Elucidation of the pathway for biosynthesis of saponin adjuvants from the soapbark tree

James Reed[1,a], Anastasia Orme[1,a, §], Amr El-Demerdash[1,a, §§], Charlotte Owen[1], Laetitia B.B. Martin[1], Rajesh Chandra Misra[1], Shingo Kikuchi[1], Martin Rejzek[1], Azahara C. Martin[1], Alex Harkess[2,3], Jim Leebens-Mack[4], Thomas Louveau[1, §§§], Michael J. Stephenson[1] and Anne Osbourn[1]*

Affiliations:

[1]John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK.

[2]Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn AL, 36849, US.

[3]HudsonAlpha Institute for Biotechnology, Huntsville AL, 35806, US.

[4]Department of Plant Biology, 4505 Miller Plant Sciences, University of Georgia, Athens, GA 30602-7271, US.

[§]Current address, Illumina Centre, Great Abington, Cambridge, CB21 6DF, UK.

[§§]Department of Chemistry, Faculty of Sciences, Mansoura University, Mansoura 35516, Egypt.

[§§§]Current address, GTP Technology, 52 L'Occitane, Immeuble Gould, 31670 Labège, France.

[a]These authors contributed equally to this work

*Correspondence to: anne.osbourn@jic.ac.uk

Abstract:

The Chilean soapbark tree *Quillaja saponaria* produces soap-like molecules (QS saponins) that are important vaccine adjuvants. These highly valuable compounds are sourced by extraction from the bark, and their biosynthetic pathway is unknown. Here we sequence the *Q. saponaria* genome. Through genome mining and combinatorial expression in tobacco, we identify 16 pathway enzymes that together enable the production of advanced QS pathway intermediates that represent a bridgehead for adjuvant bioengineering. We further identify the enzymes needed to make QS-7, a saponin with excellent therapeutic properties and low toxicity that is present in low abundance in *Q. saponaria* bark extract. Our results enable production of *Q. saponaria* vaccine adjuvants in tobacco and open the way for new routes to access and engineer natural and new-to-nature immunostimulants.

One-Sentence Summary:

Uncovering the biosynthetic route to vaccine adjuvants produced by the soapbark tree, *Quillaja saponaria.*

Introduction

Vaccination is a huge success story in the fight against infectious diseases. Vaccines frequently require an adjuvant component (an immunostimulant) to enhance the immune response to the antigen. However, to date only a few adjuvants are licensed for human use. Triterpene glycosides (saponins) from the Chilean soapbark tree (*Quillaja saponaria*) have proven to be highly effective adjuvants due to their ability to elicit both antibody and cellular immune responses (*1*). These saponins are collectively known as QS saponins. The QS-21 fraction, comprising isomeric forms of a complex triterpene saponin, is an immune-potentiator used in the adjuvant 'AS01'. AS01 has been licensed for use in two human vaccines (the GSK vaccines Shingrix and Mosquirix, for shingles and malaria, respectively). A mixture of QS saponins, including QS-21, QS-17 and QS-7, is also included in 'Matrix-M', a combination adjuvant used in the NVX-CoV2373 COVID-19 vaccine produced by Novavax (*2*). QS-17 is a glycosylated derivative of QS-21. QS-7 has the same core structure as QS-21 but the two saponins differ in the nature of their modifications at the C-28 position. QS-7 has a simple acetyl group instead of the long (C-18) acyl chain present in QS-21, and there are also differences in the C-28 sugar moieties (Fig. 1A). Due to their chemical

complexity, the only current commercial source of these saponins is from the bark of the soapbark tree itself. However, these key immunogenic saponins represent only a portion of >100 structurally related molecules produced by *Q. saponaria* (*3*), and so require extensive purification. This issue is further exacerbated by variation in saponin content and composition between individual trees due to environmental and genetic factors (*4-6*). Although a number of saponin biosynthetic enzymes have been characterized in recent years from taxonomically diverse plant species (e.g. *7-10*), much remains to be learned about the enzymes that generate the enormous structural diversity of saponins. Indeed, QS-21 has a total of seven different types of glycosidic moieties, including unusual sugars such as D-fucose, D-apiose and L-arabinofuranose, for which the cognate enzymes are not yet known. Understanding the biosynthetic pathways for QS saponins will therefore provide new insights into how these molecules are made and diversified. It will further open up opportunities to produce saponins optimized for their immunostimulatory properties and low toxicity in heterologous hosts for use in the vaccines of the future.

Results

Biosynthesis of the quillaic acid scaffold

Triterpenes are biosynthesized from the linear isoprenoid precursor 2,3-oxidosqualene (**1**), which can be cyclized into >100 different diverse scaffold products (*11*). The most common of these scaffolds is *β*-amyrin (**2**). The core structure of QS-7, QS-21 and QS-17 is quillaic acid (QA) (**5**), which is based on β-amyrin, but with oxidized groups at the C-16*α*, C-23 and C-28 positions (Fig. 1B). We therefore initiated our investigations of saponin biosynthesis in *Q. saponaria* by searching for the enzymes required for *β*-amyrin biosynthesis and oxidation.

QS saponins are normally extracted commercially from bark. At the time of starting this work, transcriptome data derived from *Q. saponaria* leaves were available through the 1000 Plants (1KP) Project (*12*). We obtained saplings of *Q. saponaria* (JIC accession S10) and verified the presence of QS-21 in the leaves, consistent with previous reports (*13*) (Fig. S1). A BLAST search against the 1KP data was conducted using a characterized *β*-amyrin synthase (GgbAS1, Genbank accession AB037203) from licorice (*Glycyrrhiza glabra*) as a query (*14*). This revealed a single full-length candidate with 88% amino acid sequence identity to GgbAS1. We used gene-specific primers (Data S1) to clone the corresponding sequence from cDNA prepared from the leaves of *Q. saponaria* S10. We then investigated the function of this candidate by *Agrobacterium*-mediated transient expression in the leaves

of *Nicotiana benthamiana*. GC-MS analysis of leaf extracts revealed a peak with the same retention time and mass spectrum as an authentic β-amyrin standard (**2**), confirming that this enzyme (hereafter named QsbAS1) is indeed a *β*-amyrin synthase (Fig. S2).

We next considered candidates for oxidation of β-amyrin (**2**). Most known triterpene oxidases are members of the cytochrome P450 monooxygenase (CYP) superfamily (*8*). Of these, the CYP716 family is commonly associated with triterpene biosynthesis and includes enzymes known to perform C-28 and C-16α oxidation (*8,15*). A BLAST search of the 1KP *Q. saponaria* transcriptome dataset was carried out using a known C-28 oxidase from *Medicago truncatula* (CYP716A12, Genbank accession FN995112) (*16*), a saponin-producing species that (like *Q. saponaria*) belongs to the Fabales order. From this, the two highest scoring hits were selected for further investigation. Transient expression of the first of these (*CYP716A224*) with *QsbAS1* in *N. benthamiana* resulted in near total conversion of *β*-amyrin (**2**) to oleanolic acid (**3**) (Fig. S3). The second enzyme (CYP716A297) showed very little activity towards *β*-amyrin. However co-expression of both *CYP716A224* and *CYP716A297* in combination with *QsbAS1* resulted in formation of a new product which we identified as echinocystic acid (**4**) using an authentic standard (Fig. S3). These two CYPs are therefore able to oxidize two (C-28 and C-16*α*) of the three positions that are oxidized in QA (**5**) (Fig. 1C). In searching for the final oxidase, we compiled a list of all CYP sequences in the 1KP *Q. saponaria* transcriptome dataset that appeared to be full length (n = 35). After eliminating enzymes that were closely related to known CYPs associated with primary metabolism we were left with 26 candidates, of which 17 were successfully cloned and transiently expressed in *N. benthamiana* (Data S2). Using this approach, a single candidate (*CYP714E52*) was identified which, when co-expressed with *QsbAS1*, *CYP716A224* and *CYP716A297*, resulted in production of QA (**5**) in *N. benthamiana* (Fig. 1B). We then carried out large-scale transient expression by vacuum agro-infiltration of 209 plants, purified ~30 mg of this product, and confirmed its structure as QA (**5**) by ^{1}H NMR (Fig. S4; Fig. 1C). A phylogenetic tree showing the relatedness of the three CYPs required for QA biosynthesis to other previously characterized triterpene modifying CYPs from plants is shown in Fig. S5.

Generation of a pseudochromosome-level genome assembly for *Q. saponaria*

Genes for plant specialized metabolic pathways are commonly co-expressed and may also be physically co-localized or ‘clustered’ within the genome (*17*). Co-expression analysis requires availability of RNA-seq data for multiple different tissues/treatments, while discovery of biosynthetic gene clusters is dependent on availability of a genome assembly,

neither of which were available for *Q. saponaria*. To facilitate discovery of the saponin biosynthetic steps downstream of QA, we therefore generated *de novo* transcriptome and genome sequence resources for *Q. saponaria* accession S10. RNA-seq data were generated for six different tissues (primordia; expanding, mature and old leaves; green stems, and roots) using Illumina HiSeq4000. QS-21 was present in all tissues examined (Fig. S1). The estimated genome size of *Q. saponaria* based on flow cytometry is 411 Mbp (*18*). PacBio long read sequencing and Hi-C (high-throughput/resolution chromosome conformation capture) were used to generate a chromosome-scale assembly (Table S1, Fig. S6, see Materials and Methods). The draft genome was annotated by RNA-seq read alignment, filtering, gene model generation and selection of final gene models (Table S1, Fig S7, see Materials and Methods). Karyotype analysis revealed 28 chromosomes, consistent with a haploid chromosome number of 14 (Fig. 2A). The 14 scaffolds therefore represent the 14 chromosomes of *Q. saponaria* S10. Synteny analysis provided evidence for a whole genome duplication event in *Q. saponaria* S10 (Fig. 2B), consistent with hypothesised polyploidy events observed across members of the Fabales (*19*).

Investigation of the expression profiles of the characterized QA biosynthesis genes in different *Q. saponaria* tissues revealed that these genes are highly significantly co-expressed (Fig. 2C), with highest absolute expression in the leaf primordia and lowest in old leaves (Fig. 2C; Fig. S8), suggesting that it may be possible to identify further candidate downstream QS pathway genes based on co-expression using these genes as bait.

We next mined the *Q. saponaria* genome assembly using plantiSMASH, an algorithm designed to predict biosynthetic gene clusters (BGC) in plant genomes (*20*). plantiSMASH predicted a total of 51 candidate clusters, of which 34 were assigned to the 'saccharide' and/or 'terpene' classes (Fig. S9; Data S3) and so may be relevant to triterpene glycoside (i.e. saponin) biosynthesis. The four QA biosynthetic genes (*QsbAS1*, *CYP714E52*, *CYP716A224*, *CYP716A297*) are not physically clustered with each other. Of note, however, the gene encoding one of the CYPs required for QA biosynthesis (*CYP716A297*) is located adjacent to a 'saccharide' biosynthetic gene cluster (cluster #45) which includes genes predicted to encode sugar transferases and other enzymes with potential functions in specialized metabolism (Fig. 2D). Some of these genes have similar expression profiles to *CYP716A297*, potentially suggesting functional association (Fig. 2D).

Addition of the C-3 sugar chain

Having discovered the biosynthetic steps to QA (**5**) (Fig. 1C), we next focused on identification of the enzymes required for addition of sugars at the C-3 and C-28 positions of the QA scaffold. The enzymes typically responsible for glycosylation of plant natural products belong to glycosyltransferase family 1 (GT1) (*21, 22*). GT1 enzymes use uridine diphosphate (UDP)-activated sugar donors to transfer sugar units onto small molecules and so are referred to as UDP-dependent glycosyltransferases (UGTs). We therefore mined the *Q. saponaria* genome annotation to find all predicted full length (>410 aa) UGT genes by searching with InterPro code IPR002213. This yielded a total of 166 predicted UGT genes, which were then prioritized based on strength of co-expression with *QsbAS1* (PCC cut-off of 0.7) and on absolute gene expression levels in primordia (TPM>1600), resulting in a shortlist of 20 UGT genes (Table S2). The two most highly co-expressed UGTs *Qs0321930* and *Qs0321920* (PCC 0.987 and 0.985, respectively) are co-located in the BGC shown in Fig. 2D, along with a third co-expressed UGT gene *Qs0321940* (PCC 0.956). This cluster also contains a gene for another class of carbohydrate-active enzyme – *Qs0321900*, which is predicted to encode a cellulose synthase-like (CSL) protein. *Qs0321900* is not co-expressed with *QsbAS1* (PCC -0.59), although it is expressed at moderate levels in primordial tissue. Interestingly, another unlinked but closely related predicted CSL gene *Qs0000870* is very highly co-expressed with *QsbAS1* (PCC 0.992), suggestive of a role in the QS pathway. We cloned all 20 UGT candidates and both CSL genes in order to evaluate their functions.

Co-expression of each of the UGT and CSL genes with the four QA pathway genes was carried out by transient expression in *N. benthamiana*, and modification of QA (**5**) monitored by untargeted LC-MS. No conversion of QA (**5**) was observed when the UGT candidates were co-expressed. However, when either of the two CSL genes were co-expressed with the QA pathway genes, LC-MS analysis of leaf extracts revealed a peak with a mass corresponding to QA plus D-glucuronic acid and a concomitant reduction in QA (**5**) levels (Fig. 3A). We then scaled up our transient plant expression experiments. Following vacuum infiltration of 104 *N. benthamiana* plants co-expressing the QA pathway genes with *CSL1*, we were able to purify 9.5 mg of product (Data S4). We also obtained 2.1 mg of the product of co-expression of the QA pathway genes with *CSL2* (from 80 *N. benthamiana* plants) (Data S4). ^{1}H NMR revealed that the spectra for the two products were identical (Fig. S10). Extensive 2D NMR analysis (COSY, HSQC, HMBC and ROESY) confirmed that both products were 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid (**6**; abbreviated to QA-Mono) (Tables S3 and S4).

Phylogenetic analysis revealed that CSL-1 and -2 belong to the CSL-M subfamily, and they are hereafter named CSLM1 and CSLM2 (Fig. S11). Although CSL proteins have not traditionally been regarded as small molecule glycosyltransferases, two other examples have recently been reported from other plant species (*9*,*23*). The strong co-expression of *CSLM2* with *QsbAS1* suggests that CSLM2 may be primarily responsible for 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid (**6**) biosynthesis in *Q. saponaria*.

We next screened our suite of cloned UGT candidates for the ability to glycosylate 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid (**6**). Co-expression of *Qs0123860* (ranked third in Table S2) based on co-expression with *QsbAS1*) with the QA pathway genes and *CSLM1* resulted in a new product with the mass of QA-GlcA plus a hexose (Fig. 3B). Following scale-up by vacuum infiltration of 104 *N. benthamiana* plants, 7.3 mg of this product was purified and its structure determined to be 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid}-quillaic acid (**7**; abbreviated to QA-Di) by NMR (COSY, HSQC, HMBC and ROESY) (Table S5 and Data S4). Thus *Qs0123860* (UGT73CU3) encodes a QA-3-*O*-glucuronoside-*β*-1,2-galactosyltransferase capable of adding the second sugar to the C-3 position of *Q. saponaria* saponins.

Another round of co-expression experiments led to the identification of two UGTs that were able to further glycosylate 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid}-quillaic acid (**7**). One of these (*Qs0283870*) generated a product with a mass consistent with addition of a pentose, while the product of the second (*Qs0283850*) had a mass consistent with addition of a deoxyhexose (Fig. 3B). It is known that saponins from *Q. saponaria* show variation in the terminal sugar of the C-3 oligosaccharide chain, and that either D-xylose or L-rhamnose can occur at this position (*3, 5, 24*). Following large-scale vacuum infiltration, the two products were purified and their structures determined by extensive 2D NMR as 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (**8**; abbreviated to QA-TriX) (21.6 mg purified) and 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (**9**; abbreviated to QA-TriR) (43.3 mg purified), respectively (Tables S6, S7 and Data S4). Thus *Qs0283870* (UGT73CX1) encodes a xylosyltransferase, and *Qs0283850* (UGT73CX2) a rhamnosyltransferase. These genes are ranked sixth and thirteenth respectively in Table S2 based on co-expression with *QsbAS1*. Phylogenetic analysis reveals that all three UGTs (UGT73CU3/UGT73CX1/UGT73CX2) belong to subgroup D of the GT1 family (Fig. S12). This subgroup includes several other enzymes that

are known to glycosylate triterpenes from both monocots and dicots (*21*). In summary, CSLM1/2 together with the three UGT enzymes characterised here collectively enable the conversion of QA (**5**) to either QA-TriX (**8**) or QA-TriR (**9**) (Fig. 3C).

Addition of the C-28 sugar chain

QS-7, QS-21 and QS-17 share a common core consisting of the QA scaffold, the C-3 sugar chain and a tetrasaccharide moiety at C-28 consisting of D-fucose, L-rhamnose, D-xylose, and D-apiose (Fig. 1A). They differ in the nature of the other sugar and acyl groups attached to the C-28 sugar chain. A survey of the structures of saponins reported from *Q. saponaria* indicates that C-3 glycosylation is likely to precede the modifications at the C-28 position (Fig. S13). Our results thus far are consistent with this hypothesis. Having successfully reconstituted the pathway for addition of the C-3 trisaccharide chain, we next turned our attention to elucidation of the steps needed for glycosylation at C-28. The sugar that is linked directly to the QA scaffold at this position is D-fucose, which is attached via an ester linkage. The UGT gene *Qs0321930* (*UGT74BX1*) shows the highest level of co-expression with *QsbAS1* (PCC 0.987) (Table S2). It is located in biosynthetic gene cluster #45 (Fig. 2D). Furthermore, *Qs0321930* is the only gene on the UGT candidate list that is predicted to encode a member of subgroup L of the GT1 family (Fig. S12), a subgroup known to contain ester-forming UGTs (*21*). Indeed, transient expression of *Qs0321930* together with the previously identified *Q. saponaria* saponin biosynthesis genes (the four QA genes, *CSLM2*, *UGT73CU3* and *UGT73CX1*, producing **8**) resulted in formation of small amounts of a new product with a mass consistent with addition of a deoxyhexose, which we anticipated to be the C-28 fucoside of **8** (abbreviated as QA-TriX-F (**10**)) (Fig. S14). UDP-α-D-fucose has been suggested to be limiting in *N. benthamiana*, which could account for the low abundance of the new product (*9*). Nevertheless, screening of additional UGT candidates against the putative QA-TriX-F (**10**) resulted in identification of a UGT in subgroup A capable of addition of a further deoxyhexose, with a mass consistent with addition of L-rhamnose as the second sugar in the C-28 sugar chain, to form QA-TriX-FR (**12**) (Fig. S14). The activity of this putative rhamnosyltransferase (*UGT91AR1*) was dependent on the presence of D-fucose. The gene encoding it (*Qs0321920*) has the second highest level of co-expression with *QsbAS1* (PCC 0.985; Table S2) and is located in biosynthetic gene cluster #45 adjacent to the putative fucosyltransferase gene (*UGT74BX1*) to which it shares only ~30% amino acid sequence identity. A further round of screening identified another subgroup A UGT encoded by *Qs0234120* (*UGT91AQ1*) that appeared to modify QA-TriR-FR by addition of a pentose,

suggesting that this may be the C-28 xylosyltransferase producing QA-TriX-FRX (**14**) (Fig. S14).

In contrast to the QA C-3 glycosides, only trace amounts of the three putative C-28 glycosides were observed, with large quantities of unconverted precursor QA-TriX (**8**) remaining (Fig. S14). It was apparent that the poor conversion from the QA-TriX (**8**) product to QA-TriX-F (**10**) was likely to represent a significant bottleneck, impeding further pathway elucidation and structural verification of the products. We noted that biosynthetic gene cluster #45 also harbors two predicted short chain dehydrogenase/reductase (SDR) genes. One of these (*Qs0321910*) is located immediately adjacent to the *UGT91AR1* gene and has a similar expression pattern to the QS enzymes characterized so far (co-expression with *QsbAS1*, PCC 0.871) (Fig. 2D). Most of the known sugar nucleotide interconverting enzymes are members of the SDR superfamily (*25, 26*). Transient co-expression of this SDR enzyme with the gene set for QA-TriX-F (**10**) biosynthesis resulted in substantial increases in the levels of the QA-TriX-F (**10**) product, suggesting that the SDR has a role in D-fucosylation, potentially by converting an endogenous UDP-sugar substrate in *N. benthamiana* to UDP-D-fucose, thereby furnishing enhanced FucT activity (Fig. 4A). Further, co-expression of the additional C-28 sugar transferases including *UGT91AR1* and *UGT91AQ1* demonstrated that the amounts of the relevant products [QA-TriX-FR (**12**) and QA-TriX-FRX (**14**), respectively] were likewise substantially increased in the presence of the *Qs0321910* SDR (Fig. S15).

We next exploited the new SDR to perform large-scale transient expression experiments in *N. benthamiana* in order to purify the new UGT products. During the previous purifications of the C-3 quillaic acid trisaccharide products, we obtained around two-fold higher yields of the the C-3 rhamnose (QA-TriR (**9**)) compared to the C-3 xylose version (QA-TriX (**8**)) (Data S4). We therefore opted to generate and purify the putative C-28 glycosides based on the QA-TriR (**9**) scaffold. Following infiltration of 100-200 *N. benthamiana* plants, the products were purified and their identities confirmed by extensive 1D- and 2D-NMR analysis as follows: UGT74BX1 product (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-fucopyranosyl ester}-quillaic acid) (**11**) (abbreviated to QA-TriR-F) (1 mg purified); UGT91AR1 product (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**13**) (abbreviated to QA-TriR-FR) (43.9 mg purified); and the UGT91AQ1 product (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-

xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**15**) (abbreviated to QA-TriR-FRX) (3.1 mg purified) (Tables S8, S9 and S10; Data S4).

The terminal sugar in the linear tetrasaccharide at C-28 in saponins such as QS-21 can be either D-xylose or D-apiose (Fig. 1A). Having identified the enzymes that add the first three sugars in the C-28 sugar chain, we carried out a final round of screening to identify the sugar transferases that add these sugars. This led to the identification of two further functional UGTs that each generated a product consistent with QA-TriX-FRX plus a pentose, but with slightly different retention times (Fig. 4B). The genes encoding these enzymes (*Qs0234130* and *Qs0234140,* ranked fourteenth and eighteenth respectively in Table S2) were both located in the chromosome 7 biosynthetic gene cluster #31 with the previously characterized C-28 xylosyltransferase *UGT91AQ1* (Data S3). This region is syntenic to the chromosome 11 biosynthetic gene cluster #45, suggesting that the two clusters may share a common evolutionary origin and may have arisen as a consequence of genome duplication (Fig. S16). We also noted that a predicted UDP-D-apiose/UDP-D-xylose synthase gene (*Qs0088320*) was highly expressed in *Q. saponaria* leaf primordia. This gene was not physically clustered with the previously characterized saponin biosynthesis genes but showed strong co-expression with them (co-expression with *QsbAS1*, PCC 0.943). Transient expression of this putative UDP-apiose/UDP-xylose synthase (QsAXS) with either *Qs0234130* and *Qs0234140* resulted in a marked increase (around 11-fold) in the amount of the *Qs0234140* product generated (Fig. S17).

We next carried out large-scale transient expression in *N. benthamiana* and purified each of the two new UGT products using the QA-TriR-FRX scaffold (**15**). Their structures were determined by extensive 1D- and 2D-NMR as (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**17**) **(***Qs0234140;* 13.2 mg purified) (abbreviated to QA-TriR-FRXX) and (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-apiofuranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**19**) (*Qs0234130;* 13.2 mg purified) (abbreviated to QA-TriR-FRXA) (Tables S11, S12 and Data S4). Thus *Qs0234130* (*UGT73CY3*) encodes the terminal xylosyltransferase, and *Qs0234140* (*UGT73CY2*) encodes the terminal C-28 apiosyltransferase (Fig. 4C). The importance of QsAXS in boosting the *Qs0234140* product suggests that UDP-α-D-apiose may be lacking in *N. benthamiana.* The fully characterized

pathway up to this point is shown in Fig. 5. Around one third of characterized QS saponins are derived from these scaffolds (*24*), making this an important branch point for saponin diversification.

The mechanism for D-fucosylation

Given the importance of the SDR encoded by *Qs0321910* for enhancing D-fucosylation and subsequent addition of the sugar chain at C-28, we sought to further understand the function of this enzyme. We hypothesized that this enzyme would be responsible for production of UDP-D-fucose. Despite the fact that plant sugar biosynthetic pathways are generally well characterized, the biosynthetic route to D-fucose in plants is unknown. In bacteria, dTDP-D-glucose is converted to dTDP-4-keto-6-deoxy-D-glucose by a dTDP-D-glucose 4,6-dehydratase. The 4-keto group of this intermediate is then reduced by dTDP-4-keto-6-deoxy-D-glucose reductase to form dTDP-D-fucose (Fig. 6A) (*27*). The first of these steps is shared with dTDP-L-rhamnose biosynthesis, and indeed it is known that plants synthesize the analogous UDP-L-rhamnose from UDP-D-glucose via UDP-4-keto-6-deoxy-D-glucose (*28*). Since UDP-4-keto-6-deoxy-D-glucose would be expected to be present in plant cells as part of UDP-L-rhamnose biosynthesis, we hypothesized that the *Qs0321910* SDR may function as a 4-ketoreductase. Furthermore, phylogenetic analysis of this SDR revealed that it is a member of the SDR114C family (Fig. S18), as previously defined by Moummou *et al.* (*29*). Several members of this family have been shown to reduce carbonyl groups to alcohols in alkaloid and terpenoid biosynthesis (*29-32*), consistent with our proposal that the *Qs0321910* SDR may carry out C-4 reduction of UDP-4-keto-6-deoxy-D-glucose to form UDP-D-fucose.

To test this hypothesis, we purified the *Qs0321910* SDR for functional analysis *in vitro* (Fig. S19). The anticipated UDP-4-keto-6-deoxy-D-glucose substrate is not commercially available. Therefore, to generate this compound from UDP-D-glucose, we cloned and purified a characterized UDP-D-glucose 4,6-dehydratase from the *Acanthocystis turfacea* chlorella virus 1 (ATCV-1 UGD, Genbank accession YP_001427025.1) (*33*) (Fig. S19). We also purified the *Q. saponaria* UGT74BX1 (Fig. S19). In a single reaction, we combined UDP-D-glucose with the purified QA-TriR (**9**) product and the *Qs0321910* SDR, ATCV-1 UGD and UGT74BX1 enzymes. Subsequent LC-MS analysis confirmed the conversion of QA-TriR (**9**) to QA-TriR-F (**11**). We also demonstrated that production of **11** was dependent on the presence of both ATCV-1 UGD and the *Qs0321910* SDR, consistent with the anticipated pathway (Fig. 6B). Interestingly, while co-incubation of UGT74BX1 and SDR alone did not result in any new products, the combination of UGT74BX1 and ATCV-1 UGD resulted in

partial conversion of **9** to a product with a mass consistent with addition of 4-keto-6-deoxy-D-glucose (hereby abbreviated to QA-TriR-4K6DG), suggesting that the UGT74BX1 can utilize UDP-4-keto-6-deoxy-D-glucose as a substrate (Fig. 6B). Furthermore, LC-MS analysis of leaf extracts from plants transiently expressing the QA-TriX-F (**10**) gene set (with the SDR excluded) also revealed a peak with a mass consistent with QA-TriX-4K6DG. This peak was larger than the QA-TriX-F (**10**) peak (Fig. S20). A medicagenic acid-3-*O*-glucuronide (MA-GlcA) C-28 D-fucosyltransferase (SOAP6 - UGT74BB2) was recently described from spinach (*9*). Transient expression of this enzyme in *N. benthamiana* along with the relevant MA-GlcA enzymes has been reported to yield a new product with the mass of MA-GlcA plus a deoxyhexose (*9*), suggesting a similar phenomenon may be occurring.

We next attempted to ascertain a direct link between the *Qs0321910* SDR and UDP-D-fucose production *in vitro.* To this end, UDP-D-glucose was first incubated with ATCV-1 UGD and the reaction was monitored by NMR. Initially, as anticipated, we observed the formation of UDP-4-keto-6-deoxy-D-glucose. However, addition of the SDR did not result in production of UDP-D-fucose (Fig. S21), despite the clear evidence that the purified *Qs0321910* SDR was functional in our initial *in vitro* experiment. Furthermore, we were unable to detect UDP-D-fucose following transient expression of the SDR in *N. benthamiana* (Fig. S22). The identification of these sugar nucleotides was confirmed with NMR-verified standards.

The failure of the *Qs0321910* SDR to convert UDP-4-keto-6-deoxy-D-glucose to UDP-D-fucose *in vitro* coupled with the observation that UGT74BX1 appears to utilize UDP-4-keto-6-deoxy-D-glucose as a substrate suggested that our initial model was incorrect. We therefore considered a new model in which UDP-4-keto-6-deoxy-D-glucose may serve as a sugar donor for UGT74BX1, forming the QA-TriR-4K6DG product. The 4-keto-6-deoxy-D-glucose (attached to the QA-TriR) would then be reduced at the C-4 position to give the observed QA-TriR-F (**11**) product. To test this, we performed a modified version of our initial *in vitro* assay by combining QA-TriR with UDP-D-glucose, ATCV-1 UGD and UGT74BX1 . As before, we observed conversion of QA-TriR to a new product consistent with addition of 4-keto-6-deoxy-hexose (QA-TriR-4K6DG). We next heat-inactivated the ATCV-1 UGD/UGT74BX1 enzyme mix prior to addition of the *Qs0321910* SDR. Subsequent LC-MS analysis showed the conversion of the putative QA-TriR-4K6DG to a new product identified as QA-TriR-F (Fig. 6C). Our results indicate that the SDR encoded by *Qs0321910* SDR does not operate at the sugar nucleotide level, but rather reduces 4-keto-6-deoxy-D-glucose to D-

fucose after transfer to the QA-TriR backbone (Fig. 6D). We therefore named this SDR as QsFucSyn.

Three further steps for production of QS-7

The enzymes discovered up until this point allow us to make the advanced saponin pathway heptasaccharide intermediates **16**, **17**, **18** and **19 (**Fig. 5**)**. We next searched for the steps needed to make QS-7 (Fig. 1A). Three additional modifications to the C-28 sugar chain are needed to convert **18** into QS-7, specifically addition of two sugars (L-rhamnose and D-glucose) and an acetyl group (Fig. 7). During our screen for the terminal C-28 glycosyltransferases, we detected putative glucosyltransferase activity for *Qs0321940* (*UGT91AP1*), and co-expression with the enzyme set for **18** resulted in a product anticipated to be the glucoside of **18** (Fig. S23). The gene encoding this enzyme is located within the chromosome 11 biosynthetic gene cluster #45 (Fig. 2D) and is co-expressed with the known QS genes (ranked seventh in Table S2; co-expression with *QsbAS1* PCC 0.956). *Qs0321940* may therefore encode a glucosyltransferase implicated in QS-7 biosynthesis (Fig. S23). Two more steps would then be required to achieve biosynthesis of QS-7, namely addition of an L-rhamnose and an acetyl group at the C-3 and C-4 positions of D-fucose, respectively. Based on the structures of known saponins from *Q. saponaria*, acetylation appears to precede rhamnosylation (Fig. S24) (*24*). We shortlisted 10 candidate *Q. saponaria* BAHD acyltransferase genes based on levels of co-expression with *QsbAS1* (PCC ≥ 0.9, TPM ≥1600) (Table S13), successfully cloned and screened seven for activity towards the full heptasaccharide scaffold **18**, and identified a single enzyme (encoded by *Qs0206480*, PCC 0.900) that generated a product with a mass consistent with addition of an acetyl group (Fig. S25). *Qs0206480* (*QsACT1*) is located on chromosome 13 and is not clustered with any of the previously characterised genes. We next screened the remaining unassigned UGT candidates for the ability to modify this putative acetylated substrate and identified two enzymes that gave products that likely corresponded to addition of either L-rhamnose or D-glucose (encoded by *Qs0023500* (*UGT73B44*, ranked seventeenth in Table S2) and *Qs0213660 (UGT73B43*, ranked twentieth in Table S2), respectively) (Fig. S26). These two UGTs belong to subgroup D of the UGT1 family and share 72% amino acid sequence identity. The genes encoding them are not located in predicted biosynthetic gene clusters. Co-expression of these two enzymes together with the gene set for **18** did not result in a product featuring both sugars, suggesting that the UGTs compete for the same position (Fig. S26). Indeed, saponins featuring either L-rhamnose or D-glucose at the C-3 position of fucose have

been isolated from *Q. saponaria* (*24,34,35*), with QS-7 featuring L-rhamnose. This therefore strongly suggested that *Qs0023500* rhamnosyltransferase is the last outstanding step for QS-7 biosynthesis. We therefore co-expressed the gene set for (**18**) with the newly discovered candidate glucosyltransferase (*Qs0321940*), acetyltransferase (*Qs0206480*), and rhamnosyltransferase (*Qs0023500*) genes. Subsequent LC-MS analysis revealed a small peak with the same retention time and mass as a QS-7 standard (Fig. 7). Quantification of the QS-7 levels in *N. benthamiana* (7.9 μg per gram dry leaf weight) revealed them to be comparable to those found in many tissues of *Q. saponaria* with the exception of bark, which was around 3-fold higher (Fig. S27). Following large scale infiltration of 410 *N. benthamiana* plants and fractionation by reversed phase HPLC (see Materials and Methods) approximately 11 mg of semi-pure (3-5%) QS-7 was obtained. Subsequent 1D- and 2D-NMR analysis enabled us to assign the structure of this compound as QS-7 (**20**) based on comparison with published data (*34*) (Figs. S28-43). Furthermore, our recorded ^{1}H-NMR spectrum showed complete superimposition with the chemical shifts of a pure QS-7 standard under identical conditions (Figs. S44, 45). Together these results demonstrate the successful elucidation of the QS-7 pathway and its reconstitution in a heterologous host.

Conclusion

Here we report the characterization of a total of 14 *Q. saponaria* enzymes that enable the biosynthesis of the advanced heptasaccharide triterpene glycoside intermediates **16** , **17** , **18** and **19**. We further identify two other enzymes required for efficient glycosylation with the rare sugars, D-fucose and D-apiose. A biosynthetic pathway for D-fucose had not previously been characterized, despite the widespread occurrence of this sugar in the plant kingdom (*21*). We initially expected the glycosyltransferase UGT74BX1 to add D-fucose to QA-TriR (**9**) (Fig. 5), yet we and others found no evidence of this sugar nucleotide in representative dicot plants (*36*). Here we provide evidence for a different route to D-fucosylation in which UDP-4-keto-6-deoxy-glucose serves as the sugar donor for UGT74BX1, the 4-keto-6-deoxy-D-glucose moiety attached at the C-28 position of the saponin scaffold then being subsequently reduced *in situ* to yield D-fucose. This discovery raises broader questions about the origin of D-fucose moieties found in other plant natural products [e.g. foxglove cardiac glycosides (*37*)].

Using our transient plant expression platform, we have been able to purify all of the QS pathway intermediates from QA to QA-TriR-FRXA in milligram quantities (in some cases tens of milligrams), demonstrating the power of transient plant expression for rapid access to

these molecules. We further demonstrate the production of the vaccine adjuvant QS-7 (**20**). QS-7, unlike QS-21, has negligible toxicity towards animal cells (*1*). However, despite its promise as an adjuvant, supply of this saponin is limited by its low abundance in *Q. saponaria* bark extracts. Although the levels of QS-7 in *N. benthamiana* were also low, our work opens up for the first time the possibility of producing QS-7 and other related QS molecules in a heterologous expression system. Clearly optimization of the biosynthetic process with the aim of attaining commercial scale production levels is beyond the scope of this current work, but our results now make this a very attractive ambition. The availability of the complete genome sequence and comprehensive transcriptome resources for *Q. saponaria* now opens up opportunities to use this 'instruction manual' to access QS-21 and a diverse array of other QS saponins. Collectively these advances will enable investigation of the poorly understood relationship between QS saponin structure and adjuvant activity, and ultimately the generation of designer saponins with optimal immunostimulatory activity and low toxicity through metabolic engineering approaches.

Materials and methods

Detailed materials and methods can be found in the supplementary materials.

Quillaja saponaria plant material and saponin quantification

A *Quillaja saponaria* sapling (approximately 1 m high) was obtained from Burncoose Nurseries, Cornwall, UK and maintained in a glasshouse (24°C, 16 h light). We named this accession S10. Extracts (80% methanol) of freeze-dried tissues (young, mature and old leaves, primordium, green stem, bark and root, with four biological replicates) were analysed using a Thermo Scientific QExactive Hybrid Quadrupole-Orbitrap Mass spectrometer HPLC and saponin content determined relative to standard curves generated using purified QS-7 and QS-21 samples obtained from Desert King (San Diego, CA, USA).

Generation of sequence resources for Q. saponaria

Genes for the biosynthesis of quillaic acid were identified by mining the assembled 1KP transcriptome derived from *Q. saponaria* leaves (downloaded from http://www.onekp.com/public_data.html) for candidate OSC and CYP sequences using BLASTP. For discovery of the remaining QS pathway genes we generated *de novo* transcriptome data for six different *Q. saponaria* tissues using Illumina HiSeq4000 PE150, and a draft genome assembly using PacBio Sequel sequencing. A Hi-C library was prepared

using the Phase Genomics Plant Hi-C 2.0 Kit (Seattle, WA) and sequenced with Illumina PE75. The draft contig assembly was scaffolded into 14 pseudomolecules by Phase Genomics Proximo software. Following RNASeq guided genome annotation, the completeness of the gene space was assessed by BUSCO analysis (*38*).

Cloning and transient expression

Oligonucleotide primers were designed based on predicted gene sequences and flanked with attB sites for Gateway cloning (Data S1). RNA extracted from primordia and young leaves was used for cDNA synthesis. RNA isolation was carried out using a Qiagen RNeasy® Plant Mini kit with the modified protocol according to (*39*). Candidate sequences were amplified, cloned into pDONR207 using BP clonase (ThermoFisher) and sequenced (Eurofins), before being introduced into the binary expression vector pEAQ-*HT*-DEST1 (*40*) for transient expression in *N. benthamiana*. For ease of performing infiltrations, in some cases, multiple genes were incorporated into a single binary vector using Golden Gate cloning (*41, 42*). For screening of candidate genes, agro-infiltrations were performed at small-scale using a needleless syringe (*43*,*44*). For purification of compounds, large-scale vacuum infiltrations were performed as described previously (*44*). Leaf material was harvested five days after infiltration and frozen at -80°C prior to lyophilization for 24-72 hours. All experiments included co-expression of the truncated feedback-insensitive mevalonate pathway enzyme 3-hydroxy-3-methylglutaryl-CoA reductase (tHMGR) to boost triterpene yield (*44*).

Metabolite analysis

Standards were obtained from the following sources: oleanolic acid (Merck); echinocystic acid (Extrasynthese); quillaic acid (Extrasynthese); QS-7 and QS-21 (Desert King). Internal standards coprostanol (GC-MS) and digitoxin (LC-MS) were obtained from Merck. Leaf extracts were analysed by GC-MS or LC-MS, depending on the polarity of the compounds under investigation. Full details of the methods used for metabolite analysis, scale-up and purification of compounds for structural determination by NMR, investigation of QsFucSyn activity and sugar nucleotide analysis are provided in the supplementary materials.

References and Notes

1. C. R. Kensil, U. Patel, M. Lennick, D. Marciani, Separation and characterization of saponins with adjuvant activity from *Quillaja saponaria* Molina cortex. *J. Immunol.* **146**, 431-437 (1991).
2. A. King, Soapbark reaches out to fill essential role in some vaccine recipes. *Chemistry World* 21 June 2022, https://www.chemistryworld.com/news/soapbark-branches-out-to-fill-essential-role-in-vaccine-recipes/4015836.article
3. G. C. Kite, M. J. Howes, M. S. Simmonds, Metabolomic analysis of saponins in crude extracts of *Quillaja saponaria* by liquid chromatography/mass spectrometry for product authentication. *Rapid Commun. Mass Spectrom.* **18**, 2859-2870 (2004).
4. A. S. Grandon, B. M. Espinosa, D. L. Rios, O. M. Sanchez, C. K. Saez, S. V. Hernandez, A. J. Becerra, Variation of saponin contents and physiological status in *Quillaja saponaria* under different environmental conditions. *Nat. Prod. Commun.* **8**, 1697-1700 (2013).
5. L. P. Iglesias, J. G. Castro, V. A. Artze-Vargas, R. O. Peredo, Production of biomass in ultra high density plantations. U.S. Patent No. 11,254,699 (2022).
6. S. Copaja, C. Blackburn, R. Carmona, Variation of saponin contents in *Quillaja saponica* Molina. *Wood Sci. Technol.* **37**, 103-108 (2003).
7. K. Miettinen K, S. Iñigo, L. Kreft, J. Pollier, C. De Bo, A. Botzki, F. Coppens, S. Bak, A. Goossens, The TriForC database: a comprehensive up-to-date resource of plant triterpene biosynthesis. *Nuc. Acids Res.* **46**, (D1), D586–D594 (2018).
8. K. Malhotra, J. Franke, Cytochrome P450 monooxygenase-mediated tailoring of triterpenoids and steroids in plants. *Beilstein J. Org. Chem.* **18**, 1289–1310 (2022).
9. A. Jozwiak, P. D. Sonawane, S. Panda, C. Garagounis, K. K. Papadopoulou, B. Abebie, H. Massalha, E. Almekias-Siegl, T. Scherf, A. Aharoni, Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery. *Nat. Chem. Biol.* **16**, 740-748 (2020).
10. Y. Li, 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. Karafiatova, J. Vrana, 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. Dolezel, T. Langdon, B. Han, A. Osbourn, Subtelomeric assembly of a multi-gene

pathway for antimicrobial defense compounds in cereals. *Nat Commun* **12**, 2563 (2021).

11. R. Xu, G. C. Fazio, S. P. Matsuda, On the origins of triterpenoid skeletal diversity. *Phytochemistry* **65**, 261-291 (2004).
12. J. H. Leebens-Mack *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679-685 (2019).
13. T. Schlotterbeck, M. Castillo–Ruiz, H. Cañon–Jones, R. S. Martín, The use of leaves from young trees of *Quillaja saponaria* (Molina) plantations as a new source of saponins. *Econ. Bot.* **69**, 262-272 (2015).
14. H. Hayashi, P. Huang, A. Kirakosyan, K. Inoue, N. Hiraoka, Y. Ikeshiro, T. Kushiro, M. Shibuya, Y. Ebizuka, Cloning and characterization of a cDNA encoding β-amyrin synthase involved in glycyrrhizin and soyasaponin biosyntheses in licorice. *Biol. Pharm. Bull.* **24**, 912-916 (2001).
15. K. Miettinen, 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, A. Goossens, The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. *Nat. Commun.* **8**, 14153 (2017).
16. M. Carelli, E. Biazzi, F. Panara, A. Tava, L. Scaramelli, A. Porceddu, N. Graham, M. Odoardi, E. Piano, S. Arcioni, S. May, C. Scotti, O. Calderini, *Medicago truncatula* CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *Plant Cell* **23**, 3070-3081 (2011).
17. G. Polturak, Z. Liu, A. Osbourn, New and emerging concepts in the evolution and function of plant biosynthetic gene clusters. *Curr. Opin. Green Sustain. Chem.* **33**, 100568 (2022).
18. S. Garcia, T. Garnatje, O. Hidalgo, G. Mas de Xaxars, J. Pellicer, I. Sánchez-Jiménez, D. Vitales, J. Vallès, First genome size estimations for some eudicot families and genera. *Collect. Bot.* **29**, 7-16 (2010).
19. S. B. Cannon, M. R. McKain, A. Harkess, M. N. Nelson, S. Dash, M. K. Deyholos, Y. Peng, B. Joyce, C. N. Stewart, Jr, M. Rolf, T. Kutchan, X. Tan, C. Chen, Y. Zhang, E. Carpenter, G. K.-S. Wong, J. J. Doyle, J. Leebens-Mack, Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.* **32**, 193-210 (2014).

20. S. A. Kautsar, H. G. Suarez Duran, K. Blin, A. Osbourn, M. H. Medema, plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res*. **W1**, W55-W63 (2017).
21. T. Louveau, A. Osbourn, The sweet side of plant-specialized metabolism. *Cold Spring Harb. Perspect. Biol.* **11**, (2019).
22. S. Rahimi, J. Kim, I. Mijakovic, K. Jung, G. Choi, S.-C. Kim, Y.-J. Kim, Triterpenoid-biosynthetic UDP-glycosyltransferases from plants. *Biotechnol. Adv.* **37**, 107394 (2019).
23. S. Y. Chung, H. Seki, Y. Fujisawa, Y. Shimoda, S. Hiraga, Y. Nomura, K. Saito, M. Ishimoto, T. Muranaka, A cellulose synthase-derived enzyme catalyses 3-*O*-glucuronosylation in saponin biosynthesis. *Nat. Commun.* **11**, 5664 (2020).
24. J. D. Fleck, A. H. Betti, F. P. da Silva, E. A. Troian, C. Olivaro, F. Ferreira, S. G. Verza, Saponins from *Quillaja saponaria* and *Quillaja brasiliensis*: Particular chemical characteristics and biological activities. *Molecules* **24**, 171 (2019).
25. Y. Yin, J. Huang, X. Gu, M. Bar-Peled, Y. Xu, Evolution of plant nucleotide-sugar interconversion enzymes. *PLoS One* **6**, e27995 (2011).
26. Y. Kallberg, U. Oppermann, B. Persson, Classification of the short-chain dehydrogenase/reductase superfamily using hidden Markov models. *FEBS J.* **277**, 2375-2386 (2010).
27. Y. Yoshida, Y. Nakano, T. Nezu, Y. Yamashita, T. Koga, A novel NDP-6-deoxyhexosyl-4-ulose reductase in the pathway for the synthesis of thymidine diphosphate-d-fucose. *J. Biol. Chem.* **274**, 16933-16939 (1999).
28. T. Oka , T. Nemoto, Y. Jigami, Functional analysis of *Arabidopsis thaliana* RHM2/MUM4, a multidomain protein involved in UDP-D-glucose to UDP-L-rhamnose conversion, *J.B.C.* **282**, 5389–5403 (2007).
29. H. Moummou, Y. Kallberg, L.B. Tonfack, B. Persson, B.Van der Rest. The plant short-chain dehydrogenase (SDR) superfamily: genome-wide inventory and diversification patterns. *BMC Plant Biol.* **12**, 219 (2012).
30. H. W. Choi, B.-G. Lee, N. Kim, Y. Park, C. W. Lim, H. K. Song, B. Hwang, A Role for a Menthone Reductase in Resistance against Microbial Pathogens in Plants. *Plant Physiol*. **148**, 383-401 (2008).
31. T. Czechowski, E. Forestier, S. H. Swamidatta, A. D. Gilday, A. Cording, T. R. Larson, D. Harvey, Y. Li, Z. He, A. J. King, G. D. Brown, I. A. Graham, Gene

discovery and virus-induced gene silencing reveal branched pathways to major classes of bioactive diterpenoids in Euphorbia peplus. *PNAS* **119**, e2203890119 (2022).

32. J. Ziegler, S. Voigtländer, J. Schmidt, R. Kramell, O. Miersch, C. Ammer, A. Gesell, T. M. Kutchan, Comparative transcript and alkaloid profiling in Papaver species identifies a short chain dehydrogenase/reductase involved in morphine biosynthesis. *Plant J.* **48**, 177-192 (2006).
33. M. P. Chothi, G. A. Duncan, A. Armirotti, C. Abergel, J. R. Gurnon, J. L. Van Etten, C. Bernardi, G. Damonte, M. Tonetti, Identification of an L-rhamnose synthetic pathway in two nucleocytoplasmic large DNA viruses. *J. Virol.* **84**, 8829-8838 (2010).
34. S. Guo, L. Kenne, Characterization of some *O*-acetylated saponins from *Quillaja saponaria* Molina. *Phytochemistry* **54**, 615-623 (2000).
35. S. Guo, L. Kenne, Structural studies of triterpenoid saponins with new acyl components from *Quillaja saponaria* Molina. *Phytochemistry* **55**, 419-428 (2000).
36. M. Pabst, J. Grass, R. Fischl, R. Léonard, C. Jin, G. Hinterkörner, N. Borth, F. Altmann, Nucleotide and Nucleotide Sugar Analysis by Liquid Chromatography-Electrospray Ionization-Mass Spectrometry on Surface-Conditioned Porous Graphitic Carbon. *Anal. Chem.* **82**, 9782-9788 (2010).
37. W. Kreis, A. Hensel, U. Stuhlemmer, Cardenolide biosynthesis in foxglove 1. *Planta Medica* **64** 491-499 (1998).
38. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, E. M. Zdobnov, BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647-4654 (2021).
39. D. J. MacKenzie, M. A. McLean, S. Mukerji, M. Green, Improved RNA extraction from woody plants for the detection of viral pathogens by reverse transcription-polymerase chain reaction. *Plant Dis.* **81**, 222-226 (1997).
40. F. Sainsbury, E.C. Thuenemann, G.P. Lomonossoff, pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682-693 (2009).
41. C. Engler, R. Kandzia, S. Marillonnet,. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3** e3647 (2008).

42. E. Weber, C. Engler, R. Gruetzner, S. Werner, S. Marillonnet, A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6**, e16765 (2011).

43. F. Sainsbury, P. Saxena, K. Geisler, A. Osbourn, G.P. Lomonossoff, Using a virus-derived system to manipulate plant natural product biosynthetic pathways. *Methods Enzymol* **517**, 185-202 (2012).

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

45. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-52 (2011).

46. B. J. Haas, 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, A. Regev, *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* **8**, 1494-512. (2013).

47. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **15**, 2114-2120 (2014).

48. D. L. Mapleson, L. Venturini, G. Kaithakottil, D. Swarbreck, Efficient and accurate detection of splice junctions from RNAseq with Portcullis. *GigaScience* **7**, (2018).

49. L. Venturini, S. Caim, G. Kaithakottil, D.L. Mapleson, D. Swarbreck, Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, giy093, doi:10.1093/gigascience/giy093 (2018).

50. M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).

51. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417-419 (2017).

52. A. Hallab, “Protein function prediction using phylogenomics, domain architecture analysis, data integration, and lexical scoring” dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn (2015).

53. S. Ou, W. Su, Y. Liao, K. Chougule, J. R. A. Agda, A. J. Hellinga, C. S. B. Lugo, T. A. Elliott, D. Ware, T. Peterson, N. Jiang, C. N. Hirsch, M. B. Hufford, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18 (2019).

54. M.-D. Rey, G. Moore, A. C. Martín, Identification and comparison of individual chromosomes of three accessions of *Hordeum chilense*, *Hordeum vulgare*, and *Triticum aestivum* by FISH. *Genome* **61**, 387–396 (2018).

55. R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

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

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

58. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

59. D. Hoffmeister, J. Yang, L. Liu, J. S. Thorson, Creation of the first anomeric D/L-sugar kinase by means of directed evolution. *PNAS* **100**, 13184 (2003).

60. M. P. Chothi, G. A. Duncan, A. Armirotti, C. Abergel, J. R. Gurnon, J. L. Van Etten, C. Bernardi, G. Damonte, M. Tonetti, Identification of an L-rhamnose synthetic pathway in two nucleocytoplasmic large DNA viruses. *J. Virol.* **84**, 8829 (2010).

61. J. C. Errey, B. Mukhopadhyay, K. P. R. Kartha, R. A. Field, Flexible enzymatic and chemo-enzymatic approaches to a broad range of uridine-diphospho-sugars. *Chem. Commun.* **23**, 2706 (2004).

62. V. Martinez, M. Ingwers, J. Smith, J. Glushka, T. Yang, M. Bar-Peled, Biosynthesis of UDP-4-keto-6-deoxyglucose and UDP-rhamnose in pathogenic fungi *Magnaporthe grisea* and *Botryotinia fuckeliana*. *J. Biol. Chem.* **287**, 879 (2012).

63. M. Rejzek, B. Mukhopadhyay, C. Q. Wenzel, J. S. Lam, R. A. Field, Direct oxidation of sugar nucleotides to the corresponding uronic acids: TEMPO and platinum-based procedures. *Carbohydr. Res.* 342, 460-466 (2007).

64. B. A. Wagstaff, M. Rejzek, S. Kuhaudomlarp, L. Hill, I. Mascia, S. A. Nepogodiev, H. C. Dorfmueller, R. A. Field, Discovery of an RmlC/D fusion protein in the microalga Prymnesium parvum and its implications for NDP-β-L-rhamnose biosynthesis in microalgae. *J. Biol. Chem.* **294**, 9172–9185 (2019).

65. R. Behmüller, I. C. Forstenlehner, R. Tenhaken, C. G. Huber, Quantitative HPLC-MS analysis of nucleotide sugars in plant cells following off-line SPE sample preparation. *Anal. Bioanal. Chem*. **406**, 3229-3237 (2014).

66. J. E. Lunn, R. Feil, Janneke H. M. Hendriks, Y. Gibon, R. Morcuende, D. Osuna, W.-R. Scheible, P. Carillo, M.-R. Hajirezaei, M. Stitt, Sugar-induced increases in trehalose 6-phosphate are correlated with redox activation of ADP-glucose pyrophosphorylase and higher rates of starch synthesis in Arabidopsis thaliana. *Biochem. J*. **397**, 139-148 (2006).

67. M. Rejzek, L. Hill, E. S. Hems, S. Kuhaudomlarp, B. A. Wagstaff, R. A. Field, Profiling of sugar nucleotides. *Methods Enzymol.* **597**, 209-238 (2017).

68. A. E. Wilson, L. Tian, Phylogenomic analysis of UDP-dependent glycosyltransferases provides insights into the evolutionary landscape of glycosylation in plant metabolism. *Plant J.* **100**, 1273-1288 (2019).

69. S. Guo, L. Kenne, L. N. Lundgren, B. Rönnberg, B. G. Sundquist, Triterpenoid saponins from *Quillaja saponaria*. *Phytochemistry.* **48**, 175-180 (1998).

70. H. Tang, J. E. Bowers, X. Wang, R. Ming, M. Alam, A. H. Paterson, Synteny and collinearity in plant genomes. *Science*. **320**, 486-488 (2008).

71. T. Czechowski, E. Forestier, S. H. Swamidatta, A. D. Gilday, A. Cording, T. R. Larson, D. Harvey, Y. Li, Z. He, A. J. King, G. D. Brown, I. A. Graham, Gene discovery and virus-induced gene silencing reveal branched pathways to major classes of bioactive diterpenoids in Euphorbia peplus. *PNAS* **119**, e2203890119 (2022).

Acknowledgments: We thank Drs Martin Stocks and Georgina Pope (PBL Technology) for advice and support; Rachel Melton and Mike Ambrose (John Innes Centre; JIC) for help in sourcing plant material; JIC Horticultural Services for assistance with plant cultivation; the JIC Metabolomics, NMR and Chemistry platforms for assistance with instruments and method development; and Norwich Bioscience Institutes (NBI) Research Computing for computational support. We thank Prof. David Nelson and to the UGT Nomenclature Committee for formal assignment of the *Q. saponaria* CYPs and UGTs, respectively. Finally,

we thank Professor Rob Field and our industrial collaborators for their comments and productive discussion.

Funding: This work has been supported by John Innes Centre Innovation Fund Award KEC IF29 2018 AO29 (AO), a Biotechnological and Biological Sciences Research Council (BBSRC) Super Follow-on-Fund award BB/R005508/1 (RCM, SK, AE-D, A Orme), Industrial funding (JR, RCM, SK, AE-D, CO, MR, A Orme), the joint Engineering and Physical Sciences Research Council/ Biotechnological and Biological Sciences Research Council (BBSRC)-funded OpenPlant Synthetic Biology Research Centre grant BB/ L014130/1 (MS, AO), the John Innes Foundation (CO, AO), and the BBSRC Institute Strategic Programme Grant 'Molecules from Nature – Products and Pathways' (BBS/E/J/000PR9790) (TL, AO).

Author contributions: JR and AO conceived and designed the project. QS-21 and QS-7 profiling of *Q. saponaria* tissues, QS-7 quantification in *N. benthamiana* and generation of transcriptome resources, LM; purification of genomic DNA, JR, LM; preparation of Hi-C library, AH; computational analysis of the genome assembly, CO, AH, JL-M; bioinformatics analysis (including gene discovery, co-expression analysis, phylogenetics), CO, A Orme, JR, LM, TL; cloning and screening of candidate enzymes, JR, A Orme, LM, TL, SK; generation of GoldenGate vectors, RCM; initial scale-up of quillaic acid production and NMR, JR, MS; full scale-up, purification and structural analysis of pathway intermediates and QS-7, AE-D; karyotyping, ACM; enzyme purification and *in vitro* glycosylation assays, SK; synthesis, purification and NMR analysis of sugar nucleotides, MR; sugar nucleotide profiling, MR, JR. JR and AO wrote the manuscript, with input from other authors.

Competing interests: JR, A Orme, TL, LM and AO are inventors of patents arising from this work.

Data and Materials Availability: The fully assembled and annotated *Q. saponaria* genome sequence has been deposited under NCBI BioProject ID PRJNA914519. RNASeq reads are deposited under NCBI BioProject ID PRJNA914309 (SRA accessions SRR22829626 - SRR22829649). The sequences of the genes characterized in this study can also be found in GenBank as the following: *QsbAS1* (*Qs0315350*), OQ107256; *CYP716A224* (*Qs0259300*), OQ107260; *CYP716A297* (*Qs0322000*), OQ107248; *CYP714E52* (*Qs0148690*), OQ107266; *CSLM1* (*Qs0321900*), OQ107253; *CSLM2* (*Qs0000870*), OQ107265; *UGT73CU3* (*Qs0123860*), OQ107259; *UGT73CX2* (*Qs0283850*), OQ107255; *UGT73CX1* (*Qs0283870*),

OQ107254; *UGT74BX1* (*Qs0321930*), OQ107250; *UGT91AR1* (*Qs0321920*), OQ107251; *UGT91AQ1* (*Qs0234120*), OQ107264; *UGT73CY3* (*Qs0234130*), OQ107263; *UGT73CY2* (*Qs0234140*), OQ107262; *UGT73B44* (*Qs0023500*), OQ107261; *UGT91AP1* (*Qs0321940*), OQ107249; *UGT73B43* (*Qs0213660*), OQ107257; *Apiose/xylose synthase* (*Qs0088320*), OQ107247; *QsFucSyn* (*Qs0321910*), OQ107252 and *QsACT1* (*Qs0206480*), OQ107258. All of the above *Q. saponaria* genes characterized in this study are available as expression constructs (either as DNA preparations or in the relevant microbial strains) from A. Osbourn under a material transfer agreement with Plant Bioscience Ltd.

Supplementary Materials

Materials and Methods

Figs. S1 to S53

Tables S1 to S16

References (*45-71*)

Data S1-S5

Figure legends

Fig. 1. Reconstitution of the steps to quillaic acid. (**A**) QS-7 and QS-21 share a core structure (shown in black) consisting of the triterpene scaffold quillaic acid, a branched trisaccharide at C-3 featuring D-glucuronic acid (D-GlcA), D-galactose (D-Gal) and D-xylose (D-Xyl) and a linear tetrasaccharide at C-28 featuring D-fucose (D-fuc), L-rhamnose (L-Rha), D-xylose and D-apiose (D-Api). This core structure is common to around a third of all reported QS saponins. Note: QS-21 variants also exist with L-rhamnose in place of D-xylose at C-3 (*) and D-xylose in place of D-apiose at C-28 (**). QS-17 is a glycosylated derivative of QS-21 (both have a D-glucose (D-Glc) attached to the L-rhamnose of the C-28 sugar chain (as shared with QS-7), while QS-17 also has an additional L-rhamnose attached to the L-arabinofuranose (L-Ara*f*) of the C-18 acyl chain). (**B**) LC-MS Extracted Ion Chromatograms (EIC) for *N. benthamiana* leaf extracts following co-expression of the β-amyrin synthase QsbAS1 with the CYPs CYP716A224 (a C-28 oxidase), CYP716A297 (a C16α oxidase) and CYP714E52 (a C-23 oxidase). The combination of all four enzymes results in the production of the QS scaffold, quillaic acid (QA) (*m/z* = 485) (**5**). Top, extract from control leaves that are not expressing the C-23 oxidase. (**C**) Biosynthetic route to QA: QsbAS1, β-amyrin synthase; CYP716A224, C-28 oxidase; CYP716A297, C16α oxidase; CYP714E52, C-23 oxidase. The structure of QA was confirmed by NMR (Fig. S4). Note that CYP714E52 was also found to be active on oleanolic acid. The resulting product is anticipated to be the C-23 aldehyde of oleanolic acid (gypsogenin) (Fig. S46).

Fig. 2. Generation of *Q. saponaria* genome and transcriptome sequences resources. (**A**) Karyotype analysis of *Q. saponaria* S10 meristem tissue at mitotic metaphase I, revealing 28 chromosomes. Scale bar = 5μm. (**B**) Circular synteny plot showing the 14 chromosomes of *Q. saponaria* S10. Syntenic blocks (indicated by the coloured lines) provide evidence of a whole genome duplication event. (**C**) Hierarchical clustering of the top 50 *Q. saponaria* genes that are co-expressed with *QsbAS1*, as calculated by Pearson Correlation Coefficient (PCC) value of Z-scores (generated from DESeq2 VST- transformed read quantification values). The four QA biosynthetic genes (labelled) show tight co-expression and are expressed most strongly in primordial tissue. PCC values for the three QA CYPs with QsbAS1 are shown to the right. (**D**) A biosynthetic gene cluster (#45) predicted by plantiSMASH is located on chromosome 11, very close to the QA biosynthesis gene *CYP716A297*. Several of the genes in this region also show high expression in primordial tissue.

Fig. 3. Addition of the C-3 sugar chain. (**A**) LC-MS Extracted Ion Chromatograms (EIC) of *N. benthamiana* leaf extracts showing that co-expression of either of the predicted cellulose synthase-like (CSL) genes *CSLM1* or *CSLM2* with the four QA genes results in the conversion of QA to a new more polar product (retention time 14 min). The mass spectra (right) indicate that the product is the same for *CSLM1* and *CSLM2*, and are consistent with addition of glucuronic acid to QA to form 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid (**6**) (Abbreviated to QA-Mono). IS, internal standard (digitoxin). (**B**) LC-MS EIC of *N. benthamiana* leaf extracts following co-expression with UGT candidates that add additional sugars at the C-3 position. A control sample from leaves expressing the QA pathway plus glucuronosyltransferase (CSLM1) is shown at the top. In the second panel, further co-expression of *Qs0123860* resulted in conversion of QA-Mono (**6**) to a new product consistent with addition of a galactose to form 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid}-quillaic acid (**7**) (Abbreviated to QA-Di). Co-expression of either *Qs0283870* (third panel) or *Qs0283850* (bottom panel) with the QA-Di gene set resulted in conversion of QA-Di (**7**) to new products. The *Qs0283870* product was consistent with addition of a xylose to form 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (**8**) (abbreviated to QA-TriX) while the *Qs0283850* product was consistent with addition of a rhamnose to form 3-*O*-{α-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (**9**) (abbreviated to QA-TriR). The mass spectra of these products and structures are shown to the right of the chromatograms. (**C**) Summary of the pathway from QA (**5**) to QA-TriX (**8**) and QA-TriR (**9**). The structures of compounds **6**-**9** were all confirmed by NMR following large-scale infiltration and purification (Tables S3-S7 and Data S4).

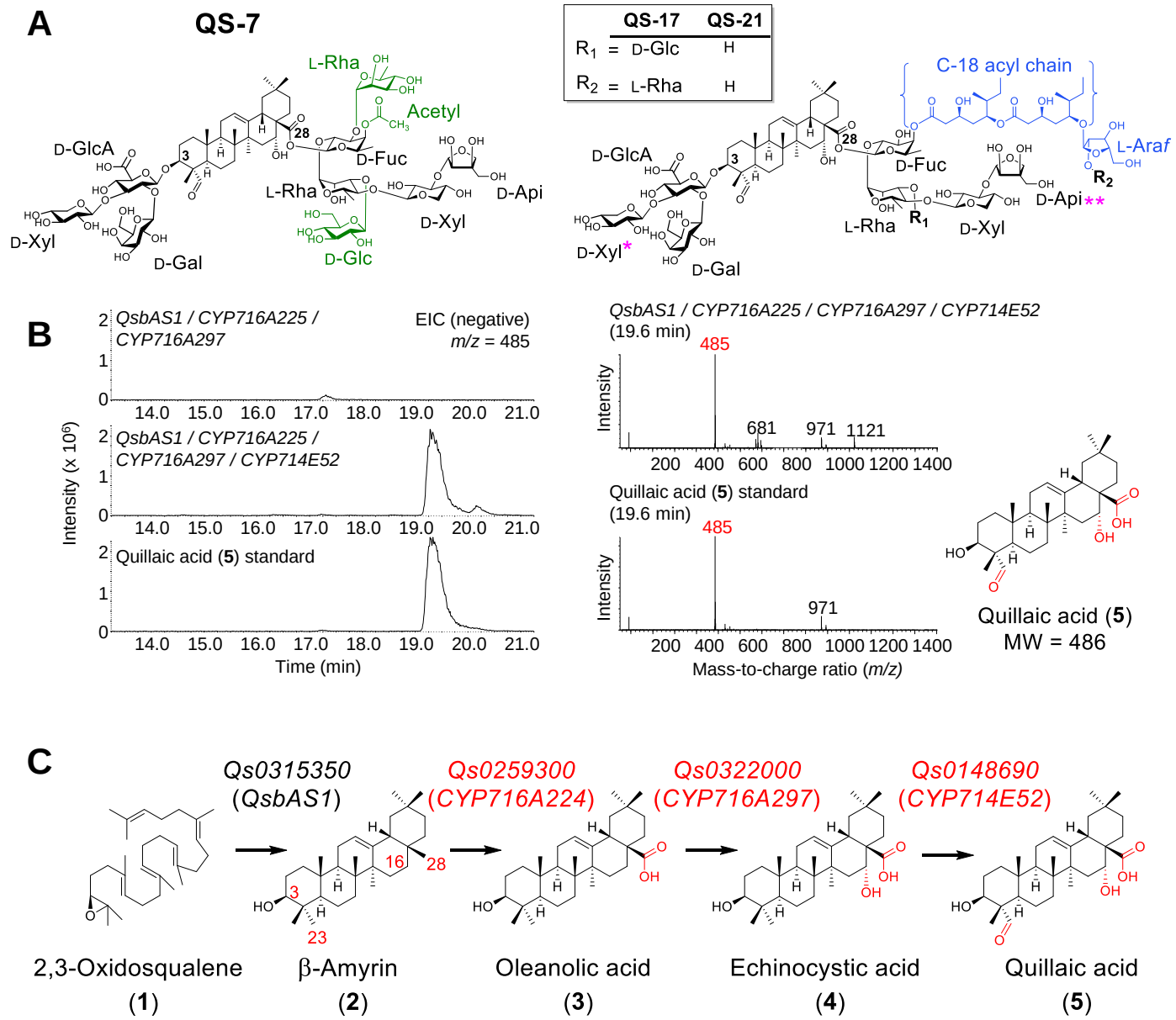
Fig. 4. Addition of the C-28 sugar chain. (**A**) LC-MS Extracted Ion Chromatograms (EIC) of *N. benthamiana* leaf extracts following transient expression of the gene set for production of the D-fucosylated saponin 3-*O*-{α-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-fucopyranosyl ester}-quillaic acid (**10**) (Abbreviated to QA-TriX-F). Only low amounts of **10** accumulate in *N. benthamiana* (top). However co-expression of the short chain dehydrogenase encoded by *Qs0321910* results in marked increases in the yield of this product (bottom), as well as increasing further downstream products (Fig. S15). IS, internal standard (digitoxin). (**B**) Identification of the terminal xylosyl- and apiosyltransferases required for synthesis of the linear tetrasaccharide at C-28. The gene set for production of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-

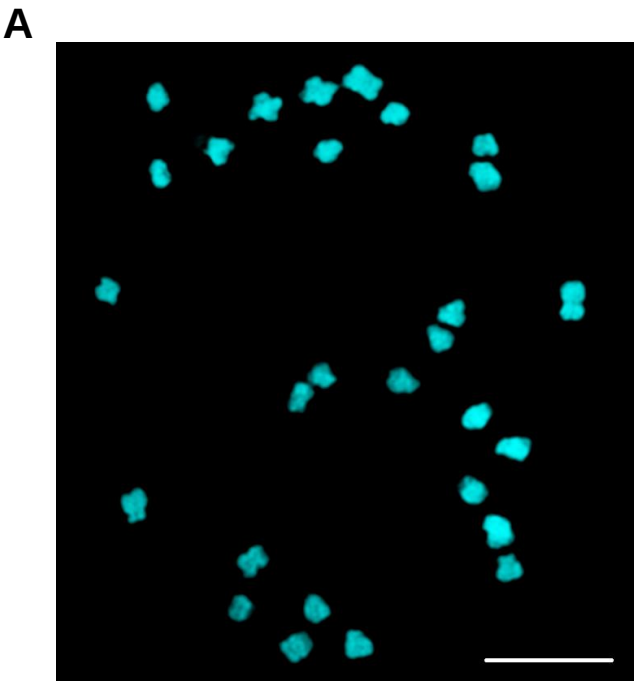
galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**14**) (abbreviated to QA-TriX-FRX) was transiently co-expressed in *N. benthamiana* (top). Further co-expression of either *Qs0234130* (middle) or *Qs0234140* (bottom) resulted in the appearance of new products with identical masses (consistent with addition of pentoses) and slight differences in retention times. Large scale infiltration, purification and NMR analysis of the products (using the QA-TriR-FRX (**15**) scaffold) determined that *Qs0234130* is the terminal xylosyltransferase, while *Qs0234140* is the terminal apiosyltransferase (Tables S9 and S10). (**C**) Summary of the biosynthetic pathway for the C-28 tetrasaccharide chain. The structures of compounds **11, 13, 15, 17** and **19** were all confirmed by NMR following large-scale infiltration and purification (Tables S8-S12).

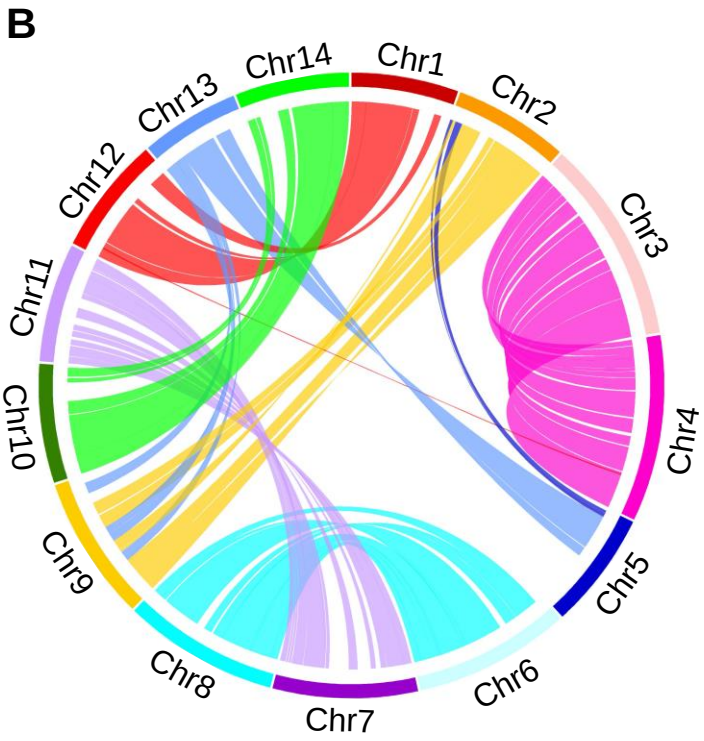
Fig. 5. The complete pathway to compounds 16-19. A summary giving full compound names, abbreviations, isolated yields, retention times and *m/z* data as well as the set of genes transiently expressed in *N. benthamiana* for each compound is provided as Data S4.

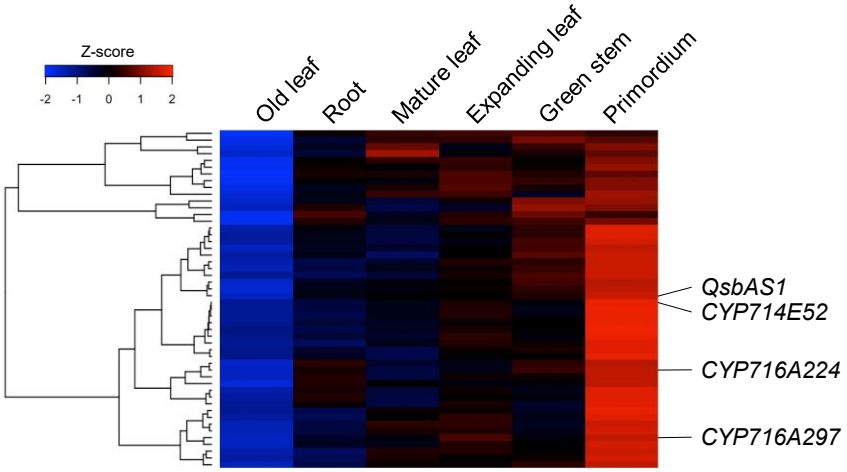
Fig. 6. Characterization of the SDR encoded by *Qs0321910*. (**A**) Biosynthesis of D-fucose in bacteria. dTDP-D-glucose is converted in a two-step process to dTDP-D-fucose via dTDP-4-keto-6-deoxy-D-glucose. (**B**) *In vitro* production of QA-TriR-F (**11**) from QA-TriR and UDP-D-glucose. Top, QA-TriR (**9**) incubated with UDP-D-glucose and UGT74BX1 only. Addition of ATCV-1 UGD (which converts UDP-D-glucose to UDP-4-keto-6-deoxy-glucose) resulted in new products anticipated to be QA-TriR-4-keto-6-deoxy-glucose (QA-TriR-4K6DG) (**) and its hydrate (*) (second from top). No conversion of QA-TriR was observed with the addition of the SDR alone (third from top). However, the combination of ATCV-1 UGD and SDR resulted in total conversion of QA-TriR (**9**) to QA-TriR-F (**11**). Mass spectra for the QA-TriR-4K6DG (**) and its hydrate (*) are shown in Fig. S47. (**C**) The SDR reduces the QA-TriR-4K6DG product to form QA-TriR-F (**11**). QA-TriR was incubated with UDP-D-glucose in the presence of ATCV-1 UGD and UGT74BX1, resulting in the formation of QA-TriR-4K6DG (**) and its hydrate (*). This enzyme mix was inactivated by boiling before addition of the SDR. LC-MS analysis of the reaction at 0 min (top), 60 min (middle) and 180 min (bottom) revealed that the formation of QA-TriR-F with consumption of QA-TriR-4K6DG (**) and hydrate form (*), demonstrating that the SDR reduces the 4-keto-6-deoxy-glucose attached to QA-TriR to form the D-fucose in QA-TriR-F (**11**). (**D**) Proposed biosynthetic pathway to QA-TriR-F (**11**) from QA-TriR (**9**) and UDP-D-glucose.

Fig. 7. Production of QS-7. The gene set for production of the core heptasaccharide QA-TriX-FRXA (**18**) was transiently co-expressed in *N. benthamiana* along with the *Qs0206480, Qs032140,* and *Qs0206480* genes. LC-HRMS of *N. benthamiana* leaf extracts revealed a peak with the exact mass and retention time of an authentic QS-7 (**20**) standard. This peak was absent if any one of the *Qs0206480, Qs032140,* and *Qs0206480* genes was omitted. Large scale infiltration and purification allowed the isolation of a small quantity of semi-pure QS-7 from *N. benthamiana* and structural confirmation by NMR (Tables S14-S15).

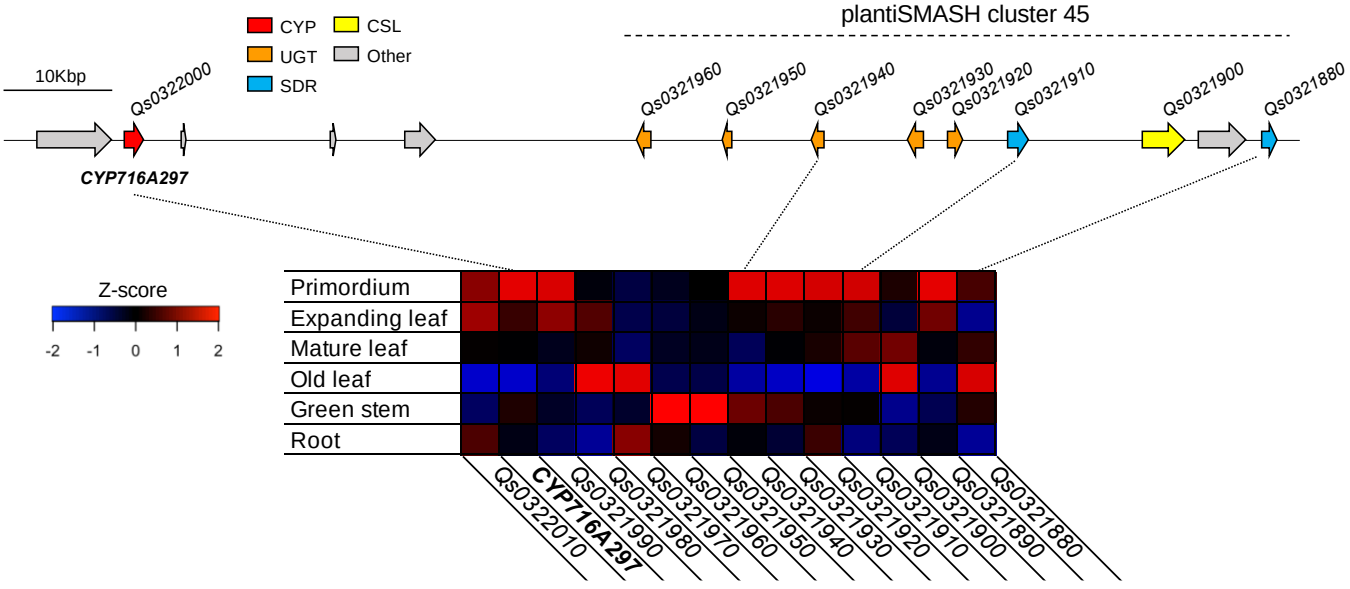

A
QS-7
L-Rha
Acetyl
D-GlcA
D-Fuc
L-Rha
D-Api
D-Xyl
D-Xyl
D-Gal
D-Glc
QS-17 QS-21
R1 = D-Glc H
R2 = L-Rha H
C-18 acyl chain
L-Araf
D-GlcA
D-Fuc
D-Api**
R2
L-Rha
R1
D-Xyl
D-Xyl*
D-Gal
B
QsbAS1 / CYP716A225 / CYP716A297
EIC (negative)
m/z = 485
QsbAS1 / CYP716A225 / CYP716A297 / CYP714E52
Quillaic acid (5) standard
Intensity (x 10^6)
Time (min)
QsbAS1 / CYP716A225 / CYP716A297 / CYP714E52
(19.6 min)
485
681
971
1121
Intensity
Quillaic acid (5) standard
(19.6 min)
485
971
Intensity
Mass-to-charge ratio (m/z)
Quillaic acid (5)
MW = 486
C
Qs0315350
(QsbAS1)
Qs0259300
(CYP716A224)
Qs0322000
(CYP716A297)
Qs0148690
(CYP714E52)
2,3-Oxidosqualene
(1)
β-Amyrin
(2)
Oleanolic acid
(3)
Echinocystic acid
(4)
Quillaic acid
(5)

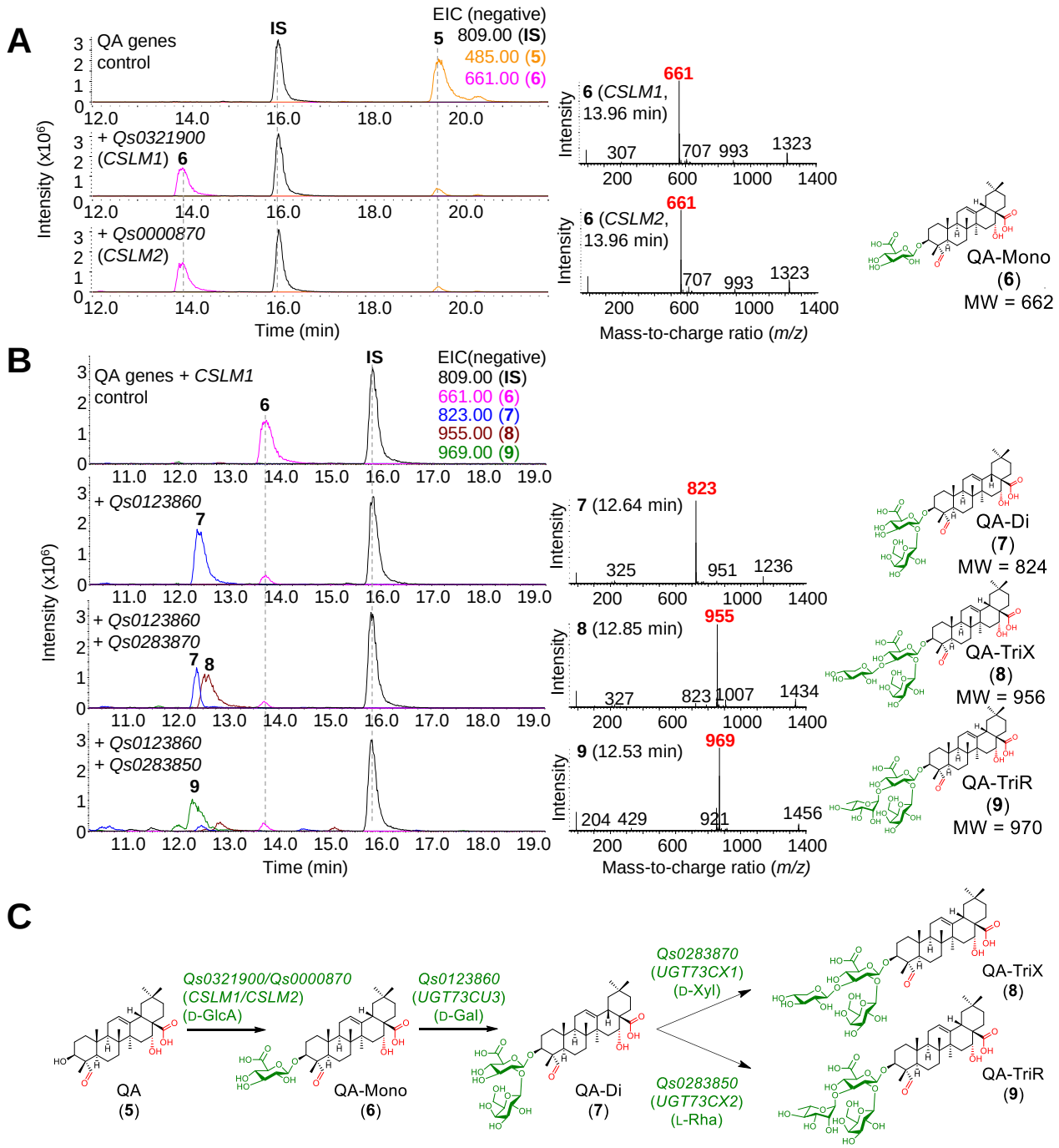


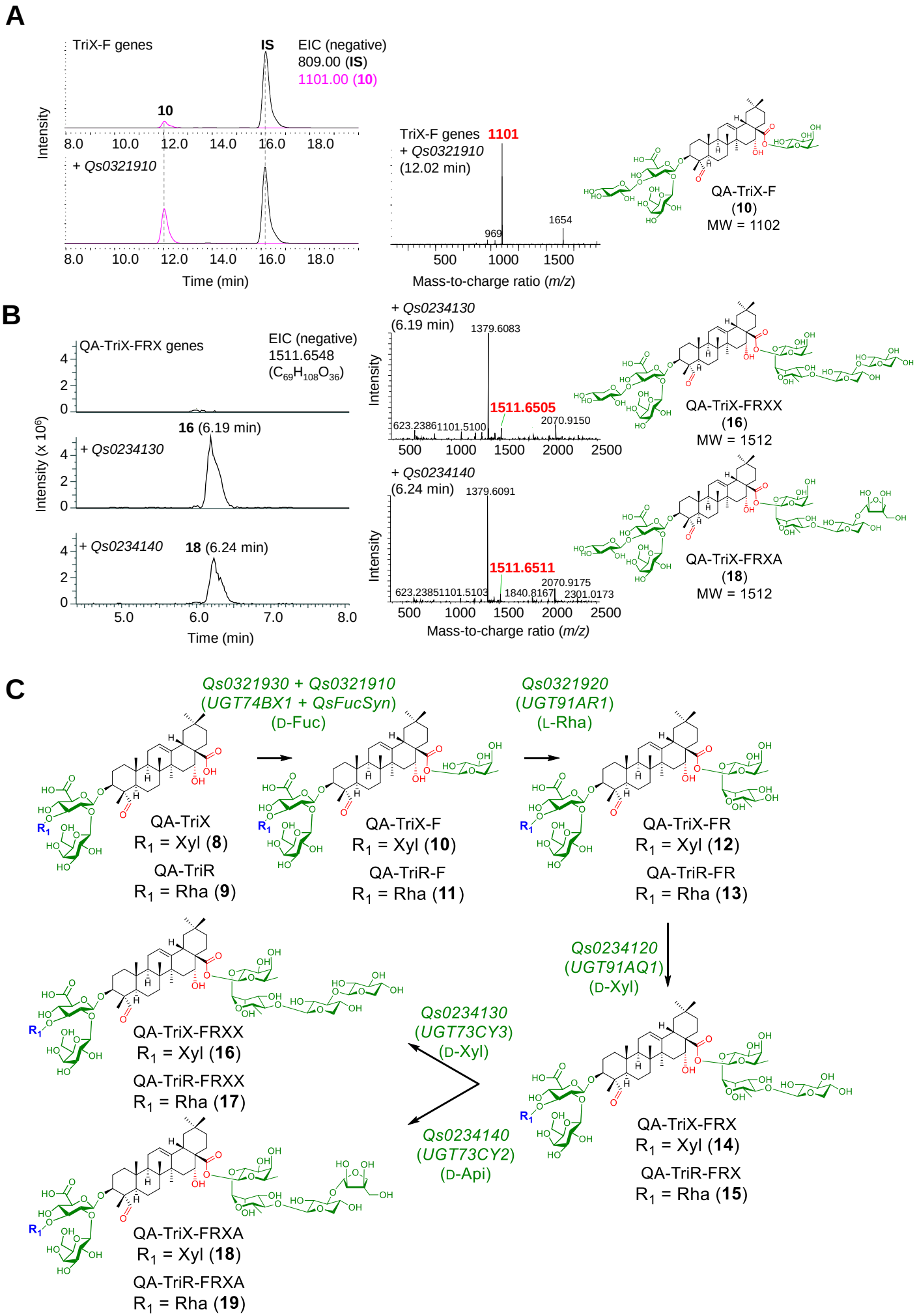


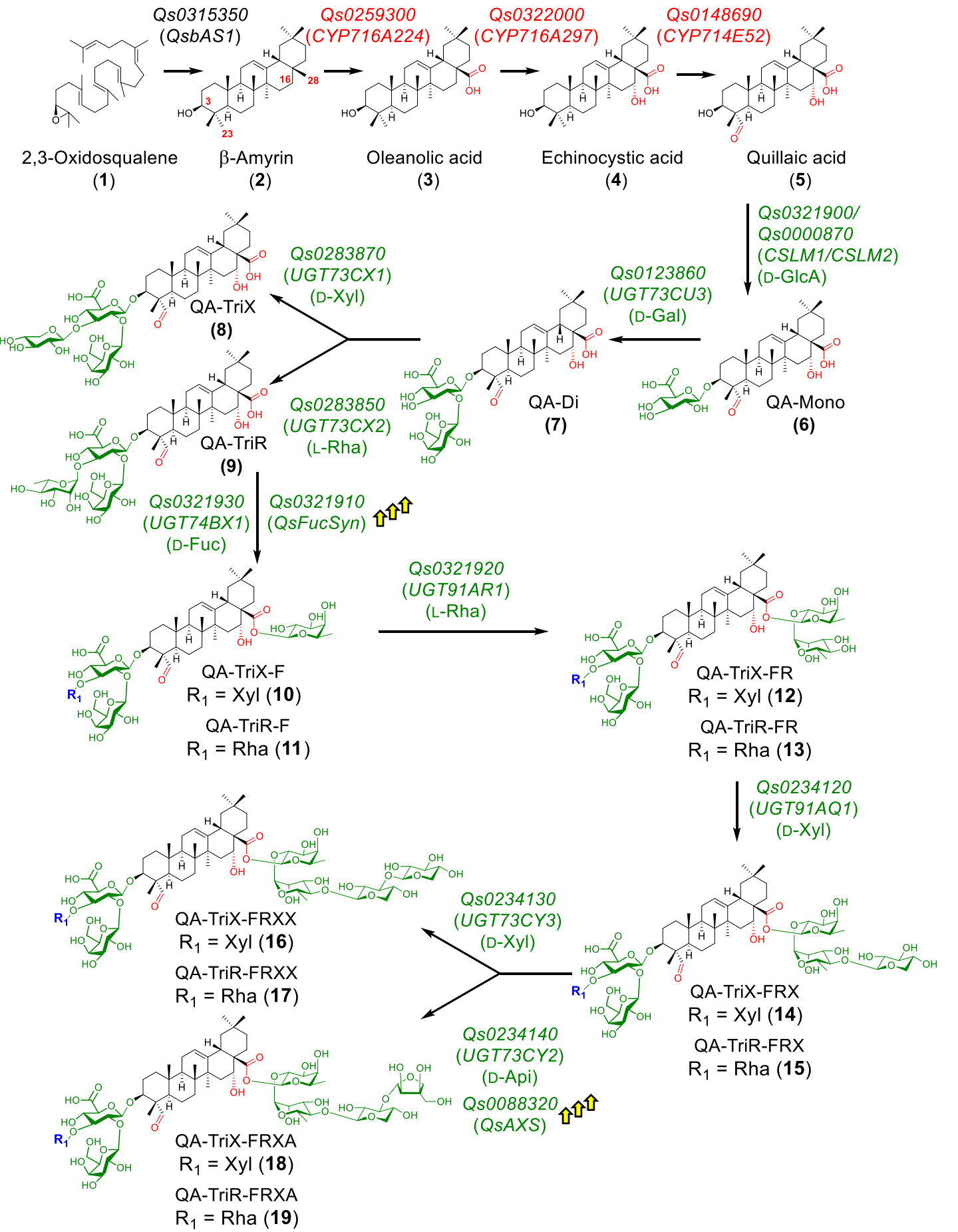


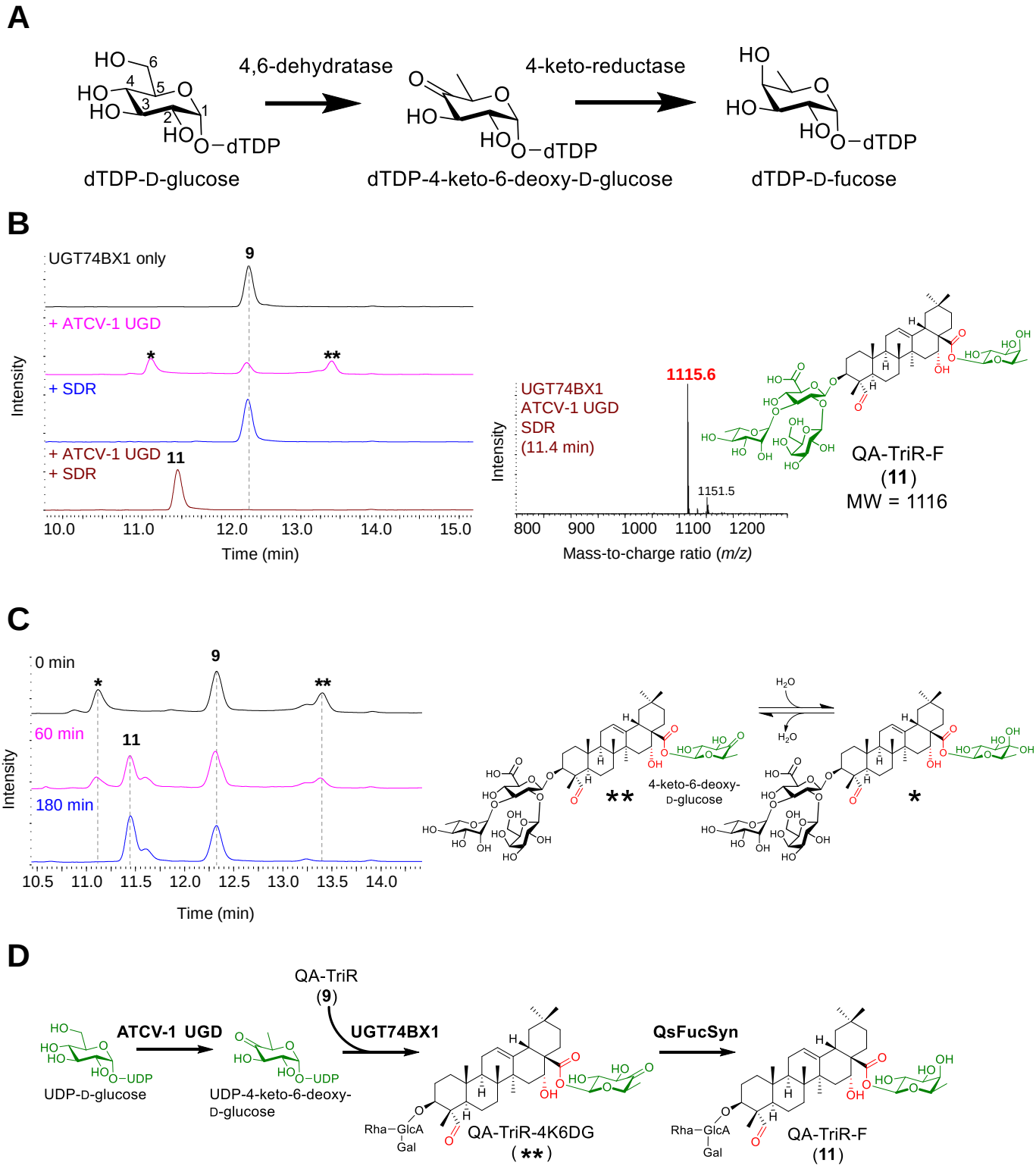
Gene ID	QsbAS1 co-expression (PCC)
CYP716A224	0.986
CYP716A297	0.990
CYP714E52	0.957

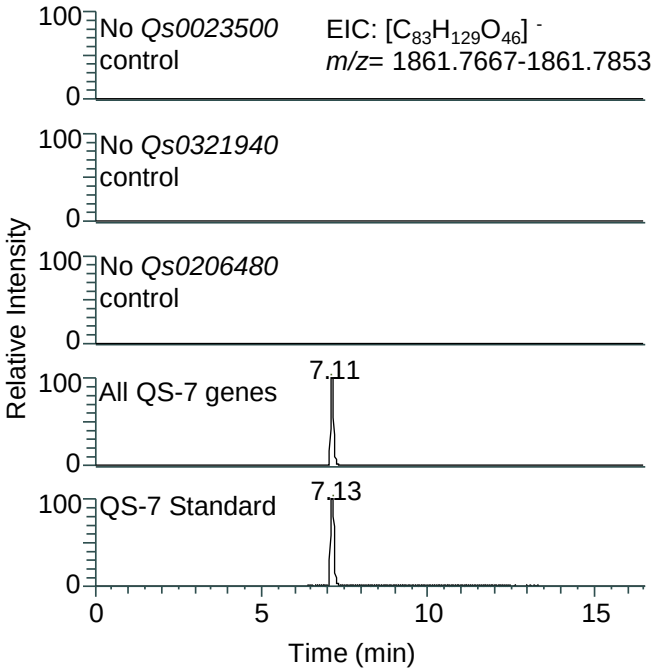


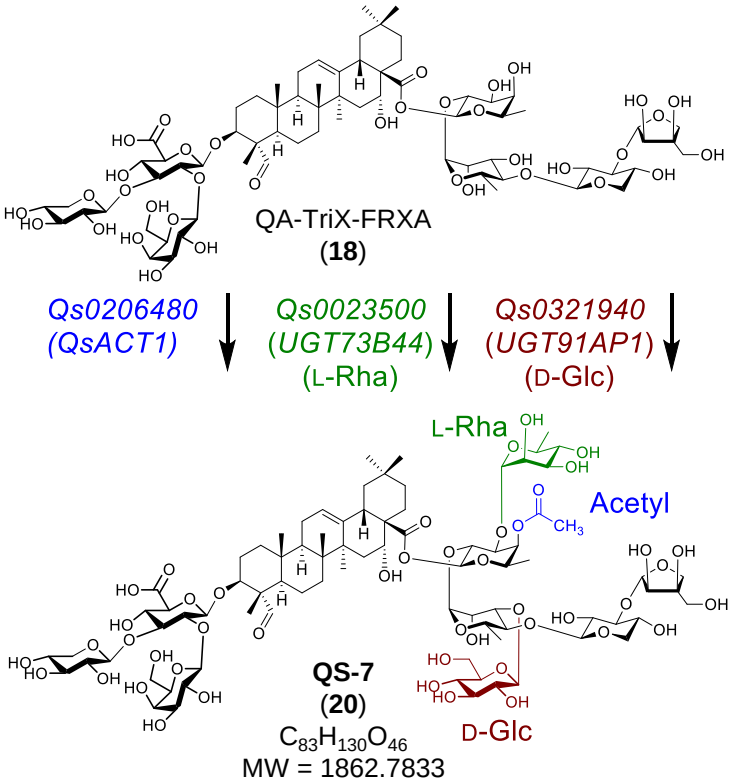

A
QA genes control
IS
EIC (negative)
809.00 (IS)
485.00 (5)
661.00 (6)
+ Qs0321900 (CSLM1)
+ Qs0000870 (CSLM2)
Intensity (x10⁶)
Time (min)
6 (CSLM1, 13.96 min)
6 (CSLM2, 13.96 min)
661
Mass-to-charge ratio (m/z)
QA-Mono (6)
MW = 662
B
QA genes + CSLM1 control
EIC(negative)
809.00 (IS)
661.00 (6)
823.00 (7)
955.00 (8)
969.00 (9)
+ Qs0123860
+ Qs0283870
+ Qs0283850
7 (12.64 min)
8 (12.85 min)
9 (12.53 min)
823
955
969
QA-Di (7)
MW = 824
QA-TriX (8)
MW = 956
QA-TriR (9)
MW = 970
C
Qs0321900/Qs0000870 (CSLM1/CSLM2) (D-GlcA)
Qs0123860 (UGT73CU3) (D-Gal)
Qs0283870 (UGT73CX1) (D-Xyl)
Qs0283850 (UGT73CX2) (L-Rha)
QA (5)
QA-Mono (6)
QA-Di (7)
QA-TriX (8)
QA-TriR (9)


A
TriX-F genes
IS
EIC (negative)
809.00 (IS)
1101.00 (10)
10
+ Qs0321910
Intensity
8.0 10.0 12.0 14.0 16.0 18.0
Time (min)
TriX-F genes 1101
+ Qs0321910
(12.02 min)
969
1654
500 1000 1500
Mass-to-charge ratio (m/z)
QA-TriX-F
(10)
MW = 1102
B
QA-TriX-FRX genes
EIC (negative)
1511.6548
(C69H108O36)
16 (6.19 min)
+ Qs0234130
18 (6.24 min)
+ Qs0234140
Intensity (x 10^6)
5.0 6.0 7.0 8.0
Time (min)
+ Qs0234130
(6.19 min)
1379.6083
1511.6505
623.2386 1101.5100
2070.9150
+ Qs0234140
(6.24 min)
1379.6091
1511.6511
623.23851101.5103
1840.8167
2070.9175
2301.0173
500 1000 1500 2000 2500
Mass-to-charge ratio (m/z)
QA-TriX-FRXX
(16)
MW = 1512
QA-TriX-FRXA
(18)
MW = 1512
C
Qs0321930 + Qs0321910
(UGT74BX1 + QsFucSyn)
(D-Fuc)
Qs0321920
(UGT91AR1)
(L-Rha)
QA-TriX
R1 = Xyl (8)
QA-TriR
R1 = Rha (9)
QA-TriX-F
R1 = Xyl (10)
QA-TriR-F
R1 = Rha (11)
QA-TriX-FR
R1 = Xyl (12)
QA-TriR-FR
R1 = Rha (13)
Qs0234120
(UGT91AQ1)
(D-Xyl)
Qs0234130
(UGT73CY3)
(D-Xyl)
Qs0234140
(UGT73CY2)
(D-Api)
QA-TriX-FRXX
R1 = Xyl (16)
QA-TriR-FRXX
R1 = Rha (17)
QA-TriX-FRX
R1 = Xyl (14)
QA-TriR-FRX
R1 = Rha (15)
QA-TriX-FRXA
R1 = Xyl (18)
QA-TriR-FRXA
R1 = Rha (19)


Qs0315350
(QsbAS1)
Qs0259300
(CYP716A224)
Qs0322000
(CYP716A297)
Qs0148690
(CYP714E52)
2,3-Oxidosqualene
(1)
β-Amyrin
(2)
Oleanolic acid
(3)
Echinocystic acid
(4)
Quillaic acid
(5)
Qs0321900/
Qs0000870
(CSLM1/CSLM2)
(D-GlcA)
Qs0123860
(UGT73CU3)
(D-Gal)
Qs0283870
(UGT73CX1)
(D-Xyl)
Qs0283850
(UGT73CX2)
(L-Rha)
QA-TriX
(8)
QA-Di
(7)
QA-Mono
(6)
QA-TriR
(9)
Qs0321930
(UGT74BX1)
(D-Fuc)
Qs0321910
(QsFucSyn)
QA-TriX-F
R1 = Xyl (10)
QA-TriR-F
R1 = Rha (11)
Qs0321920
(UGT91AR1)
(L-Rha)
QA-TriX-FR
R1 = Xyl (12)
QA-TriR-FR
R1 = Rha (13)
Qs0234120
(UGT91AQ1)
(D-Xyl)
Qs0234130
(UGT73CY3)
(D-Xyl)
QA-TriX-FRX
R1 = Xyl (14)
QA-TriR-FRX
R1 = Rha (15)
QA-TriX-FRXX
R1 = Xyl (16)
QA-TriR-FRXX
R1 = Rha (17)
Qs0234140
(UGT73CY2)
(D-Api)
Qs0088320
(QsAXS)
QA-TriX-FRXA
R1 = Xyl (18)
QA-TriR-FRXA
R1 = Rha (19)


A
4,6-dehydratase
4-keto-reductase
dTDP-D-glucose
dTDP-4-keto-6-deoxy-D-glucose
dTDP-D-fucose
B
UGT74BX1 only
+ ATCV-1 UGD
+ SDR
+ ATCV-1 UGD
+ SDR
Intensity
Time (min)
UGT74BX1
ATCV-1 UGD
SDR
(11.4 min)
1115.6
1151.5
Mass-to-charge ratio (m/z)
QA-TriR-F
(11)
MW = 1116
C
0 min
60 min
180 min
Intensity
Time (min)
4-keto-6-deoxy-
D-glucose
D
ATCV-1 UGD
UDP-D-glucose
UDP-4-keto-6-deoxy-
D-glucose
QA-TriR
(9)
UGT74BX1
QA-TriR-4K6DG
(**)
QsFucSyn
QA-TriR-F
(11)


No Qs0023500
control
EIC: [C83H129O46] -
m/z= 1861.7667-1861.7853
No Qs0321940
control
No Qs0206480
control
All QS-7 genes
7.11
QS-7 Standard
7.13
100
0
Relative Intensity
0
5
10
15
Time (min)


QA-TriX-FRXA
(18)
Qs0206480
(QsACT1)
Qs0023500
(UGT73B44)
(L-Rha)
Qs0321940
(UGT91AP1)
(D-Glc)
L-Rha
Acetyl
D-Glc
QS-7
(20)
C83H130O46
MW = 1862.7833

Science
AAAS

Supplementary Materials for

Elucidation of the pathway for biosynthesis of saponin adjuvants from the soapbark tree

James Reed, Anastasia Orme, Amr El-Demerdash, Charlotte Owen, Laetitia B.B. Martin, Rajesh Chandra Misra, Shingo Kikuchi, Martin Rejzek, Azahara C. Martin, Alex Harkess, Jim Leebens-Mack, Thomas Louveau, Michael J. Stephenson and Anne Osbourn

Corresponding author: anne.osbourn@jic.ac.uk

The PDF file includes:

Materials and Methods
Figs. S1 to S53
Tables S1 to S16

Other Supplementary Materials for this manuscript include the following:

Data S1 to S5

Materials and Methods

Evaluation of QS-7 and QS-21 content in *Quillaja saponaria* accession S10

Seven different tissue types were analyzed (young, mature and old leaves, primordium, green stem, bark and root, with four biological replicates). Samples were freeze-dried for 24 h, and 15 mg aliquots disrupted in 2 mL graduated conical tubes (Starlab, E1420-0304) using two 3 mm tungsten beads (1000 rpm, 1 min) with a Geno/Grinder (Spex). After addition of 600 µL of 80% methanol, samples were incubated at 70°C for 30 min with shaking at 1000 rpm and then centrifuged at 13,000 x g for 10 min. The supernatants were transferred to fresh microcentrifuge tubes and defatted using 400 µL of hexane. The lower aqueous methanol phase was dried under a nitrogen flow at 60°C in a Provair sciences MiniVap®. After drying, samples were resuspended in 130 µL 80% methanol and filtered using a Spin-X 0.2 µm filter column (Costar). The filtrates were transferred into glass inserts placed in 1.5 mL sample vials (Agilent). Analysis of QS7 and QS-21 content was carried out using a Thermo Scientific QExactive Hybrid Quadrupole-Orbitrap Mass spectrometer HPLC system, calibrated using Pierce +ve/-ve calibration standards according to the manufacturer's instructions. Detection: MS (ESI ionization), scan range of 400-2500 m/z in negative mode, 70,000 resolution. Data dependent MS^2, isolation window of 4.0 m/z, collision energy of 30, resolution of 17,500, dynamic exclusion of 5.0 s. Method: Solvent A: [H_2O + 0.1 % formic acid] Solvent B: [acetonitrile (CH_3CN)]. Injection volume: 10 µL. Gradient: 15% [B] from 0 to 0.75 min, 15% to 60% [B] from 0.75 to 13 min, 60% to 100% [B] from 13 to 13.25 min, 100% to 15% [B] from 13.25 to 14.5 min, 15% [B] from 14.5 to 16.5 min. Method was performed using a flow rate of 0.6 $mL.min^{-1}$ and a Kinetex column 2.6 µm XB-C18 100 Å, 50 x 2.1 mm (Phenomenex). Analysis was performed using Xcalibur and FreeStyle softwares (Thermo Scientific). Purified QS-7 and QS-21 samples from Desert King (San Diego, CA, USA) were used to generate standard curves to determine the absolute amounts of these molecules in the samples.

Assembly of the 1KP transcriptome data and identification of QsbAS1 and oxidases

The assembled 1KP transcriptome derived from *Q. saponaria* leaves was downloaded from http://www.onekp.com/public_data.html (OQHZ; also available in the CyVerse Data Store, https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/oneKP_capstone_2019) and BLASTP searches were carried out to identify candidate OSC and CYP genes. To ensure that all relevant sequences were recovered, the source SRA data was also downloaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/ERR706840) and re-assembled using Trinity (*45*, *46*). Adapter sequences were trimmed with Trimmomatic (*47*). This dataset was mined using HMMER with the SqHop cyclase Pfams (PF13243, PF13249) and CYP Pfam (PF00067) for candidate OSC and CYP genes, respectively. To shortlist potential candidates that may oxidize the C-23position of echinocystic acid, the identified CYP sequences (164 total) were analyzed to check for sequence similarity. Where multiple sequences were found with ≥98% identity they were removed, leaving a single representative transcript, yielding a total of 151 sequences. Any truncated sequences (under 450 amino acids) were removed, leaving 37 sequences, of which a further two were removed (one due to missing start/stop codons, the other due to lacking the conserved cysteine residue) to give 35 candidates. A BLAST search was used to identify the closest homologues from *Arabidopsis thaliana* for each CYP. Candidates were deprioritized if the *A. thaliana* homologues were known to be involved in primary metabolic functions, leaving a final shortlist of 26 candidates. The candidate enzymes and their predicted sequences are provided in Data S2.

Cloning of *Q. saponaria* genes for transient plant expression

Oligonucleotide primers were designed based on predicted gene sequences and flanked with attB sites for Gateway cloning (Data S1). RNA extracted from primordia and young leaves (see above) was used for cDNA synthesis. The harvested tissues were flash frozen in liquid nitrogen and ground to a fine powder using a pestle and mortar (also in liquid nitrogen). RNA isolation was carried out using a Qiagen RNeasy® Plant Mini kit with the modified protocol according to (*39*). Following cleanup of the purified RNAs as per the protocol of the RNeasy® Mini Handbook (Qiagen), RNA quality was assessed by using nanodrop ratios and 1% agarose gel. cDNA synthesis was performed using Superscript III (Thermo Fisher) with oligo dT primers according to the manufacturer's instructions. Candidate sequences were amplified from cDNA of either primordia or young leaves using iProof polymerase (Bio-Rad), cloned into pDONR207 using BP clonase (ThermoFisher) and sequenced (Eurofins), before being introduced into the binary expression vector pEAQ-*HT*-DEST1 (*40*). The expression constructs were transformed into *A. tumefaciens* strains LBA4404 or GV3101. For ease of performing infiltrations, in some cases, multiple genes incorporated into a single binary vector using Golden Gate cloning were used (*41, 42*). The coding sequence of each gene was domesticated by removal of *BpiI* and/or *BsaI* restriction sites as needed and assembled into Golden Gate entry vector pL0-pICH41308. Genes were further assembled into level 1 expression cassettes consisting of the flanking modified 5' and 3' UTRs from Cowpea mosaic virus (*40*) under control of the CaMV35S promoter and Nos terminator. To enhance the expression of recombinant proteins in *N. benthamiana*, the P19 viral suppressor of gene silencing was also assembled under the control of CaMV35S promoter and CaMV35S terminator. Finally, multiple genes were incorporated into level 2 and/or a set of level M binary expression vectors (Figs. S48-53) and the vectors transformed into *A. tumefaciens* strain LBA4404 or GV3101. The Golden gate constructs were used interchangeably with the pEAQ constructs.

Preparation of *Q. saponaria* genomic DNA for genome sequencing

Young leaves of *Q. saponaria* S10 (2.6 g) were flash frozen in liquid N_2 and ground to a powder using a pestle and mortar. 10 mL of extraction buffer (2% w/v cetyl trimethylammonium bromide (CTAB), 100 mM Tris-HCl (pH 8.0) 1.4 M NaCl, 20 mM EDTA, 10 mg/mL proteinase K) was added, and the sample incubated at 55°C for 30 min in a 50mL falcon tube with intermittent shaking. After incubation on ice for 5 min, 5 mL chloroform was added and the tube was gently inverted several times. The sample was centrifuged at 2,100 x g for 30 min and the upper aqueous phase (approx. 7.5 mL) transferred to a fresh tube containing an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1 v/v). Following mixing, a further centrifugation step (2,100 x g for 30 min) was performed. The upper aqueous phase was then transferred to a fresh tube and mixed with 10% v/v of 3 M NaOAc, (pH 5.2 with HCl) followed by 2.5 volumes of ice cold 100% ethanol and incubated on ice for 30 min to precipitate nucleic acids. The sample was centrifuged (2,100 x g for 30 min at 4°C) and the supernatant discarded. The pellet was washed with 70% ice cold ethanol, the tube centrifuged (2,100 x g for 10 min at 4°C), and the washing step repeated twice more. The tubes were then inverted over a paper towel to dry the pellet. Finally, the dried pellet was resuspended in 1 mL of H_2O containing 0.1 mg/mL RNase A.

Transcriptome analysis

RNA was extracted from six different tissue types of *Q. saponaria* S10 (young, mature and old leaves, primordium, stem, and root) with four biological replicates per tissue. The harvested

tissues were flash-frozen in liquid N_2 and ground to a fine powder using a pestle and mortar (in liquid N_2). RNA isolation was performed as described above. Aliquots (4 μg) of each sample were sent to the Earlham Institute for quality assessment. A Tecan plate reader was used to quantify the material using fluorimetry. The RQS was measured using a Perkin Elmer GX II. Following NEXTflex directional RNA-seq library preparation, sequencing was carried out using Illumina HiSeq4000 PE150. The data were run through several QC assessment tools: firstly PAP, which consists of; FastQC, Centrifuge, FastQ Screen -> MultiQC; then EI-MAP; QualiMap, RSeQC, Picard, BamTools, Bowtie 2 / HiSAT 2, CutAdapt, SortMeRNA, FastQC (after adaptor trimming with CutAdapt), followed by Mikado.

Genome sequencing, assembly and annotation

PacBio Sequel sequencing of genomic DNA (using 11 SMRT cells) and sequence assembly by HGAP4 was carried out by the Earlham Institute. This resulted in a draft genome assembly of 769 polished contigs, with a total length of 354.9 Mbp, a maximum contig length of 18.2 Mbp and an N50 of 5.5 Mbp. A Hi-C library was prepared using the Phase Genomics Plant Hi-C 2.0 Kit (Seattle, WA), with 1 gram of flash-frozen leaf tissue as input. Fifteen PCR cycles were used to amplify the library to a final concentration of 45 ng/ul. The library was sequenced with Illumina PE75 reads to a depth of 279 million read pairs. Finally, the draft contig assembly was scaffolded into 14 pseudomolecules by Phase Genomics Proximo software. The final genome version consisted of 14 ordered clusters (i.e. pseudochromosomes) made up of 147 contigs and totaling 346.9 Mbp in length (97.4% of the genome assembly), with a remaining 8 Mbp comprised of 624 unplaced contigs and an N50 of 26.44 Mbp.

RNASeq guided genome annotation and transcript quantification was carried out by the Earlham Institute. Transcriptome reads were mapped to the draft genome and quality filtered using Portcullis (*48*). Following repeat identification, transcript classification was carried out in Mikado (*49*), which was then used to generate hints for Augustus (*50*). Multiple annotation runs were carried out in Augustus using varying parameter weightings, and these outputs were then integrated through Mikado to produce protein-coding genes (genes with supported ORFs or high coding potential), 'predicted' genes (genes with limited homology support (< 30%) and with low coding potential) and transposable element genes (genes with > 40% overlap with interspersed repeats). Genes were also classified into 'high' and 'low' confidence, with high confidence genes having >80% coverage to reference proteins or >60% protein coverage and with at least 40% of the structure supported by transcriptome data. This resulted in 30,780 high-confidence protein coding genes and 5,247 remaining low-confidence or transposable element encoding genes, with a total of 41,850 transcripts and a mean transcript CDS length of 1,219 bp (Table S1). Transcripts were re-aligned and quantified against final gene models using Salmon (*51*). Functional annotation was carried out using InterProScan and summarized by AHRD (*52*), giving 40,011 transcripts with a putative functional description. The completeness of the gene space was assessed by BUSCO analysis (*38*) using the embryophyta dataset (n = 1,440), of which 93.9% were fully represented (Fig. S7). These annotations were subsequently transferred to the final, pseudochromosome-level assembly. Repeats were annotated de novo with EDTA v2.0.0 using default parameters (*53*).

Data availability

The fully assembled and annotated *Q. saponaria* genome sequence has been deposited under NCBI BioProject ID PRJNA914519. RNASeq reads are deposited under NCBI BioProject ID

PRJNA914309 (SRA accessions SRR22829626 - SRR22829649). The sequences of the genes characterized in this study can also be found in GenBank as the following: *QsbAS1 (Qs0315350*), OQ107256; *CYP716A224* (*Qs0259300*), OQ107260; *CYP716A297* (*Qs0322000*), OQ107248; *CYP714E52* (*Qs0148690*), OQ107266; *CSLM1* (*Qs0321900*), OQ107253; *CSLM2* (*Qs0000870*), OQ107265; *UGT73CU3* (*Qs0123860*), OQ107259; *UGT73CX2* (*Qs0283850*), OQ107255; *UGT73CX1* (*Qs0283870*), OQ107254; *UGT74BX1* (*Qs0321930*), OQ107250; *UGT91AR1* (*Qs0321920*), OQ107251; *UGT91AQ1* (*Qs0234120*), OQ107264; *UGT73CY3* (*Qs0234130*), OQ107263; *UGT73CY2* (*Qs0234140*), OQ107262; *UGT73B44* (*Qs0023500*), OQ107261; *UGT91AP1* (*Qs0321940*), OQ107249; *UGT73B43* (*Qs0213660*), OQ107257; *Apiose/xylose synthase* (*Qs0088320*), OQ107247; *QsFucSyn* (*Qs0321910*), OQ107252 and *QsACT1* (*Qs0206480*), OQ107258.

Karyotyping

Shoot apical meristems were obtained from two-year-old cuttings of *Q. saponaria* S10. The preparation of mitotic metaphase spreads was carried out as described previously (*54*) with minor modifications. Briefly, excised shoot meristems were treated for 2 h with nitrous oxide gas to accumulate metaphase cells. Meristems were later fixed in 90% acetic acid for 30 min and digested in 1% pectolyase Y23 and 4% cellulose Onozuka R-10 (Yakult Pharmaceutical, Tokyo, Japan) solution in 1x citrate buffer for 45 min at 37ºC. Digested meristems were washed in 70% EtOH, macerated into a fine cell suspension with a dissection needle, and centrifuged to eliminate the EtOH. Finally, cells were resuspended in 100% acetic acid and used to prepare the chromosome spreads. Chromosomes were counterstained with DAPI (1 μg/ml) and mounted in Prolong Diamond (Thermo Fisher Scientific Molecular Probes, Eugene, OR, USA). Images were acquired using a Leica DM5500B microscope equipped with a Hamamatsu ORCA-FLASH4.0 camera and controlled by Leica LAS X software v2.0.

Phylogenetic analysis

Protein sequences were extracted from the genome via Interpro annotation generated by AHRD output (UGT: IPR002213, CSL: IPR005150, GH1: IPR001360). Alignments of gene families were carried out using protein sequences in MUSCLE (*55*), with a maximum of 100 iterations. Phylogenetic trees were generated from alignments with RaXML (*56*) using the PROTGAMMAAUTO model and 100 bootstraps. Trees were cross-referenced with the established nomenclature in each case to ensure accurate tree generation and gene classification, and clades were labelled accordingly.

plantiSMASH analysis and co-expression analysis

The fully annotated *Q. saponaria* S10 genome was analyzed for the presence of putative biosynthetic gene clusters using the plantiSMASH 1.0 algorithm (*20*). Salmon quantification outputs were processed using DESeq2 and mean size-factor normalized read counts were generated for each tissue (*51,57*). Variance stabilizing transformed read counts were used to generate Pearson's correlation coefficients (PCC) for each gene versus *QsbAS1*. Hierarchical clustering of expression data was performed on subsets of genes by hclust in R (*58*).

Metabolite analysis

GC-MS analysis

Ten mg aliquots of lyophilized leaf material were disrupted using two 3 mm tungsten carbide beads (Qiagen) by shaking at 1000 rpm for 60 sec in a Geno/Grinder (Spex). Ethyl acetate (500 µL) containing 50 µg/mL coprostanol internal standard (Sigma) was added and the samples incubated at room temperature with occasional shaking for 10 min. They were then centrifuged for 1 min at 13,500 x g and 100µL of supernatant was transferred to 1.5 mL sample vials (Agilent) and dried at 42°C using an EZ-2 centrifugal evaporator (Genevac). Samples were derivatized using 25 µL of Tri-Sil Z reagent prior to analysis. GC-MS analysis was performed using an Agilent 7890B fitted with a ZB5-HT column (Zebron) coupled to an Agilent 5977A mass selective detector. Injections were performed in split mode using a 20:1 split (split flow = 20mL/min) with the injection pulse pressure set to 30 psi. The GC temperature program was set to 170°C for 2 min, followed by a gradient to 300°C at 20°C per minute and held at 300°C for an additional 11.5 min (20 min total). The mass spectrometer was set to scan from 60 to 800 mass units with an initial 8-minute solvent delay. Data analysis was performed using MassHunter Qualitative Software (Agilent).

LC-MS/CAD analysis of N. benthamiana leaf extracts

Ten mg aliquots of lyophilized leaf material were disrupted as above, and 550 µL of 80% methanol containing 20 µg/mL digitoxin internal standard (Sigma-Aldrich) added to each sample. Samples were then incubated at 40°C for 20 minutes with shaking at 1000 rpm, before defatting by partitioning twice with 300 µL hexane. The lower aqueous methanol phase was transferred to a fresh microcentrifuge tube and dried at 42°C in an EZ-2 centrifugal evaporator (Genevac). Samples were resuspended in 75 µL methanol and filtered using a Spin-X 0.2 µm filter column (Costar) before transfer to 1.5 mL sample vials (Agilent).

For analysis of compounds from quillaic acid as far as QA-Tri[X/R]-FRX, analysis was performed using a Prominence HPLC system (Shimadzu) connected to an LCMS-2020 single quadrupole mass spectrometer (Shimadzu) and a Corona Veo RS Charged Aerosol Detector (Dionex). The MS detector was set to dual ESI/APCI ionization mode scanning from masses 100-2000. Chromatography was performed using a Kinetex 2.6 µm XB-C-18 100 Å, column (50 x 2.1 mm) (Phenomenex, part number 00B-4496-AN) with a flow rate of 0.3 mL min^{-1} injecting 10 µL per run. The mobile phase consisted of H_2O with 0.1% formic acid (solvent A) and acetonitrile (solvent B) and began at 15% [B] for 1.5 min, followed by a gradient from 15-60% [B] until 26.0 min and 60-100% [B] to 26.5 min and held at 100% [B] until 28.5 min. The column was re-equilibrated from 100-15% [B] until 29.0 min. Data analysis was performed using LabSolutions software (Shimadzu).

For analysis of compounds downstream of QA-Tri[X/R]-FRX, analysis was carried out using a Thermo Scientific Q Exactive Hybrid Quadrupole-Orbitrap Mass spectrometer HPLC system, calibrated using Pierce +ve/-ve calibration standards according to the manufacturer's instructions. Detection: MS (ESI ionization), scan range of 400-2500 m/z in negative mode, 70,000 resolution. Data dependent MS2, isolation window of 4.0 m/z, collision energy of 30, resolution of 17,500, dynamic exclusion of 5.0 s. Method: Solvent A: [H_2O + 0.1 % formic acid] Solvent B: [acetonitrile (CH_3CN)]. Injection volume: 10 µL. Gradient: 15% to 60% [B] from 0.75 to 13 min, 60% to 100% [B] from 13 to 13.25 min, 100% to 15% [B] from 13.25 to 14.5

min, 15% [B] from 14.5 to 16.5 min. The method was performed using a flow rate of 0.6 mL.min-1 and a Kinetex column 2.6 µm XB-C18 100 Å, 50 x 2.1 mm (Phenomenex). Analysis was performed using Xcalibur and FreeStyle softwares (Thermo Scientific)

Semi-quantification of AXS

The gene sets for production of QA-TriX-FRXX (**16**) and QA-TriX-FRXA (**18**) were transiently expressed in *N. benthamiana*. Six plants in total were used for each of the two compounds and for half of the plants in each group, *QsAXS* was also transiently expressed. After five days, leaves were harvested, lyophilized and extracts were analyzed as detailed above using the Thermo Scientific Q Exactive Hybrid Quadrupole-Orbitrap Mass spectrometer HPLC system. Semi-quantification of the products was performed using Xcalibur software (Thermo) by dividing the peak area of the **16** and **18** products (*m/z* = 1511.5616) versus that of the internal standard (digitoxin, *m/z* = 809.4337, 1.1 µg/mg leaf tissue). These values were averaged across the three samples for each test group.

Large-scale agro-infiltration of *N. benthamiana*, compound purification and NMR analysis

Large-scale vacuum infiltration of *N. benthamiana* plants was carried out as described previously (*44*). Leaf material was harvested five days after infiltration and frozen at -80°C prior to lyophilization for 24-72 hours. Detailed methods for extraction, purification and structural analysis are provided below.

General procedures for purification of compounds

Organic solvents used for extraction and flash chromatography were reagent grade and used directly without further distillation. Extraction of the compounds **5**-**9** was performed using a Speed Extractor E-914 (Büchi). Briefly, lyophilized leaf material was dispersed with 1 part 0.3-0.9 mm quartz sand (by volume) and layered over 3 cm quartz sand within a 120 mL extraction cell. Unless otherwise stated, the program consisted of one cycle of defatting with hexane followed by two cycles of methanol. Cycles were performed at 100°C and 130 bar. The first cycle had zero hold time, and the second cycles two and three had 5 min hold times. The run finished with a 1 min methanol flush and 12 min N_2 flush. For compounds **11**, **13**, **15**, **17**, **19** and QS-7 (**20**), extraction was performed using a combination of MeOH/H_2O (90/10 (**11, 13**), or 80/20 (**15, 17, 19** and **20**)) under refluxing at 95°C for two days. Flash column chromatography (FCC) was performed using an Isolera One (Biotage). Columns were (Biotage Sfär C-18, 60 g, 50 mL/min), using a general gradient system of [(95/5→25/75) of H_2O/ACN + 0.1 FA]. Analytical TLC experiments were performed on silica gel precoated aluminum plates (F254, 20 × 20 cm, Merck KGaA, Germany). TLC plates were visualized under UV light (254 nm) followed by staining with *p*-anisaldehyde (2% v/v *p*-anisaldehyde, 2% v/v, Conc. H_2SO_4). Semi-preparative HPLC experiments were performed on Ultimate 3000 (Thermo Fisher Scientific) fitted with Luna C_{18} column (250 × 10 mm i.d.; 5 µm; (Phenomenex)). Preparative HPLC experiments were performed using a 1290 preparative analytical system (Agilent) fitted with a Luna C_{18} (250 × 21.2 mm i.d.; 5 µm; (Phenomenex)). LC-MS analysis of fractions was conducted using the systems and methods as described above. 1D and 2D NMR spectra were recorded on a Bruker Avance 400 and 600 MHz spectrometers equipped with a BBFO Plus Smart probe and a triple resonance TCI cryoprobe, respectively (JIC, UK) and were analyzed used Mestrenova software. The chemical shifts are relative to the residual signal solvent (MeOH-d_4: δ_H 3.31; δ_C 49.15).

Isolation of quillaic acid (5)

A total of 209 *N. benthamiana* plants were infiltrated as described above with *A. tumefaciens* carrying the genes as detailed in Data S4. Lyophilized leaf material (70 g) was extracted with the SpeedExtracter using four cycles (100°C, 130 bar). The first cycle was performed using hexane with zero hold time, while cycles two to four were performed in ethanol with a five-minute hold time. The run terminated with a two-minute solvent flush and six-minute N_2 flush.

The ethanol portion of the extraction containing quillaic acid was dried onto silica gel 60 (Material Harvest) and used for flash chromatography. The collected fractions were assessed by GC-MS and thin layer chromatography (TLC) and quillaic acid-containing fractions were pooled. Column 1 consisted of a SNAP Ultra 50 g (Biotage) using a flow rate 100 mL/ min and collecting 90 mL fractions. The mobile phase consisted of hexane (solvent A) and ethyl acetate (solvent B), with a gradient from 5-100% [B] over 10 column volumes (CV) and held at 100% for 5 CV. Column 2 consisted of a SNAP Ultra 50 g (Biotage) using a flow rate 100 mL/ min and collecting 90 mL fractions. The mobile phase consisted of dichloromethane (solvent A) and ethyl acetate (solvent B), with a gradient from 10-60% [B] over 10 column volumes (CV) and held at 100% for 2 CV. These conditions were repeated for a third column using a SNAP Ultra 10 g (Biotage) with a flow rate 36 mL/ min and collecting 17 mL fractions. Quillaic acid-containing fractions were treated with activated charcoal (Sigma). Column 4 consisted of a SNAP Ultra 10 g (Biotage) with a flow rate of 36 mL/min, 17 mL fractions. The column used an isocratic mobile phase consisting of 15% ethyl acetate in dichloromethane over 20 CV. A small amount of HCl (400 μL of concentrated HCl in 40 mL ethanol) was added to the quillaic acid-containing fractions which helped to reduce streaking. Finally, a fifth column was run using the same conditions as column 4 but with 30 CV. The purest fractions were pooled, yielding 30 mg of quillaic acid (**5**) as a white powder with trace amounts of yellow impurities.

Isolation of the CSLM-1/CSLM-2 product (3-O-{β-D-glucopyranosiduronic acid}-quillaic acid) (6)

For the CSLM-1 product, 104 *N. benthamiana* plants were infiltrated, affording 68 g of dry leaves. Dry leaf powder was defatted with hexane and then treated with exhaustive pressurized solvent extraction using methanol. The methanolic extract was collected and dried under reduced pressure before re-dissolving in a minimal amount of methanol and adding an equivalent volume of water. The extract was partitioned with ethyl acetate (3 L). The upper organic layer was collected, dried over anhydrous $MgSO_4$ and evaporated under reduced pressure to afford 560 mg of reddish-brown material. This was re-dissolved in methanol and saturated with cold acetone to give a pale-yellow precipitate containing crude saponins. This enriched saponin fraction was further purified by semi-preparative C_{18} HPLC. The mobile phase consisted of H_2O + 0.1% FA [solvent A] and acetonitrile + 0.1% FA [solvent B] at 4 mL/min. The run consisted of an initial isocratic run in 10% [B] for 5 min, followed by a gradient from 10-100% [B] over 50 minutes and held at 100% [B] for 5 minutes. The column was re-equilibrated from 100-90% [B] over 1 min and held at 90% [B] for 4 minutes. This method enabled isolation of 9.5 mg of 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid as white amorphous material. For the CSLM-2 product, 80 *N. benthamiana* plants were infiltrated affording 54 g of dry leaves. Extraction and purification were performed as above for the CSLM-1 product, affording 2.1 mg of 3-*O*-{*β*-D-glucopyranosiduronic acid}-quillaic acid (**6**).

Isolation of the C3-GalT product (3-O-{β-D-galactopyranosyl-(1→2)-β-D-glucopyranosiduronic acid}-quillaic acid) (7)

102 *N. benthamiana* plants were infiltrated, affording 32 g of dry leaves. The dried leaf powder was then treated with pressurized solvent extraction and first defatted by using hexane followed by exhaustive extraction using methanol. The methanolic extract was collected and dried under reduced pressure before re-dissolving in the least amount of methanol and adding an equivalent volume of water. The extract was partitioned with ethyl acetate (4 L). The upper organic layer was collected, dried over anhydrous $MgSO_4$ and evaporated under reduced pressure to afford 500 mg of reddish-brown material. Afterwards, this saponin-containing fraction was further purified by semi-preparative C_{18} HPLC according to the conditions used for the QA-GlcA product above. This enabled isolation of 7.3 mg of the product 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid}-quillaic acid (**7**) as a white amorphous material.

Isolation of the C3-XylT product (3-O-{β-D-xylopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-quillaic acid) (8)

100 *N. benthamiana* plants were infiltrated affording 78 g of dry leaves. Dry leaf powder was treated to pressurized solvent extraction and first defatted by using hexane followed by exhaustive extraction using methanol. The methanolic extract was collected and dried under reduced pressure before re-dissolving in the least amount of methanol and adding an equivalent volume of water. A series of liquid-liquid partitions were performed using hexane, dichloromethane, ethyl acetate and *n*-butanol. The butanol layer was dried over anhydrous $NaSO_4$, evaporated under reduced pressure, and re-dissolved in the least amount of methanol and subjected to semipreparative C_{18}-HPLC according to the conditions used for the QA-GlcA product above. Theis enabled isolation of 21.6 mg of 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (**8**) as a pale brown amorphous material.

Isolation of the C3-RhaT product (3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-quillaic acid) (9)

A total of 100 *N. benthamiana* plants were infiltrated, affording 70 g of dry leaves. Dry leaf powder was extracted and treated as previously described for compound **8**. The butanol layer was dried over anhydrous $NaSO_4$, evaporated under reduced pressure, and re-dissolved in the least amount of methanol and subjected to semipreparative C_{18}-HPLC according to the conditions used for the QA-GlcA product above. This enabled isolation of 43.3 mg of product 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-quillaic acid (**9**) as a pale brown amorphous material.

Isolation of the C28-FucT product (3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-O-{β-D-fucopyranosyl ester}-quillaic acid) (11)

A total of 124 *N. benthamiana* plants were infiltrated, affording 40 g dry leaves. Dry leaf powder was chemically extracted under refluxing and crude aqueous-methanolic extract was treated according to the previously mentioned protocol as for compound **9**. The *n*-butanol layer was collected, evaporated under reduced pressure, re-dissolved in the least amount of methanol and saturated with cold acetone to precipitate a saponin-enriched crude fraction. This fraction was

subjected to semipreparative C_{18}-HPLC column (Luna C_{18} column (250 × 10 mm i.d.; 5 μm; USA). The mobile phase consisted of H_2O + 0.1% FA [solvent A] and acetonitrile + 0.1% FA [solvent B]. The run consisted of a gradient from 10%-70% [B] over 35 min at 3 mL/min before re-equilibrating by 10% [B] for 5 min. A further C-18 semi-preparative purification column run was performed under isocratic conditions using 40% [B] at 1mL/min. This afforded 1 mg of (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-fucopyranosyl ester}-quillaic acid) (**11**) as pale-yellow amorphous material (85-90% purity).

Isolation of C-28-RhaT product (3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-O-{α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranosyl ester}-quillaic acid) (13)

A total of 112 *N. benthamiana* plants were infiltrated, affording 46 g dry leaves. Dry leaf powder was processed as above for compound **11**. The saponin-enriched crude fraction was subjected to semi-preparative C_{18}-HPLC purification. The mobile phase consisted of H_2O + 0.1% FA [solvent A] and acetonitrile + 0.1% FA [solvent B]. The run consisted of a gradient from 10%-70% [B] over 50 min at 3 mL/min. This afforded 43.9 mg of (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**13**) as a pale-yellow amorphous material.

Isolation of the C28-XylT3 product (3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-O-{β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranosyl ester}-quillaic acid)(15)

A total of 202 *N. benthamiana* were infiltrating, affording 65 g dry leaves. Dry leaf powder was chemically processed via the aforementioned protocol as compound **11**. The saponin-enriched crude fraction was subjected to semipreparative C_{18}-HPLC purifications as for compound **11**. This afforded 3.1 mg of (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**15**) as a pale-yellow amorphous material.

Isolation of C28-XylT4 (3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-O-{β-D-xylopyranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranosyl ester}-quillaic acid) (17)

A total of 105 *N. benthamiana* plants were infiltrated, affording 58 g dry leaves. Dry leaf powder was processed as per the aforementioned protocol applied for compound **11**. The saponin-enriched crude fraction was subjected to preparative C_{18}-HPLC (Luna, 250 x 21.2 mm, 5 μm; C18 (2), USA). The mobile phase consisted of H_2O + 0.1% FA [solvent A] and acetonitrile + 0.1% FA [solvent B] at 25 mL/min. The run consisted of a gradient from 10-17% [B] over 17 min. This afforded 13.2 mg of (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-xylopyranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**17**) as a pale-yellow amorphous material.

Isolation of C28-ApiT4 product (3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-O-{β-D-apiofuranosyl-(1→3)-β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-β-D-fucopyranosyl ester}-quillaic acid) (19)

A total of 105 *N. benthamiana* plants were infiltrated, affording 57 g dry leaves. Dry leaf powder was proceeded as described above for compound **11**. The saponin-enriched crude fraction was purified using semipreparative C_{18}-HPLC. The mobile phase consisted of H_2O + 0.1% FA [solvent A] and acetonitrile + 0.1% FA [solvent B] at 25 mL/min. The run consisted of a gradient from 10-70% [B] over 17 min followed by a further semipreparative purification stage using an isocratic system of 40% [B] over 30 min at 2 mL/min. This afforded 13.2 mg of (3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{*β*-D-apiofuranosyl-(1→3)-*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-fucopyranosyl ester}-quillaic acid) (**19**), a pale-yellow amorphous material of two rotameric conformers (1:1).

Isolation of QS-7 (20) from Nicotiana benthamiana

A total of 410 *N. benthamiana* plants were infiltrated, affording 104 g dry leaves. Dry leaf powder was extracted and processed as previously mentioned as applied for compound **11**. The crude saponin-enriched fraction was then subjected to semipreparative C_{18}-HPLC purification using an eluent system of water containing 20 mM of NH_4HCO_3 [A], adjusted to pH 8.6 (adjusted with ammonium hydroxide solution) along with acetonitrile [B]. The run consisted of a gradient from 25-70% [B] over 30 minutes at 4 mL/min. This afforded a semi-purified fraction (10-12 mg, as pale brown amorphous material) that contained 3-5% QS-7 (**20**) molecule based on ^{1}H-NMR analysis.

Preparation of proteins for *in vitro* studies

GalK Y371H

Escherichia coli galactokinase GalK with the Y371H mutation was used for enzymatic synthesis of UDP-α-D-fucose. The Y371H mutant variant has been shown to have enhanced activity towards D-fucose (*59*). Site-directed mutagenesis of GalK (in pET-15b vector) was introduced using a Q5® site-directed mutagenesis kit (New England Biolabs) with primers GalK_Y371K_F/R (Data S1) to make a T to C transition at position 1111 (codon TAC to CAC). The expression plasmid was transformed into *E. coli* Rosetta (DE3) competent cells (Novagen) and the N-terminally His-tagged GalK Y371H was purified by nickel affinity chromatography and Superdex 200 gel filtration chromatography. The eluant had a concentration of 4.6 mg protein/ml. Unit definition: one unit will convert 1.0 µmol of galactose to galactose-1-phosphate per minute at pH 7.4 at 30°C.

ATCV-1 UGD

The template DNA of UDP-D-glucose 4,6-dehydratase (UGD, Genbank accession YP_001427025.1) (*60*) from the *Acanthocystis turfacea* chlorella virus 1 (ATCV-1) was codon-optimised and synthesized by IDT (USA). ATCV-1 UGD was expressed as a fusion protein with an amino-terminally added large and highly soluble trigger factor (TF, 48 kDa) using pCold-TF expression vector (TaKaRa Bio). The ATCV-1 UGD was amplified by PCR using

oligonucleotides Cold_NdeI_AtUGD_FW and Cold_XhoI_AtUGD_RV (Data S1), and the amplified fragment was inserted into pCold-TF vector between NdeI and XhoI sites by In-fusion cloning (TaKaRa Bio/Clontech). The resultant fusion protein had 6x His-tag at its amino-terminal end. The expression construct was transformed into *E. coli* Rosetta (DE3) competent cells. The protein was induced by cold shock treatment on ice-cold water and supplemented with 0.5 mM IPTG to the culture medium, and incubating the cells for overnight at 16°C. After disruption of the cells by sonication, the His-tagged fusion protein was captured with TALON metal affinity resin (TaKaRa Bio/Clontech), and then the eluant was subjected to Superdex 200 gel filtration chromatography. The peak fractions were concentrated with Vivaspin 20, 50,000 MWCO PES (Sartorius, VS2031), and the concentration was determined to be 15.3 mg protein/ml.

QsFucSyn

Q. saponaria FucSyn (*Qs0321910* SDR) was expressed as a fusion protein with an amino-terminal trigger factor using pCold-TF expression vector. The QsFucSyn was amplified by PCR using oligonucleotides Cold_NdeI_QsFucSyn_FW and Cold_XhoI_QsFucSyn_RV (Data S1), and the amplified fragment was inserted into pCold-TF vector between NdeI and XhoI sites by In-fusion cloning. The expression and purification of TF-QsFucSyn was done by the same method for ATCV-1 UGD except for use of Ni Sepharose 6 Fast Flow (Cytiva) for capturing TF-QsFucSyn. The concentrated fraction was determined to be 14.7 mg protein/ml.

UGT74BX1

Q. saponaria UGT74BX1 was expressed with a carboxy-terminal hexahistidine tag in *N. benthamiana* using the *Agrobacterium*-infiltrated transient expression (*44*). The His-tag was added by PCR using oligonucleotides EAQ-QsUGT-Q-His_FW and EAQ-QsUGT-Q-His_RV encoding six histidine residues (Data S1), and the amplified fragment was inserted into a unique NruI site of linearised pEAQ-*HT* vector (*40*) by In-fusion cloning. The expression construct was transformed into *Agrobacterium tumefaciens* strain GV3101 and infiltrated into 3-week-old *N. benthamiana* leaves (*44*). After 6 days of incubation to allow sufficient accumulation of the enzyme, infiltrated parts of the fresh leaves were collected (1 g) and was ground in 5 ml of grinding buffer (50 mM HEPES-KOH, pH 7.8, 330 mM sorbitol, 1% polyvinylpolypyrrolidone, 7 mM 2-mercaptoethanol, and cOmplete EDTA-free protease inhibitor cocktail [Roche, 11 873 580 001]) using a mortar and pestle on ice. The homogenate was filtered through two layers of Miracloth (Calbiochem), centrifuged at 3,220 x *g* for 10 min to remove debris, and then centrifuged at 30,000 x *g* for 20 min to obtain cleared lysate without microsomes. The lysate (1.5 ml) was incubated with 50 µl slurry of TALON metal affinity resin in the presence of 5 mM imidazole and 0.1% (w/v) Triton X-100 for 2 h in a cold room with end-over-end mixing. The resin was washed four times with TBS-TX-Imi buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% Triton X-100 and 5 mM imidazole) and once with buffer A4 (20 mM HEPES, pH 7.5, and 150 mM NaCl). His-tagged UGT74BX1 was eluted twice with 250 µl of elution buffer (20 mM HEPES, pH 7.5, 150 mM NaCl, and 150 mM imidazole). The eluant was subjected to two cycles of dilution in buffer A4 and concentration with Vivaspin 20, 50,000 MWCO PES to minimize imidazole content. The concentration of UGT74BX1 was adjusted to 0.3 mg protein/ml.

Enzymatic synthesis and *in vitro* transformations of UDP-sugars

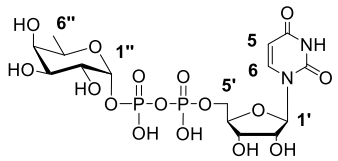
General procedures for NMR and HR-MS

NMR spectra were recorded on a Bruker Avance III 400 MHz spectrometer. Chemical shifts of ^{1}H NMR signals recorded in D_2O are reported with respect to residual solvent peak at δ_H 4.79 ppm. Chemical shifts of ^{31}P NMR signals recorded in D_2O are reported with respect to external 85% H_3PO_4 at δ_P 0 ppm. High resolution accurate mass spectra were obtained using a Synapt G2 Q-Tof mass spectrometer using negative electrospray ionisation.

Strong anion exchange (SAX) HPLC on Poros HQ 50

The chromatography was performed on a Dionex Ultimate 3000 (Thermo-Fisher) instrument equipped with UV/vis detector. Aqueous solution of a sample was applied on a Poros HQ 50 column (50 × 10 mm, column volume (CV) = 3.9 ml, Perceptive Biosystems). The column was first equilibrated with 5 CV of 5 mM ammonium bicarbonate buffer, followed by linear gradient of ammonium bicarbonate from 5 mM to 250 mM in 15 CV, then hold for 5 CV, and finally back to 5 mM ammonium bicarbonate in 1 CV and hold for 3 CV at a flow rate of 8 ml/min and an online UV detection to monitor A_{265}. After multiple injections, the column was washed with 3 CV of 1 M ammonium bicarbonate followed by 3 CV of Milli-Q water.

UDP-α-D-Fucose

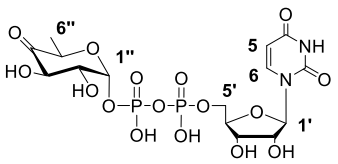


UDP-α-D-Fuc

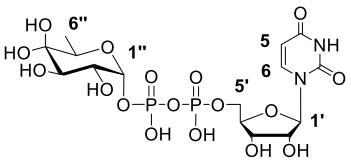
UDP-α-D-fucose was prepared by enzymatic transformation using GalK Y371H as described previously with minor modifications(*61*). In brief, D-fucose (3 mg, 18.3 μmol), UTP (10 mg, 18 μmol), ATP (0.1 mg, 0.18 μmol), PEP (3.4 mg, 18 μmol) and UDP-glucose (0.1 mg, 0.16 μmol) were dissolved in buffer (500 μl, 50 mM HEPES, pH 8.0, 5 mM KCl, 10 mM $MgCl_2$). Pyruvate kinase (50 U), glucose-1-phosphate uridylytransferase (5 U), inorganic pyrophosphatase (5 U), GalK Y371H (5 U) and GalPUT (5 U) were added, and the volume of the mixture was adjusted to 1000 μl with the buffer. The reaction was flushed with nitrogen and stirred at 30°C whilst being monitored by SAX HPLC. After 22 hours UTP (R_f 11.5 min) was fully consumed and a new peak (R_f 7.5 min) of the title compound was detected. The reaction was quenched by addition of MeOH (1 ml) and precipitated enzymes were removed by centrifugation. The supernatant was filtered (0.45 μm PTFE disc filter) and chromatographed by SAX HPLC. Pooled fractions were freeze dried to give title compound as diammonium salt (5.3 mg, 49.7%). ^{1}H NMR (D_2O, 400 MHz): δ_H 7.97 (1H, d, $^{3}J_{5,6}$ = 8.0 Hz, H6), 6.01-5.97 (2H, m, H5, H1'), 5.57 (1H, dd, $^{3}J_{1'',P\beta}$ = 6.8 Hz, $^{3}J_{1'',2''}$ = 3.5 Hz, H1''), 4.40-4.18 (6H, m, H2', H3', H4', H5', H5''), 3.93 (1H, dd, $^{3}J_{3'',2''}$ = 10.4 Hz, $^{3}J_{3'',4''}$ = 3.3 Hz, H3''), 3.84 (1H, bd, $^{3}J_{4'',5''}$ ~0 Hz, $^{3}J_{4'',3''}$ = 3.3 Hz, H4''), 3.76 (1H, ddd, $^{3}J_{2'',1''}$ = 3.5 Hz, $^{3}J_{2'',3''}$ = 10.4 Hz, $^{4}J_{2'',P\beta}$ = 3.5 Hz, H2"), 1.23 (3H, d, $^{3}J_{6'',5''}$ = 6.6 Hz, H6''). The ^{1}H NMR spectrum is in good agreement with the literature (*62*). ^{31}P

NMR (D_2O, 162 MHz): δ_P -11.2 (d, 1P, $J_{P\alpha,P\beta}$ = 20.7 Hz, P_β), -12.8 (d, 1P, $J_{P\alpha,P\beta}$ = 20.7 Hz, P_α). ESIMS: *m/z* Calcd $[C_{15}H_{23}N_2O_{16}P_2]^-$: 549.0528. Found: 547.0527.

UDP-4-keto-6-deoxy-α-D-Glucose and its hydrate form



UDP-4k,6d-α-D-Glc



UDP-4k,6d-α-D-Glc hydrate

UDP-4-keto-6-deoxy-α-D-glucose and its hydrate form were prepared from UDP-α-D-glucose by enzymatic transformation of TF-ATCV-1-UGD. UDP-α-D-glucose (9.8 mg, 16.1 µmol) and NAD^+ (15.9 mg, 24.0 µmol) were dissolved in buffer (4 ml, 50 mM HEPES, 2 mM $MgCl_2$, pH 7.5). The mixture was split equally into 8 tubes. To each tube TF-ATCV-1-UGD (61 µg, 0.9 nmol) was added and the mixture was incubated at 25°C with shaking at 300 rpm. An analytical sample (500 µl) was separated, mixed with D_2O (100 µl) and the progress of the reaction was monitored by NMR. ^{1}H NMR spectra showed 100% conversion after 22 hrs. The reaction was quenched by addition of equivalent volume of MeOH, mixed by vortexing, filtered through 0.45 µm disc filter (PTFE), and stored at -80°C. The product was purified by SAX HPLC with UV detection to monitor A_{265} to give the target compound (R_f = 8.2 min) as a diammonium salt (5.32 mg, 57 % yield) after freeze drying. In aqueous solutions UDP-4-keto-6-deoxy-α-D-glucose exists in equilibrium with its hydrate form.^{1}H NMR (D_2O, 400 MHz) a mixture of keto and hydrate forms in a ratio 1:4 was observed: δ_H 7.98 (1H, d, $^3J_{5,6}$ = 8.1 Hz, H6 hydrate), 7.97 (1H, d, $^3J_{5,6}$ = 8.1 Hz, H6 keto), 6.01-5.97 (2H, m, H5, H1'), 5.75 (1H, dd, $^3J_{1'',P\beta}$ = 7.3 Hz, $^3J_{1'',2''}$ = 3.4 Hz, H1'' keto), 5.56 (1H, dd, $^3J_{1'',P\beta}$ = 6.9 Hz, $^3J_{1'',2''}$ = 3.7 Hz, H1'' hydrate), 4.43-4.35 (2H, m, H2', H3'), 4.32-4.19 (3H, m, H4', H5'), 4.11 (1H, q, $^3J_{5'',6''}$ = 6.5 Hz, H5"), 3.86 (1H, ddd, $^3J_{2'',3''}$ = 10.2 Hz, $^3J_{2'',1''}$ = 3.4 Hz, $^4J_{2'',P\beta}$ ~ 3.2 Hz, H2'' keto), 3.80 (1H, d, $^3J_{3'',2''}$ = 10.2 Hz, H3''), 3.64 (1H, ddd, $^3J_{2'',3''}$ = 10.0 Hz, $^3J_{2'',1''}$ = 3.7 Hz, $^4J_{2'',P\beta}$ ~ 3.2 Hz, H2'' hydrate), 1.28 (3H, d, $^3J_{6'',5''}$ = 6.5 Hz, H6'' keto), 1.23 (3H, d, $^3J_{6'',5''}$ = 6.5 Hz, H6'' hydrate). ^{1}H NMR was in good agreement with the literature data (*62*).

Sources of other UDP-sugars

UDP-GlcNAcA (UDP-2-acetamido-2-deoxy-α-D-glucuronic acid) was prepared according to published procedure (*63*). UDP-β-L-rhamnose was prepared as described previously (*64*). UDP-α-D-glucose was obtained commercially from Merck (94335).

Attempt to produce UDP-α-D-fucose *in vitro* using ATCV-1 UGD and *QsFucSyn*

UDP-α-D-glucose (4 mM), NAD^+ (6 mM), and NADPH (6 mM) were dissolved in deuterated buffer (pD 7.5, 50 mM HEPES, 2 mM $MgCl_2$) to give the indicated final concentrations in total volume of 600 µl. At first, the ^{1}H NMR (400 MHz) spectrum of no enzyme control was acquired. TF-ATCV-1-UGD (2.3 nmol, 3.8 µM final concentration) was then added, and the reaction progress monitored by ^{1}H NMR at 294 K. After UDP-Glc was fully consumed

following overnight incubation, TF-QsFucSyn was added (1.8 nmol, 3.0 µM final concentration) to the mixture and the reaction progress was monitored by ^{1}H NMR.

In vitro glycosylation assays

The sugar acceptor QA-TriR (0.1 mM), the sugar donor UDP-α-D-glucose (1 mM), cofactors NAD^+ (1.5 mM) and NADPH (1.5 mM) were mixed in a buffer containing 50 mM HEPES, pH 7.5, 2 mM $MgCl_2$ and 0.5% (v/v) 2-mercaptoethanol in a final volume of 50 µl. Reactions were initiated by addition of purified TF-ATCV-1-UGD (0.3 mg/ml final), UGT74BX1 (0.01 mg/ml), and TF-QsFucSyn (0.3 mg/ml) to the mixture as indicated in the figure legends and incubated at 25°C for 14 hours. After quenching with methanol (final 50%), the filtered product (10 µl) was analysed with the Prominence HPLC system (Shimadzu) connected to the LCMS-2020 single quadrupole mass spectrometer (Shimadzu) equipped with the Corona Veo RS Charged Aerosol Detector (Dionex). Chromatography was performed using the RP-C18 column (Kinetex 2.6 µm 100 Å, 50 x 2.1 mm, Phenomenex) by the same method for *N. benthamiana* leaf extracts.

Extraction and analysis of sugar nucleotides from *N. benthamiana*

N. benthamiana plants were infiltrated with *A. tumefaciens* LBA4404 transformed with pEAQ-*HT*-DEST1 vectors harboring either green fluorescent protein (*GFP*) or *QsFucSyn.* After four days, leaf material was harvested and a total of 2 g (fresh weight) was taken for each test and flash frozen in liquid N_2. Samples were then spiked with 2 µg (10 µL of a 200 µg / mL solution) a non-plant sugar nucleotide (UDP-GlcNAcA) for use as an internal standard. The leaves were ground to a powder in liquid N_2 using a pestle and mortar and transferred into a 50 mL falcon tube while still frozen.

Extraction of the sugar nucleotides was adapted from a previously published protocol (*65,66*). Briefly, 10 mL ice-cold $CHCl_3$:CH_3OH (3:7) was added to the frozen leaf powder along with several 2 mm tungsten beads and shaken vigorously before transfer to $^{-}$20°C for two hours. During this period, every 30 mins, a round of disruption was performed at 1000 rpm for 60 seconds using a Geno/Grinder (Spex). After two hours, the sugar nucleotides were extracted by adding 8 mL water and samples were centrifuged at 29, 000 x g for 20 minutes at 4°C to pellet insoluble material and separate phases. The upper aqueous phase was transferred to a fresh vessel and a second extraction with 8 mL water was performed and this was combined with the first extract. The extract was evaporated at 40°C in an EZ-2 centrifugal evaporator for approximately 2 hours (Genevac). After this period, the remaining sample (mostly water) was frozen and lyophilized overnight. The following day, the dried extract was resuspended in 500 µL 5 mM ammonium bicarbonate for solid phase extraction (SPE). The SPE cartridges (Supelclean ENVICarb graphitized carbon columns (250 mg, 3 mL) (Supelco)) were first conditioned by washing with 3 mL 80% acetonitrile containing 0.1 % trifluoroacetic acid and 2 mL of water. The sample was adsorbed onto the column and the column was washed with a further 2 mL water, followed by 2 mL 25% acetonitrile and 2 mL 50mM triethylammonium acetate (TEAA) buffer (pH 7.0). Finally, sugar nucleotides were eluted with a buffer consisting of 25% acetonitrile in 50 mM TEAA buffer (pH 7.0). Samples were filtered through 0.45 µm PTFE disc filters (Whatman) before freezing and lyophilization overnight.

Sugar nucleotide analysis was performed as previously described (*67*). Briefly, analysis was carried out using a Xevo TQ-S system (Waters) equipped with a Hypercarb porous graphitic carbon (PGC) column 1 x 100 mm, particle 5 μm, CV ~ 78.5 μL (Thermo Fisher). The mobile phase consisted of [A] 0.3% formic acid (pH 9.0 with NH_4OH) and [B] acetonitrile. A gradient elution was used at flow rate 80 μL/min starting with 2% [B] which increased to 15% [B] over 20 minutes. This was followed by gradient to 50% [B] over 6 minutes, followed by a further increase to 90% [B] over 1 minute and held at 90% [B] for 3 mins. This was followed by a linear gradient to 2% [B] over 1 minute and finally held at 2% [B] for 19 minutes to equilibrate. Sugar nucleotides were detected by electrospray using multiple reaction monitoring (MRM). The MRM transitions for UDP-GlcNAcA and UDP-deoxyhexoses (UDP-D-fucose and UDP-L-rhamnose) are listed in Table S16. The identity of sugar nucleotides was verified using authentic standards as described above.

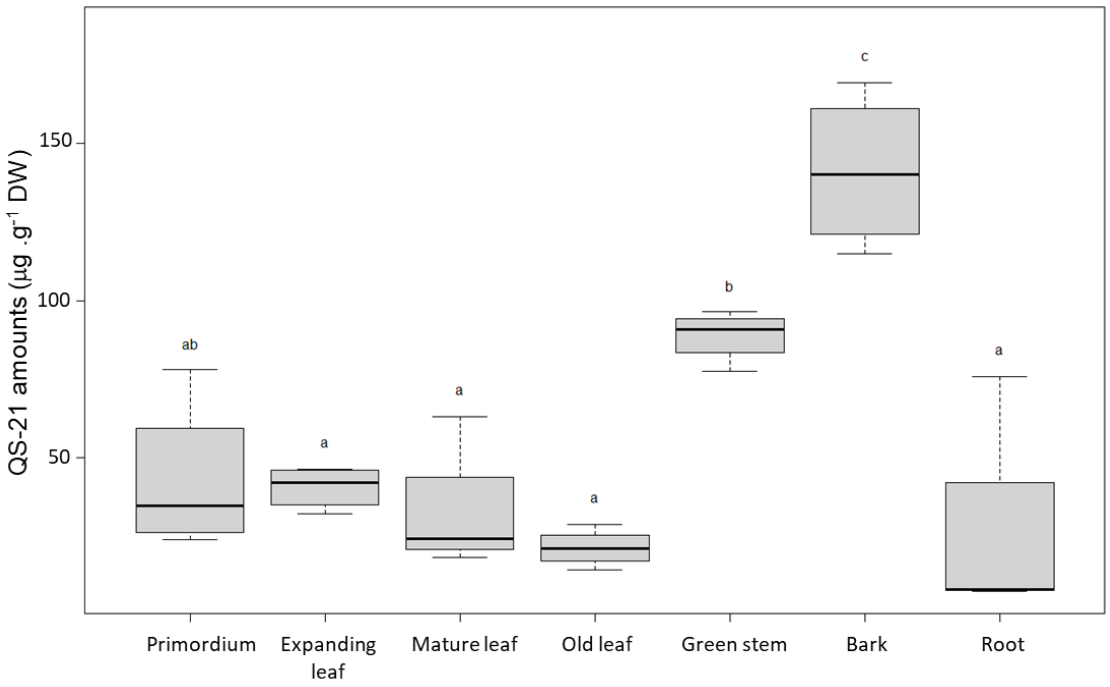


Fig. S1. QS-21 content of different *Q. saponaria* tissues. QS-21 was quantified relative to a QS-21 external standard curve and normalized based on the individual sample dry weight (15 mg dry material). The tissues are as follows: Primordium (the tip of the branch that includes the meristem and 1 leaf smaller than 0.5 cm); expanding leaf (leaf that has reached about half its mature size); mature leaf (first leaf on the branch that has reached its mature size); old leaf (leaf at the base of the branch that has not started to senesce); green stem (part of the branch that is still green in color with no sign of lignification); bark (lignified tissue covering a branch);root (roots from various developmental stages growing out of the bottom of the pot). Bars, standard error (four biological replicates). Statistical analyses comprised ANOVA and Tukey tests and were performed in R using the multcompView package

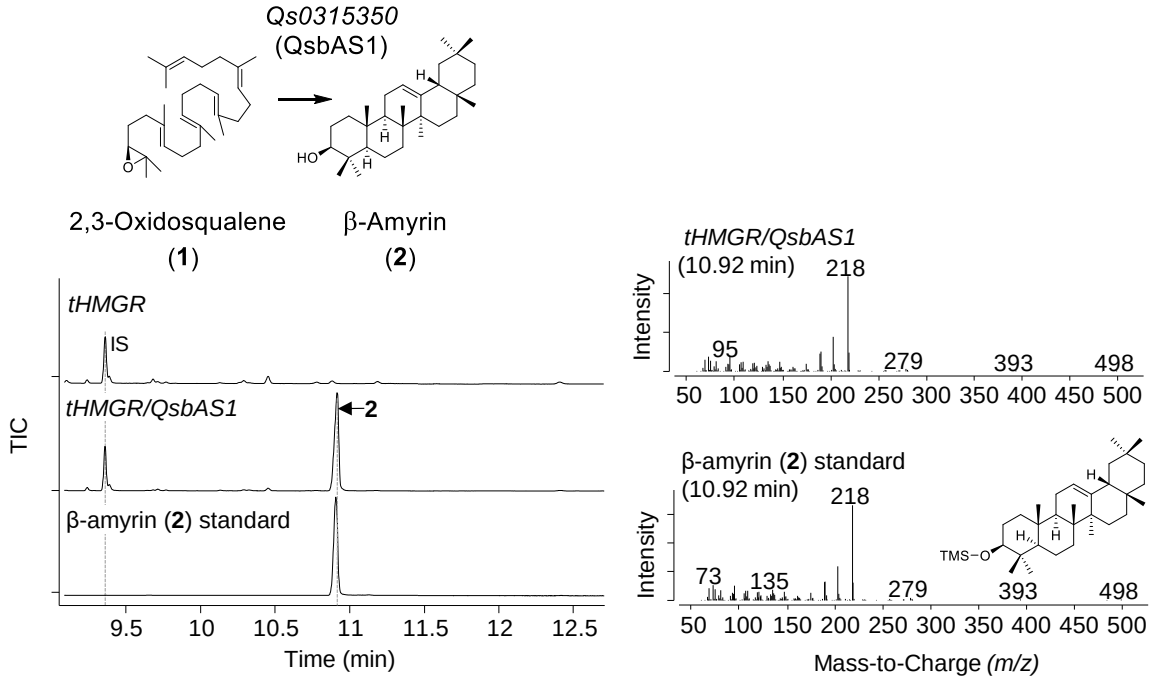


Fig. S2. *QsbAS1* encodes a functional β-amyrin synthase. GC-MS analysis of leaf extracts of *Nicotiana benthamiana* following *Agrobacterium*-mediated transient expression. Leaves were agroinfiltrated with expression constructs for *tHMGR* (control) or *tHMGR* and *QsbAS1*. IS, internal standard (coprostanol). Total ion chromatograms (TIC) are shown on the left, and mass spectra on the right. The retention time and mass spectrum for the QsbAS1 product are identical to that of a *β*-amyrin (**2**) standard.

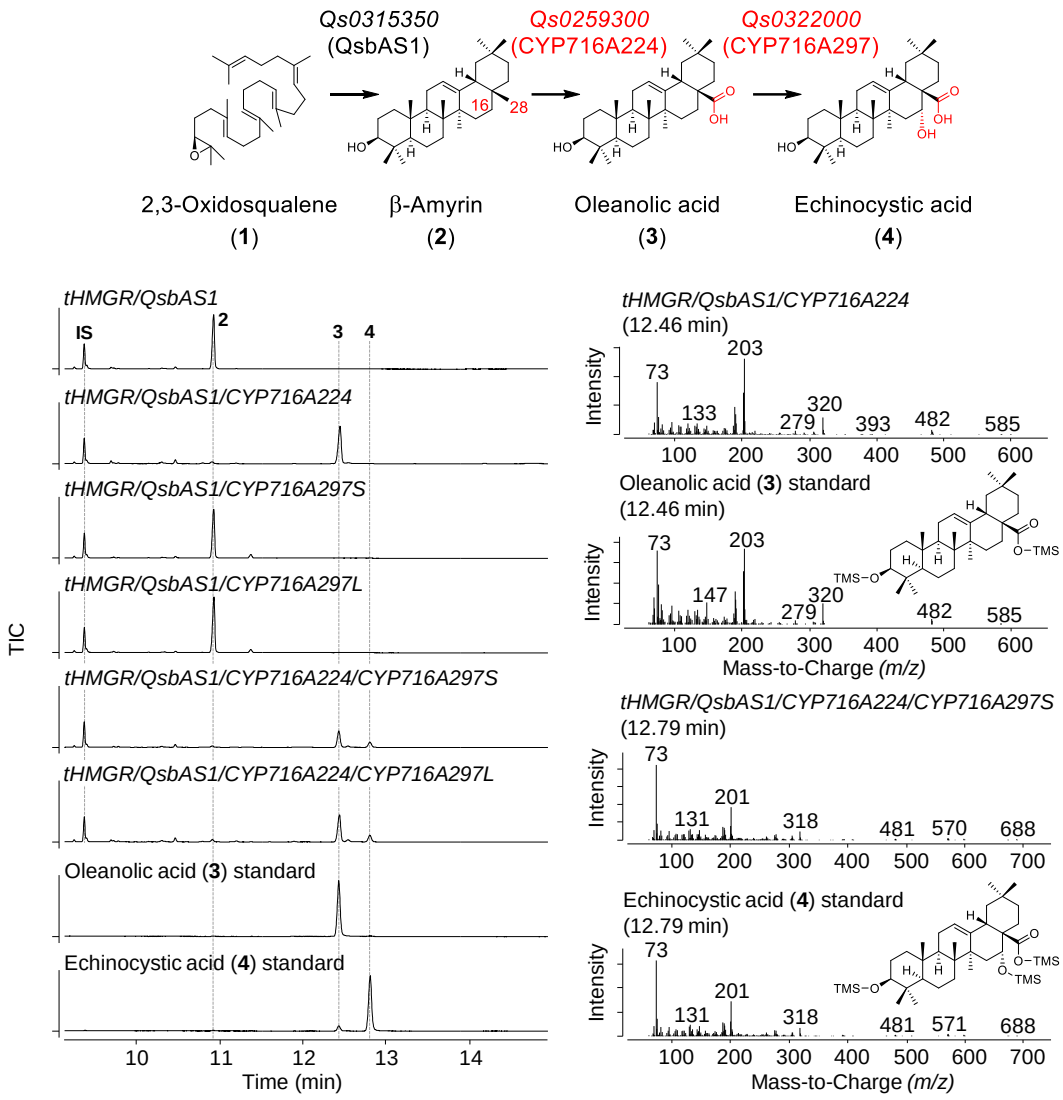


Fig. S3. CYP716A224 and CYP716A297 convert *β*-amyrin (2) to echinocystic acid (4). GC-MS analysis of leaf extracts of *N. benthamiana* following *Agrobacterium*-mediated transient expression. Leaves were agro-infiltrated with expression constructs for *tHMGR/QsbAS1* (control), *tHMGR/QsbAS1/CYP716A224*, *tHMGR/QsbAS1/CYP716A297* or *tHMGR/QsbAS1/CYP716A224/CYP716A297*. Total ion chromatograms (TIC) are shown on the left, and mass spectra on the right. Co-expression of *CYP716A224* with *QsbAS1* lead to complete conversion of *β*-amyrin (**2**) to a new product with a retention time and mass spectrum identical to that of an oleanolic acid (**3**) standard. Little conversion of β-amyrin (**2**) was observed from the combination of *tHMGR/QsbAS1/CYP716A297*, however co-expression of *tHMGR/QsbAS1/CYP716A224/CYP716A297* resulted in a new product identified as echinocystic acid (**4**) based on comparison to an authentic standard. Note, two variants of *CYP716A297* were cloned, labelled "long" and "short" (*CYP716A297L* and *CYP716A297S*, respectively). The long variant has an extra 21 amino acids at the *N*-terminus which aligns poorly with other members of the CYP716 family. Both variants were found to be functional and the short variant was used for all experiments thereafter. IS, internal standard (coprostanol).

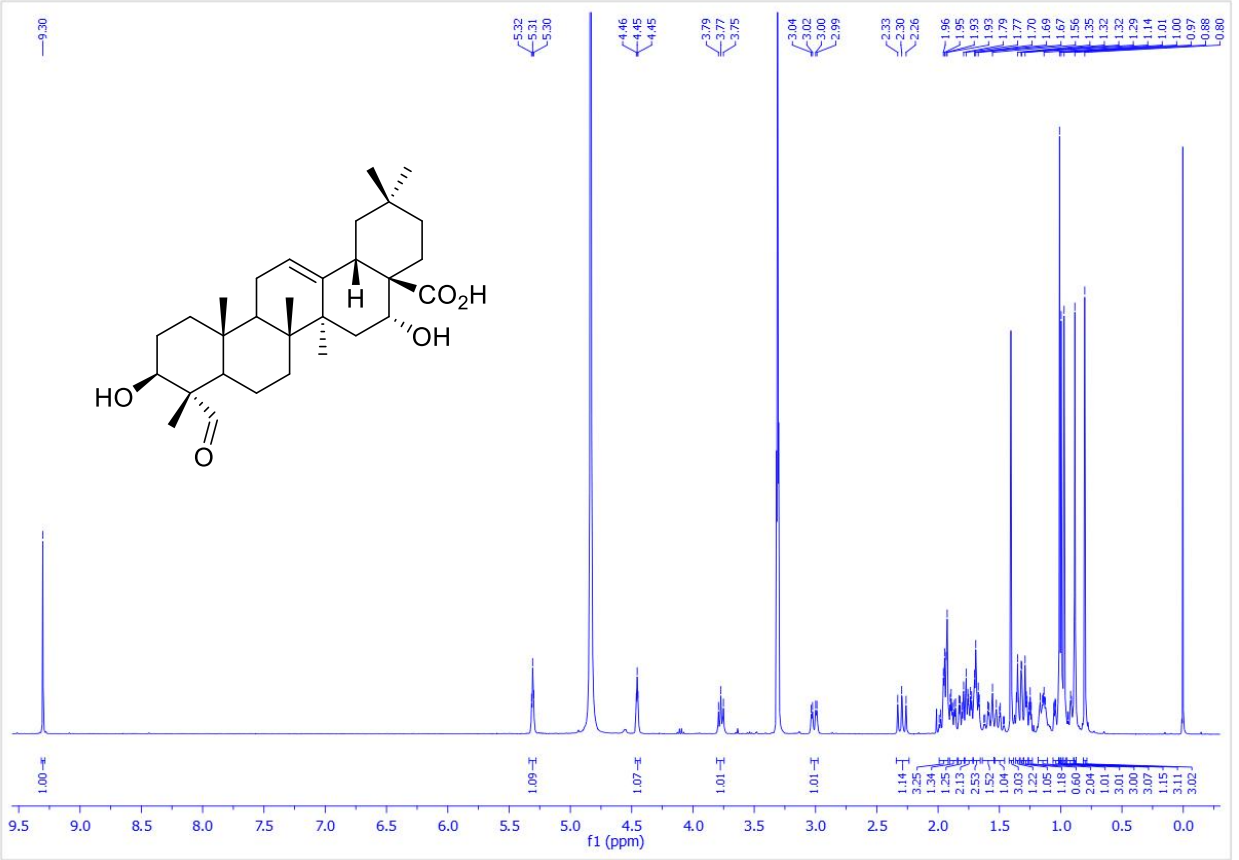


Fig. S4. Confirmation of the product of co-expression of QsbAS1, QsCYP716A224, QsCYP716A297 and QsCYP716A224 as quillaic acid (5) by ^{1}H-NMR. Recorded in MeOH-d_4, 600 MHz.

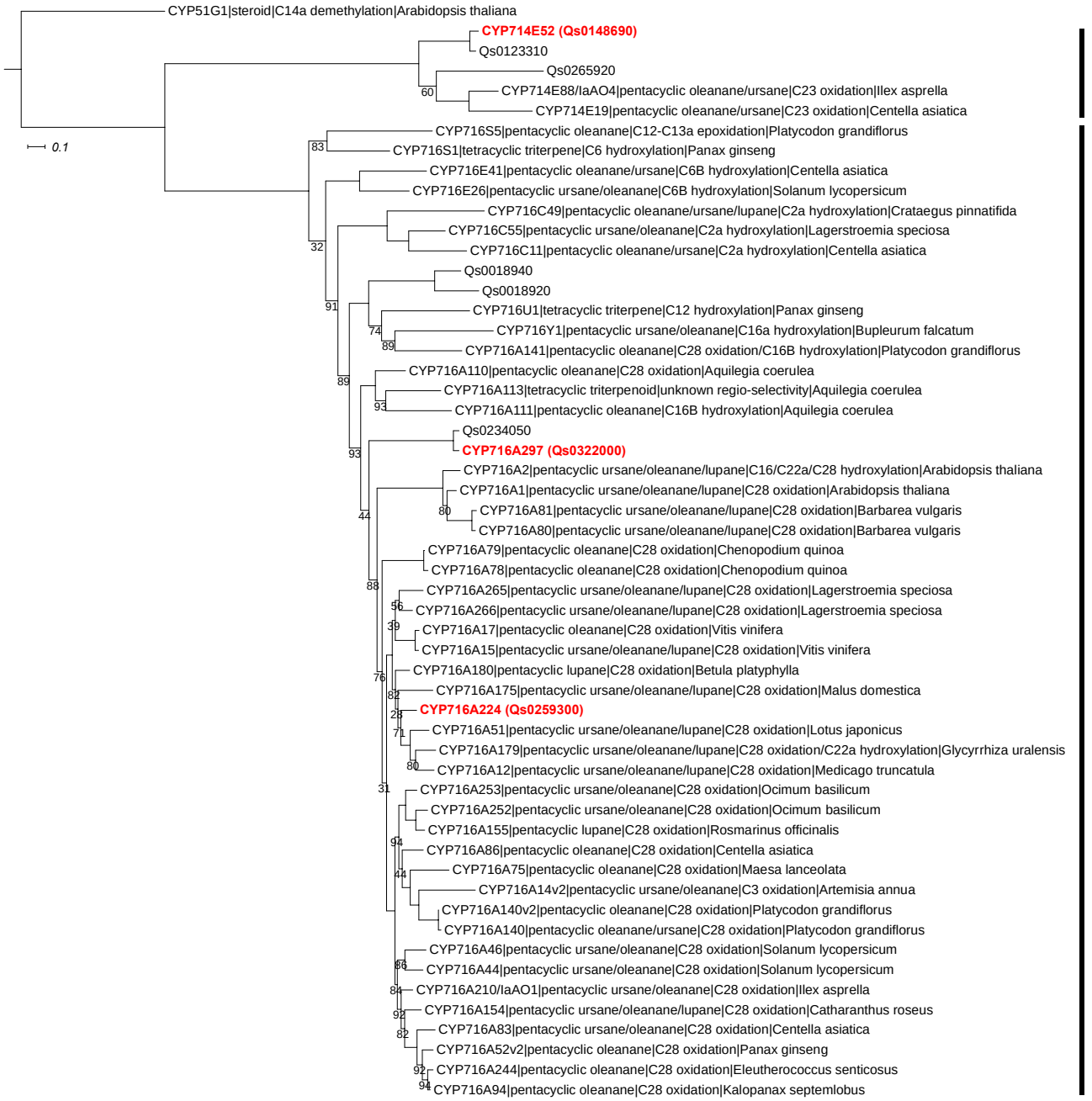


Figure S5. Maximum likelihood phylogeny of CYP716 and CYP714 sequences. All *Q. saponaria* CYP714 and CYP716 protein sequences were aligned with triterpene active CYP714 and CYP716 sequences as designated in (*8*). CYP51G1 (*Arabidopsis thaliana*) was used as an outgroup. The three CYPs identified in this study required to make quillaic acid are highlighted in red. The remaining triterpene functional CYPs are labelled with CYP name, scaffold class, reaction and species according to (*8*). Bootstrap values above 95 are not shown.

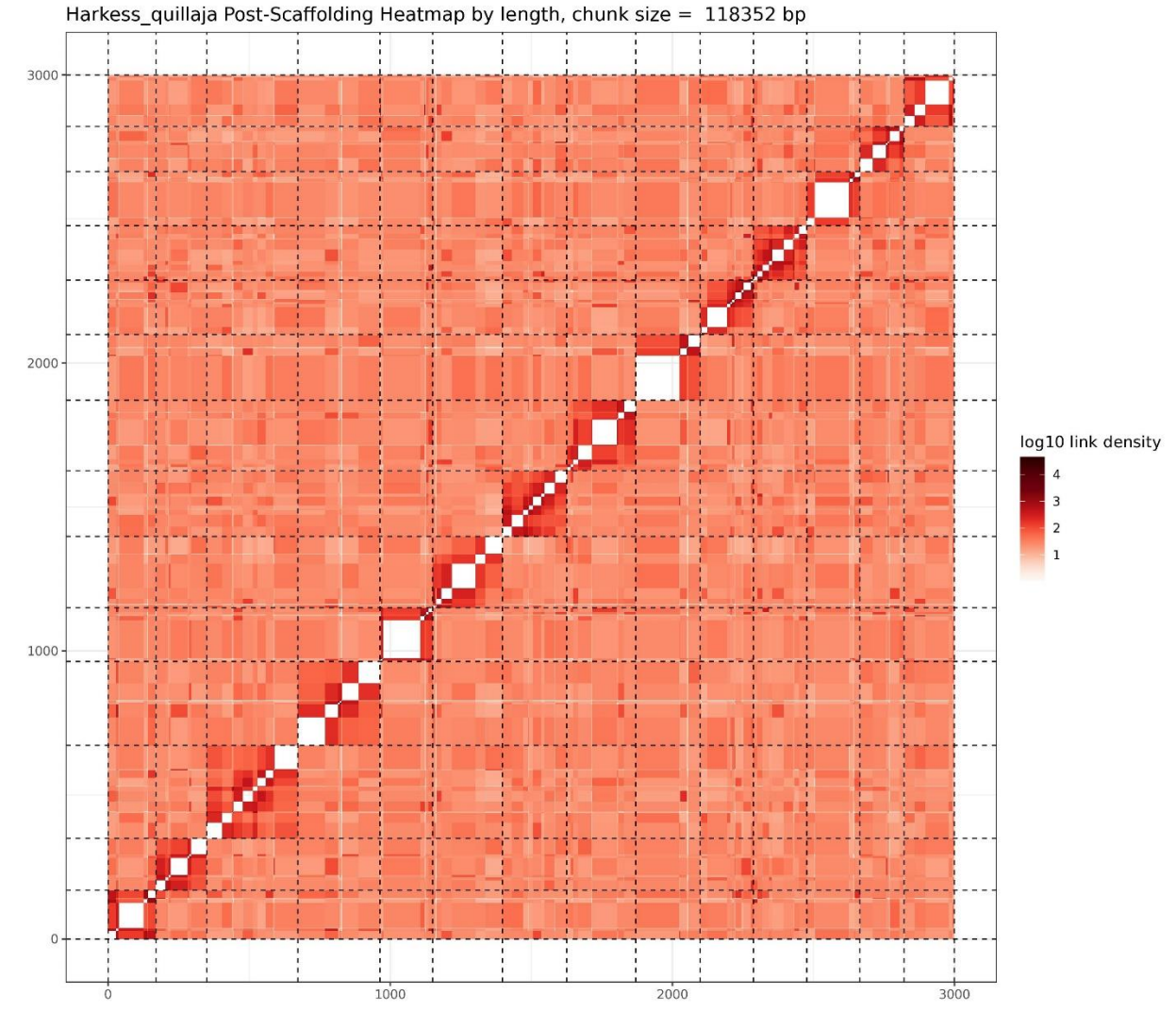


Fig. S6. Hi-C contact map showing 14 chromosomes. The axes show the 14 scaffolds generated by the scaffolding of the Hi-C links within and between contigs. Hi-C links between contigs are marked on corresponding coordinates in color: the darker the color, the higher is the linkage density between the contigs, indicating a higher probability of pertaining to the same chromosome. The linkages occurring within a same contig are marked in white as they are uninformative.

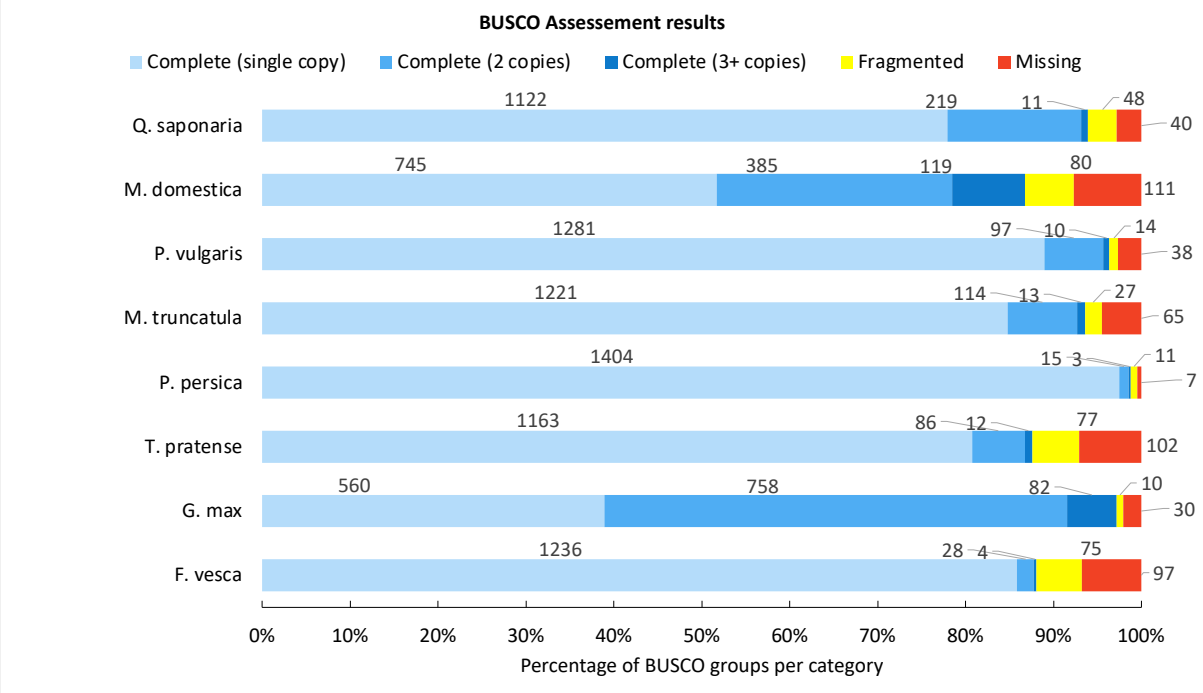


Fig. S7. BUSCO assessment of *Q. saponaria* protein complement compared to other species. BUSCO assessment carried out using embyrophyta dataset (odb9). RefSeq genomes of species used for comparison were: *F. vesca*, GCF_000184155.1*; G. max*, GCF_000004515.6*; T. pratense*, GCF_020283565.1*; P. persica*, GCF_000346465.2*; M. truncatula*, GCF_003473485.1*; P. vulgaris*, GCF_000499845.1*; M. domestica* GCF_002114115.1.

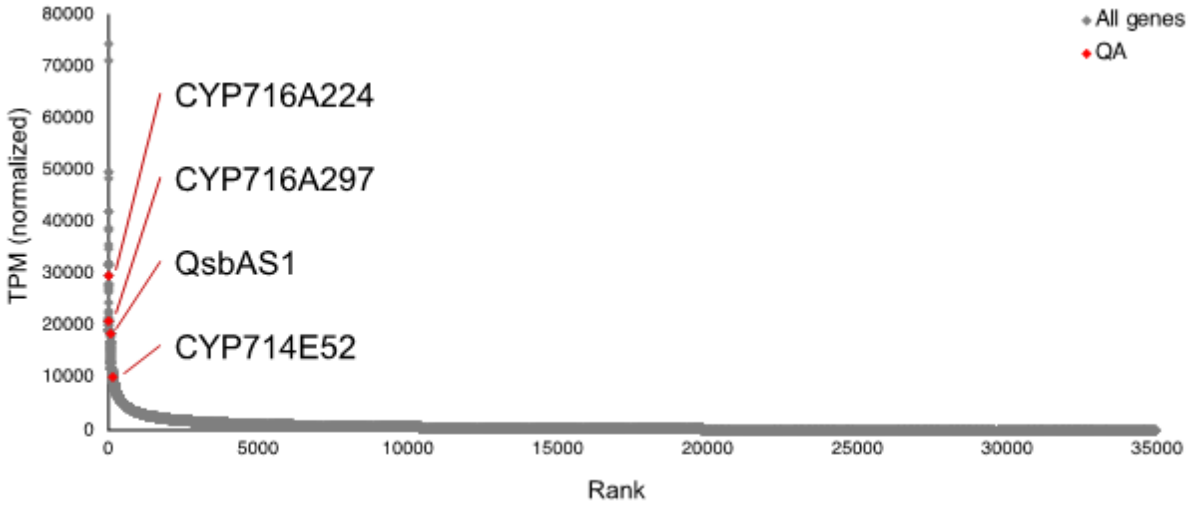


Fig. S8. Mean absolute expression levels of genes in primordia tissue. Mean TPM counts were normalized using size-factor estimates in DESeq2 (*57*). The four genes required to produce QA are within the top 160 genes when ranked by absolute transcript abundance in primordia tissue.

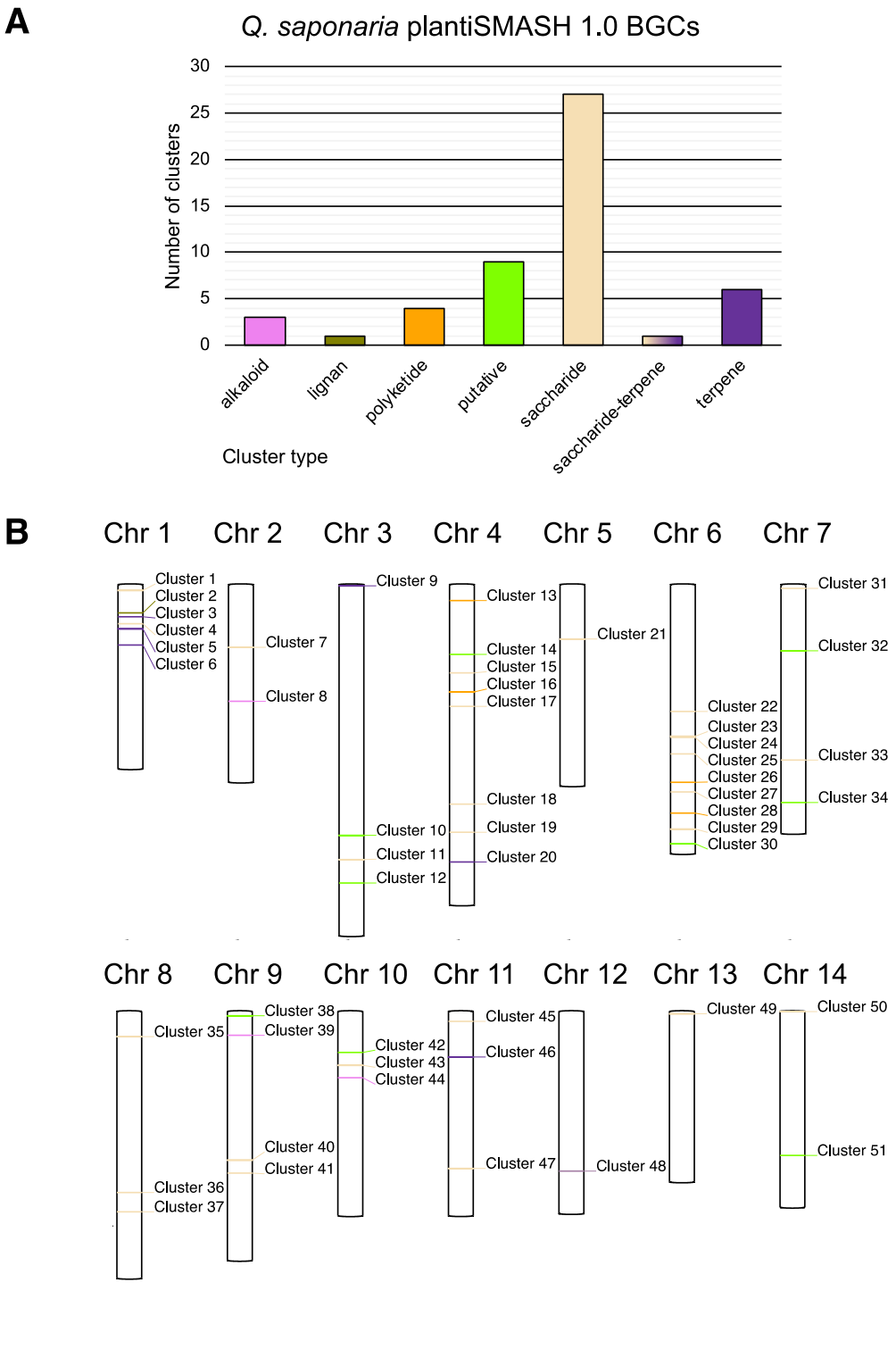


Fig. S9. Results for plantiSMASH mining of the *Q. saponaria* S10 genome assembly. (A) Counts of putative BGC type classifications. The majority of BGCs annotated are of the 'saccharide' type. (**B**) Chromosomal locations of putative BGCs. Colors of bands indicate BGC type as in (**A**).

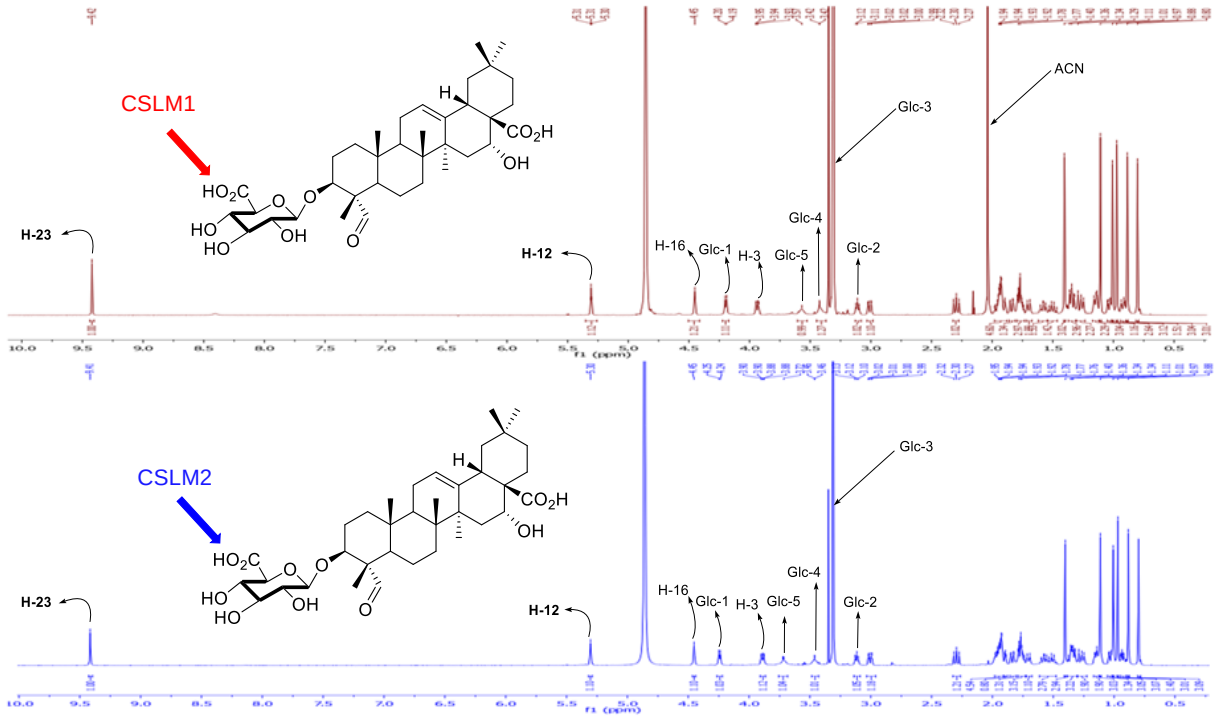


Fig. S10. ^{1}H NMR spectra comparison of quillaic acid 3-*O*-*β*-D-glucopyranosiduronic acid (6) produced by CSLM1 (Top) and CSLM2 (Bottom), MeOH-d_4, 600 MHz

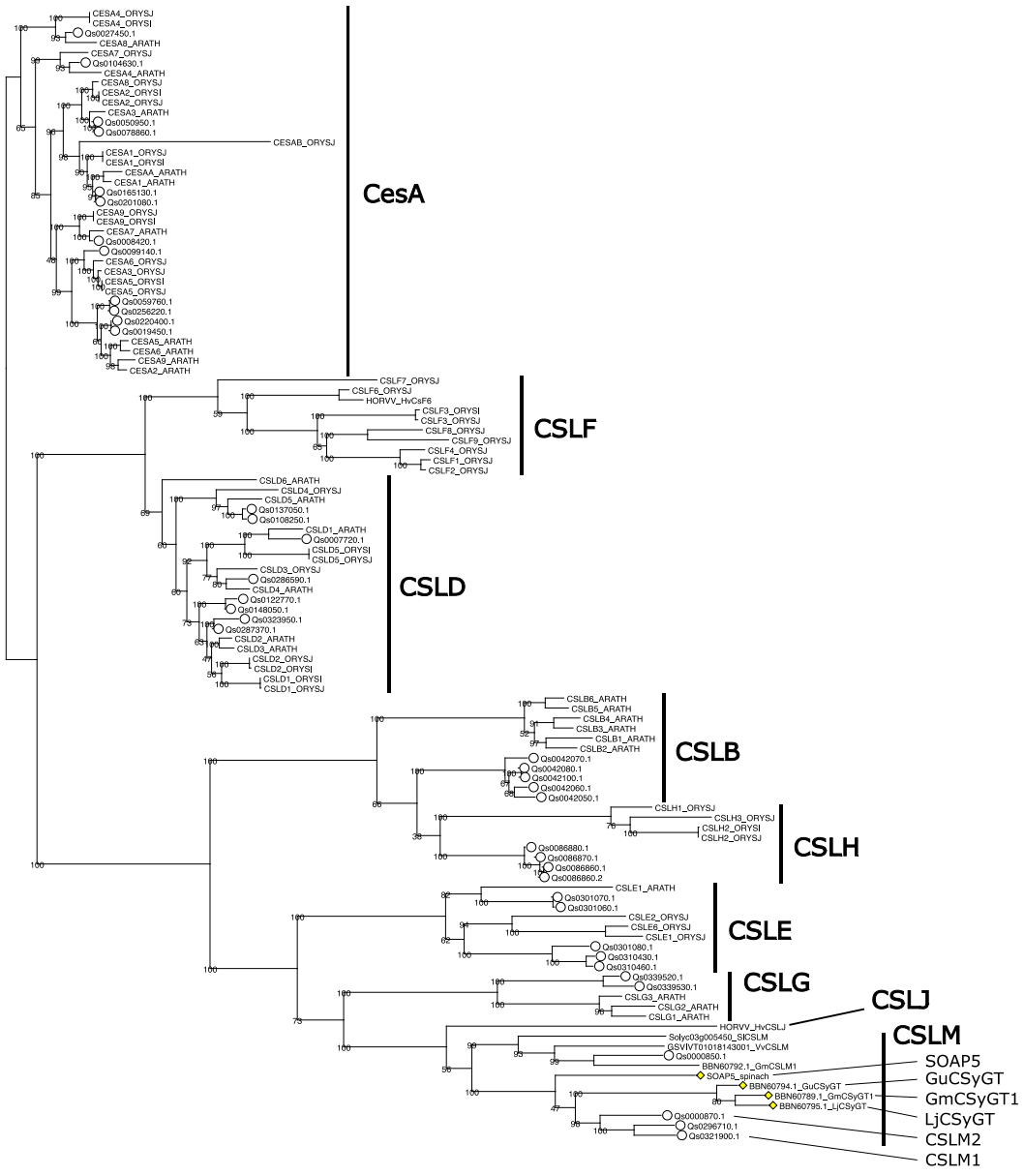


Fig. S11. Phylogeny of cellulose synthase (CesA) and cellulose synthase-like (CSL) genes. Genes from *Q. saponaria* are indicated with white circles. All UniProt sequences of this gene family were included from *Arabidopsis thaliana*, *Oryza sativa* Japonica and *Oryza sativa* Indica (UniProt names displayed on tree). GenBank IDs for additional genes included in the alignment and phylogeny in are from *Vitis vinifera*: *GSVIVT01018143001_VvCSLM* (CBI26389.3); *Hordeum vulgare*: *HORVV_HvCsF6* (XP_044960146.1), *HORVV_HvCsLJ* (XP_044974896.1); *Solanum lycopersicum*: *Solyc03g005450_SlCSLM* (XP_004234035.1); *Glycine max*: *BBN60792.1_GmCslM1* (XP_003536256.1), *GmCSyGT1* (XP_006582441.1); *Glycyrrhiza uralensis*: *GuCSyGT* (BBN60794.1); *Lotus japonicus*: *LjCSyGT* (BBN60795.1) and *Spinacia oleracea*: *SOAP5* (XP_021842158.1). Genes marked with yellow diamonds indicate characterised GlcA transferase activity. *QsCSLM1* and *QsCSLM2* encode functional GlcA transferase enzymes as discussed in the main text.

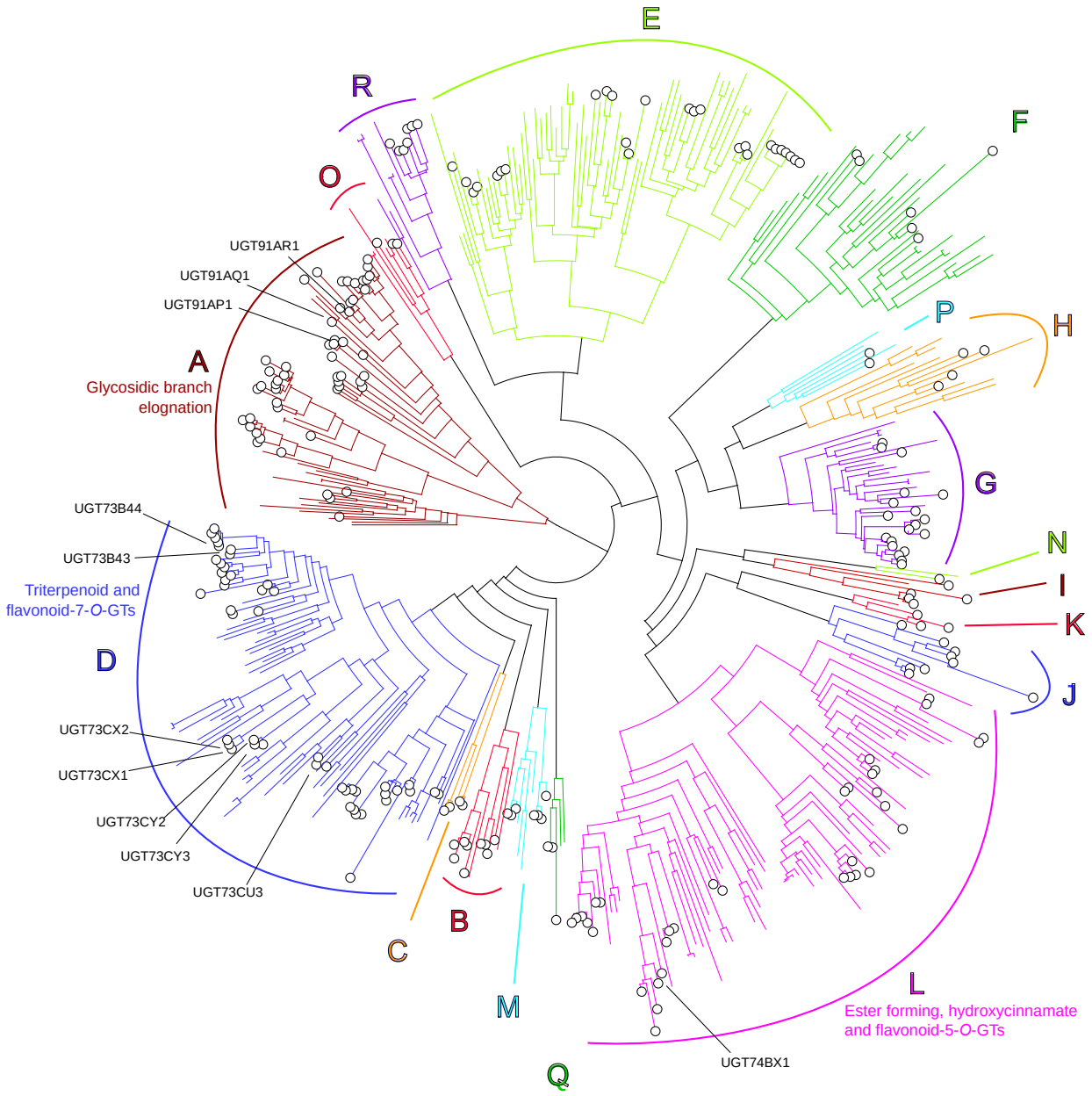
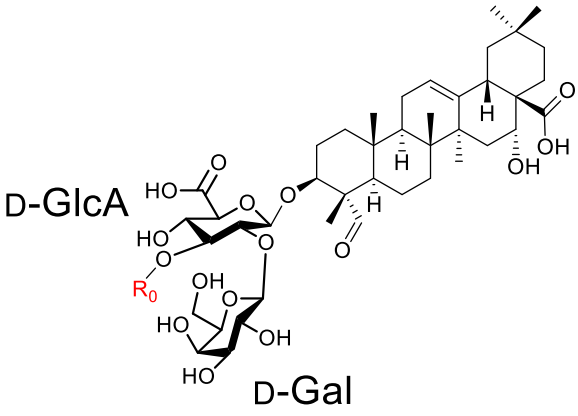


Fig. S12. Phylogeny of UGT genes. Genes from *Q. saponaria* are indicated with white circles. Other genes included in the alignment and phylogeny are from (*21*), in order to classify the UGTs into the labelled groups (*68*). UGTs indicated with names are functional saponin biosynthetic genes as discussed in the main text. Typical activates of UGT clades are labelled for group A, D and L as per *(21)*.



Compound	R_0
1	H
2	α-L-Rha*p*
3	β-D-Xyl*p*

Fig. S13. Reported structures of *Q. saponaria* saponins indicate that C-3 glycosylation is likely to precede modification at C-28. Isolation of C-3 glycosylated saponins from *Quillaja* bark extract has previously been reported (*69*). Figure adapted from Guo et al. (*69*) and compound numbers denoted here are taken from this paper. Compounds **1**, **2** and **3** above correspond to compounds **7** (QA-Di), **9** (QA-TriR) and **8** (QA-TriX) in the present study, respectively. To our knowledge, there are no reported saponins isolated from *Quillaja* sp. featuring glycosylation at C-28 in the absence of C-3 glycosylation.

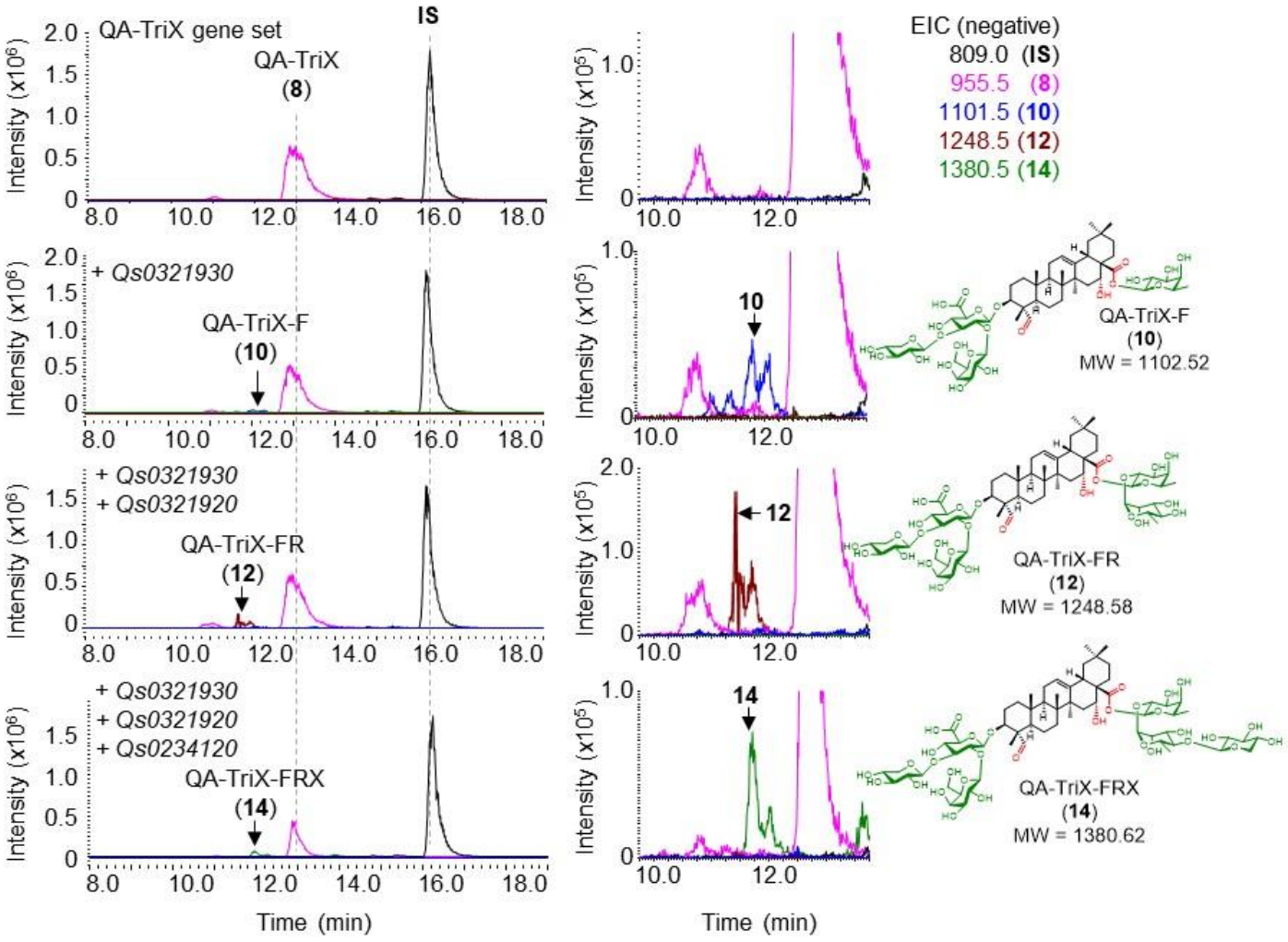


Fig. S14. Identification of the first three glycosyltransferases at C-28. LC-MS extracted ion chromatograms (EIC) of *N. benthamiana* leaf extracts following transient expression of the QA-TriX (**8**) (*m/z 955*) gene set. Co-expression of *Qs0321930* resulted in appearance of trace amounts of a new product consistent with addition of D-fucose, anticipated to be QA-TriX-F (**10**) (*m/z 1101*). Further co-expression of *Qs0321920* resulted in conversion of **10** to a new product consistent with addition of L-rhamnose, anticipated to be QA-TriX-FR (**12**) (*m/z 1248*). Finally, co-expression of *Qs0234120* resulted in conversion of **12** to a new product consistent with addition of a xylose and anticipated to be QA-TriX-FRX (**14**) (*m/z 1380*). Only trace amounts of these compounds were observed, with most of the QA-TriX (**8**) still present in the extracts, suggesting that D-fucosylation is a limiting step. IS, internal standard (digitoxin). The portion of the chromatogram showing the new compounds is expanded to the right.

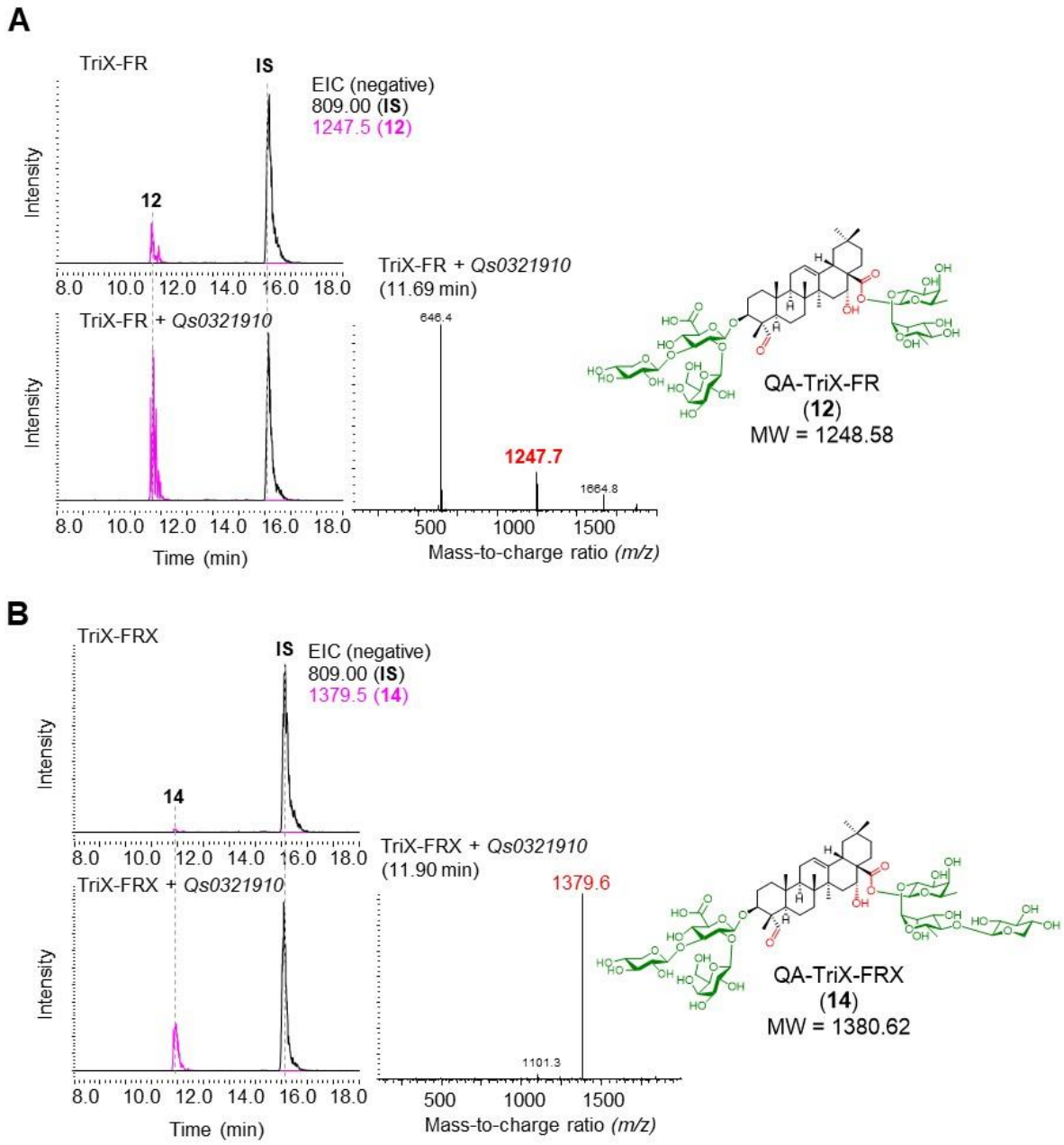


Fig. S15. The SDR encoded by Qs0321910 allows increased production of C-28 glycosides. LC-MS Extracted Ion Chromatograms (EIC) of *N. benthamiana* leaf extracts following transient expression of the gene set for production of the D-fucosylated saponin QA-TriX-FR (**12**) (**A**) and QA-TriX-FRX (**14**) (**B**) in the absence (top) or presence (bottom) of *Qs0321910.* The presence of *Qs0321910* also results in substantial increases to these products, likely arising from the increased abundance of the precursor QA-TriX-F (**10**). IS, internal standard (digitoxin)

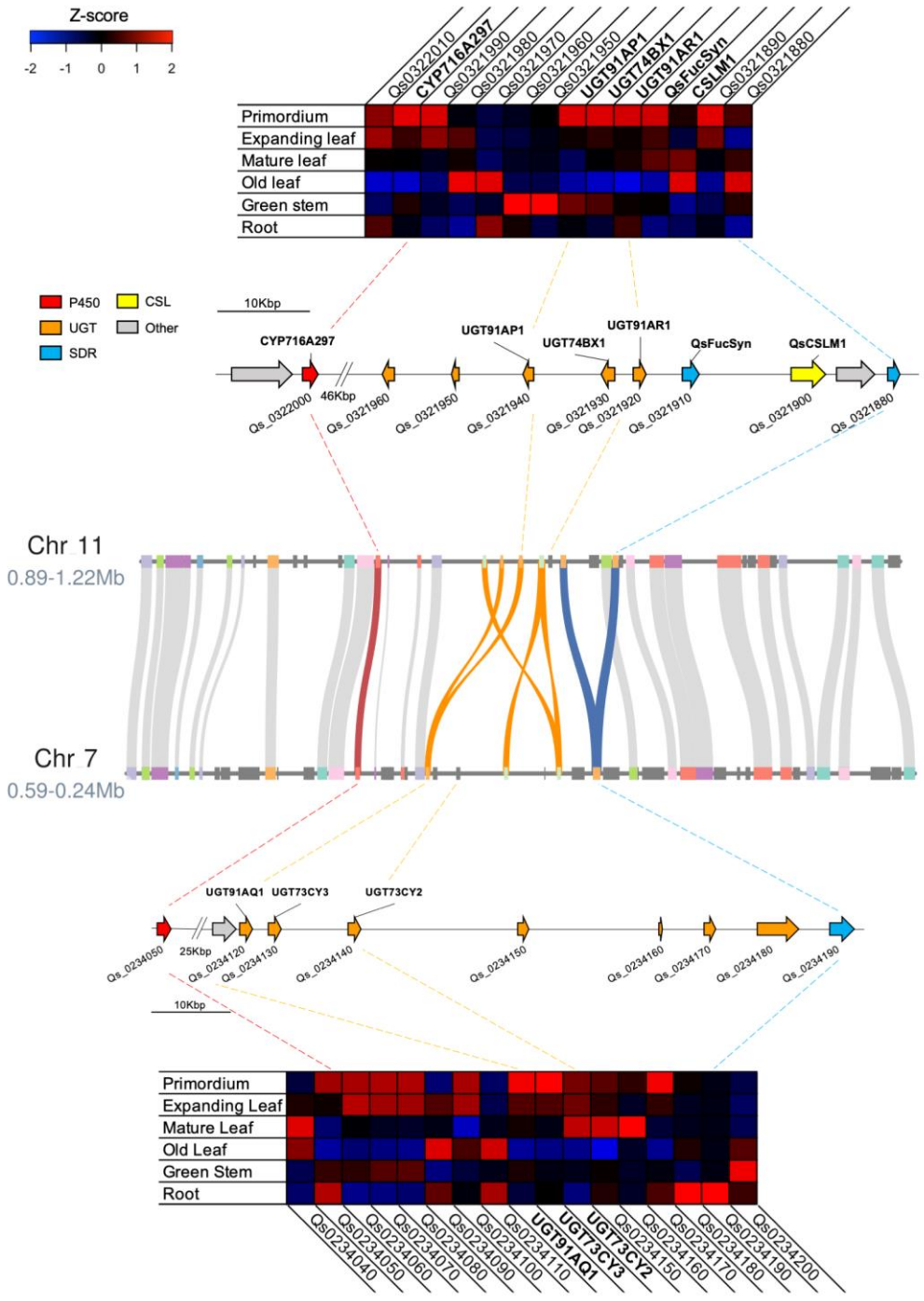


Fig. S16. Microsynteny and expression of functional BGCs in *Q. saponaria*. Syntenic regions encompassing plantiSMASH clusters #45 (chromosome 11, above) and #31 (chromosome 7, below). Synteny analysis was carried out using Python-implemented MCScan (*70*). Heatmaps show *Z*-scores (generated from DESeq2 VST- transformed read quantification values).

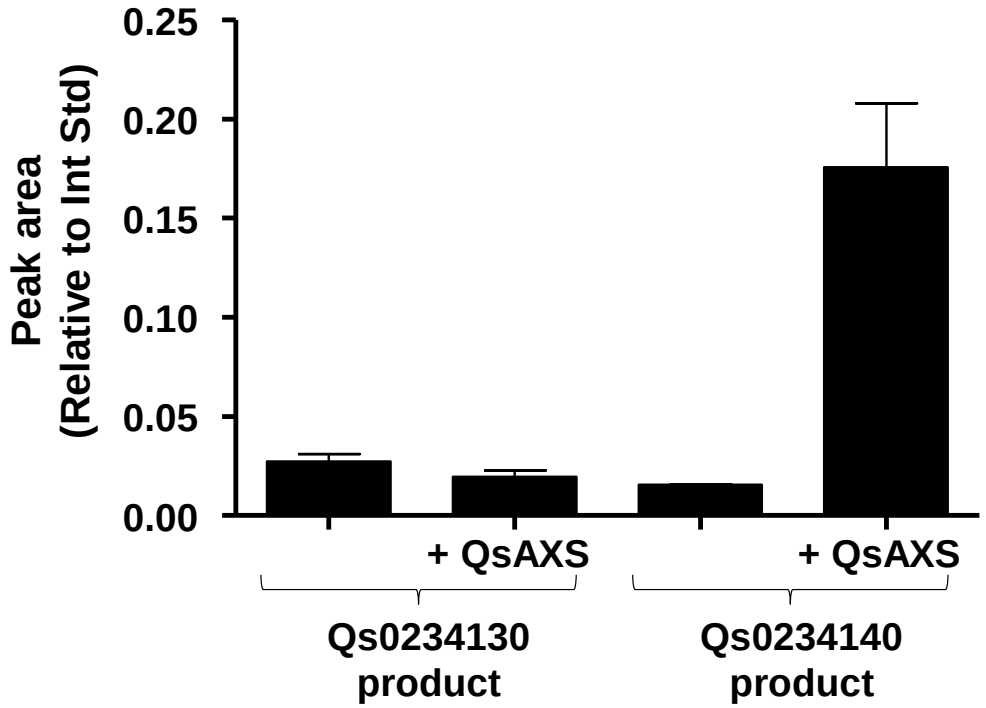


Fig. S17. UDP-apiose/UDP-xylose synthase (QsAXS) substantially increases the Qs0234140 product. Transient expression of the gene set for QA-TriX-FRX (**14**) plus either Qs0234130 or Qs0234140 was performed. Each of these two gene sets were tested in the presence or absence of QsAXS and the relevant products were quantified based on relative peak area versus the internal standard (digitoxin). The presence of AXS resulted in approximately 11-fold increases to the *Qs0234140* product. AXS did not increase the abundance of the *Qs0234130* product, however. Three separate leaves from different plants were infiltrated for each of the four conditions. Error bars show standard deviation.

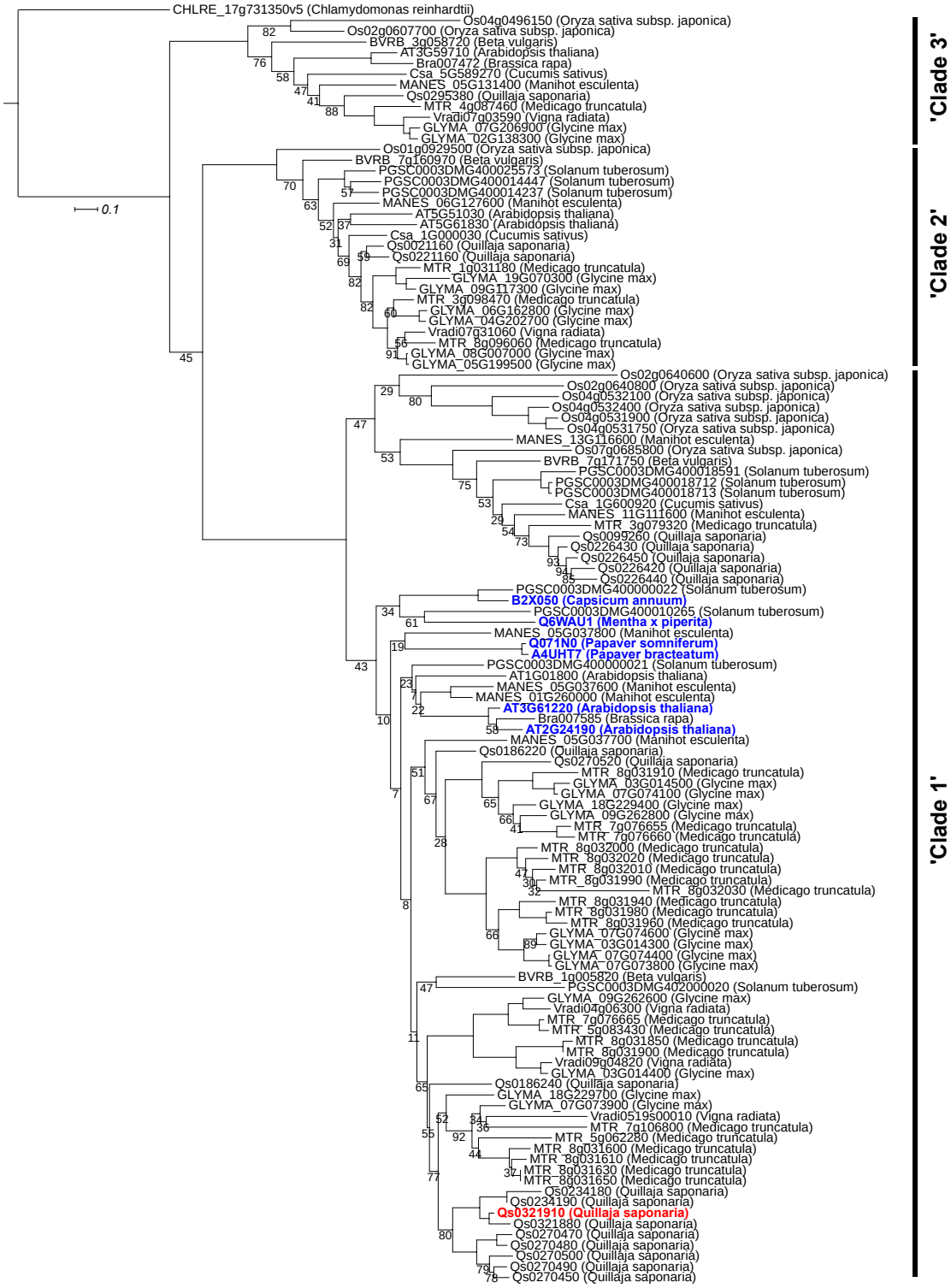


Fig. S18. Maximum likelihood phylogeny of SDR114C family genes. SDR114C genes were extracted from representative plant genomes using HMMER with a profile derived from the representative genes identified in (*29*). Gene locus IDs and species names are displayed on the

tree. Labelled clades are according to (*71*). Functional SDRs are highlighted in blue, with two (+)-neomenthol dehydrogenases from *Arabidopsis thaliana* (AT3G61220, AT2G24190). In addition to the full SDR114C complements from representative species, four additional functional SDRs are included: a (+)-neomenthol dehydrogenase (Uniprot ID B2X050, *Capsicum annum*), an (-)-isopiperitenone reductase (Uniprot ID Q6WAU1, *Mentha x piperita*) and two salutaridine reductases (Uniprot ID Q071N0, *Papaver somniferum*; Uniprot ID A4UHT7, *Papaver bracteatum*). Bootstrap values above 95 are not shown.

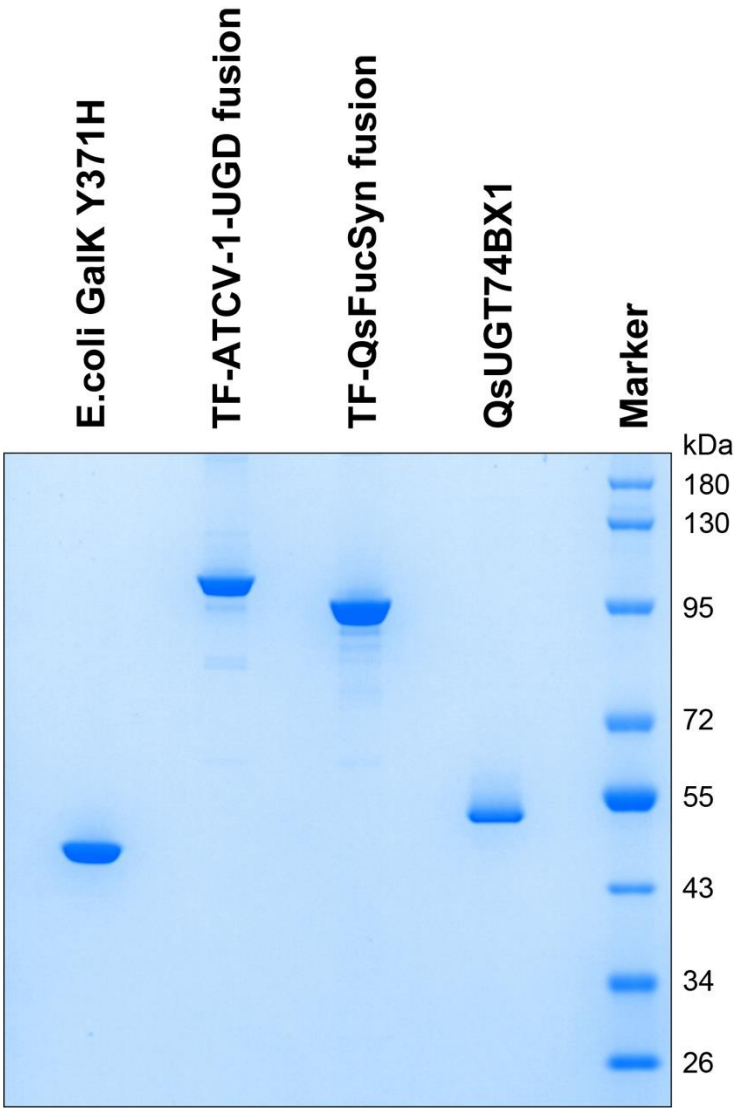


Fig. S19. Purified enzymes for *in vitro* studies. GalK Y371H, *Escherichia coli* galactokinase with a Y371H mutation. TF-ATCV-1-UGD, a fusion protein of *E. coli* trigger factor (TF) and UDP-D-glucose 4,6-dehydratase from the *Acanthocystis turfacea* chlorella virus 1. TF-QsFucSyn, a fusion protein of *E. coli* trigger factor and *Qs0321910* SDR. QsUGT74BX1, *Q. saponaria* UGT74BX1. The rightmost lane is blue pre-stained protein standard, broad range (New England Biolabs, P7718).

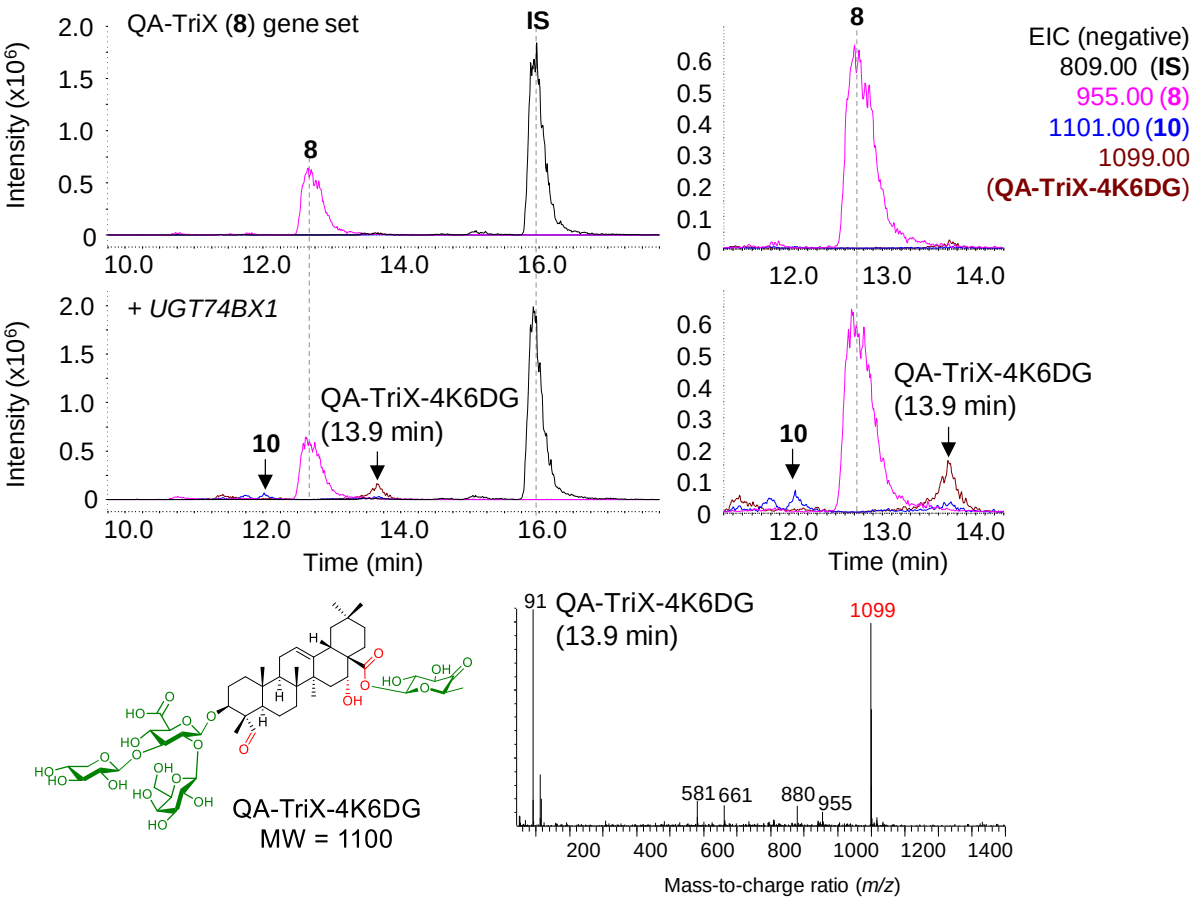


Fig. S20. Evidence of TriX-4K6DG production in *N. benthamiana*. The gene set for QA-TriX-F production was expressed in *N. benthamiana* without the yield-boosting *QsFucSyn*. A peak with a mass consistent with QA-TriX-4K6DG (*m/z* = 1099) can be seen at 13.9 mins.

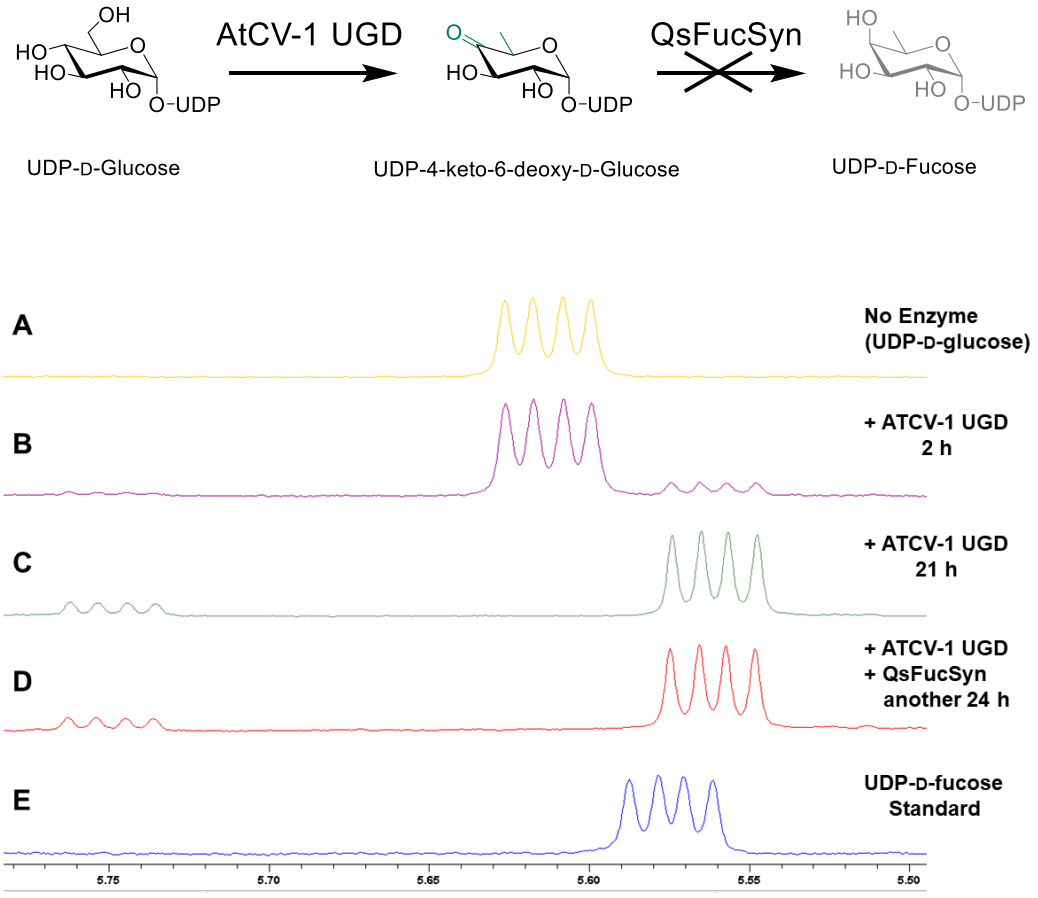


Fig. S21. QsFucSyn does not convert UDP-4-keto-6-deoxy-D-glucose to UDP-D-fucose *in vitro*. Expanded ^{1}H NMR spectra to show the anomeric proton 1" (H1") resonances of starting material and products. No enzyme control shows UDP-D-glucose (panel **A**). Two hours after addition of TF-ATCV-1-UGD showing appearance of tiny peaks of H1" of UDP-4-keto-6-deoxy-D-Glucose at 5.75 ppm and H1" of its hydrate form at 5.56 ppm (panel **B**). Twenty-one hours incubation achieved complete conversion from UDP-D-glucose into UDP-4-keto-6-deoxy-D-glucose and its hydrate form (panel **C**). Another 24 h incubation after addition of TF-QsFucSyn did not cause any changes in the H1" resonance (panel **D**). H1" resonance of standard UDP-D-fucose at 5.57 ppm (panel **E**).

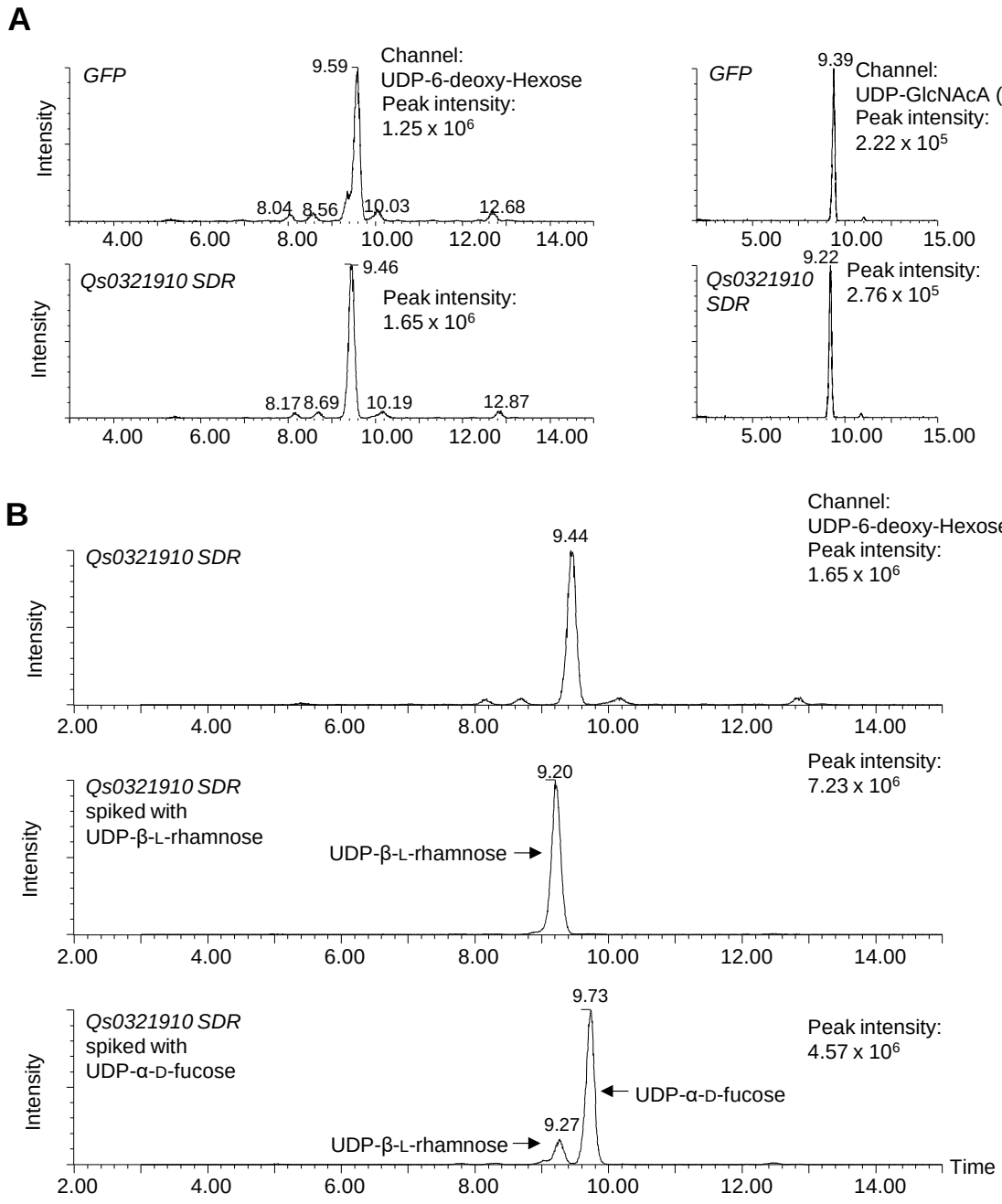


Fig. S22. Sugar nucleotide profiling of *N. benthamiana* following transient expression of *Qs0321910* SDR. (**A**) Sugar nucleotide analysis of extracts of leaves expressing either green fluorescent protein (GFP, top) or *Qs0321910* SDR (bottom). Both extracts look comparable with only a single peak observed in the UDP-deoxyhexose channel, likely corresponding to UDP-β-L-rhamnose. To the right, the channel for the internal standard (UDP-2-acetamido-2-deoxy-α-D-glucuronic acid (UDP-GlcNAcA) is shown to demonstrate equivalent amounts of sample being analyzed). (**B**) The *Qs0321910* SDR leaf extract sample (top) was further spiked with standards of either UDP-L-rhamnose (middle) or UDP-D-fucose (bottom). UDP-L-rhamnose co-eluted with the peak observed in the SDR leaf extract sample, demonstrating that the UDP-6-deoxy-hexose

extracted from *N. benthamiana* is UDP-L-rhamnose. In contrast, UDP-D-fucose elutes as a separate peak at 9.73 minutes (bottom), suggesting that this sugar nucleotide is not present in *N. benthamiana.* There was a significant shift in retention time of the target analytes between individual runs, a phenomenon well documented for porous graphitic carbon stationary phase (*67*). Thus, it was necessary to verify the identity of the UDP-sugar observed in *N. benthamiana* by spiking samples with authentic standards.

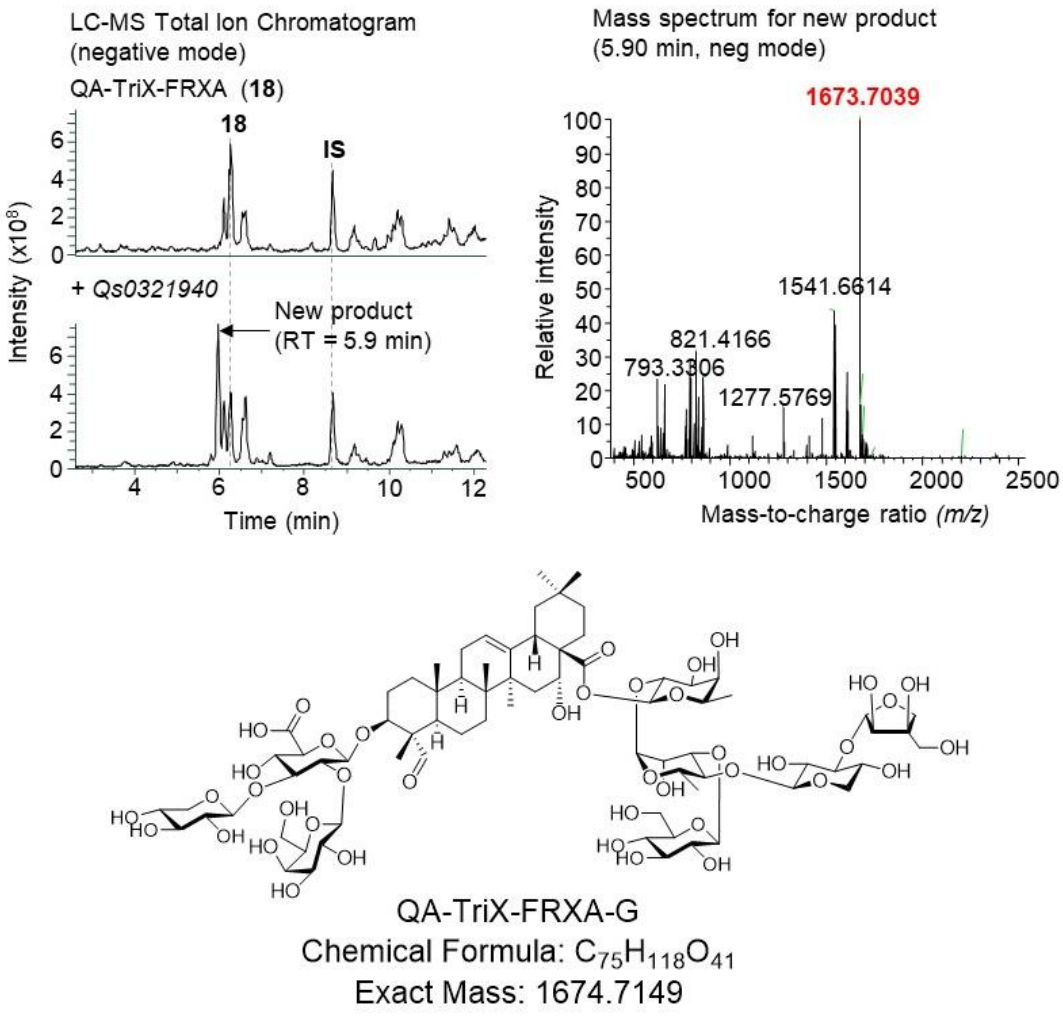
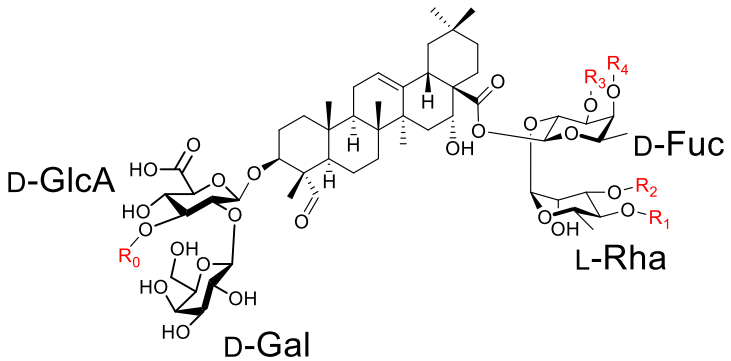


Fig. S23. Discovery of a probable glucosyltransferase encoded by *Qs0321940*. LC-HRMS total ion chromatograms of *N. benthamiana* leaf extracts following transient expression of the gene set for production of QA-TriX-FRXA (**18**). Further co-expression of *Qs0321940* resulted in appearance of a new product. The mass of this was consistent with addition of a hexose, anticipated to be glucose. The mass spectrum of this product is shown to the right, while the predicted structure (QA-TriX-FRXA-G) is shown below. IS, internal standard (digitoxin).

A



B

Name	Glycosylated at R3 of D-Fucose?	Acylated at R4 of D-Fucose?
4	Yes	Yes
5	Yes	Yes
6	Yes	Yes
7	Yes	Yes
8	Yes	Yes
9	Yes	Yes
10	Yes	Yes
11a/b	Yes	Yes
12a/b	Yes	Yes
13a/b	No	Yes
14a/b	Yes	Yes
15a/b	Yes	Yes
16a/b	Yes	Yes
17a/b	Yes	Yes
18a/b (QS-7)	Yes	Yes
19	Yes	Yes
20a/b	No	Yes
21a/b	No	Yes
22a/b	Yes	Yes
23	Yes	Yes
B1/2	No	Yes
B3/4	No	Yes
B5/6	No	Yes
B7/8	No	Yes
S1/2	No	Yes
S3/4 (QS-21)	No	Yes
S5/6 (QS-21)	No	Yes
S7/8	No	Yes
S9/10	No	Yes
S11/12	No	Yes
S13	Yes	Yes

Fig. S24. Glycosylation at R_3 of D-fucose is dependent on prior acylation at R_4. **(A)** The core structure commonly found in many *Q. saponaria* saponins. **(B)** Table of selected *Q. saponaria* saponins. Note that no saponins have been isolated which display glycosylation at the R_3 group of D-Fucose in the absence of an acyl group at R_4. Table adapted from Fleck et al. (*24*). The compound and R- numbers here correspond to those in that paper.

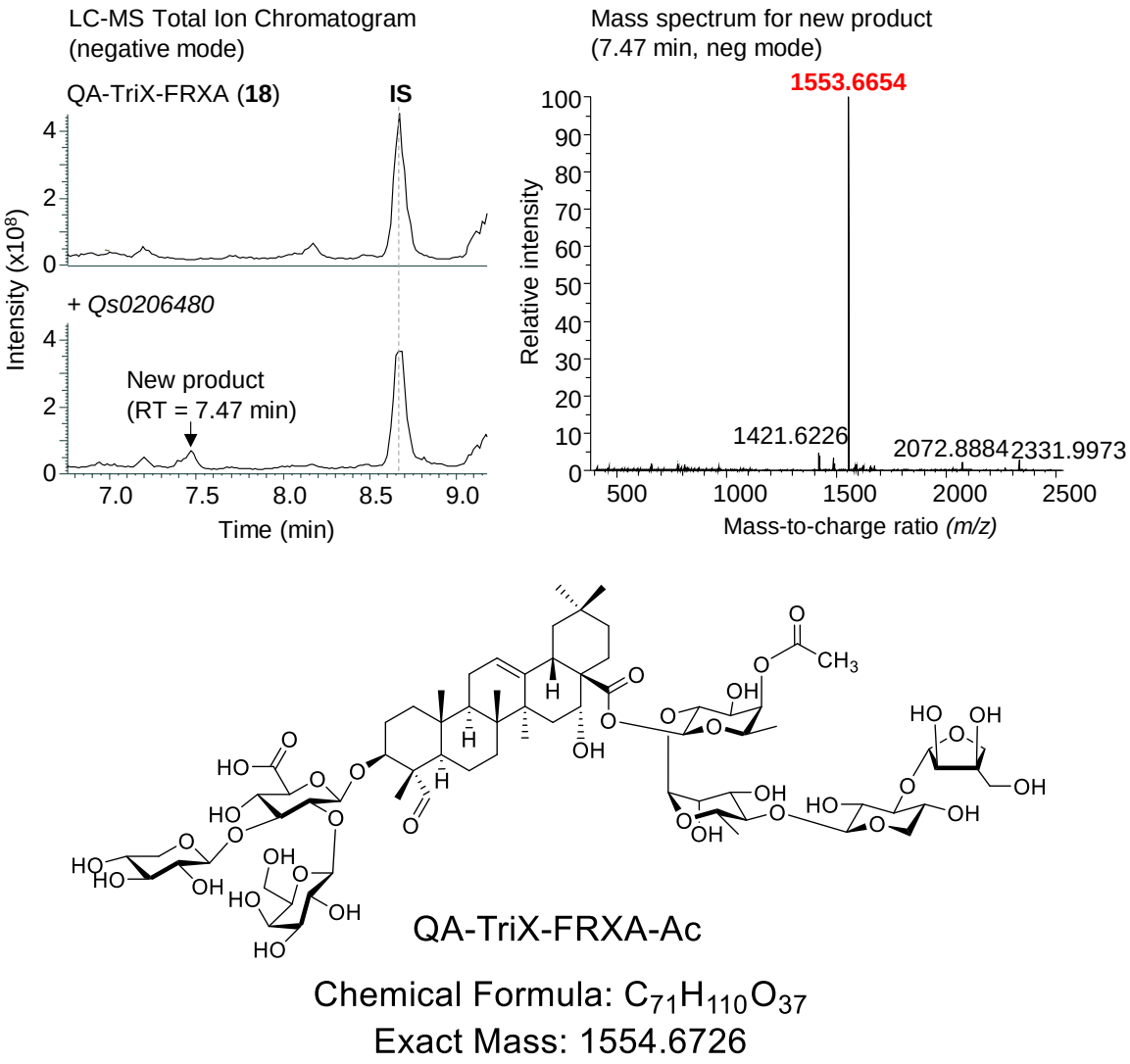
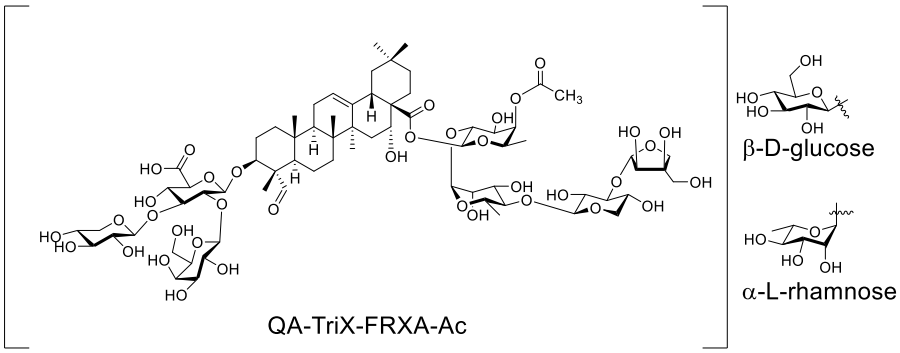


Fig. S25. Discovery of an acetyltransferase encoded by *Qs0206480*.
LC-HRMS total ion chromatograms of *N. benthamiana* leaf extracts following transient expression of the gene set for production of QA-TriX-FRXA (**18**). Further co-expression of *Qs0206480* generated a product with a mass consistent with **18** plus addition of an acetyl group. The mass spectrum of the product is shown to the right while the predicted structure of this compound (QA-TriX-FRXA-Ac) is shown below. IS, internal standard (digitoxin).



Compound	Chemical formula	Exact mass [M – H]$^-$
QA-TriX-FRXA-Ac	$C_{71}H_{110}O_{37}$	1553.6653
QA-TriX-FRXA-Ac + L-Rha	$C_{77}H_{120}O_{41}$	1699.7232
QA-TriX-FRXA-Ac + D-Glc	$C_{77}H_{120}O_{42}$	1715.7181
QA-TriX-FRXA-Ac + D-Glc + L-Rha	$C_{83}H_{130}O_{46}$	1861.7760

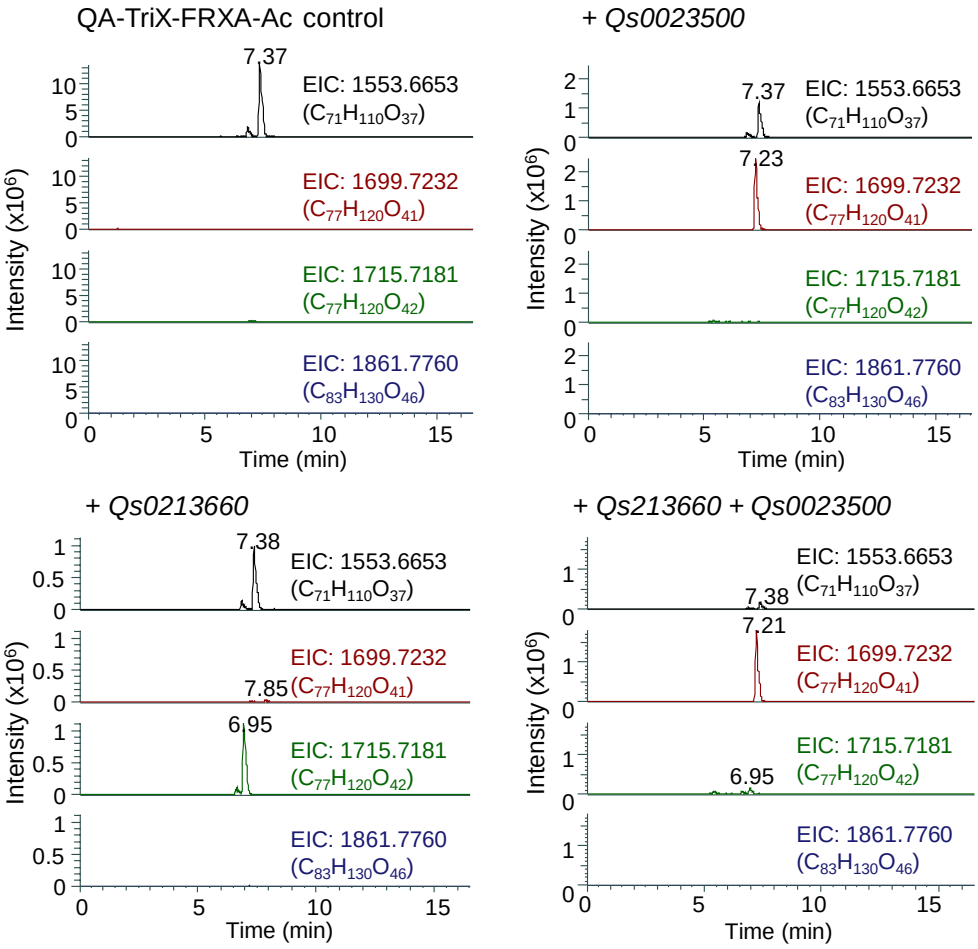


Fig. S26. Discovery of probable L-rhamnosyl and D-glucosyltransferases encoded by *Qs0023500* and *Qs0213660*, respectively. LC-HRMS extracted ion chromatograms (EIC) of *N.*

benthamiana leaf extracts following transient expression of the gene set for production of QA-TriX-FRXA-Ac (*m/z* 1553.6653) (top left). Further co-expression of *Qs0023500* generated a product with a mass consistent with QA-TriX-FRXA-Ac plus addition of a deoxyhexose, anticipated to be L-rhamnose (EIC *m/z* 1699.7232) (upper right). In contrast, co-expression of *Qs0213660* generated a product with a mass consistent with QA-TriX-FRXA-Ac plus addition of a hexose, anticipated to be D-glucose (EIC *m/z* 1715.7181) (lower left). Co-expression of both *Qs0023500* and *Qs0213660* did not result in a new product featuring both sugars (anticipated *m/z* 1861.7760), suggesting that the enzymes encoded by these genes are glycosylating the same position (lower right).

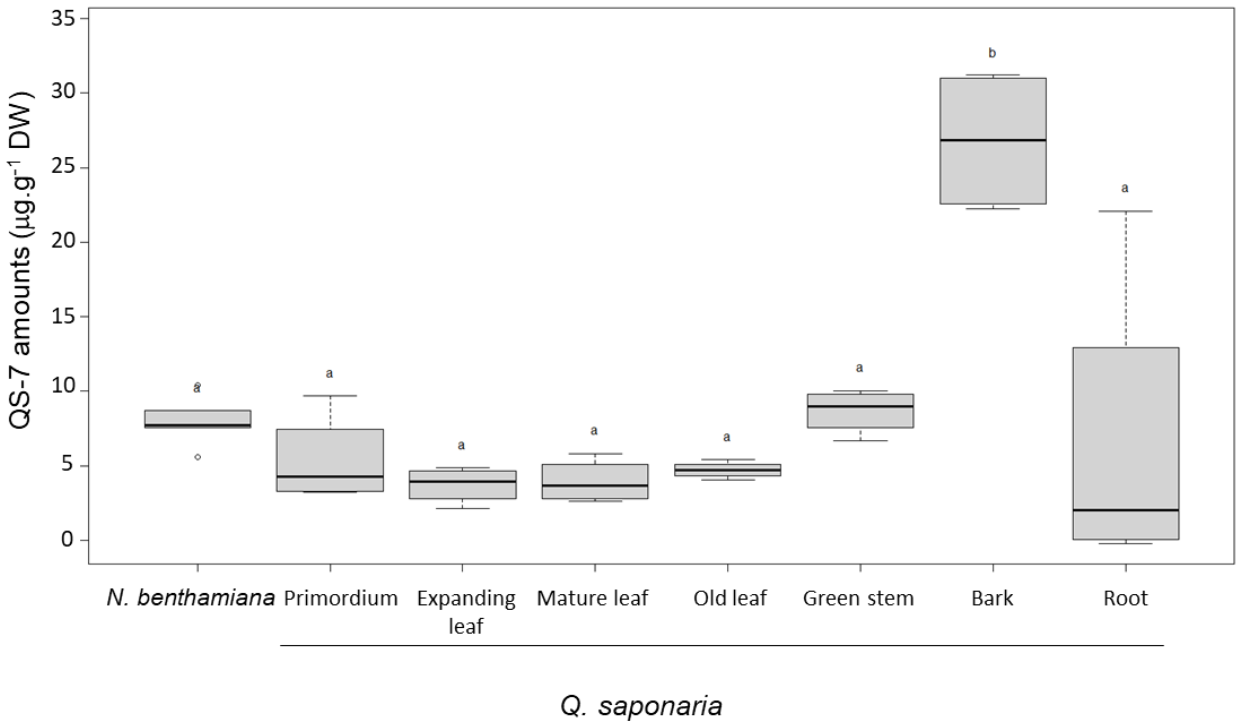


Fig. S27. QS-7 quantification in *Nicotiana benthamiana* and *Quillaja saponaria.* QS-7 was quantified relative to a QS-7 external standard curve and normalized based on the individual sample dry weight (15 mg dry material). The *N. benthamiana* samples are dried leaves from plants expressing the *Q. saponaria* gene set necessary to produce QS-7. The *Q. saponaria* tissues are as follows: Primordium (the tip of the branch that includes the meristem and 1 leaf smaller than 0.5 cm); expanding leaf (leaf that has reached about half its mature size); mature leaf (first leaf on the branch that has reached its mature size); old leaf (leaf at the base of the branch that has not started to senesce); green stem (part of the branch that is still green in color with no sign of lignification); bark (lignified tissue covering a branch); root (roots from various developmental stages growing out of the bottom of the pot). Bars, standard error (four biological replicates). Statistical analyses comprised ANOVA and Tukey tests and were done in R using the multcompView package.

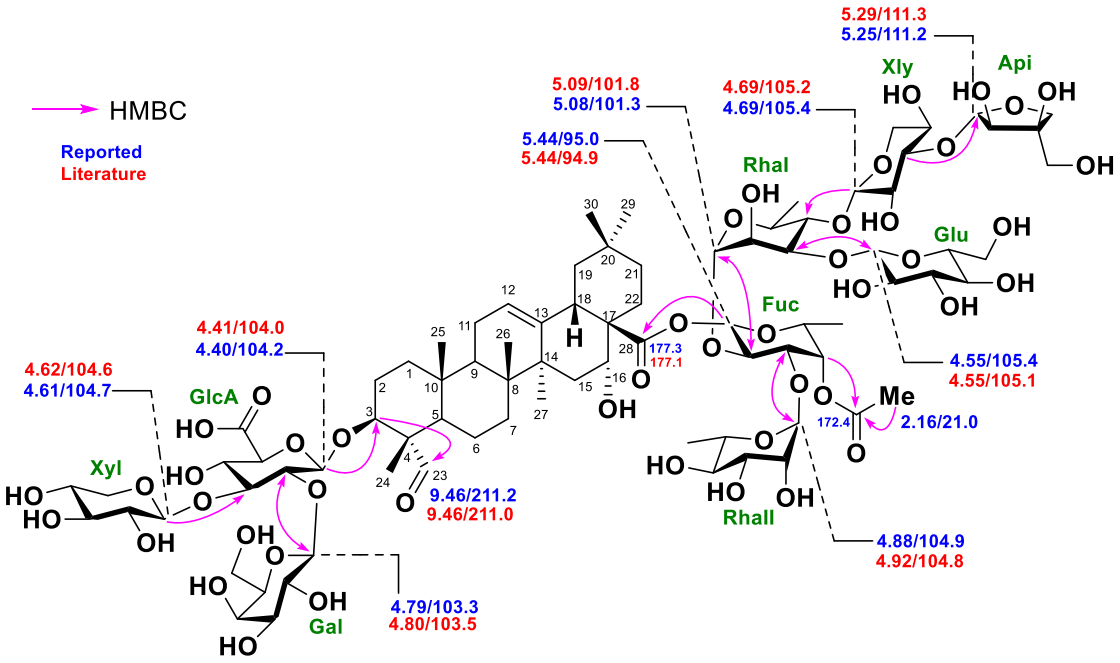


Fig. S28. Key HMBC (H→C) observed for semi-purified QS-7 (20) produced in *N. benthamiana.* NMR carried out in MeOH-d_4 (600, 150 MHz).

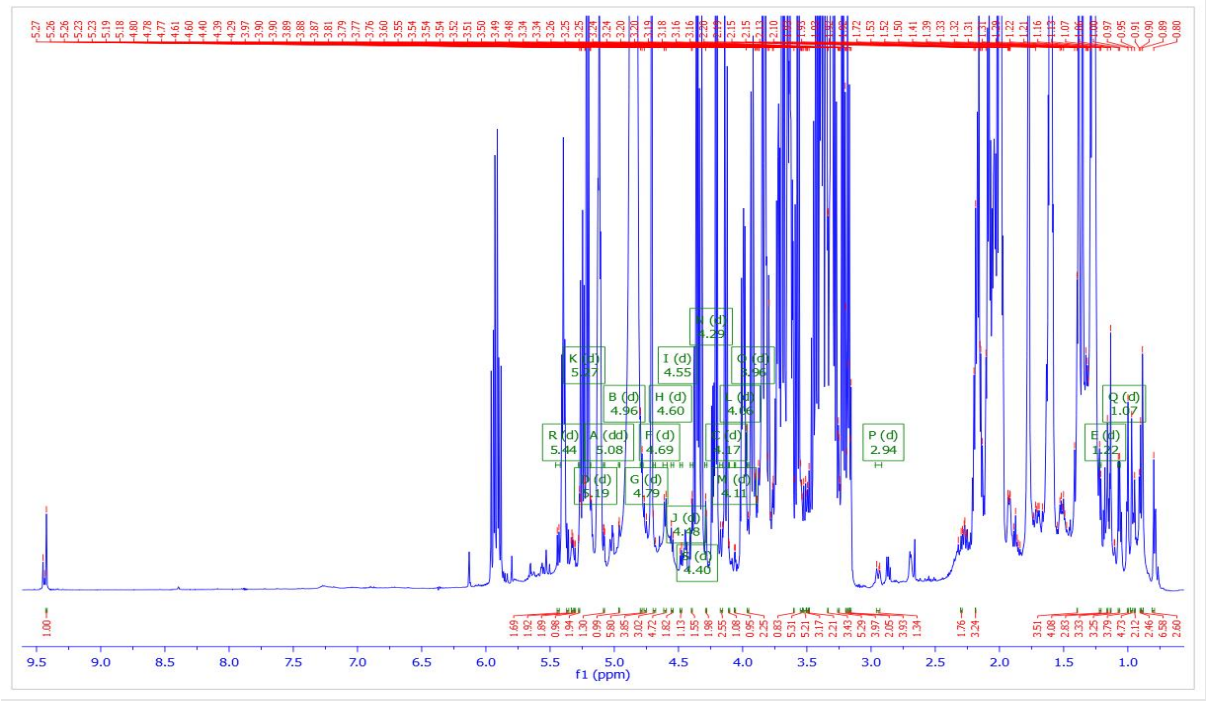


Fig. S29. Full ^{1}HNMR spectrum of semi-purified QS-7 (20) produced in *N. benthamiana.* NMR recorded in MeOH-d_4, 600 MHz.

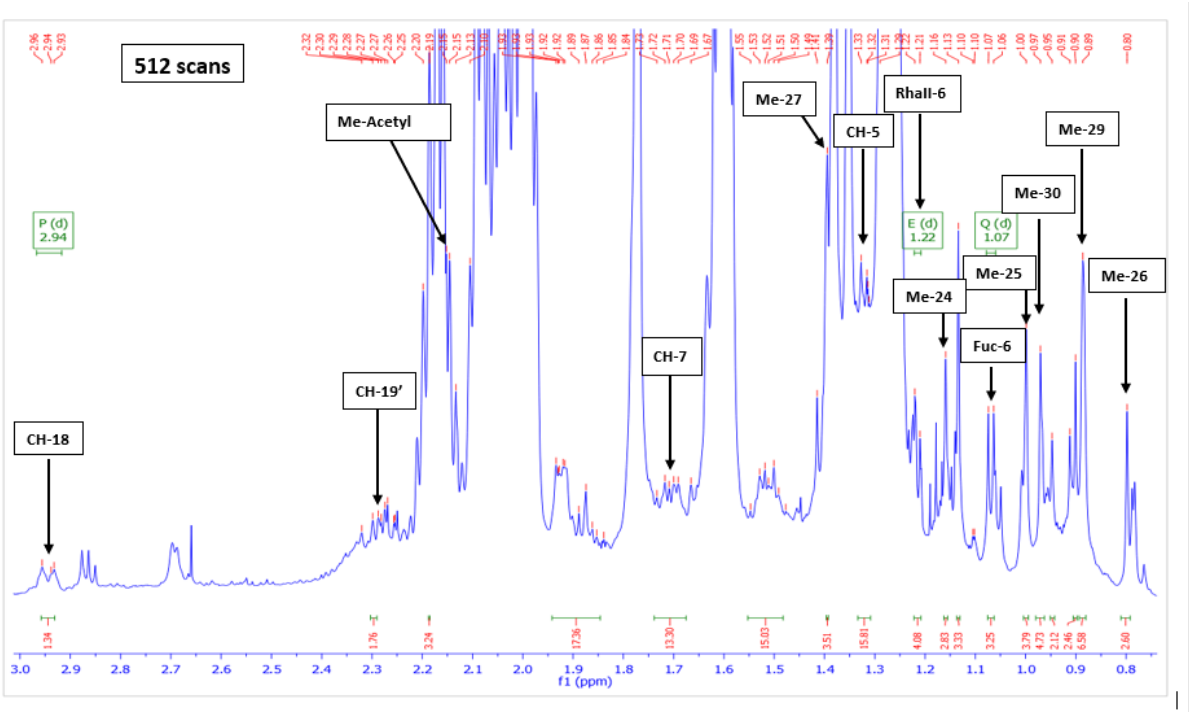


Fig. S30. Expanded ^{1}HNMR spectrum (0-3 ppm) of semi-purified QS-7 (20) produced in *N. benthamiana.* NMR recorded in MeOH-d_4, 600 MHz.

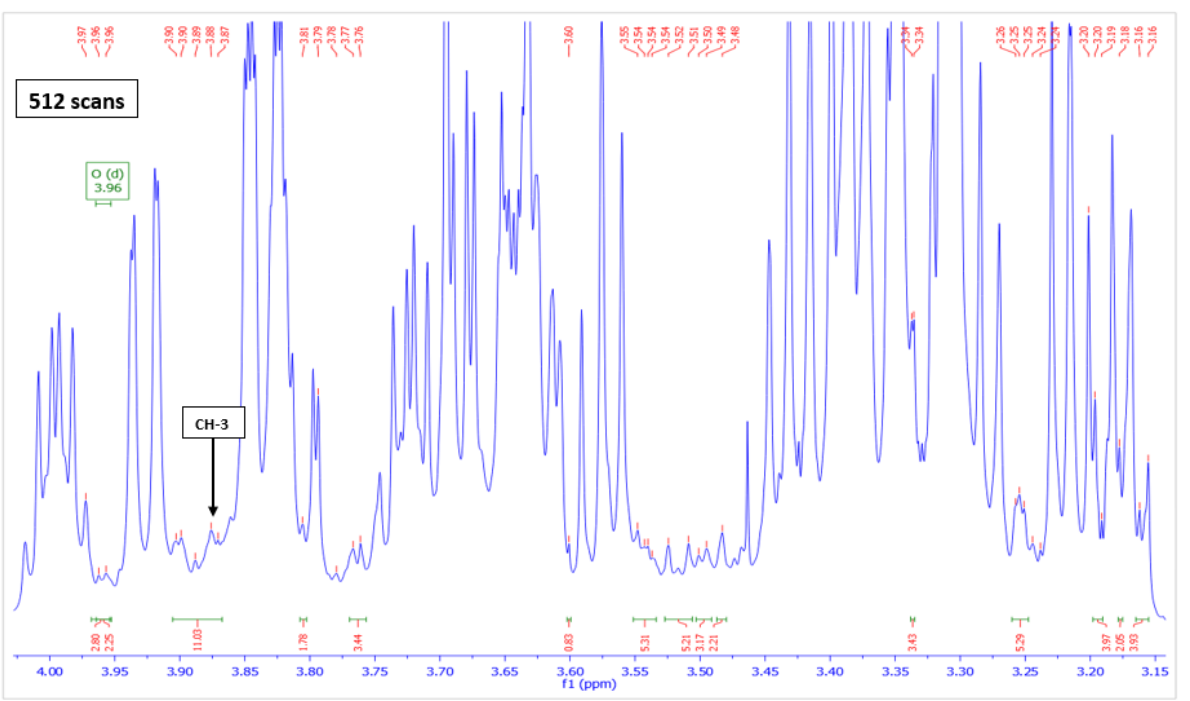


Fig. S31. Expanded ^{1}HNMR spectrum (non-anomeric region, 3.15-4.0 ppm) of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 600 MHz.

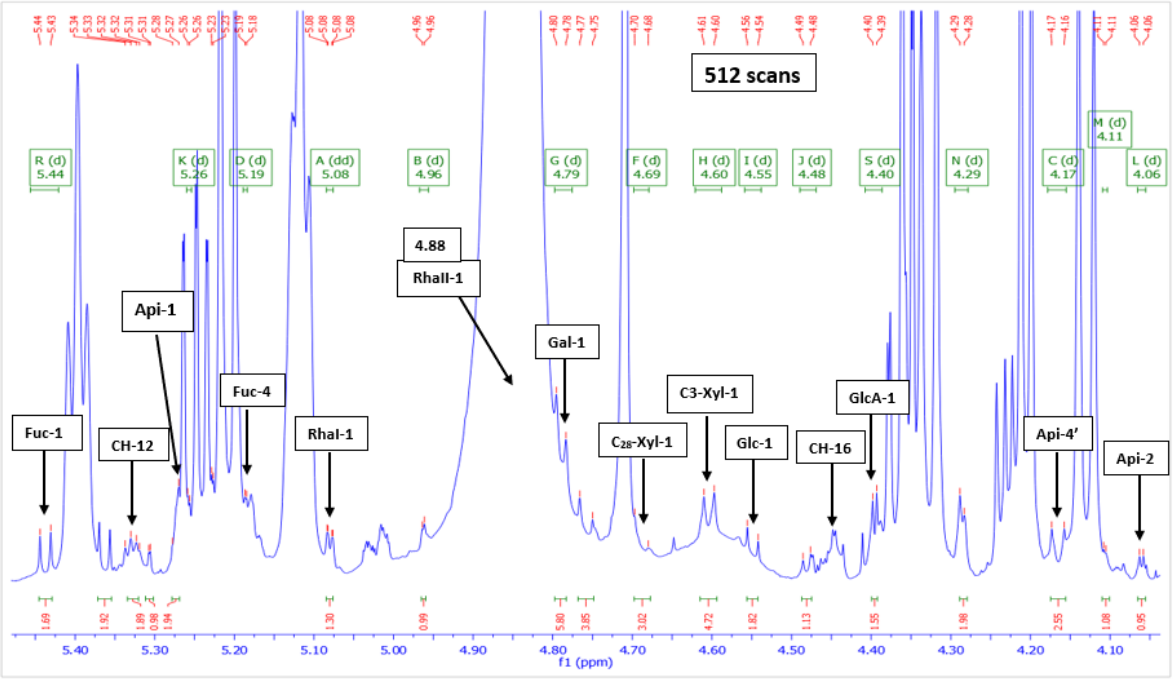


Fig. S32. Expanded ^{1}HNMR spectrum (anomeric region, 4.0-5.50 ppm) of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 600 MHz.

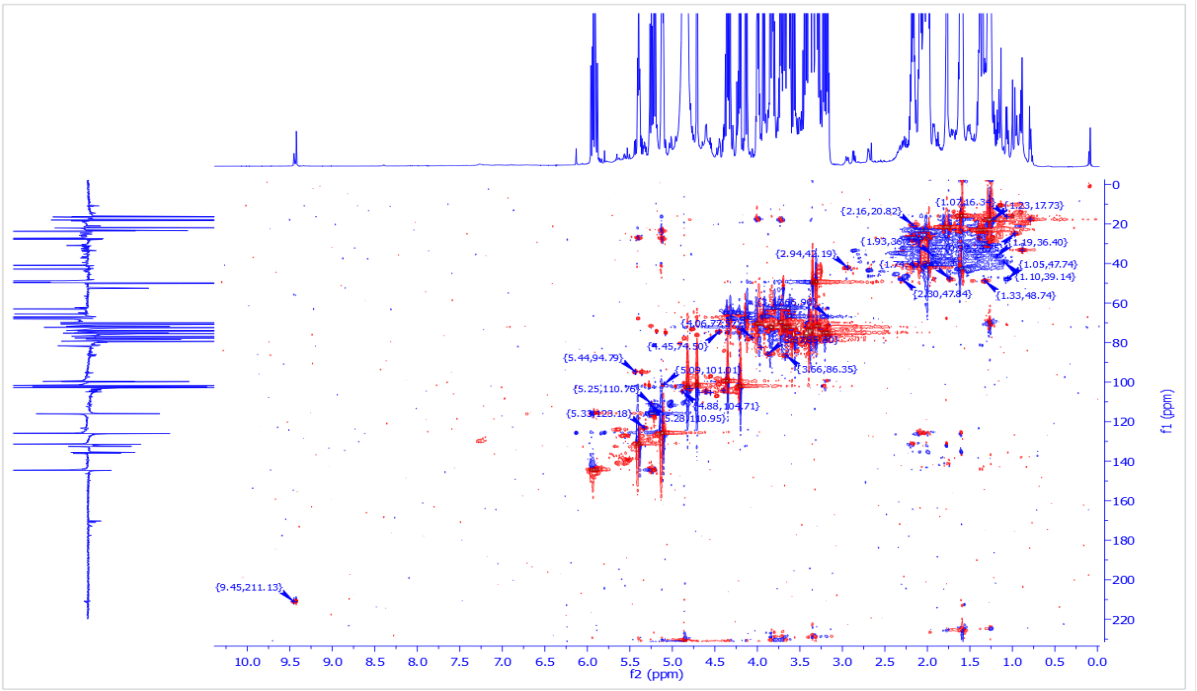


Fig. S33. ^{1}H-^{13}C HSQC spectrum of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 600, 150 MHz.

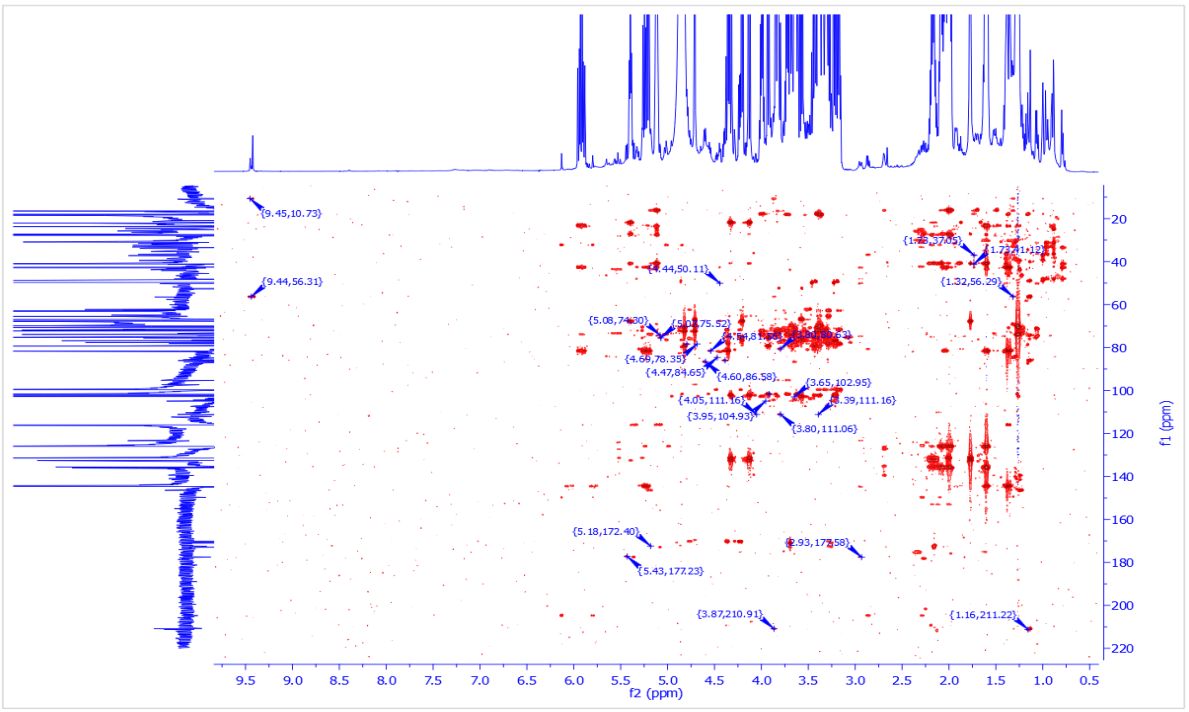


Fig. S34. ^{1}H-^{13}C HMBC spectrum of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 600, 150 MHz.

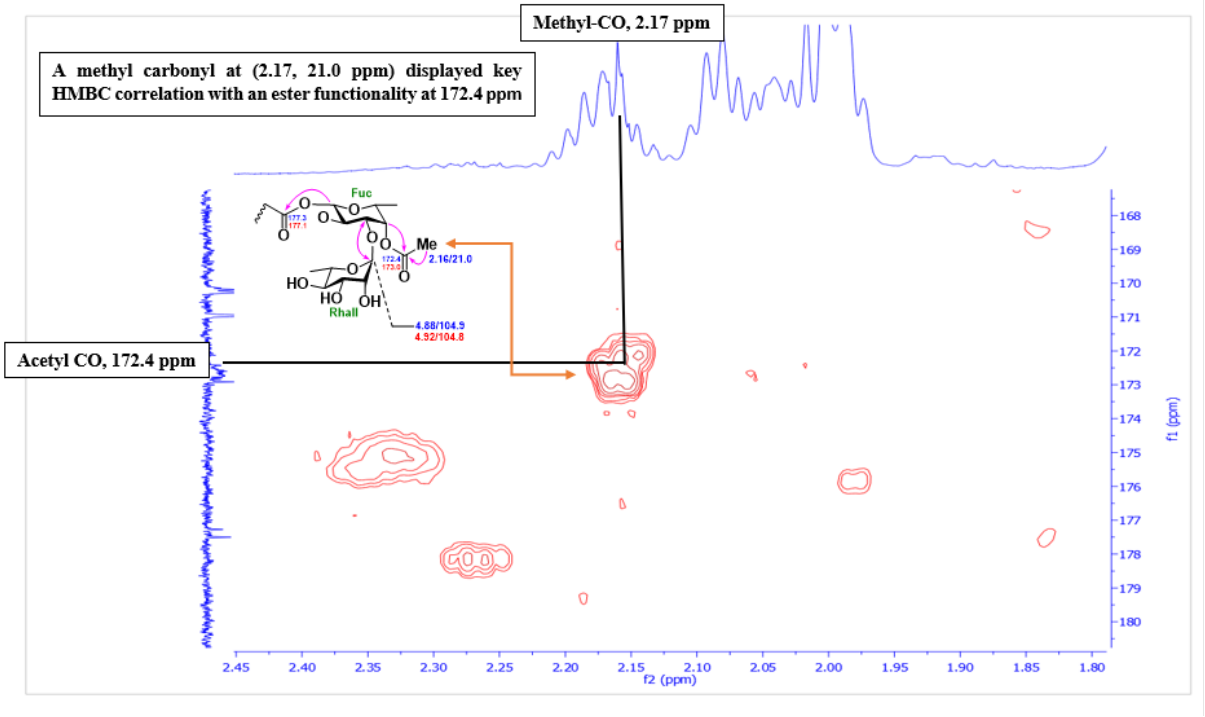


Fig. S35. Expanded ^{1}H-^{13}C HMBC spectrum of semi-purified QS-7 (20) produced in *N. benthamiana.* NMR recorded in MeOH-d_4, 600, 150 MHz.

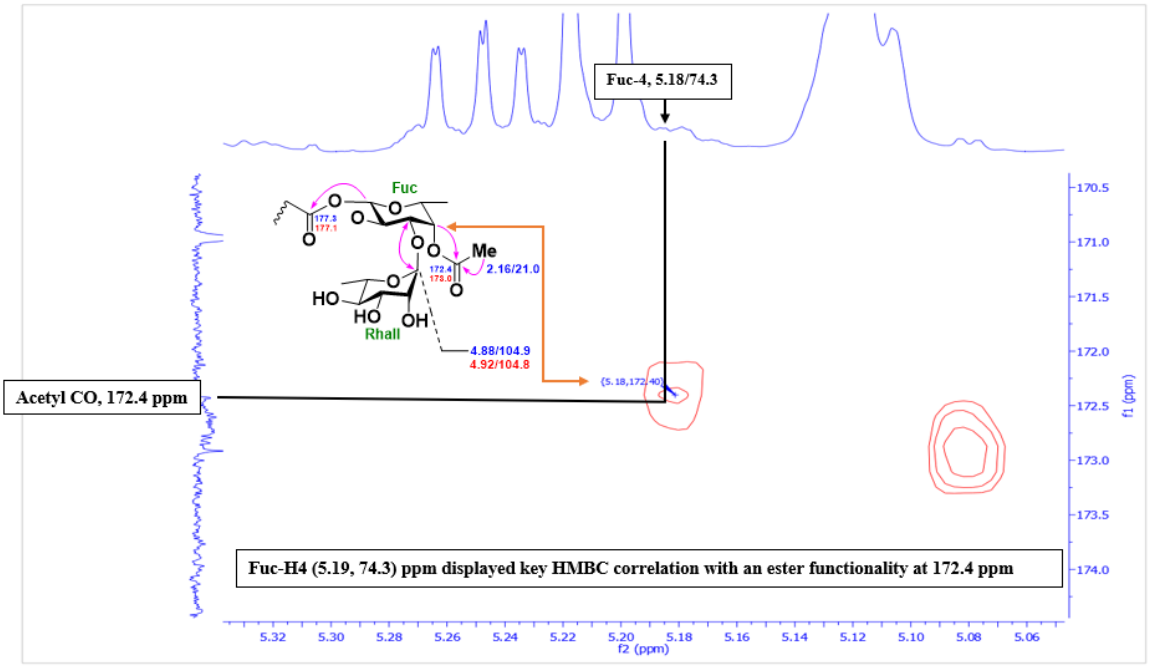


Fig. S36. Expanded ^{1}H-^{13}C HMBC spectrum of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 600, 150 MHz.

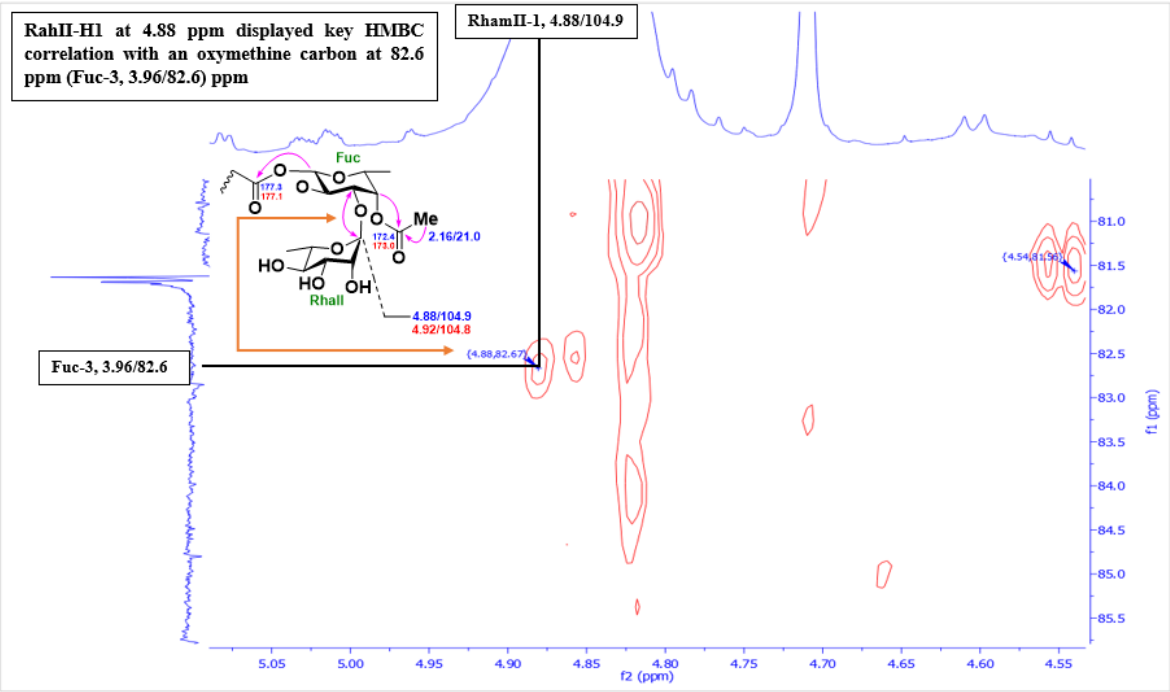


Fig. S37. Expanded ^{1}H-^{13}C HMBC spectrum of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 600, 150 MHz.

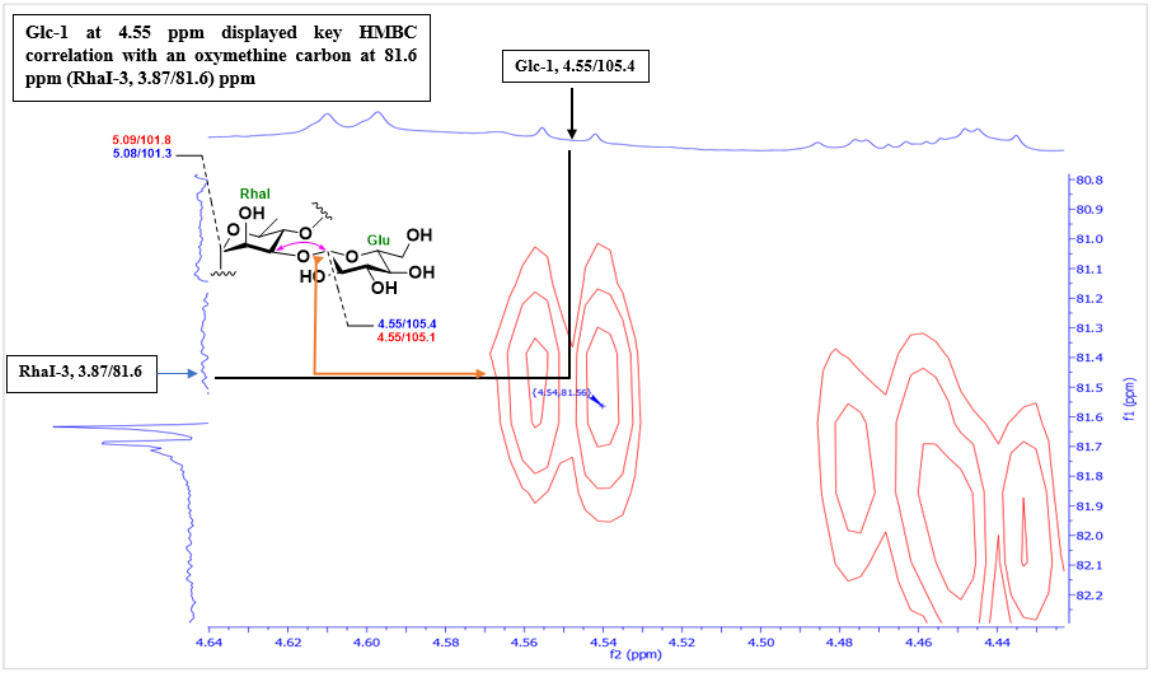


Fig. S38. Expanded ^{1}H-^{13}C HMBC spectrum of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 600, 150 MHz.

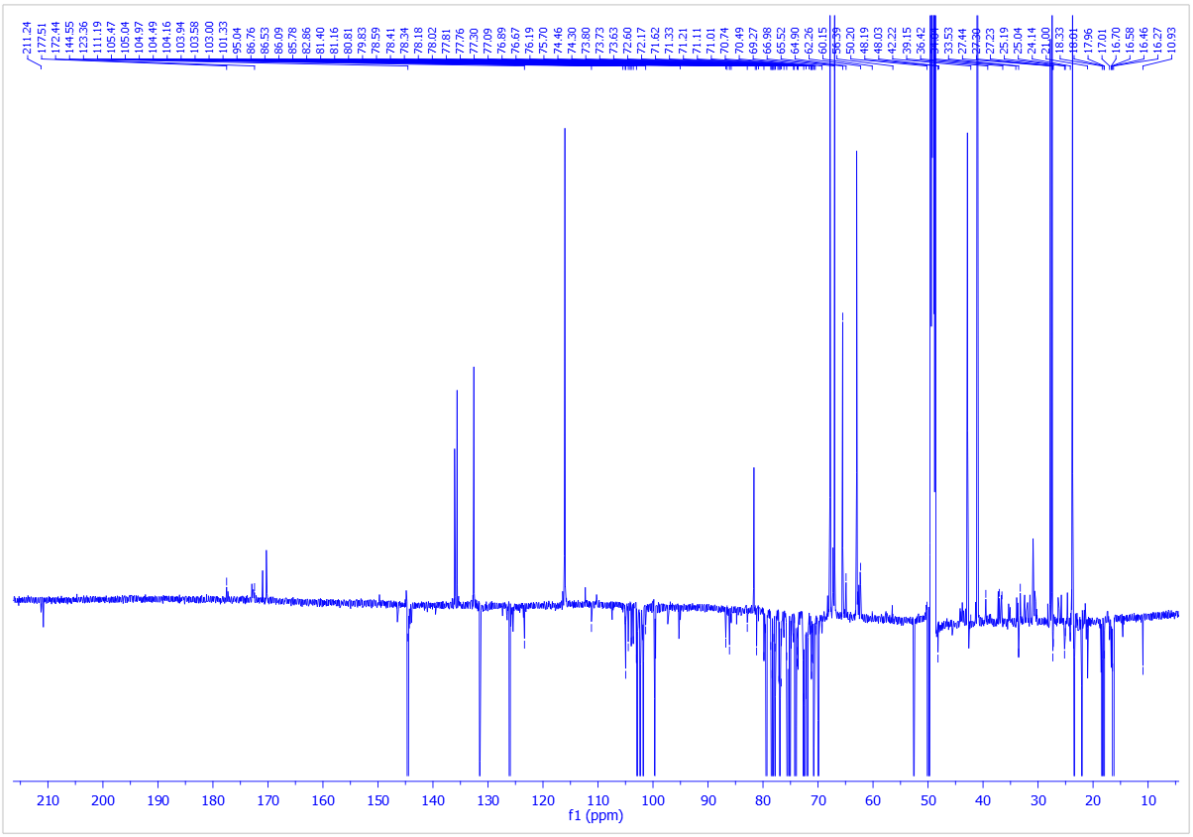


Fig. S39. Full DEPTQ-135 NMR spectrum of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 150 MHz.

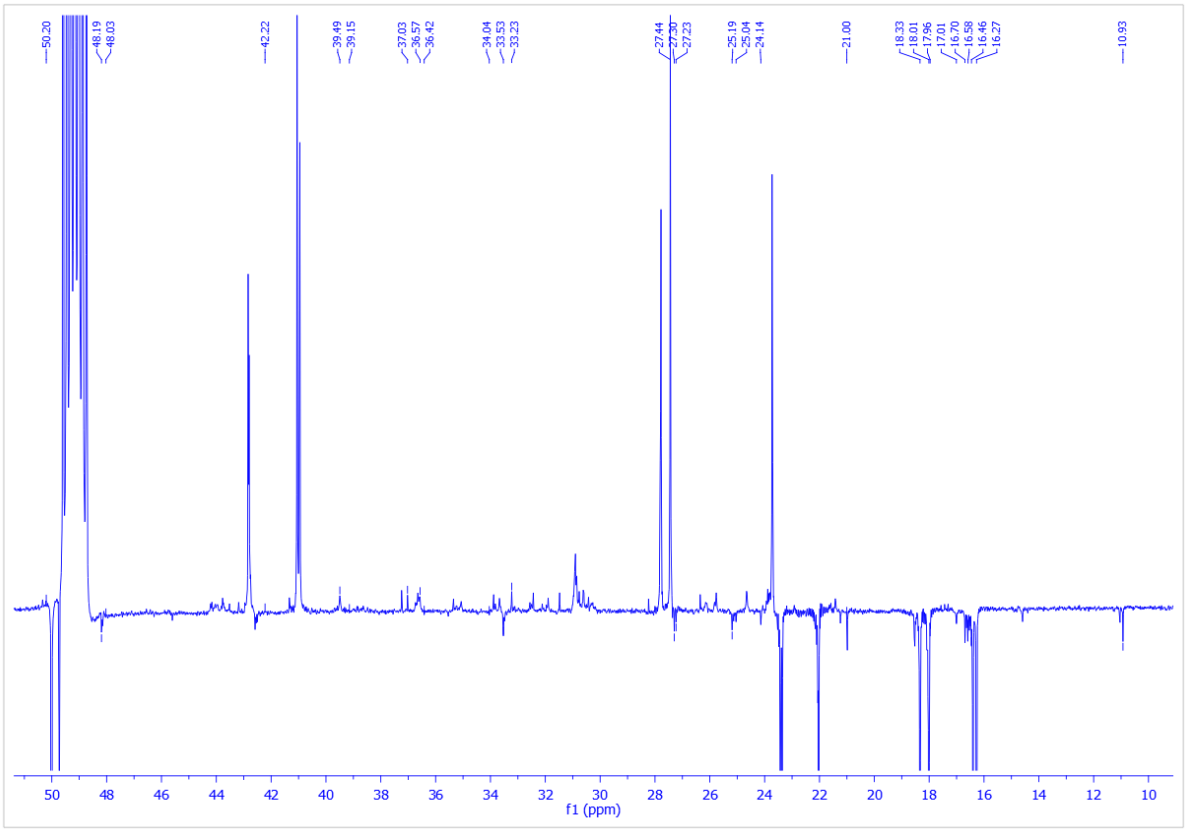


Fig. S40. Expanded DEPTQ-135 NMR spectrum (10-50 ppm) of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 150 MHz.

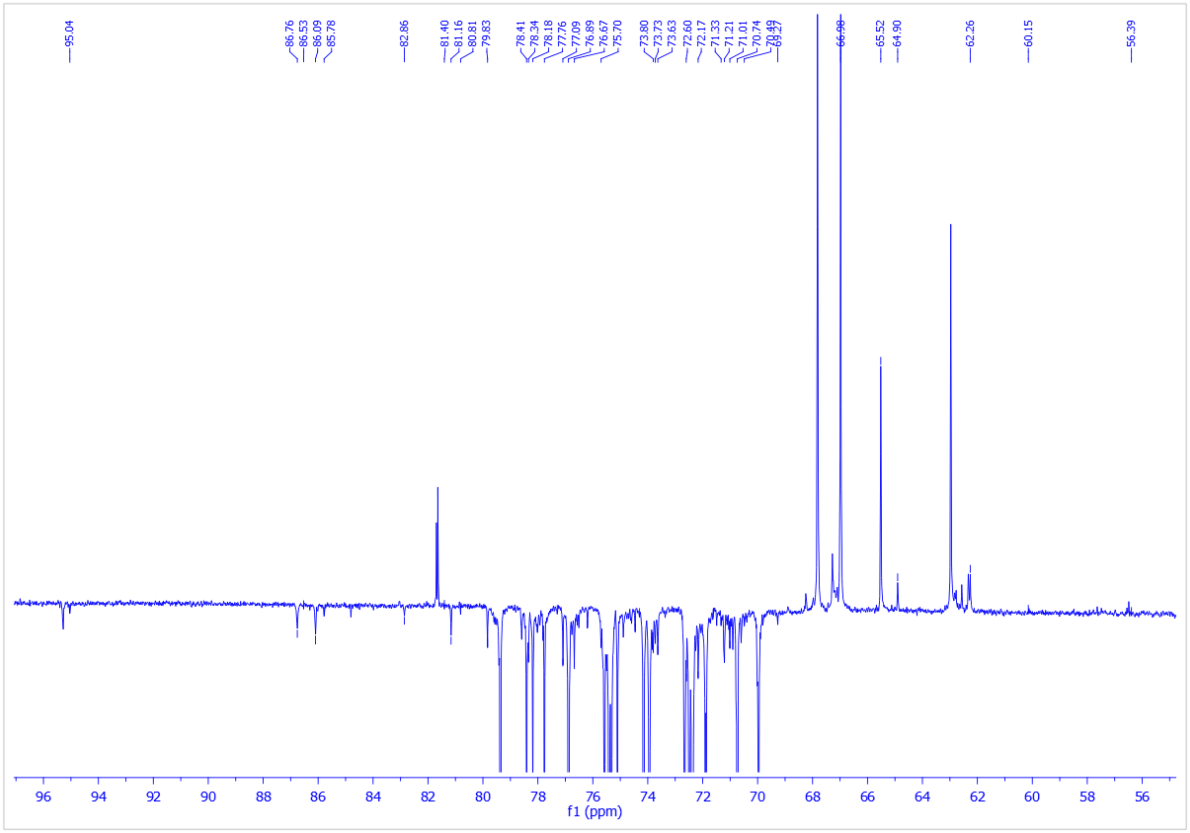


Fig. S41. Expanded DEPTQ-135 NMR spectrum (56-96 ppm) of semi-purified QS-7 (20) produced in *N. benthamiana.* NMR recorded in MeOH-d_4, 150 MHz.

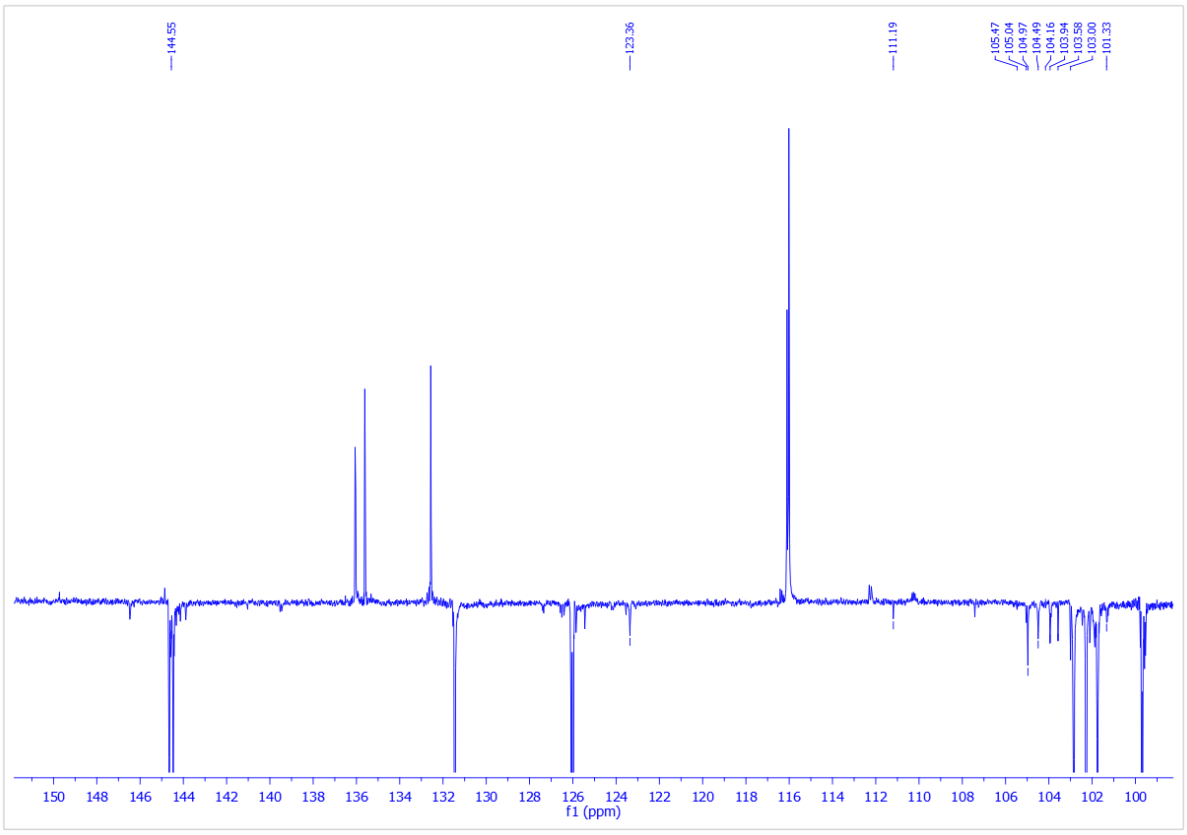


Fig. S42. Expanded DEPTQ-135 NMR spectrum (100-150 ppm) of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 150 MHz.

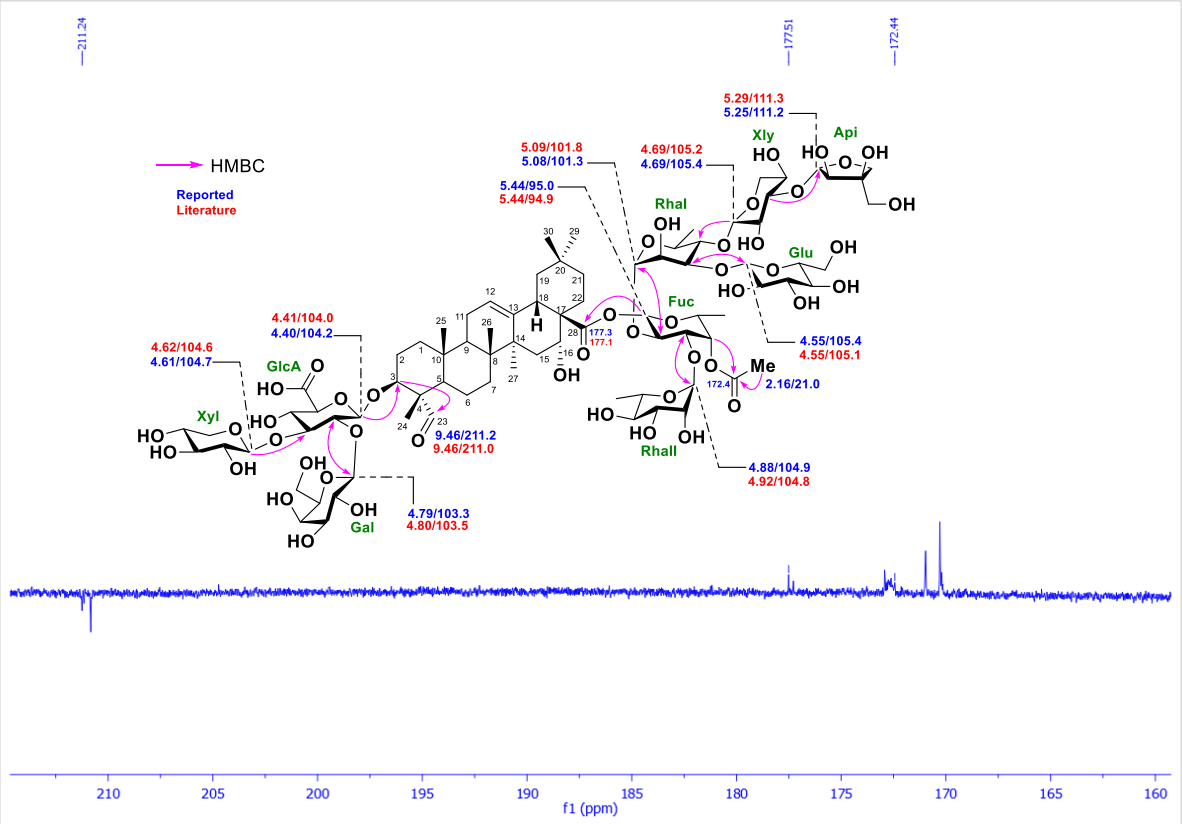


Fig. S43. Expanded DEPTQ-135 NMR spectrum (160-220 ppm) of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR recorded in MeOH-d_4, 150 MHz.

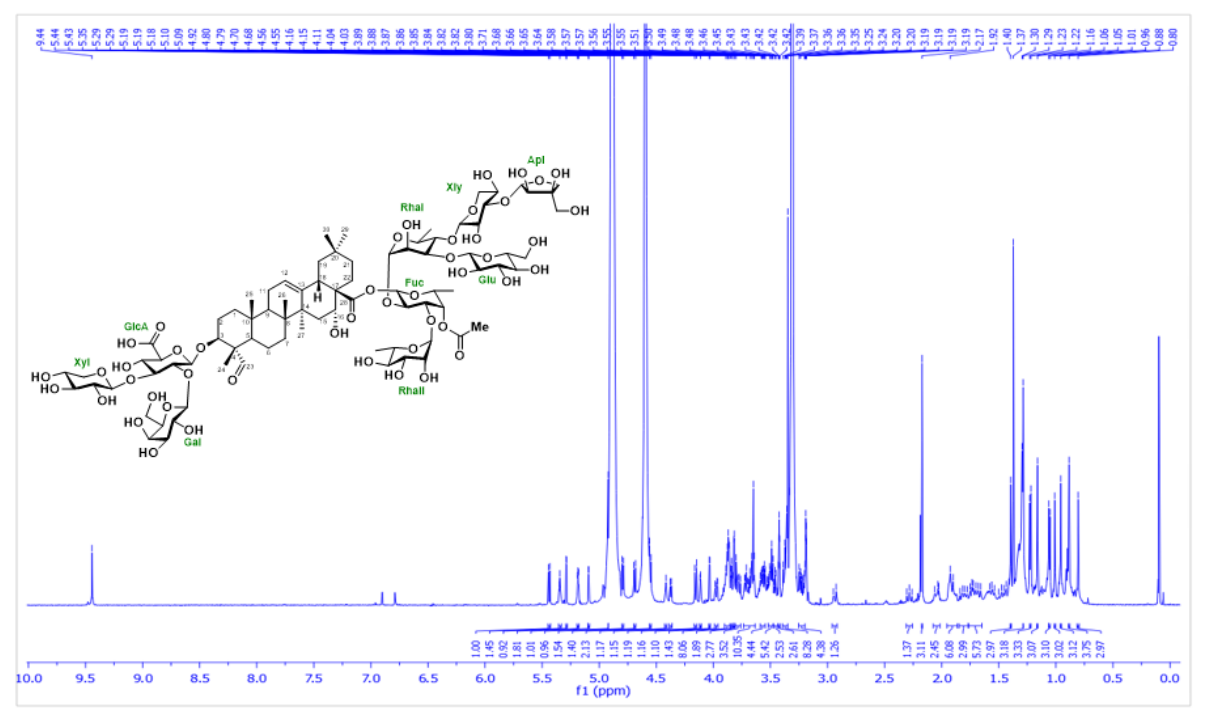


Fig. S44. ^{1}HNMR spectrum of QS-7 standard (Desert king). NMR recorded in MeOH-d_4, 600 MHz.

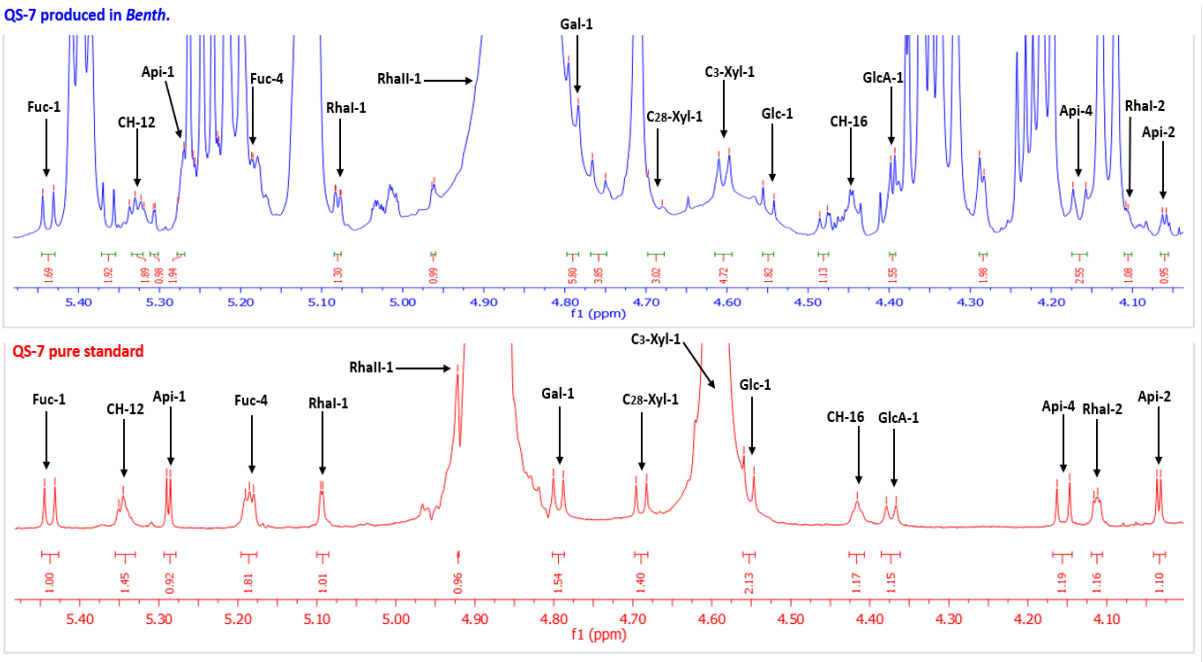


Fig. S45. ^{1}HNMR comparison between QS-7 (20) produced in *N. benthamiana* (top) and QS-7 standard (bottom) (anomeric region, 4.0-5.50 ppm). NMR recorded in MeOH-d_4, 600 MHz.

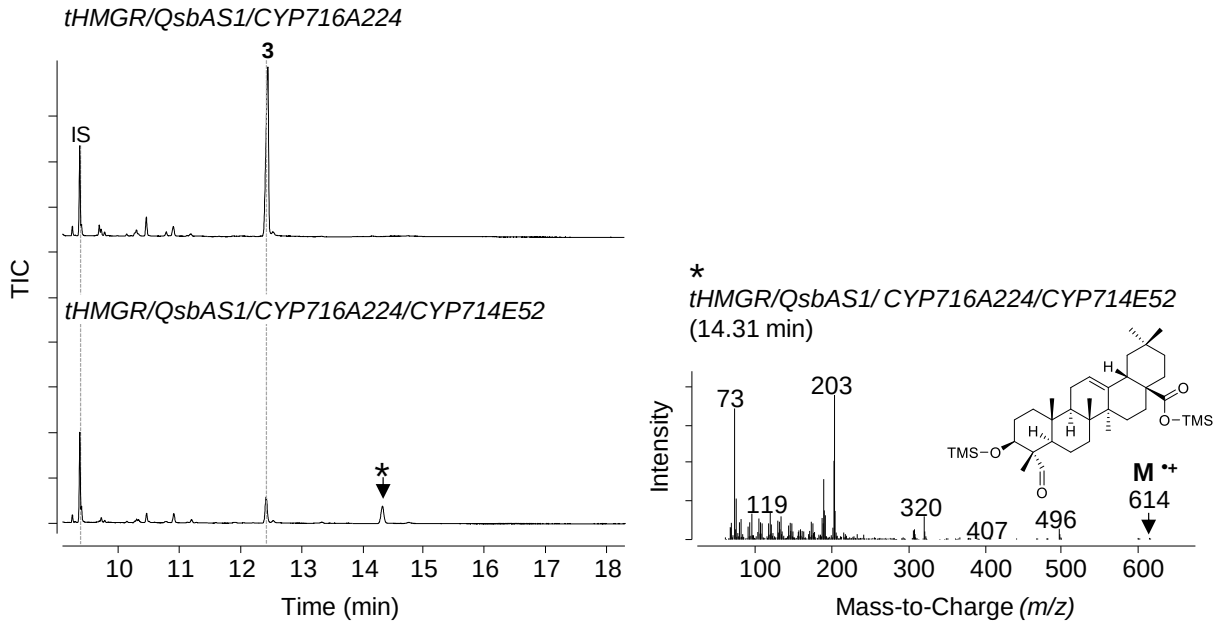


Fig. S46. CYP714E52 converts oleanolic acid (3) to a new product likely to be gypsogenin. GC-MS analysis of leaf extracts of *N. benthamiana* following *Agrobacterium*-mediated transient expression. Total ion chromatograms are shown on the left, and a mass spectrum on the right. Leaves were agro-infiltrated with expression constructs for *tHMGR/QsbAS1/CYP716A224* (control), or *tHMGR/QsbAS1/CYP716A224/CYP714E52*. A new product was observed at 14.31 mins in leaves expressing *tHMGR/QsbAS1/CYP716A224/CYP714E52*. This is consistent with the addition of an aldehyde to oleanolic acid (forming gypsogenin). IS, internal standard (coprostanol).

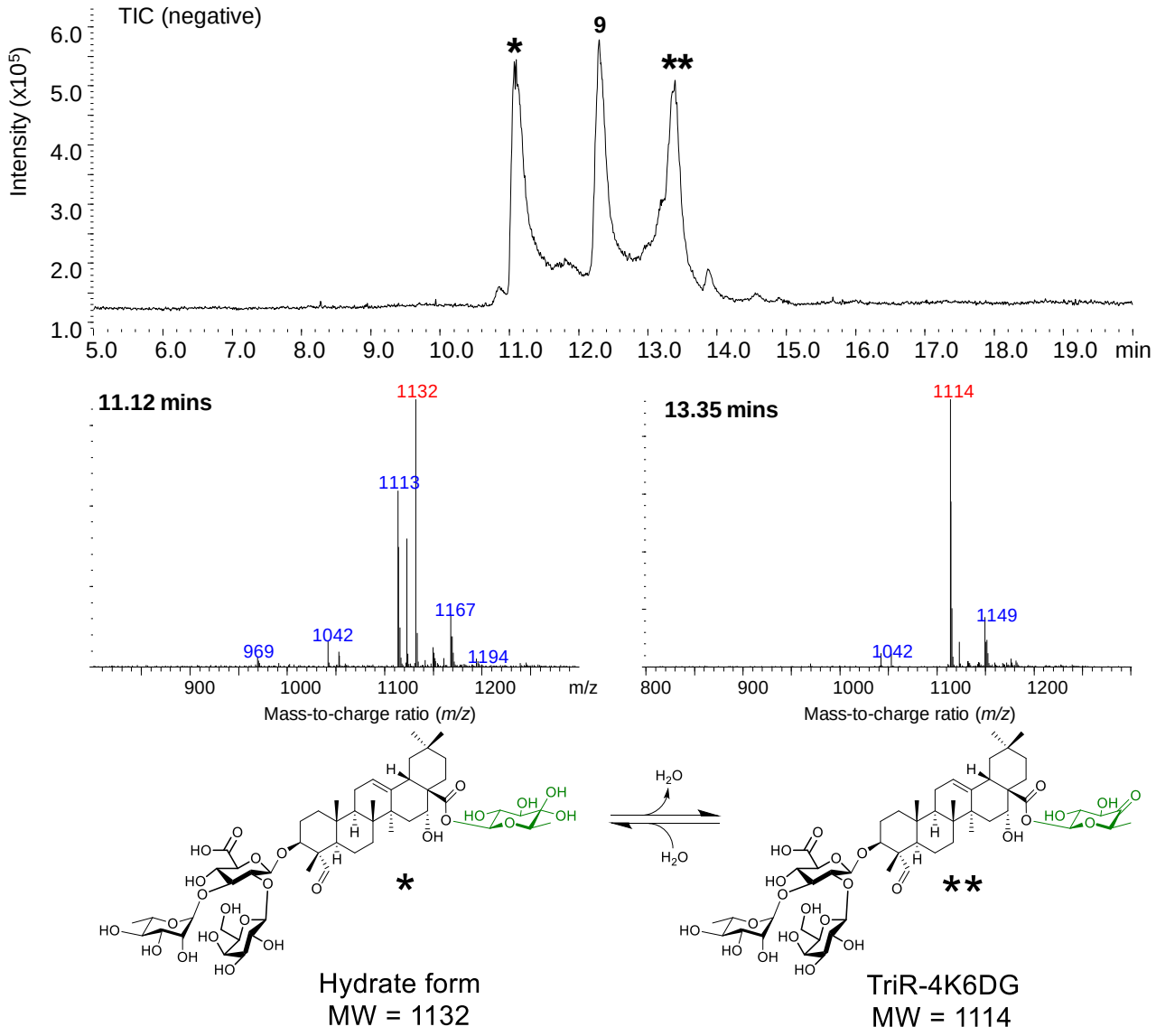


Fig. S47. Mass spectra for the putative QA-TriR-4-keto-6-deoxy-glucose (QA-TriR-4K6DG) and hydrate form. When incubated with UDP-D-glucose, ATCV-1 UGD and UGT74BX1 *in vitro*, QA-TriR (**9**) is converted to a new product which is likely to be QA-TriR-4K6DG (** 13.35 mins). This product exists in equilibrium with its hydrate form, as seen by the peak at 11.12 mins (*).

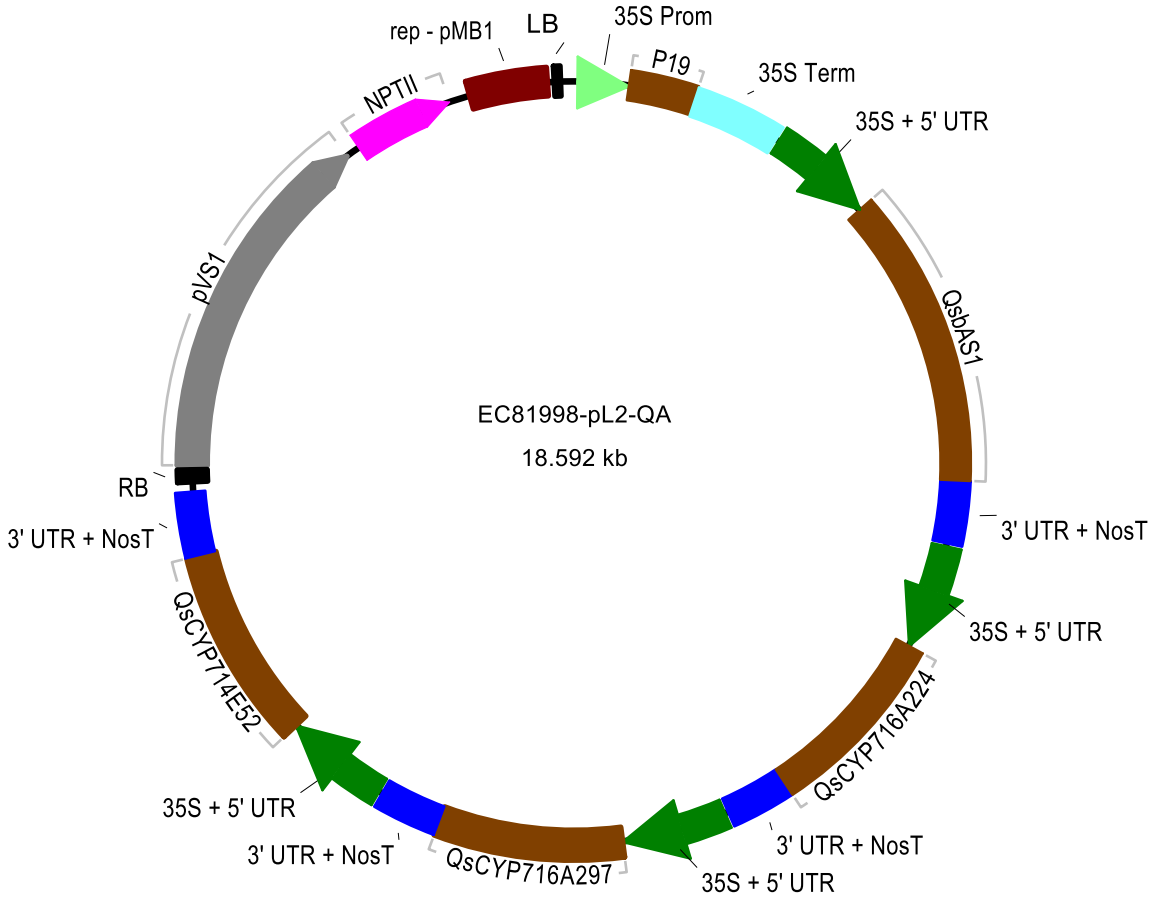


Fig. S48. Plasmid map of Golden Gate vector EC81998-pL2-QA. This contains the four genes (*QsbAS1*, *CYP716A224*, *CYP716A297* and *CYP741E52*) required for biosynthesis of quillaic acid (**5**). All genes are flanked upstream by a module consisting of the cauliflower mosaic virus (CaMV) 35S promoter and modified cowpea mosaic virus (CPMV) 5' UTR (35S + 5'UTR, green) and downstream by the CPMV 3' UTR and noscaline synthase terminator (3'UTR + NosT, blue). In addition a copy of the P19 silencing suppressor flanked by the CaMV 35S promoter and terminator is included.

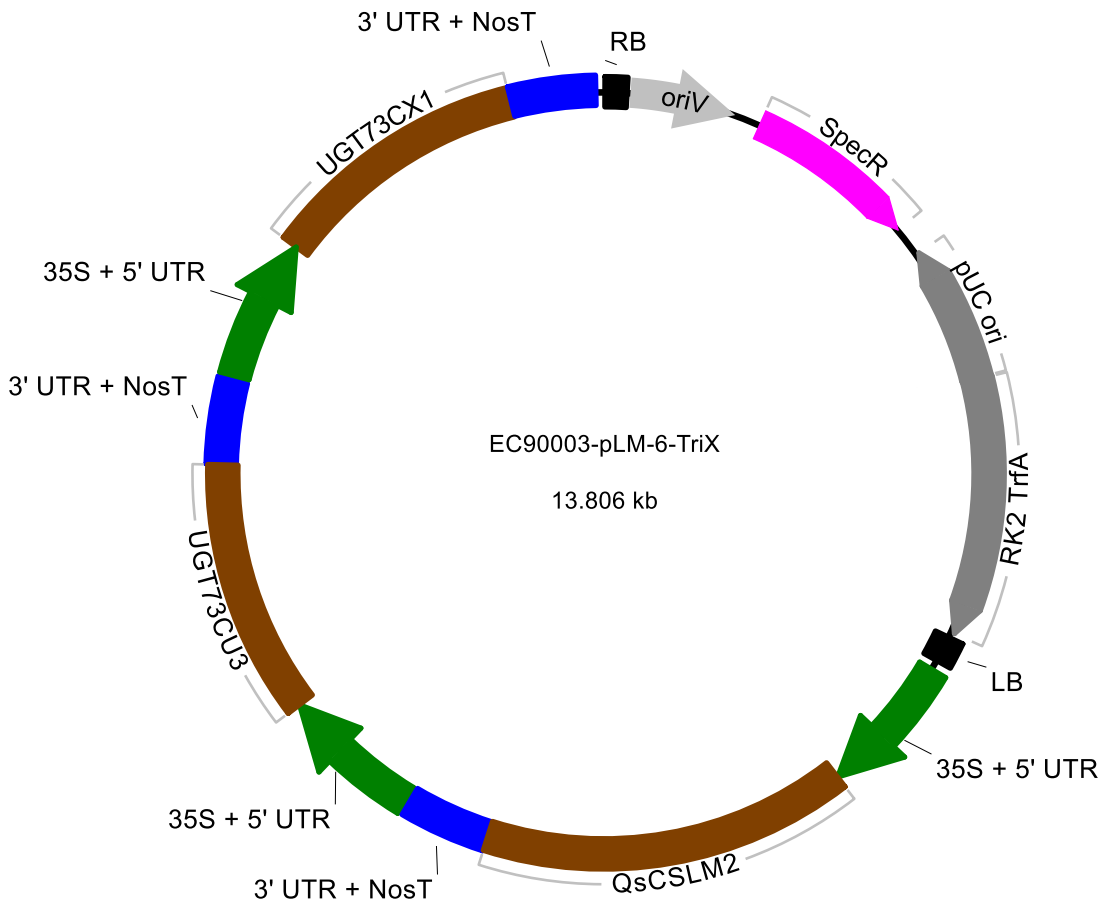


Fig. S49. Plasmid map of Golden Gate vector EC90003-pLM-6-TriX. This contains the three genes (*QsCSLM2*, *UGT73CU3* and *UGT73CX1*) required for addition of the branched trisaccharide featuring xylose to the C-3 position of quillaic acid (**5**). All genes are flanked upstream by a module consisting of the cauliflower mosaic virus (CaMV) 35S promoter and modified cowpea mosaic virus (CPMV) 5' UTR (35S + 5'UTR, green) and downstream by the CPMV 3' UTR and noscaline synthase terminator (3'UTR + NosT, blue).

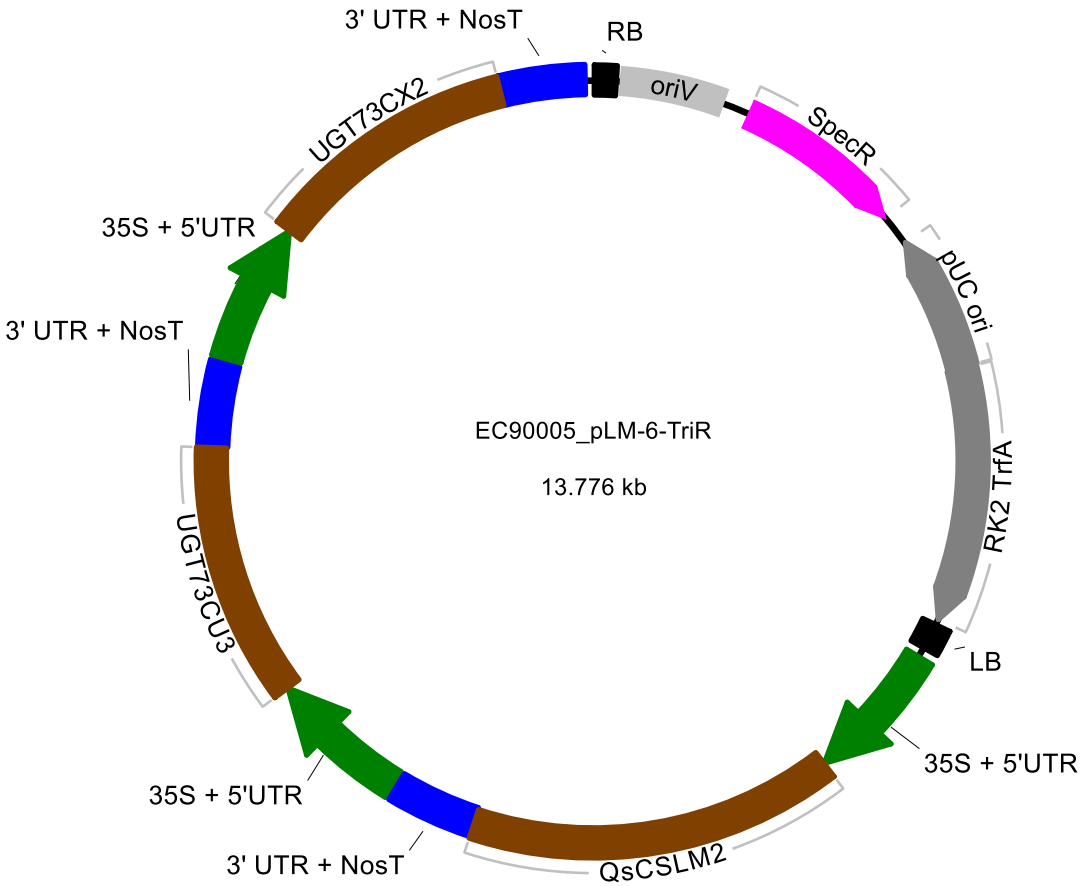


Fig. S50. Plasmid map of Golden Gate vector EC90005_pLM-6-TriR. This contains the three genes (*QsCSLM2*, *UGT73CU3* and *UGT73CX2*) required for addition of the branched trisaccharide featuring rhamnose to the C-3 position of quillaic acid (**5**). All genes are flanked upstream by a module consisting of the cauliflower mosaic virus (CaMV) 35S promoter and modified cowpea mosaic virus (CPMV) 5' UTR (35S + 5'UTR, green) and downstream by the CPMV 3' UTR and noscaline synthase terminator (3'UTR + NosT, blue).

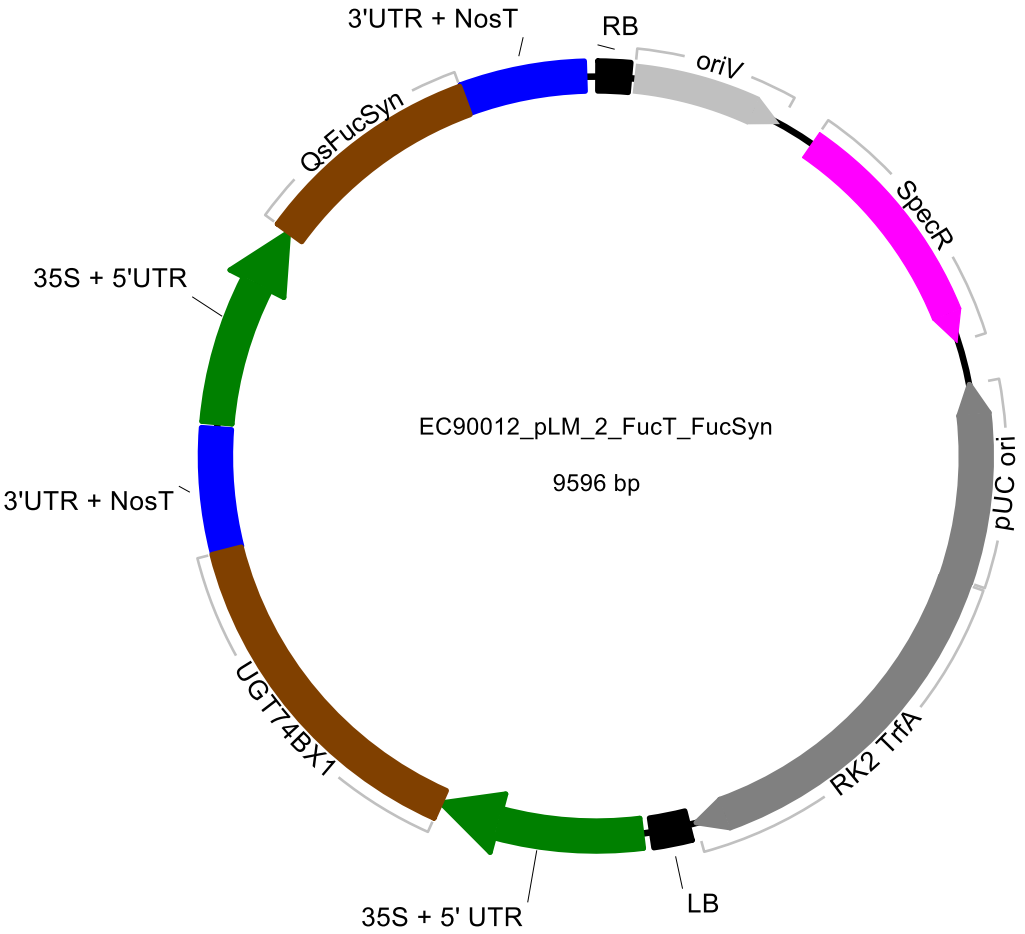


Fig. S51. Plasmid map of Golden Gate vector EC90012_pLM_2_FucT_FucSyn. This contains the gene encoding the glycosyltransferase (*UGT74BX1*) required for addition of the D-fucose at the C-28 position of QA-TriX (**8**) or QA-TriR (**9**) to form QA-TriX-F (**10**) or QA-TriR-F (**11**). The vector also includes the short chain dehydrogenase (*QsFucSyn*) which substantially increases the yields of the D-fucosylated products. All genes are flanked upstream by a module consisting of the cauliflower mosaic virus (CaMV) 35S promoter and modified cowpea mosaic virus (CPMV) 5' UTR (35S + 5'UTR, green) and downstream by the CPMV 3' UTR and noscaline synthase terminator (3'UTR + NosT, blue).

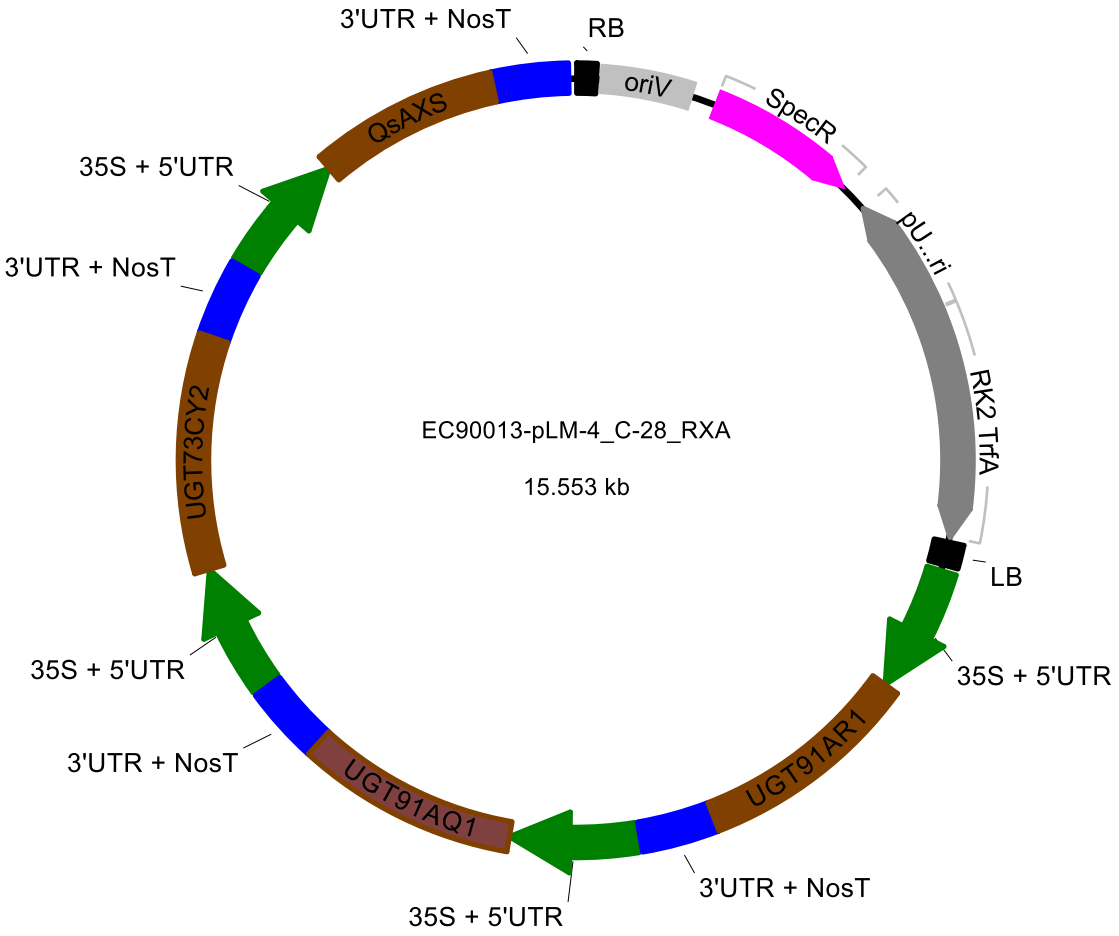


Fig. S52. Plasmid map of Golden Gate vector EC90013-pLM-4_C-28_RXA. This contains the three genes (*UGT91AR1*, *UGT91AQ1* and *UGT73CY2*) required for addition of the final three sugars (L-rhamnose, D-xylose and D-apiose) in the linear tetrasaccharide at C-28 of QA-TriX-F (**10**) or QA-TriR-F (**11**) to form QA-TriX-FRXA (**18**) or QA-TriR-FRXA (**19**), respectively. The plasmid also contains QsAXS gene which boosts the apiosylated product. All genes are flanked upstream by a module consisting of the cauliflower mosaic virus (CaMV) 35S promoter and modified cowpea mosaic virus (CPMV) 5' UTR (35S + 5'UTR, green) and downstream by the CPMV 3' UTR and noscaline synthase terminator (3'UTR + NosT, blue).

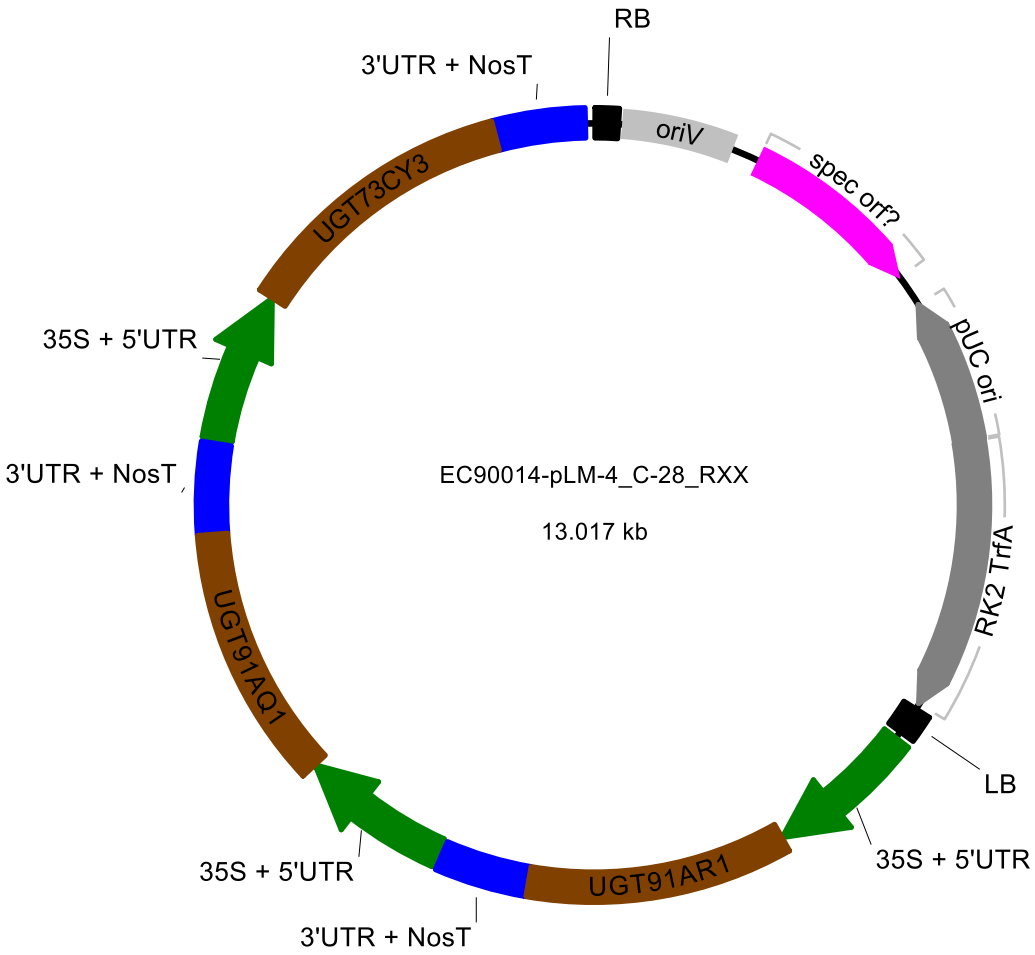


Fig. S53. Plasmid map of Golden Gate vector EC90014-pLM-4_C-28_RXX. This contains the three genes (*UGT91AR1, UGT91AQ1* and *UGT73CY3*) required for addition of the final three sugars (L-rhamnose, D-xylose and D-xylose) in the linear tetrasaccharide at C-28 of QA-TriX-F (**10**) or QA-TriR-F (**11**) to form QA-TriX-FRXX (**16**) or QA-TriR-FRXX (**17**). All genes are flanked upstream by a module consisting of the cauliflower mosaic virus (CaMV) 35S promoter and modified cowpea mosaic virus (CPMV) 5' UTR (35S + 5'UTR, green) and downstream by the CPMV 3' UTR and noscaline synthase terminator (3'UTR + NosT, blue).

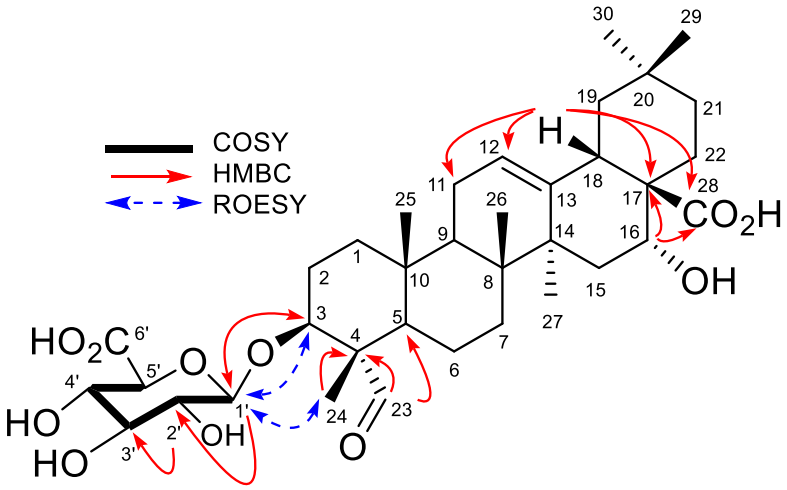
	No. /%	Size (bp)
Assembly feature		
Estimated genome size		411,000,000*
Assembled sequences (contig)	769	354,911,093
L50/N50 length (contig)	19	5,518,683
Longest contig		18,205,868
L50/N50 length (scaffolds)	6	26,440,503
L90/N90 length (scaffolds)	13	19,561,049
Assembled pseudochromosomes	14	346,890,757
Longest scaffold		37,318,896
GC content	32.96%	
Transposable elements		
Class I Retrotransposons	12.3%	43,683,674
Class II DNA transposons	13.6%	48,387,773
Other (inc. MITEs)	0.5%	1,876,770
Total	26.5%	93,948,217
Genome annotation		
Gene models/mean model length (high confidence)	30,780	4,199
Gene models/mean model length (lower confidence)	3,125	2,388
Noncoding RNAs/mean model length	467	1,828

*Garcia et al. (2010) (*18*)

Table S1. *Q. saponaria* accession S10 genome statistics and gene predictions

Gene ID	QsbAS1 co-expression (PCC)	Primordia transcript quantity (TPM)
Qs_0321930	0.987	15790.25
Qs_0321920	0.985	4375.86
Qs_0123860	0.975	8265.02
Qs_0131010	0.965	29496.29
Qs_0233700	0.961	4636.16
Qs_0283870	0.957	4110.90
Qs_0321940	0.956	6209.41
Qs_0152180	0.956	4380.37
Qs_0213710	0.955	3828.36
Qs_0234150	0.949	2489.59
Qs_0082400	0.946	3485.36
Qs_0234120	0.944	9140.54
Qs_0283850	0.931	3577.73
Qs_0234130	0.908	1646.18
Qs_0098610	0.907	12046.56
Qs_0187000	0.842	3451.64
Qs_0023500	0.835	2418.88
Qs_0234140	0.796	4127.94
Qs_0055340	0.786	6191.23
Qs_0213660	0.712	2673.27

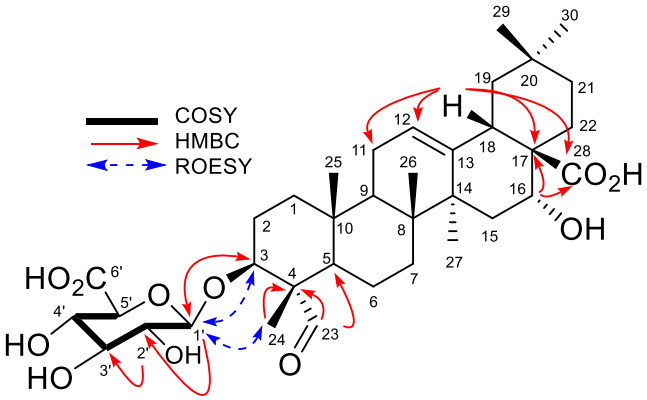
Table S2. Candidate *Q. saponaria* UGT genes. Candidates were prioritized based on co-expression with *QsbAS1* (PCC cut-off 0.7) and expression levels in primordial tissue (TPM >1600).



Quillaic acid 3-*O*-*β*- D-glucopyranosiduronic acid (6) (CSL-M1)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.70, d (13.3)/1.13, m	19	47.8, CH_2	2.30/1.04, m
2	25.8, CH_2	2.03/1.78, m	20	31.6, Cq	-
3	83.0, CH	3.94, dd (12, 4.4)	21	36.7, CH_2	1.96/1.15, m
4	56.4, Cq	-	22	32.9, CH_2	1.91/1.77, m
5	49.1, CH	1.35, m	23	209.3, CH	9.42, s
6	21.6, CH_2	1.50/0.90, m	24	10.5, CH_3	1.11, s
7	33.7, CH_2	1.58/1.26, m	25	16.3, CH_3	1.01, s
8	41.1, Cq	-	26	17.9, CH_3	0.80, s
9	48.2, CH	1.77, m	27	27.4, CH_3	1.40, s
10	37.2, Cq	-	28	181.3, Cq	-
11	24.6, CH_2	1.94/1.94, m	29	33.6, CH_3	0.89, s
12	123.3, CH	5.31, t (3.3)	30	25.0, CH_3	0.97, s
13	145.3, Cq	-	GlcA-1	104.8, CH	4.20, d (7)
14	42.9, Cq	-	GlcA-2	75.3, CH	3.11, t (8.3)
15	36.3, CH_2	1.84/1.34, m	GlcA-3	77.9, CH	3.32, overlapped with methanol
16	75.4, CH	4.45, t (3.5)	GlcA-4	73.7, CH	3.42, m
17	50.0, Cq	-	GlcA-5	76.6, CH	3.57, br s
18	42.2, CH	3.01, dd (14.0, 4.3)	GlcA-6	Not observed	-

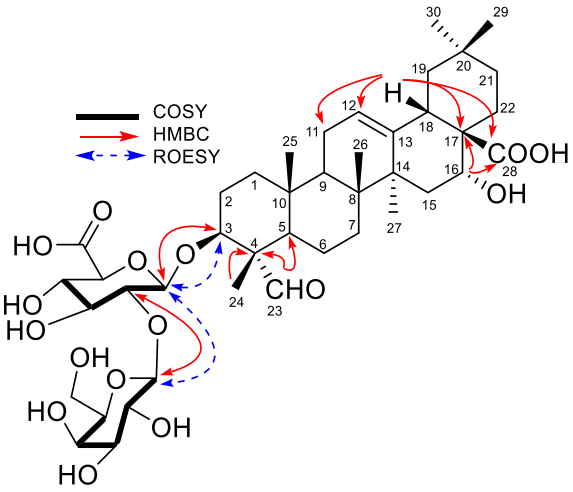
Table S3. Full NMR data showing that the product generated by co-expression of CSLM1 with the QA pathway genes is quillaic acid 3-*O*-*β*-D-glucopyranosiduronic acid (6). NMR carried out in MeOH-d_4 (600, 150 MHz)



Quillaic acid 3-*O*-*β*-D-glucopyranosiduronic acid (6) (CSL-M2)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.70, d (13.3)/1.12, m	19	47.9, CH_2	2.30/1.02, m
2	25.9, CH_2	1.97/1.78, m	20	31.6, Cq	-
3	83.6, CH	3.89, dd (11.5, 3.8)	21	36.7, CH_2	1.96/1.15, m
4	56.3, Cq	-	22	32.9, CH_2	1.90/1.76, m
5	49.2, CH, overlapped	1.34, m	23	209.2, CH	9.41, s
6	21.5, CH_2	1.52/0.91, m	24	10.6, CH_3	1.11, s
7	33.7, CH_2	1.57/1.25, m	25	16.3, CH_3	1.01, s
8	41.1, Cq	-	26	17.9, CH_3	0.80, s
9	48.2, CH	1.76, m	27	27.4, CH_3	1.40, s
10	37.2, Cq	-	28	181.2, Cq	-
11	24.6, CH_2	1.93/1.93, m	29	33.6, CH_3	0.88, s
12	123.3, CH	5.30, t (3.3)	30	25.0, CH_3	0.97, s
13	145.3, Cq	-	GlcA-1	104.8, CH	4.24, d (7.6)
14	42.9, Cq	-	GlcA-2	75.1, CH	3.12, t (8.2)
15	36.3, CH_2	1.84/1.34, m	GlcA-3	77.7, CH	3.31, overlapped with methanol
16	75.4, CH	4.45, br s	GlcA-4	73.3, CH	3.46, m
17	50.0, Cq	-	GlcA-5	76.7, CH	3.72, br s
18	42.2, CH	3.01, dd (14.3, 4.2)	GlcA-6	Not observed	-

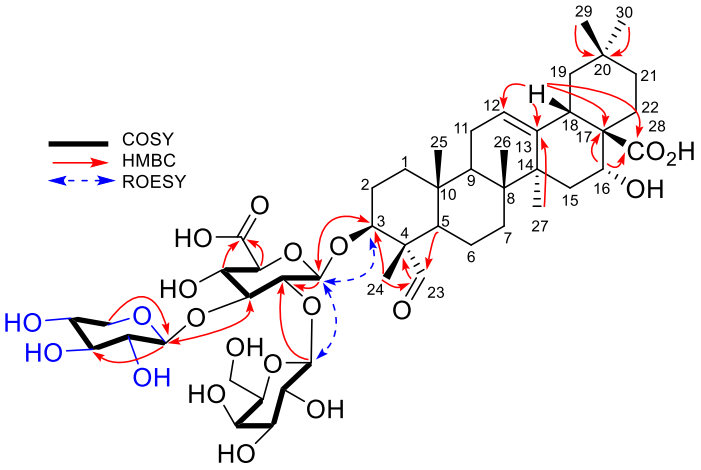
Table S4. Full NMR data showing that the product generated by co-expression of CSLM2 with the QA pathway genes is quillaic acid 3-*O*-*β*- D-glucopyranosiduronic acid (6). NMR carried out in MeOH-d_4 (600, 150 MHz)



Quillaic acid 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid} (7)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.70, d (13.1)/1.10, m	23	201.9, CH	9.46, s
2	25.6, CH_2	2.00/1.78, m	24	10.9, CH_3	1.13, s
3	84.9, CH	3.91, dd (11.2, 2.3)	25	16.4, CH_3	1.0, s
4	56.4, Cq	-	26	17.9, CH_3	0.80, s
5	49.2, CH	1.33, m	27	27.4, CH_3	1.40, s
6	21.4, CH_2	1.48/0.91, m	28	181.3, Cq	-
7	33.7, CH_2	1.55/1.24, m	29	33.6, CH_3	0.88, s
8	41.1, Cq	-	30	25.0, CH_3	0.97, s
9	48.2, CH	1.75, m	GlcA-1	103.7, CH	4.36, d (6.1)
10	37.3, Cq	-	GlcA-2	81.4, CH	3.46, m
11	24.6, CH_2	1.92/1.92, m	GlcA-3	78.1, CH	3.54, m
12	123.3, CH	5.30, br s	GlcA-4	Not observed	3.47, m
13	145.3, Cq	-	GlcA-5	77.0, CH	3.74, m
14	42.9, Cq	-	GlcA-6	Not observed	-
15	36.3, CH_2	1.83/1.33, m	Gal-1	105.4, CH	4.49, d (7.3)
16	75.4, CH	4.45, br s	Gal-2	74.0, CH	3.53, m
17	49.7, Cq	-	Gal-3	75.0, CH	3.46, m
18	42.2, CH	3.01, dd (14.2, 3.1)	Gal-4	70.6, CH	3.82, m
19	47.8, CH_2	2.29/1.02, m	Gal-5	77.1, CH	3.51, m
20	31.6, Cq	-	Gal-6	62.5, CH_2	3.80/3.73, dd (10.9, 5.5)
21	36.7, CH_2	1.94/1.13, m			
22	32.9, CH_2	1.90/1.76, m			

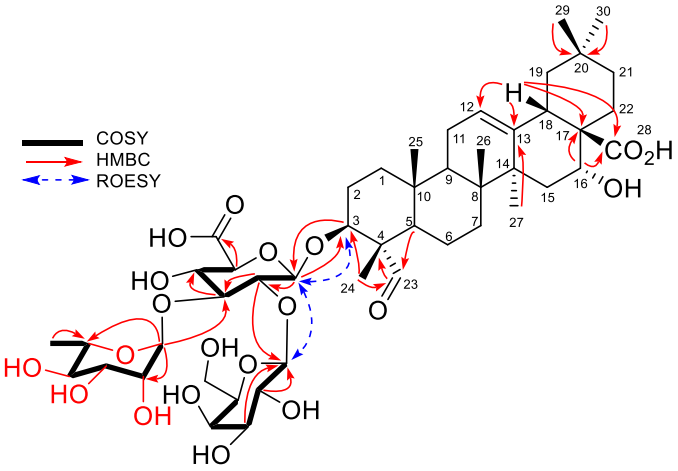
Table S5. Full NMR data for quillaic acid 3-*O*-{*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid} (7). NMR carried out in MeOH-d_4 (600, 150 MHz)



Quillaic acid 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid} (8)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.70, d (13.3)/1.12, m	24	10.9, CH_3	1.15, s
2	25.9, CH_2	1.97/1.78, m	25	16.4, CH_3	1.00, s
3	86.5, CH	3.87, dd (11.7, 4.9)	26	17.9, CH_3	0.79, s
4	56.4, Cq	-	27	27.4, CH_3	1.39, s
5	49.2, CH, overlapped	1.32, m	28	181.2, Cq	-
6	21.4, CH_2	1.51/0.91, m	29	33.6, CH_3	0.88, s
7	33.7, CH_2	1.54/1.23, m	30	25.0, CH_3	0.97, s
8	41.1, Cq	-	GlcA-1	104.6, CH	4.48, d (2.9)
9	48.2, CH	1.75, m	GlcA-2	78.3, CH	3.64, m
10	37.3, Cq	-	GlcA-3	86.7, CH	3.69, m
11	24.6, CH_2	1.92/1.92, m	GlcA-4	71.5, CH	3.56, m
12	123.3, CH	5.30, t (3.3)	GlcA-5	76.6, CH	3.80, m
13	145.3, Cq	-	GlcA-6	172.3, Cq	-
14	42.9, Cq	-	Gal-1	103.9, CH	4.79, d (7.3)
15	36.3, CH_2	1.82/1.33, m	Gal-2	73.7, CH	3.44, m
16	75.3, CH	4.45, d (3.2)	Gal-3	75.5, CH	3.41, m
17	50.0, Cq	-	Gal-4	70.9, CH	3.80, m
18	42.2, CH	3.01, dd (14.3, 4.2)	Gal-5	76.9, CH	3.48, m
19	47, 8, CH_2	2.29, t (13.6)/1.02	Gal-6	62.4, CH_2	3.76/3.69, m
20	31.6, Cq	-	Xyl-1	105.1, CH	4.58, d (7.6)
21	36.7, CH_2	1.94/1.14, m	Xyl-2	75.4, CH	3.24, m
22	32.9, CH_2	1.90/1.76, m	Xyl-3	78.4, CH	3.30, overlapped
23	210.8, CH	9.44, s	Xyl-4	71.2, CH	3.53, m
			Xyl-5	67.3, CH_2	3.90/3.25, m

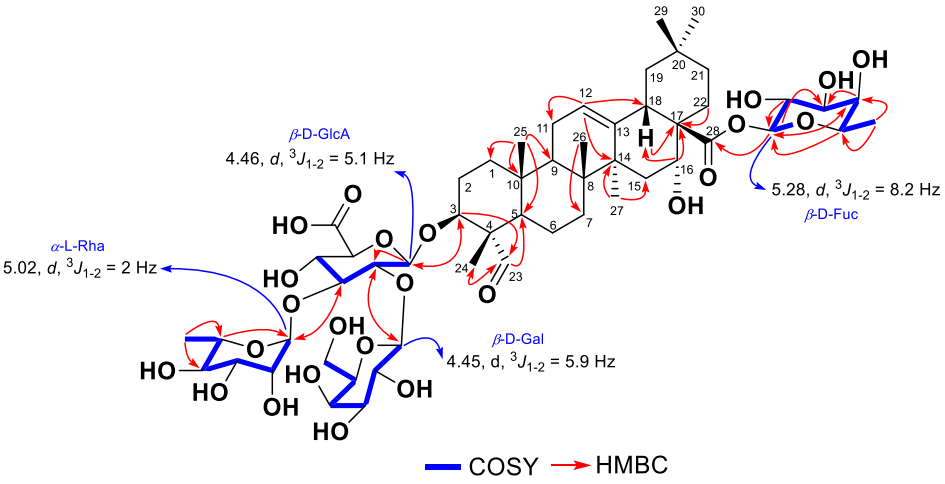
Table S6. ^{1}H, ^{13}C NMR spectral data for quillaic acid 3-*O*-{*β*-D-xylopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid} (8). NMR carried out in MeOH-d_4, (400, 100 MHz)



Quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid} (9)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.70/1.11, m	24	11.0, CH_3	1.16, s
2	25.9, CH_2	1.98/1.77, m	25	16.4, CH_3	1.00, s
3	86.2, CH	3.87, dd (12.3, 7.7)	26	18.0, CH_3	0.79, s
4	56.4, Cq	-	27	27.4, CH_3	1.40, s
5	49.2, CH, overlapped	1.33, m	28	181.2, Cq	-
6	21.5, CH_2	1.51/0.91, m	29	33.6, CH_3	0.88, s
7	33.7, CH_2	1.54/1.24, m	30	25.0, CH_3	0.97, s
8	41.1, Cq	-	GlcA-1	104.3, CH	4.48, d (6.8)
9	48.2, CH	1.75, m	GlcA-2	78.3, CH	3.64, m
10	37.2, Cq	-	GlcA-3	85.9, CH	3.65, m
11	24.6, CH_2	1.93/1.93, m	GlcA-4	73.2, CH	3.49, m
12	123.3, CH	5.30, t (3.3)	GlcA-5	76.7, CH	3.83, m
13	145.3, Cq	-	GlcA-6	172.6, Cq	-
14	42.9, Cq	-	Gal-1	104.4, CH	4.46, d (1.6)
15	36.3, CH_2	1.83/1.33, m	Gal-2	73.2, CH	3.48, m
16	75.4, CH	4.45, d (1.6)	Gal-3	75.2, CH	3.48, m
17	50.0, Cq	-	Gal-4	70.8, CH	3.81, m
18	42.2, CH	3.00, dd (14.3, 4.1)	Gal-5	77.2, CH	3.48, m
19	47.9, CH_2	2.29/1.02	Gal-6	62.5, CH_2	3.79/3.73, m
20	31.6, Cq	-	Rha-1	103.5, CH	5.03, d (1.6)
21	36.7, CH_2	1.94/1.14, m	Rha-2	72.2, CH	4.02, dd (3.3, 1.8)
22	32.9, CH_2	1.91/1.76, m	Rha-3	72.3, CH	3.65, m
23	210.9, CH	9.44, s	Rha-4	73.9, CH	3.49, m
			Rha-5	70.7, CH	3.92, m
			Rha-6	17.9, CH_3	1.24, d (6.2)

Table S7. ^{1}H, ^{13}C NMR spectral data for quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid} (9). NMR carried out in MeOH-d_4, (400, 100 MHz)

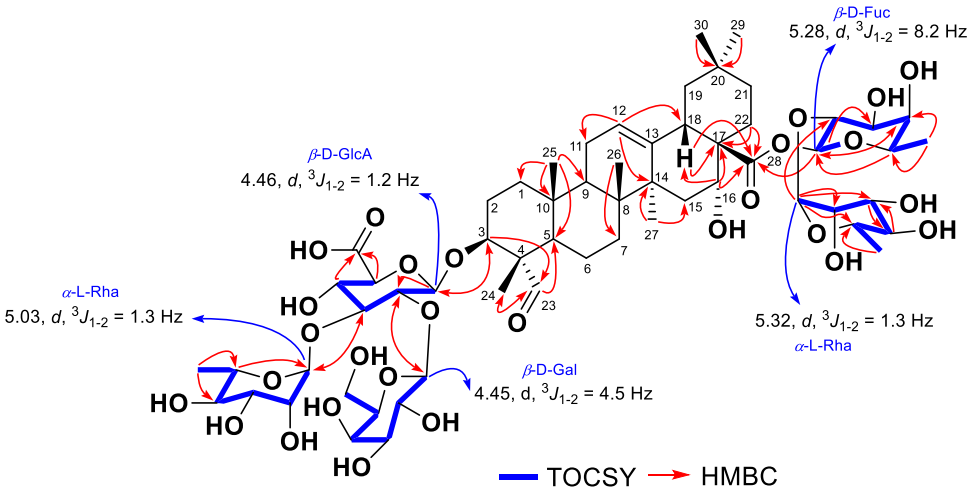


Quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-[*β*-D-fucopyranosyl] (11)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.70/1.10, m	**28**	177.5, Cq	-
2	25.8, CH_2	1.99/1.78, m	**29**	33.5, CH_3	0.88, s
3	86.2, CH	3.84, dd (3.5, 1.8)	**30**	25.1, CH_3	0.97, s
4	56.4, Cq	-	**GlcA-1**	104.3, CH	4.46, d (5.1)
5	49.4, CH, overlapped	1.33, m	**GlcA-2**	78.4, CH	3.63, m
6	21.5, CH_2	1.52/0.92, m	**GlcA-3**	86.0, CH	3.64, m
7	33.6, CH_2	1.53/1.24, m	**GlcA-4**	73.2, CH	3.48, m
8	41.1, Cq	-	**GlcA-5**	76.9, CH	3.79, m
9	48.2, CH	1.74, m	**GlcA-6**	Not detected	-
10	37.2, Cq	-	**Gal-1**	104.5, CH	4.45, d (5.9)
11	24.6, CH_2	1.92/1.92, m	**Gal-2**	73.2, CH	3.48, m
12	123.5, CH	5.31, t (3.8)	**Gal-3**	75.2, CH	3.48, m
13	144.9, Cq	-	**Gal-4**	70.8, CH	3.81, m
14	42.8, Cq	-	**Gal-5**	77.2, CH	3.47, m
15	36.4, CH_2	1.88/1.34, m	**Gal-6**	62.5, CH_2	3.78/3.72, m
16	75.1, CH	4.53, t (4.2)	**Rha-1**	103.5, CH	5.02, d (2)
17	50.0, Cq	-	**Rha-2**	72.2, CH	4.02, dd (3.3, 1.8)
18	42.2, CH	3.01, dd (14.2, 5)	**Rha-3**	72.3, CH	3.65, m
19	47.9, CH_2	2.30/1.05	**Rha-4**	73.2, CH	3.48, m
20	31.5, Cq	-	**Rha-5**	70.7, CH	3.92, m
21	36.6, CH_2	1.94/1.16, m	**Rha-6**	18.0, CH_3	1.24, d (2.2)
22	32.1, CH_2	1.93/1.78, m	**Fuc-1**	96.2, CH	5.28, d (8.2)

23	210.9, CH	9.44, s	**Fuc-2**	79.5, CH	3.58, m
24	11.0, CH_3	1.16, s	**Fuc-3**	75.5, CH	3.38, m
25	16.4, CH_3	1.00, s	**Fuc-4**	73.0, CH	3.69, m
26	18.0, CH_3	0.77, s	**Fuc-5**	70.8, CH	3.94, dd (4, 2.3)
27	27.4, CH_3	1.39, s	**Fuc-6**	17.9, CH_3	1.23, d (2.4)

Table S8. ^{1}H, ^{13}C-NMR spectral data for quillaic acid 3-*O*-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-*O*-[β-D-fucopyranosyl] (11). NMR carried out in MeOH-d_4 (600, 150 MHz)

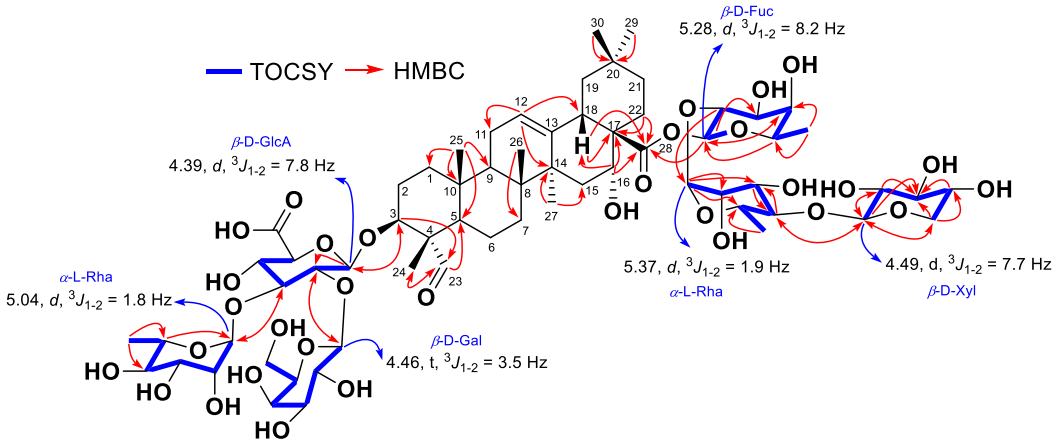


Quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{[*α*-L-rhamnopyranosyl-(1→2)-[*β*-D-fucopyranosyl]} (13)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.71/1.10, m	**GlcA-1**	104.3, CH	4.46, br d (1.2)
2	25.8, CH_2	1.99/1.78, m	**GlcA-2**	78.3, CH	3.63, m
3	86.2, CH	3.86, dd (11.9, 4.6)	**GlcA-3**	86.0, CH	3.64, m
4	56.5, Cq	-	**GlcA-4**	73.2, CH	3.48, m
5	49.3, CH, overlapped	1.33, m	**GlcA-5**	76.9, CH	3.79, m
6	21.5, CH_2	1.51/0.91, m	**GlcA-6**	173.4, Cq	-
7	33.6, CH_2	1.54/1.34, m	**Gal-1**	104.5, CH	4.46, br d (4.5)
8	41.3, Cq	-	**Gal-2**	73.2, CH	3.48, m
9	48.2, CH	1.74, m	**Gal-3**	75.2, CH	3.47, m
10	37.2, Cq	-	**Gal-4**	70.8, CH	3.81, m
11	24.6, CH_2	1.92/1.92, m	**Gal-5**	77.2, CH	3.47, m
12	123.4, CH	5.32, m	**Gal-6**	62.5, CH_2	3.78/3.72, m
13	144.8, Cq	-	**C_3-Rha-1**	103.5, CH	5.03, d (1.3)
14	42.9, Cq	-	**C_3-Rha-2**	72.2, CH	4.02, dd (3.3, 1.8)
15	36.6, CH_2	1.90/1.36, m	**C_3-Rha-3**	72.3, CH	3.65, m
16	74.8, CH	4.46, br d (2.4)	**C_3-Rha-4**	73.2, CH	3.48, m
17	50.2, Cq	-	**C_3-Rha-5**	70.7, CH	3.94, dd (10.9, 4.6)
18	42.6, CH	2.95, dd (14.3, 4.1)	**C_3-Rha-6**	17.9, CH_3	1.24, d (6.3)
19	48.2, CH_2	2.29, t, (13.6)/1.04	**Fuc-1**	95.4, CH	5.30, d (8.1)
20	31.5, Cq	-	**Fuc-2**	75.2, CH	3.79, m
21	36.6, CH_2	1.94/1.18, m	**Fuc-3**	76.5, CH	3.65, m
22	31.9, CH_2	1.93/1.83, m	**Fuc-4**	72.1, CH	3.59, m

23	210.9, CH	9.44, s	**Fuc-5**	72.8, CH	3.67, m
24	11.0, CH_3	1.15, s	**Fuc-6**	16.5, CH_3	1.22, d (6.4)
25	16.5, CH_3	1.00, s	**C_{28}-Rha-1**	101.9, CH	5.32, br d (1.3)
26	18.0, CH_3	0.78, s	**C_{28}-Rha-2**	72.1, CH	3.91, dd (3.4, 1.9)
27	27.3, CH_3	1.39, s	**C_{28}-Rha-3**	72.2, CH	3.62, m
28	177.5, Cq	-	**C_{28}-Rha-4**	73.7, CH	3.38, m
29	33.5, CH_3	0.88, s	**C_{28}-Rha-5**	70.5, CH	3.72, m
30	25.1, CH_3	0.96, s	**C_{28}-Rha-6**	18.5, CH_3	1.26, d (6.3)

Table S9. ^{1}H, ^{13}C-NMR spectral data for quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{[*α*-L-rhamnopyranosyl-(1→2)-[*β*-D-fucopyranosyl]} (13). NMR carried out in MeOH-d_4 (600, 150 MHz)

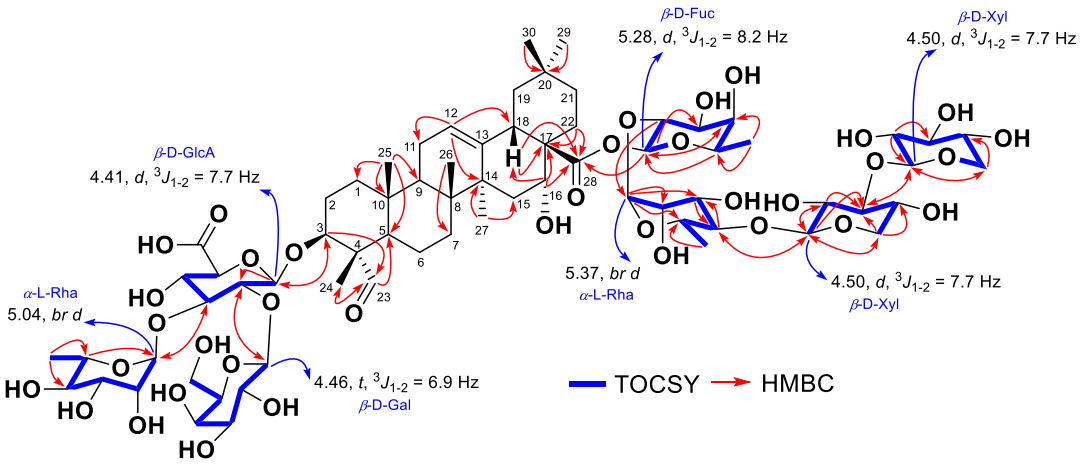


Quillaic acid 3-*O*-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-*O*-{[β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-fucopyranosyl]} (15)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.71/1.10, m	**Gal-1**	104.4, CH	4.46, d (3.5)
2	25.8, CH_2	2.01/1.78, m	**Gal-2**	73.2, CH	3.48, m
3	86.2, CH	3.87, m	**Gal-3**	75.3, CH	3.48, m
4	56.6, Cq	-	**Gal-4**	70.9, CH	3.82, m
5	49.3, CH, overlapped	1.32, m	**Gal-5**	77.1, CH	3.47, m
6	21.7, CH_2	1.50/0.91, m	**Gal-6**	62.4, CH_2	3.80/3.70, m
7	33.6, CH_2	1.50/1.32, m	**C_3-Rha-1**	103.4, CH	5.04, d (1.8)
8	41.1, Cq	-	**C_3-Rha-2**	72.3, CH	4.01, m
9	48.1, CH	1.74, m	**C_3-Rha-3**	72.3, CH	3.67, m
10	37.2, Cq	-	**C_3-Rha-4**	73.2, CH	3.49, m
11	24.6, CH_2	1.92/1.92, m	**C_3-Rha-5**	70.5, CH	4.01, m
12	123.4, CH	5.31, t (3.8)	**C_3-Rha-6**	18.0, CH_3	1.24, d (6.3)
13	144.8, Cq	-	**Fuc-1**	95.2, CH	5.28, d (8.2)
14	42.8, Cq	-	**Fuc-2**	74.8, CH	3.80, m
15	36.6, CH_2	1.92/1.45, m	**Fuc-3**	76.7, CH	3.68, m
16	74.8, CH	4.49, d (4.4)	**Fuc-4**	72.0, CH	3.53, m
17	50.2, Cq	-	**Fuc-5**	72.9, CH	3.68, m
18	42.3, CH	2.94, dd (14.4, 4.5)	**Fuc-6**	16.7, CH_3	1.22, d (6.4)
19	48.1, CH_2	2.29/1.04, m	**C_{28}-Rha-1**	101.4, CH	5.37, br d (1.9)
20	31.4, Cq	-	**C_{28}-Rha-2**	72.0, CH	3.94, dd (5.3, 3.4)
21	36.6, CH_2	1.92/1.18, m	**C_{28}-Rha-3**	72.2, CH	3.83, m
22	32.1, CH_2	1.92/1.75, m	**C_{28}-Rha-4**	84.4, CH	3.54, m
23	211.6, CH	9.44, s	**C_{28}-Rha-5**	69.0, CH	3.79, m

24	11.1, CH_3	1.16, s	**C_{28}-Rha-6**	18.5, CH_3	1.31, d (6.1)
25	16.5, CH_3	1.00, s	**Xyl-1**	107.1, CH	4.49, d (7.7)
26	17.8, CH_3	0.75, s	**Xyl-2**	76.3, CH	3.23, m
27	27.3, CH_3	1.38, s	**Xyl-3**	78.2, CH	3.35, m
28	177.5, Cq	-	**Xyl-4**	71.0, CH	3.50, m
29	33.5, CH_3	0.88, s	**Xyl-5**	67.3, CH_2	3.85/3.20, m
30	24.9, CH_3	0.94, s			
GlcA-1	104.3, CH	4.39, d (7.8)			
GlcA-2	78.3, CH	3.62, m			
GlcA-3	86.2, CH	3.63, m			
GlcA-4	73.2, CH	3.49, m			
GlcA-5	76.7, CH	3.80, m			
GlcA-6	Not detected	-			

Table S10. ^{1}H, ^{13}C-NMR spectral data for quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{[*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-[*β*-D-fucopyranosyl]} (15). NMR carried out in MeOH-d_4 (600, 150 MHz)

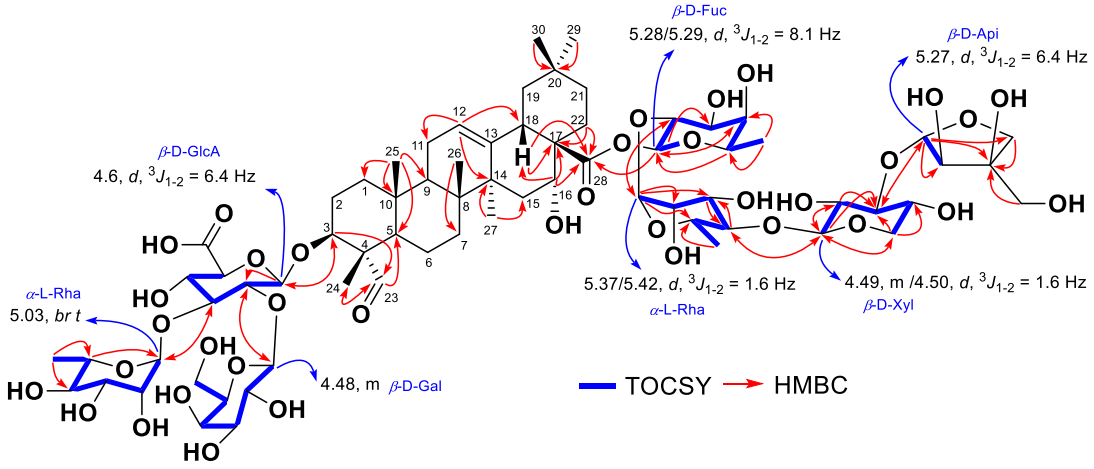


Quillaic acid 3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-*O*-{[β-D-xylopyranosyl-(1→3)-[β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-fucopyranosyl]} (17)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.71/1.10, m	**Gal-1**	104.4, CH	4.46, d (6.9)
2	25.8, CH_2	2.07/1.80, m	**Gal-2**	73.2, CH	3.49, m
3	86.3, CH	3.89, dd (12.7, 5.3)	**Gal-3**	75.2, CH	3.48, m
4	56.6, Cq	-	**Gal-4**	70.9, CH	3.82, m
5	49.6, CH, overlapped	1.32, m	**Gal-5**	77.1, CH	3.48, m
6	21.7, CH_2	1.50/0.91, m	**Gal-6**	62.4, CH_2	3.80/3.71, m
7	33.6, CH_2	1.50/1.33, m	**C_3-Rha-1**	103.4, CH	5.04, br s
8	41.2, Cq	-	**C_3-Rha-2**	72.3, CH	4.02, m
9	48.1, CH	1.75, m	**C_3-Rha-3**	72.3, CH	3.67, m
10	37.2, Cq	-	**C_3-Rha-4**	73.2, CH	3.49, m
11	24.6, CH_2	1.92/1.92, m	**C_3-Rha-5**	70.6, CH	4.01, m
12	123.4, CH	5.32, br s	**C_3-Rha-6**	18.0, CH_3	1.24, d (6.2)
13	144.8, Cq	-	**Fuc-1**	95.2, CH	5.28, d (8.2)
14	42.8, Cq	-	**Fuc-2**	74.8, CH	3.80, m
15	36.6, CH_2	1.92/1.45, m	**Fuc-3**	76.7, CH	3.68, m
16	74.8, CH	4.50, d (7.7)	**Fuc-4**	72.0, CH	3.53, m
17	50.2, Cq	-	**Fuc-5**	72.9, CH	3.68, m
18	42.3, CH	2.94, br d (15.9)	**Fuc-6**	16.7, CH_3	1.22, d (6.7)
19	48.1, CH_2	2.27, t (13.6)/1.05, m	**C_{28}-Rha-1**	101.4, CH	5.36, br s
20	31.4, Cq	-	**C_{28}-Rha-2**	72.0, CH	3.94, m
21	36.6, CH_2	1.92/1.18, m	**C_{28}-Rha-3**	72.2, CH	3.83, m
22	32.1, CH_2	1.92/1.76, m	**C_{28}-Rha-4**	84.4, CH	3.54, m
23	211.7, CH	9.44, s	**C_{28}-Rha-5**	69.0, CH	3.79, m

24	11.1, CH_3	1.17, s	**C_{28}-Rha-6**	18.5, CH_3	1.31, d (6.4)
25	16.5, CH_3	1.00, s	**Xyl (1)-1**	107.1, CH	4.50, d (7.7)
26	17.8, CH_3	0.74, s	**Xyl (1)-2**	75.2, CH	3.38, m
27	27.3, CH_3	1.37, s	**Xyl (1)-3**	87.4, CH	3.51, m
28	177.5, Cq	-	**Xyl (1)-4**	69.5, CH	3.55, m
29	33.5, CH_3	0.87, s	**Xyl (1)-5**	67.3, CH_2	3.85/3.22, m
30	24.9, CH_3	0.94, s	**Xyl (2)-1**	105.7, CH	4.54, d (7.9)
GlcA-1	104.3, CH	4.41, d (7.7)	**Xyl (2)-2**	75.2, CH	3.39, m
GlcA-2	78.3, CH	3.63, m	**Xyl (2)-3**	77.8, CH	3.40, m
GlcA-3	86.3, CH	3.63, m	**Xyl (2)-4**	71.1, CH	3.59, m
GlcA-4	73.2, CH	3.50, m	**Xyl (2)-5**	67.3, CH_2	3.95/3.29 (overlapped), m
GlcA-5	76.7, CH	3.80, m			
GlcA-6	Not detected	-			

Table S11. ^{1}H, ^{13}C-NMR spectral data for quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{[*β*-D-xylopyranosyl-(1→3)- [*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-[*β*-D-fucopyranosyl]} (17). NMR carried out in MeOH-d_4 (600, 150 MHz)



Quillaic acid 3-O-{α-L-rhamnopyranosyl-(1→3)-[β-D-galactopyranosyl-(1→2)]-β-D-glucopyranosiduronic acid}-28-*O*-{[β-D-apiofuranosyl-(1→3)-[β-D-xylopyranosyl-(1→4)-α-L-rhamnopyranosyl-(1→2)-[β-D-fucopyranosyl]} (19)

No.	δ_C, Type	δ_H mult, (*J* in Hz)	No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.4, CH_2	1.71/1.10, m	**Gal-1**	104.3/104.3, CH	4.48, m
2	25.8, CH_2	2.01/1.80, m	**Gal-2**	73.1, CH	3.49, m
3	86.3/86.4, CH	3.88, m	**Gal-3**	74.9, CH	3.50, m
4	56.5/56.6, Cq	-	**Gal-4**	70.8, CH	3.83, m
5	49.1, CH, overlapped	1.33, m	**Gal-5**	77.1, CH	3.50, m
6	21.6, CH_2	1.50/0.91, m	**Gal-6**	62.4/62.5, CH_2	3.81/3.71, m
7	33.6, CH_2	1.50/1.34, m	**C_3-Rha-1**	103.3/103.4, CH	5.03, br t
8	41.1, Cq	-	**C_3-Rha-2**	71.9, CH	4.03, m
9	48.0, CH	1.73, m	**C_3-Rha-3**	72.2, CH	3.67, m
10	37.2, Cq	-	**C_3-Rha-4**	73.1, CH	3.49, m
11	24.6, CH_2	1.92/1.92, m	**C_3-Rha-5**	70.6, CH	3.97, m
12	123.4/123.4, CH	5.31, m	**C_3-Rha-6**	18.0, CH_3	1.24, d (6.2)
13	144.7/144.8, Cq	-	**Fuc-1**	95.1/95.2, CH	5.28/5.29, d (8.1)
14	42.8, Cq	-	**Fuc-2**	74.5/74.9, CH	3.80, m
15	36.6, CH_2	1.92/1.45, m	**Fuc-3**	76.8/77.1, CH	3.68/3.69, m
16	74.4, CH	4.48, m	**Fuc-4**	72.1/72.2, CH	3.54/3.53, m
17	50.1/50.2, Cq	-	**Fuc-5**	73.0/72.9, CH	3.67/3.68, m
18	42.3, CH	2.94, br d (14.2)	**Fuc-6**	16.7, CH_3	1.22, d (6.7)
19	48.1, CH_2	2.27, td (13.7, 5.6)/1.05, m	**C_{28}-Rha-1**	101.4/101.3, CH	5.36/5.42, d (1.6)
20	31.4, Cq	-	**C_{28}-Rha-2**	71.9, CH	3.94/3.95, m
21	36.6, CH_2	1.92/1.17, m	**C_{28}-Rha-3**	72.2/72.1, CH	3.80/3.81, m

22	32.1, CH_2	1.92/1.74, m	**C_{28}-Rha-4**	84.4, CH	3.54/3.55, m
23	211.8, CH	9.45/9.46, s	**C_{28}-Rha-5**	69.0/68.9, CH	3.79/3.78, m
24	11.1, CH_3	1.17/1.18, s	**C_{28}-Rha-6**	18.4, CH_3	1.32, d (6.4)
25	16.5, CH_3	1.00, s	**Xyl-1**	107.1/107.4, CH	4.49, m/4.50, d (1.6)
26	17.8, CH_3	0.74, s	**Xyl-2**	75.8, CH	3.33/3.34, m
27	27.3, CH_3	1.38, s	**Xyl-3**	85.2/85.5, CH	3.43/3.44, m
28	177.5/177.5, Cq	-	**Xyl-4**	69.6, CH	3.51/3.52, m
29	33.5, CH_3	0.88, s	**Xyl-5**	67.3/67.0, CH_2	(3.85/3.22)/(3.89/3.24), m
30	24.8/24.9, CH_3	0.93/0.94, s	**Api-1**	111.0, CH	5.27, d (6.4)
GlcA-1	104.2/104.3, CH	4.46, d (6.4)	**Api-2**	78.0, CH	4.07, d (2.9)
GlcA-2	78.2, CH	3.63, m	**Api--3**	80.8, Cq	-
GlcA-3	85.8/85.9, CH	3.66, m	**Api-4**	75.1, CH_2	4.16, d (9.7)/3.83, m
GlcA-4	73.1, CH	3.50, m	**Api-5**	65.4, CH_2	3.68, m
GlcA-5	76.6, CH	3.81, m			
GlcA-6	Not detected	-			

Table S12. ^{1}H, ^{13}C-NMR spectral data for quillaic acid 3-*O*-{*α*-L-rhamnopyranosyl-(1→3)-[*β*-D-galactopyranosyl-(1→2)]-*β*-D-glucopyranosiduronic acid}-28-*O*-{[*β*-D-apiofuranosyl-(1→3)-[*β*-D-xylopyranosyl-(1→4)-*α*-L-rhamnopyranosyl-(1→2)-[*β*-D-fucopyranosyl]} (19). NMR carried out in MeOH-d_4 (600, 150 MHz). Chemical shifts marked in blue are reported for the two non-separable rotamers.

Gene ID	QsbAS1 co-expression (PCC)	Primordia transcript quantity (TPM)
Qs0264740	0.970	3604.71
Qs0072520	0.964	5217.33
Qs0322030	0.955	5578.42
Qs0179550	0.950	2386.09
Qs0098630	0.950	5624.47
Qs0307390	0.942	5437.64
Qs0206480	0.940	3999.06
Qs0264720	0.934	8901.26
Qs0264710	0.908	8728.14
Qs0302420	0.906	1891.92

Table S13. Shortlisted acyltransferases

Quillaic acid triterpene

No.	δ_C, Type	δ_H mult, (*J* in Hz)
1	39.1, CH_2	1.70/1.10, m
2	Not detected	
3	85.8, CH	3.87, m
4	56.3, Cq HMBC	-
5	48.9, CH	1.33, m
6	Not detected	
7	33.4, CH_2	1.36/1.58
8	41.1, Cq HMBC	-
9	48.0, CH	1.74, m
10	37.1, Cq HMBC	-
11	24.2, CH_2	1.92/1.92, m
12	123.2, CH	5.33, m
13	144.5, Cq	-
14	42.8, Cq HMBC	-
15	36.5, CH_2	1.58/1.45, m
16	74.5, CH	4.45, d (2)
17	50.1, Cq HMBC	-
18	42.2, CH	2.94, br d (14.3)
19	48.0, CH_2	2.30/1.05, m
20	31.3, Cq HMBC	-
21	36.4, CH_2	1.93/1.18, m
22	34.1, CH_2	1.67/1.67, m
23	211.2, CH	9.46, s
24	10.9, CH_3	1.16, s
25	16.3, CH_3	1.00, s
26	17.8, CH_3	0.80, s
27	27.1, CH_3	1.39, s
28	177.3, Cq, HMBC	-
29	33.3, CH_3	0.89, s
30	24.9, CH_3	0.97, s

Table S14. ^{1}H, ^{13}C NMR spectral data for quillaic acid (QA) triterpene core of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR carried out in MeOH-d_4 (600, 150 MHz)

Position	Semi-purified QS-7 ^{1}H, ^{13}C-NMR	Literature ^{1}H, ^{13}C-NMR
GlcA-1	4.40/104.2	4.37/103.9
GlcA-2	3.65/78.2	3.65/78.4
GlcA-3	3.67/86.5	3.68/86.4
GlcA-4	3.52/71.6	3.53/71.7
GlcA-5	3.58/77.3	3.59/77.5
GlcA-6	Not detected	172.7
Gal-1	4.79/103.3	4.80/103.4
Gal-2	3.47/73.3	3.47/73.2
Gal-3	3.46/75.2	3.46/75.1
Gal-4	3.85/70.5	3.84/70.5
Gal-5	3.50/76.3	3.50/76.3
Gal-6	3.73, 3.78/60.1	3.73, 3.77/62.0
C_3-Xyl-1	4.61/104.7	4.62/104.5
C_3-Xyl-2	3.22/75.2	3.23/75.2
C_3-Xyl-3	3.32/78.0	3.32/78.0
C_3-Xyl-4	3.50/71.1	3.51/71.0
C_3-Xyl-5	3.23, 3.93/66.8	3.23/3.90/66.9
Fuc-1	5.44/95.0	5.44/94.9
Fuc-2	3.92/75.5 HMBC	3.86/76.2
Fuc-3	3.96/82.7 HMBC	3.96/82.3
Fuc-4	5.18/74.3	5.18/74.3
Fuc-5	3.88/71.1	3.87/71.1
Fuc-6	1.07/16.4	1.05/16.4
RhaI-1	5.08/101.3	5.09/101.8
RhaI-2	4.11/71.2	4.11/71.3
RhaI-3	3.87/81.5	3.86/83.2
RhaI-4	3.65/78.4	3.66/78.9
RhaI-5	3.80/69.3	3.79/69.4
RhaI-6	1.29/18.4	1.29/18.7
RhaII-1	4.88/104.9	4.92/104.8
RhaII-2	3.88/72.1	3.87/72.1
RhaII-3	3.55/72.2	3.55/72.2
RhaII-4	3.38/73.8	3.37/73.8
RhaII-5	3.58/71.0	3.58/71.1
RhaII-6	1.22/17.9	1.20/17.9
C_{28}-Xyl-1	4.69/105.4	4.69/105.2
C_{28}-Xyl-2	3.19/75.4	3.19/75.3
C_{28}-Xyl-3	3.39/86.1	3.34/86.0
C_{28}-Xyl-4	3.48/70.7	3.48/70.7
C_{28}-Xyl-5	3.17, 3.88/66.9	3.17, 3.87/66.9
Api-1	5.25/111.2	5.29/111.3
Api-2	4.06/77.6	4.03/78.0
Api-3	80.6, HMBC	80.2
Api-4	3.81, 4.17/75.1	3.80, 4.14/75.0
Api-5	3.67/65.5	3.64/65.6
Glc-1	4.55/105.4	4.55/105.1
Glc-2	3.30/75.3	3.29/75.4
Glc-3	3.35/77.8	3.35/77.7
Glc-4	3.33/71.3	3.34/71.2
Glc-5	3.37/77.8	3.36/77.9
Glc-6	3.72, 3.86/62.2	3.71, 3.85/62.2
Acetyl-CO	172.4, HMBC	NR
Acetyl-Me	2.16/21.0, HMBC	NR

Table S15. ^{1}H, ^{13}C NMR spectral data for C_3, C_{28} oligosaccharides of semi-purified QS-7 (20) produced in *N. benthamiana*. NMR carried out in MeOH-d_4 (600, 150 MHz). NR – Not reported

Sugar Nucleotide	MRM transitions	Fragment
UDP-α-D-fucose	549 → 323	$[NMP-H]^-$
(UDP-D-Fuc)	549 → 159	$[H_4P_2O_7-H_3O]^-$
UDP-β-L-rhamnose	549 → 323	$[NMP-H]^-$
(UDP-L-Rha)	549 → 159	$[H_4P_2O_7-H_3O]^-$
UDP-GlcNAcA	620 → 403	$[NDP-H]^-$
	620 → 159	$[H_4P_2O_7-H_3O]^-$

Table S16. Relative retention times and MRM transitions of sugar nucleotides

Data S1. (separate file) Full list of primers for *Q. saponaria* genes cloned and described in this study. Primers were designed with 5' attB overhangs for Gateway® cloning (denoted in red). The names of genes found to be involved in QS biosynthesis are given to the right. Primers for making protein expression constructs are also listed (blue text denotes initiation codon and purple text denotes the sequence encoding hexahistidine and stop codon).

Data S2. (separate file) Full list of 35 full-length P450s identified in the *Q. saponaria* 1KP transcriptome. Gene IDs for the re-assembled SRA data are shown in column B (with sequences in columns F-H) and the top hit from the original assembled 1KP data set (prefix OQHZ) are given in column C (available from http://www.onekp.com/public_data.html). The closest *Arabidopsis thaliana* matches identified through BLAST searches are given in column D. Column E indicates whether an identified gene was successfully amplified by PCR and tested for C-23 oxidase activity.

Data S3. (separate file) Summary of *Q. saponaria* biosynthetic gene clusters as predicted by plantiSMASH. Results of genome analysis by plantiSMASH are summarized and ordered according to cluster number. Functional gene annotations for the clustered genes are also included.

Data S4. (separate file) Compound names, numbers, gene sets, isolated yields, retention times and *m/z* values. The full and abbreviated names and numbers for the major compounds identified in this study along with isolated yields. Additionally, the specific set of genes which were transiently expressed in *N. benthamiana* in order to produce each compound (for both analytical and large-scale experiments) are provided, along with the *m/z* values and retention times for each product.

Data S5. (separate file) Full NMR spectra for each compound isolated from *N. benthamiana*. Copies of 1D (^{1}H, ^{13}C, DEPTQ-135, DEPT-135 NMR) and 2D NMR (COSY, TOCSY, HSQC, HMBC, ROESY) spectra for the C-3 left hand side intermediates (**6**-**9**) and the C-28 right hand side pathway intermediates **11**, **13**, **15**, **17** and **19.**