAGATGGGTTGAAGAGAAGGACATTGTCAAC
CTCCCTTACGTGGACGCCATCGTCAAAGAGACCATGCGTTTGCACCCGGTCGCACCCTTGCTCGTGC
CGCGGATGACCCGGGAGGACTGCAAGGTCGCAGGCTACGACATCCCGAAGAACACTCGTGCCCTG
GTCAGTGTCTGGACGATCGGACGAGACCCCGAGTTGTGGGAAAAACCCGAGGAGTTTTCGCCCGA
GAGATTCCTAGGTAAATCCATCGACGTGAAGGGACACGACTTCGAGTTGCTACCGTTCGGCGCTGG
TAGGAGGATGTGTCCAGGGTACAGTCTAGGGTTGAAGGTGATCCAGGCGAGCTTGGCAAATCTTC
TGCATGGTTTTACATGGAAATTGCCGGAAAATATGACGGTGGAGGATCTGAATATGGAGGAGATT
TTCGGGCTTTCTACTCCTAGAAAGTACCCACTTGAGGTTGTGCTTCAGCCTAGGCTTCCTGCTCATCT
TTATGAGATGTAATACCCAGCTTTCTTGTACAAAGTGGTCCCC

>Ah3545c0g2i4

GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGATCTTTACAGTCTGCCCATGTGGCTCCTTCTT
CTTCTGGTCCCTCTGCTGTTTCTCCTGAAAAAGATCATGGGAGTTAAAGGGCAGAAATTGCATCTCC
CACCAAGCCCTCCTCAGCTTCCATTTCTGGGAAATTTACGCCAACTTGGAGCTCTGCCTCACCAGTCT
TTCACGAAGCTCTCTAAAAAATACGGTCCCGTCATACTCTTCAAATACGGCCGTATACCTTTTGTTAT
AATCAATTCTGCTGAGGCAGCTAGAGAGGTCTTGAAGGTTCATGATATCGACTCTTGCAGCAGAGC
ACAACTGACTGGGACTCGGAAGCTCACGTACAATTTCGTAGACGTAGCTTTTGCACCGTATGGTGA
TTACTGGAGGCAGATAAGGAAGATATGCGTTCTCGAGCTTTTCAGCTTGAAGAGGGCGCAGTCATT
TCGGTTCATTAGGGAAGAAGAAGTTGCCTTGCTCGTGAAATCCATCTCTGAATCAGCATCTTCTGCA
ACTCCCGTTGATCTGTCTGAGAAGATATATGCTCTCACCGGAAGCATAGTGTTCAGGATGGCTTTCG
GGCAGATATTCCGAGGCAGCGGTTTGGATAACGAGAAGTTTGAGAAAATAGTTCGTGAAACAGAG
TCTGTTGTGGGAGGATTCACTTCAGAGGAGTGCATTCCTTATGTTGGATGGATAATCGACAGGCTT
AATGGTTACCATGCGAAGCTCGAGAAAGTTTGCCATGATCTGGATCGTTTCTTCCAGAAAGCAATC
GATGATCATCTTAAACCCGAGAGGCCATTACAAGACCAGGAAGACATTATTGATGTTATGCTCAAA
ATAGTGAAGGAGCCAACTAAACCTTGCGAAACTCAGATCACTCATGATAACATTAAGGCAGTACTC
CTGAACATATTTTTGGGTGGAGTAGACACCAGCTCCATTACAGTAATCTGGGCAATGGCAGAGCTT
GCTAGGAATCCGAGACTAATGAAAAAGGCACAAGATGGAGTCAGAAACTGCATTGGAGAGAGAG
GAAAAATTACAGAAGATGATCTCGATCAACTTCCATTCGTGAAGATGATTATCAAAGAAACTCTGA
GATTGCATCCTCCTGCTCCTCTGCTACTCGCTAGAGAAACTATTTCCCACTTCAAAGTTGATGATTAC

GACATAACCCCCAAAACCATGATTCAAGTCAATGTTTGGGCAATCGGAAGAGATCCTAAATACTGG
GAGAATCCCGAAGAATATTACCCAGAAAGGTTTATGGATAGCAGTATTGATTTCAAAGGACAGAAT
TTCGAGTTTTTGCCGTTTGGGTCAGGTAGAAGAGTTTGTCCTGGAATTCATCTGGGAATAAGAACAT
CGGAGTTTGCACTTGCAAATCTCCTGTACTGTTTCAATTGGAAATTACCCAGTGGAATGAAAGAGG
AAGACATCAACATGGAGGAGGGTGTTGGTGTGAGTCTCAACCTCTCTAAAAAGACCCCTCTCTACC
TTGTTCCAGTCAGTTATCTCCATTAATACCCAGCTTTCTTGTACAAAGTGGTCCCC

>Ah15779c0g1i1

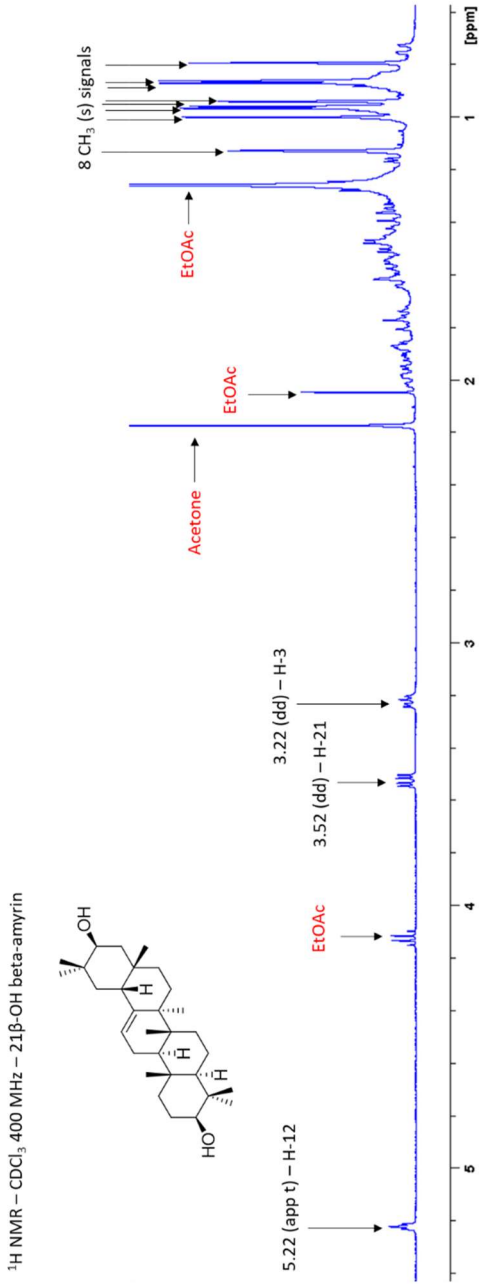
GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAATTTCCTTCTTCAATCTCTAAACCCCACCGCA
GTTGGAGTCGTCATTGCCATACTTGCTTTTTCCTATTATTTACTAAGGAGGTCCAACTCCAAATCCAC
CAAGGGGCCAAAACCACCAGAAGCCGCCGGTAGATGGCCAGTAATCGGCCATCTCCGCCTCTTAAC
TGGAACCGGAACCCGACTCCCTCACTTTGTCTTGGGAGAGTTAGCCGACAAGCATGGACCAGTCTT
CTCCATCCGGATCGGTGTGCACCCAGCTTTGGTGGTCAGCAGCTGGGAGGTGGCCAAGGAATTGTT
CACCAAATATGATGTGGATGTAGCATCTCGCCCCAAGCTCACGGTTGGGAAAATATTAGGATACAA
CTACGCCAACTTCGGCTTCGCTCCCTACAATAACTACTGGCGTGAGATGCGCAAGATTACTGCTTCA
GATTTACTTTCGAACCGCAGGCTTGAGTTACTCAAACATATTCGAGCCAATGAAGTAGTAAACACAG
TGAAAGAAACGTGCAAGCTTTGGAAAAAGAGCGGAGACGGGTCGGGAGGTATTCTGGTCGACAT
GCAACAATGGTTTGGGGACATTAATTTGAATGTGATACTGAGAATGGTTGCTGGGAAGAGGTATTT
TGGTGGTGGCGGCGCTGCTGAGAGTGATGAAAAAGAAGTGAGGAGATGCAGGAAGGCGATGAG
GGAATTCTTCGACCTCGCTGGGATTTTTGTGCTGAGAGATGCTGTACCTTATCTGGGCTGGCTTGAT
TTGGGTGGATACGAAAAGGCTATGAAGAAAACAGCCAAGGAATTGGACAGTCTTGGTCAGGAATG
GTTAGAGCAGCATCGCCGGAAGAAAGAATCCGGTGAACCGGCAGCTACTGAAGATCAAGACTTCA
TGGACGTGTTGATATCGGCCATTGATGGTACAGATTTTCTAGGCTATGATGCTGATACCATCATCAA
AGCCACATCCATGACTTTAATCGTAGGAGGCACTGATACCACATCAGTTAGTCTGCAATGGACACTC
TCGCTATTACTGAACAATCGCCACACGTTGGAAAAGGTTCAGGAAGAATTGGACAATATAGTTGGC
AAGGAAAGACTGTTGAAAGAGTCGGACATTGATAAACTTGTCTATCTCCAAGCCGTAGTGAAGGA
GTCAATGAGATTGTGCCCGGCTGGACCGCTCGCAGGAGCCCGTGAATTCAGGGAAGACTGCACCG
TTGGCGGCTACCATGTCAGGAAAGGAACCCGGCTAGTGATCAACCTCTGGAAGCTGCAAACGGAC
CCGCGGGTGTGGTCTGATCCGCTGGAGTTCAAGCCAGAGAGGTTCCTCACTACCCACATAGATGTG
GATTTAAAAGGTCAGCATTTTGAGCTGATCCCGTTTGGTACCGGTAGAAGAGCCTGTCCTGGAATA
GCGTTCGGATTACAAATGACGCATTTAGCGTTGGCCAGTTTGCTTCAAGCGTACGAGGTTTCAACTC
TTGGGAATGCAGAGGTGGACATGACTGGAACCCCTGGATTGACCAACAACAGAGCTAATCCACTTC
GAATTCTTCTTAAACCCCGCTTGCCCCTTCATCTCTACCAATAATACCCAGCTTTCTTGTACAAAGTG
GTCCCC

>Ah14918c0g5i1 (AhUGT1)

GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGAAAGCGAGAATCCTCCCCTGACAGGCTACTT
CATCCCCTACTCGACAACAAGCCACTTGATGCCACAAGTAGACATGGCCATGATCTTTGCAGCTCAT
GGTGTTGACGTTACCATCATCACCACTTACTACAACTCCCTCCTCTTCCAAAAATCCATAGACAAAGC
CACCACTTGTGGCCACCATATCCGTGTTCACAACATCAAGTTTCCATTCAAGGAGGTTGGTCTCCCT
GAAGGAATTGAGAACTTCCGTGATGGTACTTCTGGTGAAATCCTTGGCAAGATAGGGAAGGGTGT
CTCCATGCTCCAACAACCCATTGAGCAATTCATCAGAGACAACAGACCCGACTTCATTGTCGCCGAC
ATGTTTCACTTTTGGGCATCTGATTTAGCTTCTGAACTCGGGATCCCGACGCTTACGTCTCACGGGA
TGGGACACTTTGGTCTTTGTGCTGAAGATTCTATTCTTCGGTGCGAGCCTCACAAGGCCGTCCAATC
CGACAAGGAGCCGTTTCTACTACCAGGGCTTCCGGACCAGATCCGAATGACAAGATCGGAGATTGC
AGATTGGATTAAAGTACCGGATGCATTCACCAGTTTAATCCACATGATCAATGAGTCCATTGGAAA
GAGCTATGGGGTTTTGGTGAATAGCTTCTACGAGTTGGAGAGAGATTATGCTCATCACTATAGAGA
TGTCGTGGGAATAAAGTCCTGGCCCATCGGATCTGTACGACTTTTTAACGTCATAAACGACTCTGAA

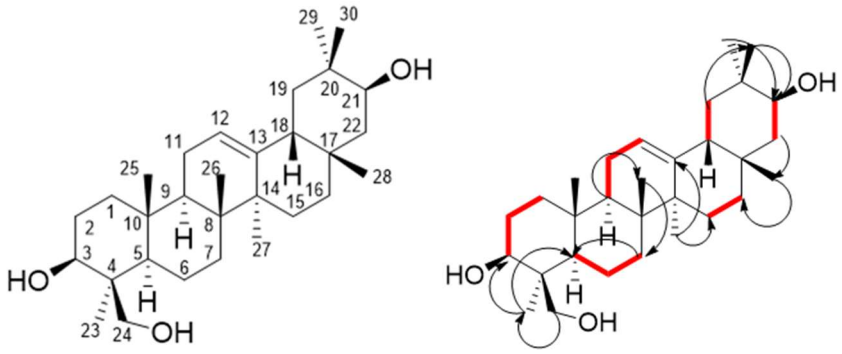
AACGATGACAAGCACGGTTGTCTAAGTTGGCTAGATTCTCAAAACCGTAACTCGGTTGTTTATGCAT
GTTTTGGAAGCTTGAGTTTTTTCAGCTCAGCTCAGCTTTCAGAGATAGCAATGGCTTTTGACTCGAT
GAATCACCCATTTATTTGGGTAGTTGGAAAACAAGACAAAACTAACAAAGAGGAATGGCTTCCAAG
AGAACTTGAGGAGAAGATAACTAAGAGCAACAAGGGCTTGATCATCAAGGGTTGGGTACCACAGA
TGTCTATCTTAGAGCATCCAGCTGTTGGTGGGTTTCTGACTCACTGTGGATGGAACTCGGTCTTGGA
GACCATTGCGACAGGAGTGCCAATGGCTACTTGGCCATTCGGGGCTGACCAATTTTACAACGAGAA
GTTGATTGAGTTGTTGGGACTCGGGGTCAAAGTTGGGACGGAGGTGTGGACTCAATTGCTGGAAG
AGAGTCGTTCGTATGTGGAAAGGGAGAAGATTGTGAAAGCAGTGAATGAGCTGATGGGTGGTGG
CGAAGAGGCGGAGAAGAGGAGGAAGAAAGTGAAAGAGCTCTCAGTGATTGCCAAGGCCGCTGTT
CAAGAAGGCGGCTCATCCTACTCAAACGTCATCTCTTTGCTTGATGAGCTTCGACAGCAGCGTCGCT
GCACGTTCAACTGATACCCAGCTTTCTTGTACAAAGTGGTCCCC

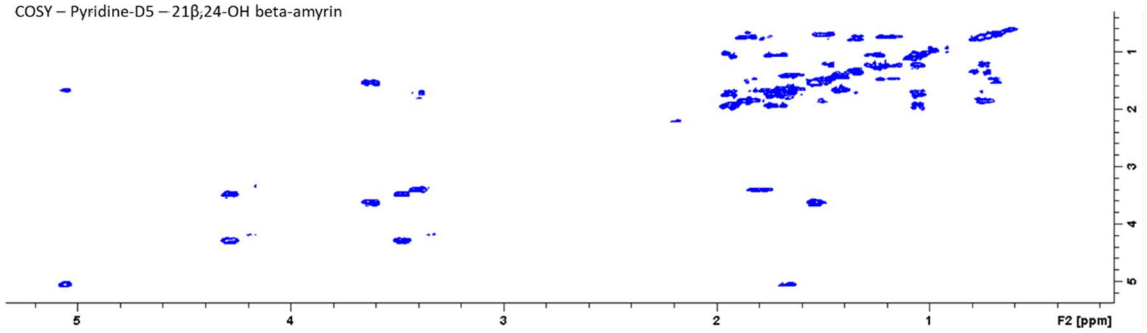
Appendix D1 – ^{1}H NMR of 21β-OH beta-amyrin

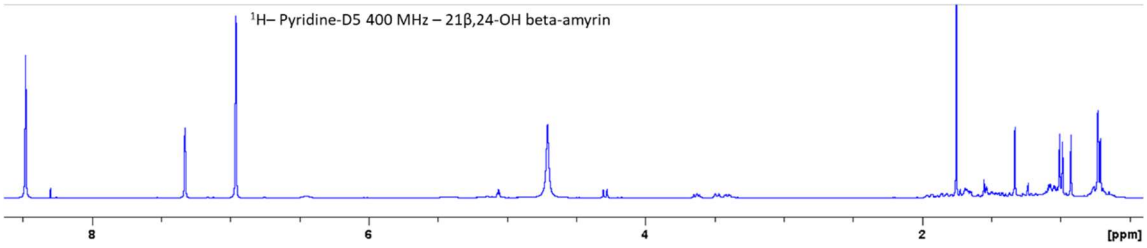


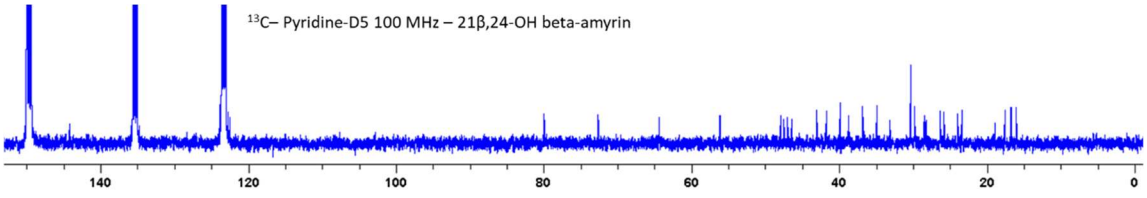
Appendix D2 – NMR of 21β,24-OH beta-amyrin

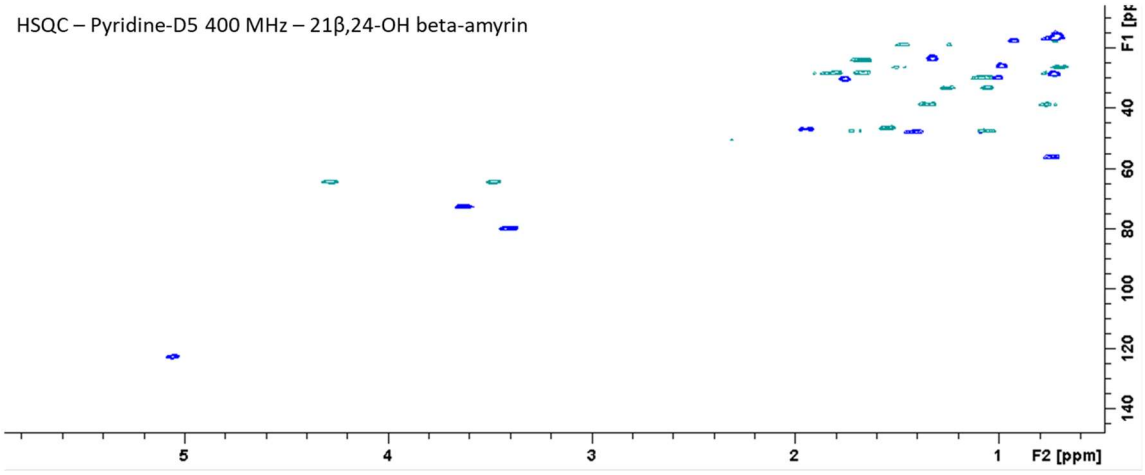
Position	*δ ^{13}C per (Ohtani, Ogawa et al. 1992) (ppm)*	*δ ^{13}C observed (ppm)*	*δ ^{1}H observed (ppm) (J in Hz)*
1	38.9	38.6	0.75 (m), 1.35 (m)
2	28.4	28.2	1.77 (m)
3	80.1	80.0	3.41 (dd) (*J* = 4.6, 11.6)
4	43.2	43.0	-
5	56.3	56.1	0.69 (m)
6	19.1	18.9	1.49 (m), 1.22 (m)
7	33.3	33.1	1.05 (m), 1.25 (m)
8	40.1	39.9	-
9	48.1	47.9	1.43 (dd) (*J* = 2.9, 9.6)
10	37.0	36.8	-
11	24.1	23.9	1.65 (m), 1.69 (m)
12	122.7	122.5	5.06 (app t) (*J* = 3.6)
13	144.3	144.1	-
14	41.9	41.7	-
15	26.5	26.3	0.69 (m), 1.49 (m)
16	29.9	28.4	1.85 (m)
17	35.1	34.9	-
18	47.2	47.0	1.95 (dd) (*J* = 3.8, 14.0)
19	46.5	47.5	1.75 (m), 1.06 (m)
20	36.9	36.8	-
21	72.8	72.6	3.63 (dd) (*J* = 7.0, 9.7)
22	47.7	46.4	2H, 1.54 (m)
23	23.5	23.4	3H, 1.33 (s)
24	64.5	64.4	3.48 (d) (*J* = 11.0), 4.29 (d) (*J* = 11.0)
25	16.2	16.0	3H, 0.70 (s)
26	16.9	16.7	3H, 0.73 (s)
27	26.0	25.8	3H, 0.98 (s)
28	28.7	28.5	3H, 0.72 (s)
29	29.9	29.8	3H, 1.00 (s)
30	17.7	17.5	3H, 0.92 (s)

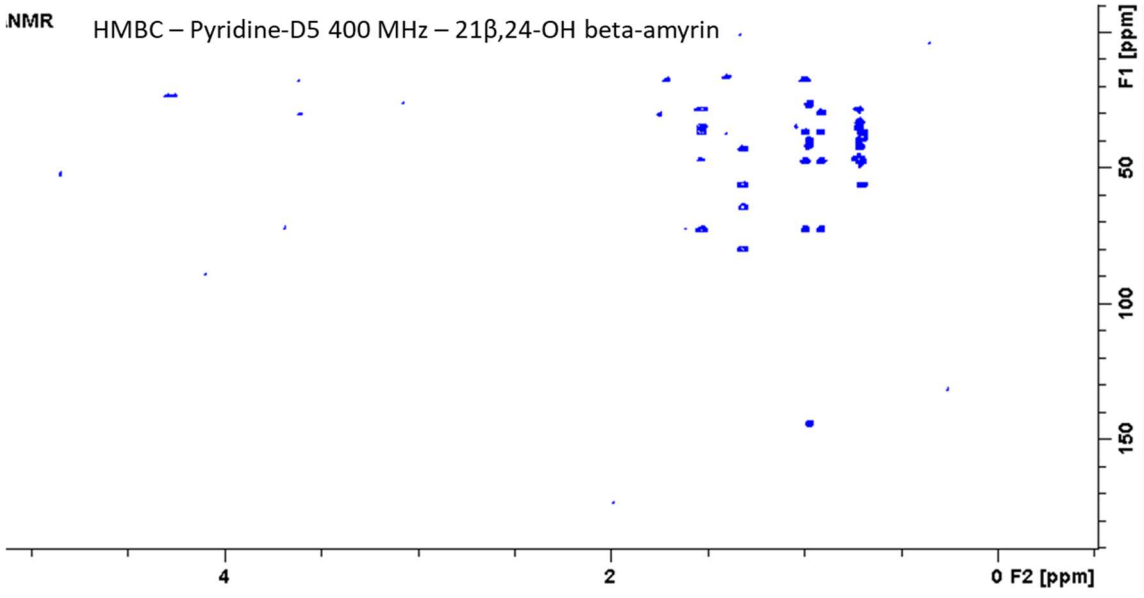

29
30
OH
19
20
21
12
11
13
18
22
17
25
26
H
28
9
14
16
15
1
10
8
2
H
7
27
3
5
4
6
HO
H
23
24
OH
OH
H
H
HO
H
OH


COSY – Pyridine-D5 – 21β,24-OH beta-amyrin
5
4
3
2
1
F2 [ppm]


¹H– Pyridine-D5 400 MHz – 21β,24-OH beta-amyrin
8
6
4
2
[ppm]

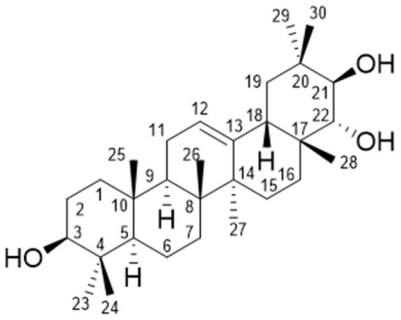

¹³C– Pyridine-D5 100 MHz – 21β,24-OH beta-amyrin
140
120
100
80
60
40
20
0

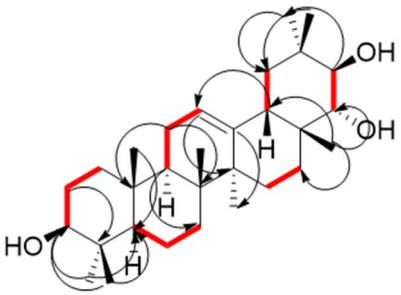

HSQC – Pyridine-D5 400 MHz – 21β,24-OH beta-amyrin
F1 [pp
40
60
80
100
120
140
5
4
3
2
1
F2 [ppm]

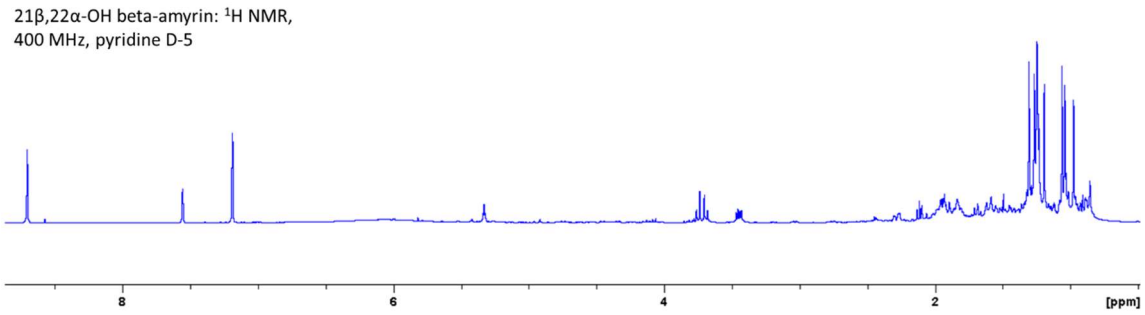

NMR
HMBC – Pyridine-D5 400 MHz – 21β,24-OH beta-amyrin
F1 [ppm]
50
100
150
4
2
0 F2 [ppm]

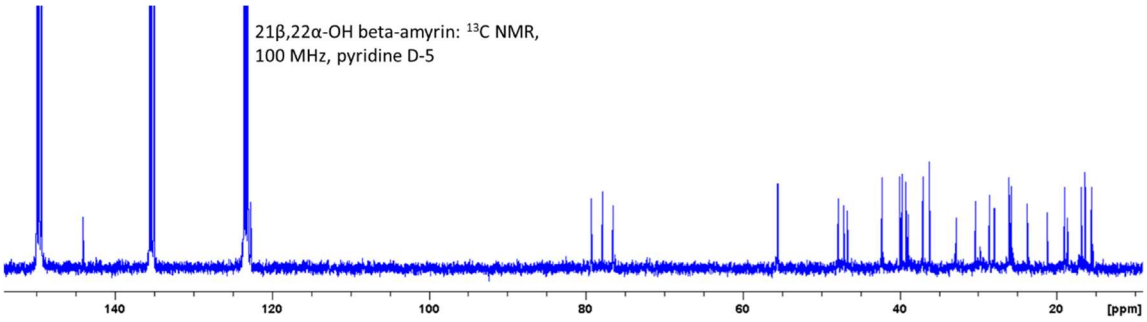
Appendix D3 – NMR of 21β,22α-OH beta-amyrin

Position	*δ ^{1}H (ppm) (J in Hz)*	*δ ^{13}C (ppm)*
1	1.08 (m)	38.9
2	1.84 (m)	27.9
3	3.45 (dd) (*J* = 10.8, 5.2)	77.9
4	-	39.2
5	0.86 (m)	55.5
6	1.44 (m)	18.6
7	1.62 (m)	32.8
8	-	40.0
9	1.69 (m)	47.8
10	-	37.0
11	1.95 (dd) (*J* = 8.9, 3.6)	23.7
12	5.33 (app t) (*J* = 3.6)	122.7
13	-	144.0
14	-	42.3
15	1.28 (m)	26.0
16	1.98 (m)	21.2
17	-	39.7
18	2.28 (dd) (*J* = 13.0, 3.8)	47.1
19	2.12 (m)	46.6
20	-	36.2
21	3.75 (d) (*J* = 10.0)	76.6
22	3.69 (d) (*J* = 10.0)	79.3
23	(3H) 1.25 (s)	28.5
24	(3H) 1.06 (s)	16.3
25	(3H) 0.98 (s)	15.6
26	(3H) 1.04 (s)	16.8
27	(3H) 1.27 (s)	26.1
28	(3H) 1.31 (s)	25.7
29	(3H) 1.24 (s)	30.3
30	(3H) 1.20 (s)	18.9

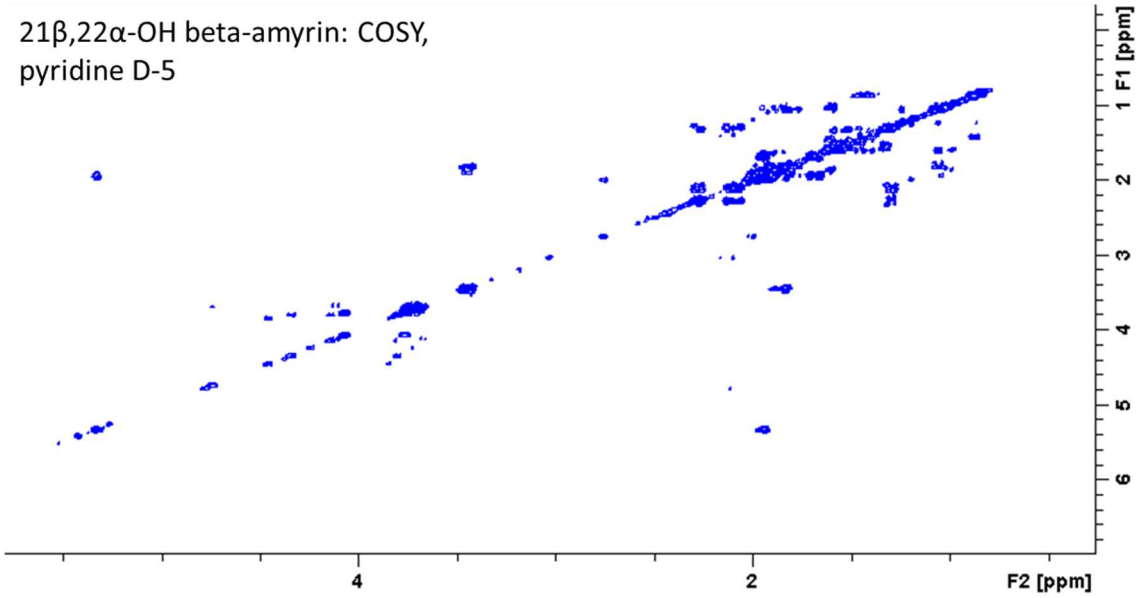




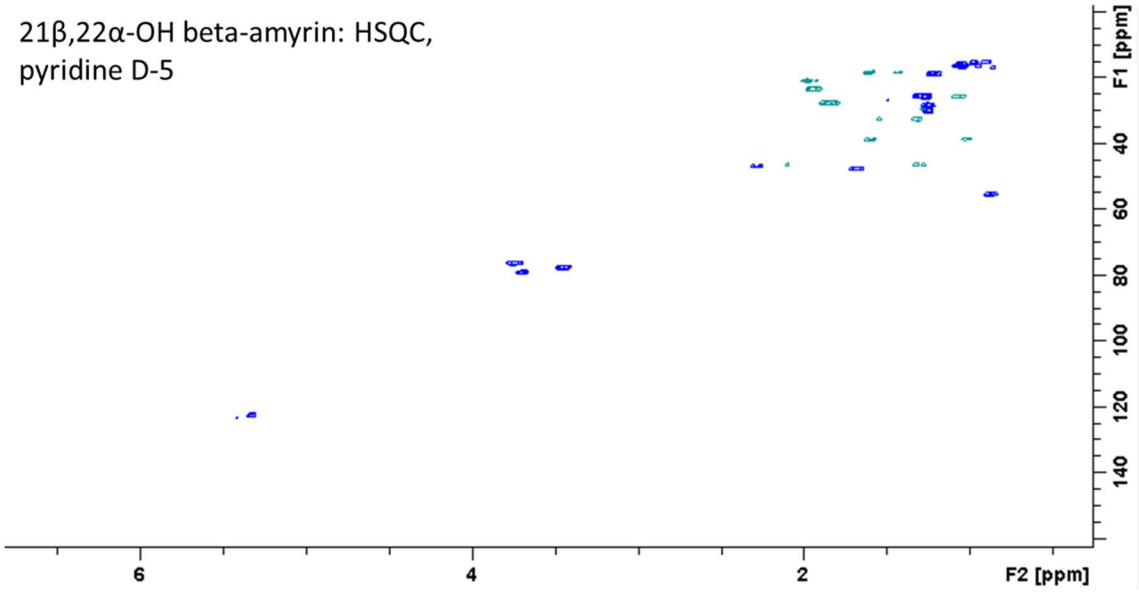


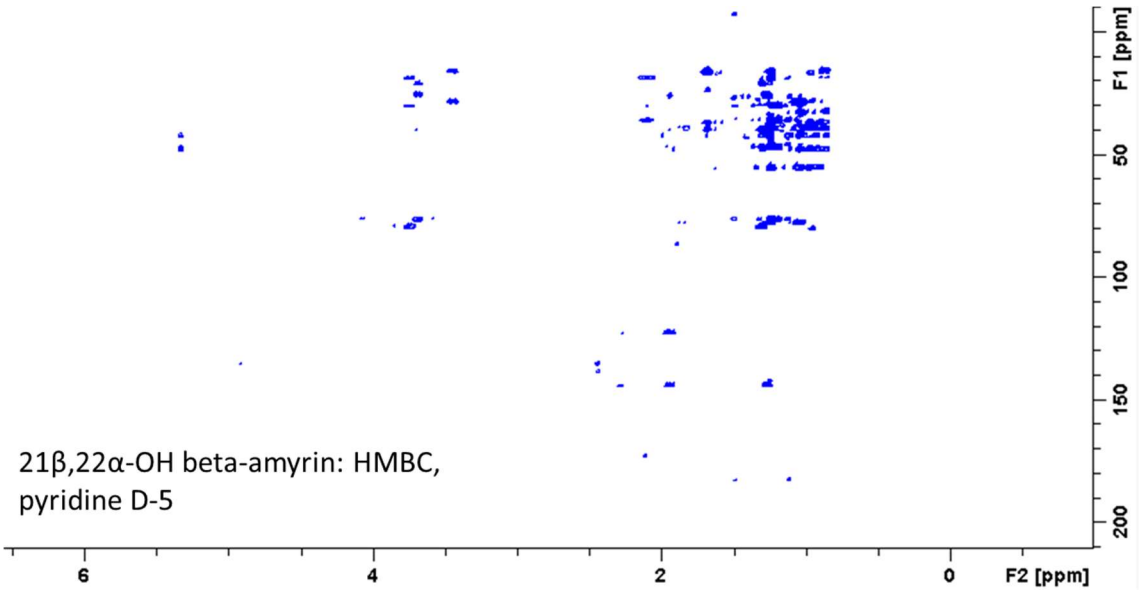


21β,22α-OH beta-amyrin: COSY,
pyridine D-5



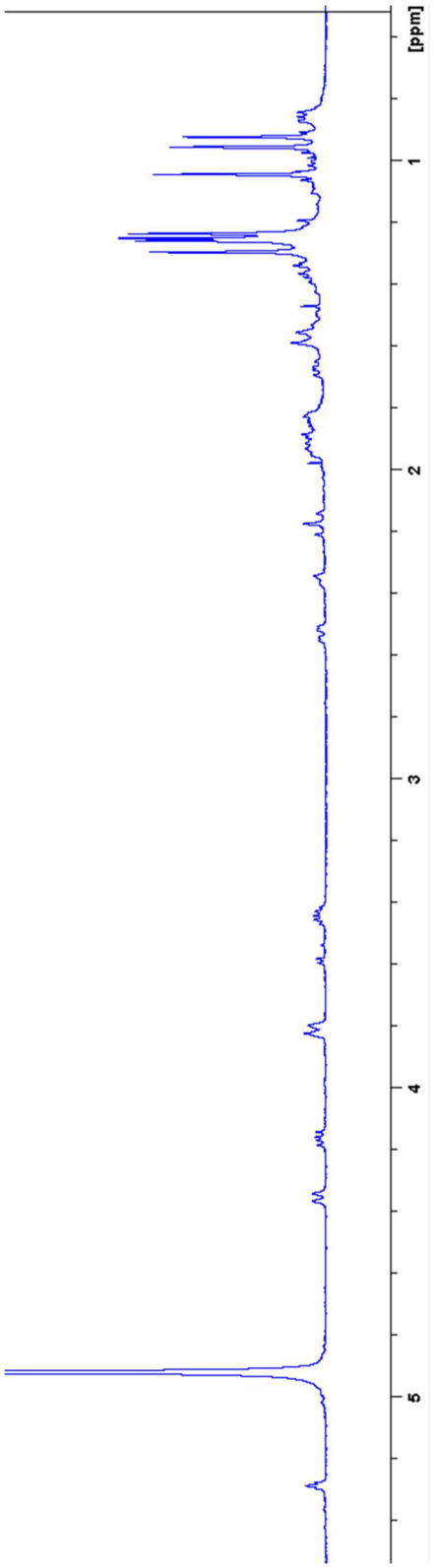
21β,22α-OH beta-amyrin: HSQC,
pyridine D-5





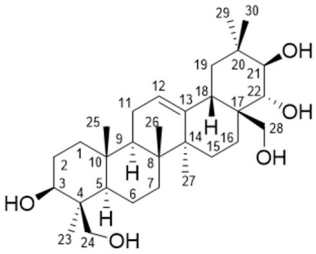
21β,22α-OH beta-amyrin: HMBC,
pyridine D-5

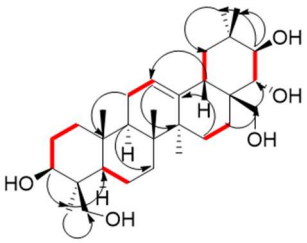
Appendix D4 – ^{1}H NMR of suspected 21β,22α,28-OH beta-amyrin



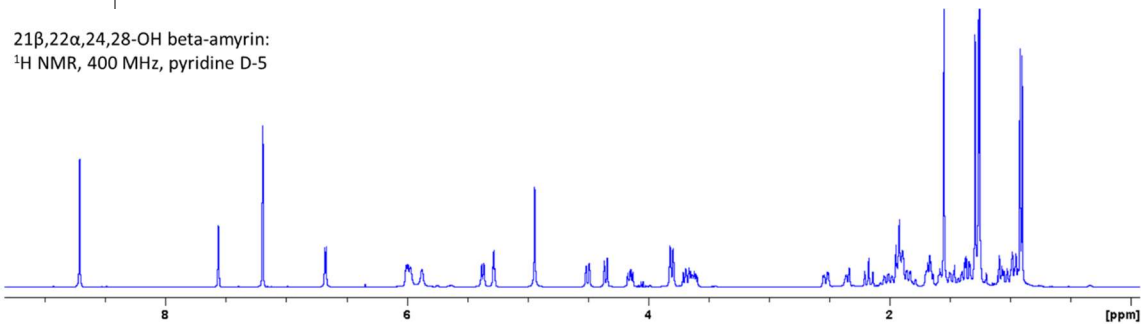
Appendix D5 – NMR of 21β,22α,24,28-OH beta-amyrin

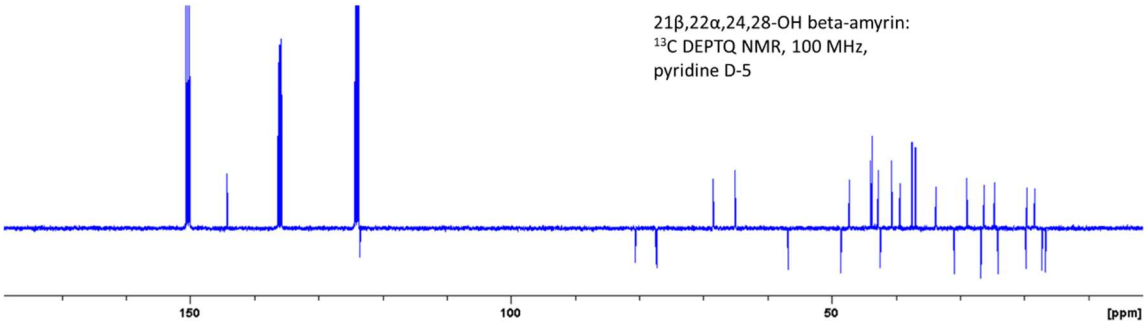
Position	*δ ^{1}H (ppm) (J in Hz)*	*δ ^{13}C Assignment (ppm)*
1	1.09 (m)	39.4 (CH_2)
2	2.04 (m)	28.9 (CH_2)
3	3.65 (m)	80.6 (CH)
4	-	43.7 (quaternary)
5	0.97 (m)	56.8 (CH)
6	1.52 (m)	19.6 (CH_2)
7	1.22 (m), 1.49 (m)	33.7 (CH_2)
8	-	39.3 (quaternary)
9	1.69 (m)	48.5 (CH)
10	-	37.4 (quaternary)
11	1.87 (m)	24.6 (CH_2)
12	5.31 (t) (J = 3.4)	123.5 (CH)
13	-	144.2 (quaternary)
14	-	42.7 (quaternary)
15	2.36 (m)	18.4 (CH_2)
16	1.92 (m)	26.2 (CH_2)
17	-	43.9 (quaternary)
18	2.55 (dd) (J = 13.5, 4.2)	42.4 (CH)
19	1.36 (m)	47.2 (CH_2)
20	-	36.9 (quaternary)
21	3.83 (d) J = 10.3 Hz	77.3 (CH)
22	4.38 (d) J = 10.3 Hz	77.1 (CH)
23	(3H) 1.58 (s)	24.0 (CH_3)
24	3.71 (m), 4.53 (dd) (J = 10.6, 1.8)	65.0 (CH_2)
25	(3H) 0.93 (s)	16.6 (CH_3)
26	(3H) 0.94 (s)	17.2 (CH_3)
27	(3H) 1.32 (s)	26.7 (CH_3)
28	3.82 (m), 4.18 (dd) (J = 10.6, 6.7)	68.4 (CH_2)
29	(3H) 1.29 (s)	30.9 (CH_3)
30	(3H) 1.28 (s)	19.6 (CH_3)

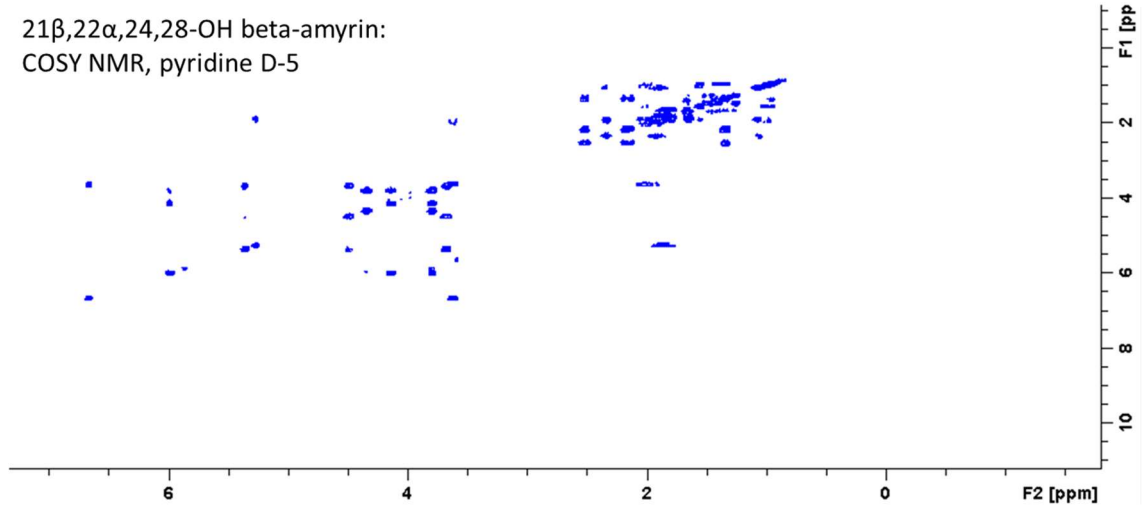


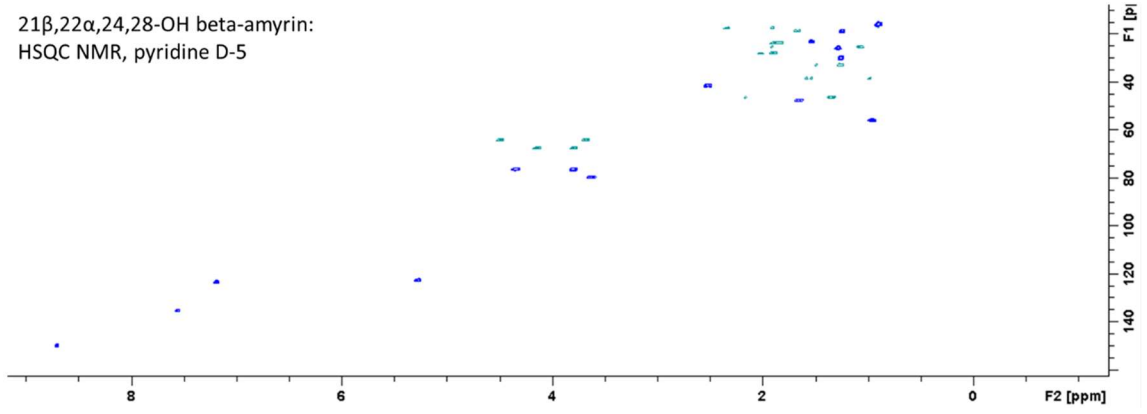


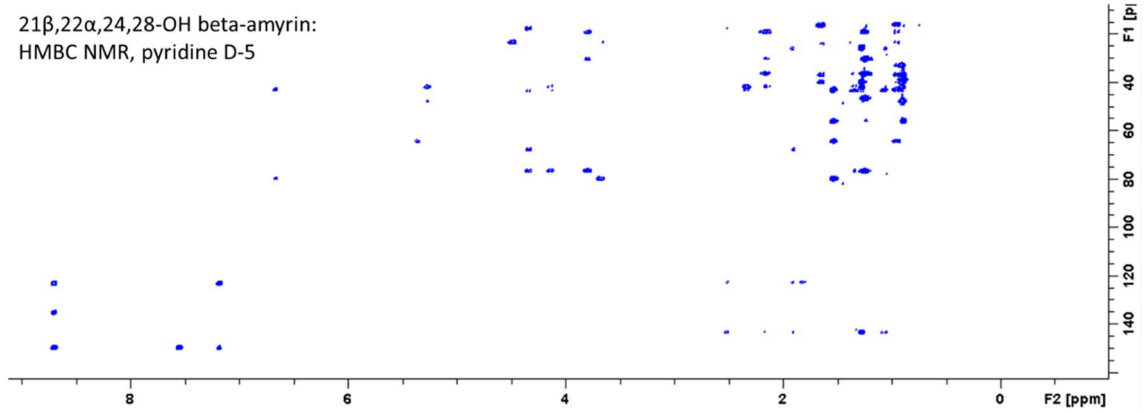
21β,22α,24,28-OH beta-amyrin:
^{1}H NMR, 400 MHz, pyridine D-5




21β,22α,24,28-OH beta-amyrin:
13C DEPTQ NMR, 100 MHz,
pyridine D-5
150
100
50
[ppm]


21β,22α,24,28-OH beta-amyrin:
COSY NMR, pyridine D-5
F1 [pp
2
4
6
8
10
6
4
2
0
F2 [ppm]


21β,22α,24,28-OH beta-amyrin:
HSQC NMR, pyridine D-5
F1 [pi
40
60
80
100
120
140
8
6
4
2
0
F2 [ppm]


21β,22α,24,28-OH beta-amyrin:
HMBC NMR, pyridine D-5
F1 [pi
40
60
80
100
120
140
8
6
4
2
0
F2 [ppm]