Supplemental Information

Methods

Metabolite analyses. For metabolomics analyses of *Siraitia* tissues, frozen tissue was ground in liquid nitrogen (IKA A11 homogenizer, IKA®-Werke GmbH & Co., Staufen, Germany). 600 µl of methanol: water (1:1) was added to 200 mg fine ground powder and the resulting mixture was vortexed for 30 seconds, sonicated for 15 min and vortexed again for 30 seconds. The sample was cleaned of debris by centrifugation (20,000g) and by filtration using PTFE 0.2 µm syringe filters (Axiva, Sigma Chemicals). For metabolic analyses of transgenic tobacco leaves expressing SgCDS under the CaMV promoter, mature leaves were freeze dried and 200 mg fine ground powder were extracted with 600 mL of 80% MeOH as described above. Chromatographic separations and identifications HPLC-DAD analysis was carried out on an Agilent 1200 HPLC system with an Agilent 1200 Diode Array Detector. The analytical column used was Zorbax Eclipse XDB - C18 (4.6x50.0 mm, 1.8 µm, Agilent Technologies, USA). The mobile phase contained A, H_2O; B, 100% HPLC grade acetonitrile. The column was equilibrated with 77% A, and the sample was injected, reaching 80% B gradient after 10 min. The mobile phase flow was 1.5 mL/min. The semi-preparative column used was: Luna 5µ C18(2) 100A, 250x10mm (Phenomenex, USA). The mobile phase contained A, H_2O; B, 100% HPLC grade acetonitrile. For semi-preparative mogroside separation, the column was equilibrated with 77% A, and then 100 µl of sample was injected, reaching 33% B after 5 min and 90% B after 12 min. For cucurbitadienol-like molecules, the column was equilibrated with 40% A, and then 100 µl of sample was injected, reaching 95% B gradient after 9 min, and then reaching 100% B after 34 minutes and running it for six minutes, before returning to 60% B. The mobile phase flow for semi-preparative column was 5 mL/min. Each substance was identified by co-migration with standards and by matching the UV spectrum of each mogroside peak against that of a standard. Portions of eluted and collected peaks were taken also for analysis in LC-MS. The LC-MS analysis was carried out on an Agilent 1290 Infinity series liquid chromatograph coupled with an Agilent 1290 Infinity DAD and Agilent 6224 Accurate Mass Time of Flight (TOF) mass spectrometer (MS). The analytical column was: Zorbax Extend-C18 Rapid Resolution HT column (2.1x50.0 mm, 1.8 µm, Agilent Technologies, Waldbronn, Germany) Mass spectrometry was performed using an Agilent 6224 Accurate Mass TOF LC-MS System equipped with dual-sprayer orthogonal ESI (for mogroside glucosylation assays) or APCI (for cucurbitadienol and hydroxylated derivatives) sources (Agilent Technologies, Santa Clara, USA). The mobile phase contained A, H_2O; B, 100% HPLC grade acetonitrile, both with 0.1% formic acid. The column was equilibrated with 100% A, and then the sample was injected, reaching 50% B gradient after 10 min. The mobile phase flow was 0.4 mL/min. Eluting compounds were subjected to dual ESI source, with one sprayer for analytical flow and one for the reference compound (Agilent Technologies, Santa Clara, USA). The ESI source was operated in positive mode with

the following settings: Gas and vaporizer temp- 300ºC; drying gas flow of 8 L/min and nebulizer set to 35 psig. VCap set to 3000 V; and Fragmentor to 110 V. Scan mode of the mass detector was set (110–1000 m/z). Each substance was identified by co-migration with commercial standards and by matching the mass spectrum of putative peak against that of a standard and expected exact mass. The chromatogram was initially analyzed by MassHunter Qualitative Analysis software v.B.06.00 (Agilent) and further analyzed by MassHunter Mass Profiler software v.B.05.00 (Agilent). Cucurbitadienol and hydroxylation products were separated by a modified program as follows. The column was equilibrated with 5 % B at a flow rate of 0.3 mL/min for 1.5 min. Eluent B was then increased to 95 % till 6 min, raised to 100 % B at 12 till 15 min and restored to 5 % by 16.5 min. The flow rate of the mobile phase was 0.3 mL/min and the column oven temperature was 40ºC. Eluting compounds were subjected to positive APCI source, with one sprayer for analytical flow and one for the reference compound (Agilent Technologies, Santa Clara, USA). The APCI source was operated in positive mode with the following settings: Gas and vaporizer temp- 350ºC; drying gas flow of 5 L/min and nebulizer set to 40 psig. VCap set to 3500 V; corona needle 7 µA and Fragmentor to 140 V. Scan mode of the mass detector was applied (110–1000 m/z).

Standards. Triterpenoids were identified by comparison of their exact mass, mass spectrum and retention times of purchased standards (squalene, 2,3-epoxysqualene, lanosterol, Sigma-Aldrich; 2,3,22,23-diepoxysqualene, Echelon Biosciences, Salt Lake City, UT, USA) and of prepared mogroside standards, as below. Standards of Mogroside VI, Mogroside V, Isomogroside V, 11-oxoMogrosideV, Mogroside IVA and Siamenoside were generously provided by The Coca Cola Company and described in (1). To obtain mogrosides with lower degree of glycosylation, as well as the aglycone mogrol, we performed enzymatic and acid hydrolysis of Mogroside V, as described below. Cellulase: 10 mg of Mogroside V were incubated (shaking at 200 rpm) together with 25mg cellulase of *Trichoderma reesii* (Sigma) for 48hours in 10 mL sodium acetate buffer (0.1M, pH4.3) at 42^{o}C. Reaction mix with accumulated Mogroside III and Mogroside II-A (M2c) was dried by lyophilisation, dissolved in 1 mL of Methanol: H_2O (1:1) and separated using semi-preparative HPLC, as above. Mild acid hydrolysis: 10 mg of Mogroside V were incubated in 0.2N HCl in methanol at 90^{o}C for 3h. Then, the reaction mix with accumulated Mogroside II-A1 (M2x), Mogroside 1 and Mogrol was lyophilized, dissolved in 1 mL of Methanol: H_2O (1:1) and separated using semi-preparative HPLC. The structures of newly acquired substances were verified by NMR, described below.

TLC. To isolate preparative amounts of cucurbitane-like substances prior to final purification on HPLC system, total extracts of yeast accumulating the products of the C11 and C19 CYP hydroxylases were applied to TLC silica gel 60 with concentrating zone 20x2.5cm (Merck KGaA, Germany). The TLC solvent system used for isolation of less polar compounds ($C_{30}H_{50}O_2$), was hexane/petroleum-ether/ethyl-acetate 10/10/10 (vol/vol/vol). The solvent system used for isolation of more polar

compounds (oxidocucurbitadienol) was hexane/petroleum-ether/ethyl-acetate 15/15/7.5 (vol/vol/vol). When the front reached middle of the plate, an additional 15 parts of ethyl-acetate were added into the solvent system. The TLC run continued until the front reached 1 cm from the upper edge of the plate. Bands were visualized using ρ-anisaldehyde/sulphuric acid/acetic acid (2) (1:1:48, vol/vol/vol), as well as by UV. The silica fractions were carefully removed from the aluminum base, and components were extracted by vortexing (30s) in 10 mL of methanol, filtered and further evaporated under a gentle flow of nitrogen gas. Each fraction was resuspended in 1 mL methanol, and checked on LC-MS for presence of the substances of interest.

NMR. NMR spectra were run in a Bruker Avance-III-700 instrument in CD_3OD as a solvent containing TMS as internal reference, at 300K. In addition to 1D ^{1}H and ^{13}C spectra (at 700.5 and 176.1 MHz, respectively), we also performed three 2D experiments: COSY ($^{1}H\times^{1}H$ correlation) HMQC (one-bond $^{1}H\times^{13}C$ correlation) and HMBC (long-range $^{1}H\times^{13}C$ correlation); this permitted the assignment of every carbon and proton signals in the molecules (see Table S8) and confirmed the molecular structures.

DNA isolation, RNA isolation, library preparation and sequencing. DNA isolation was performed using the GenElute™ Plant Genomic DNA miniprep kit (Sigma, St.Louis, MO). The quality of the DNA was analyzed by ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE) and by electrophoresis on agarose gel. The concentration of DNA was estimated using Qubit® 2.0 Fluorometer (Life technologies, Singapore) and Qubit® dsDNA BR Assay Kit. Genomic DNA samples were sent to the W.M. Keck Center for Comparative and Functional Genomics (University of Illinois, USA) for the preparation of DNA libraries for sequencing. Construction of shotgun genomic, mate-pair and TSLR DNA libraries and sequencing on the HiSeq2500 were carried out at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign (UIUC). The shotgun genomic DNA libraries were constructed from 1μg of DNA after sonication with a Covaris M220 (Covaris, MA) with the Library Preparation Kit from Kapa Biosystems (Kapa Biosystems, MA). The libraries were loaded onto a 2% agarose gel and fragments 120bp to 330bp and 360 to 510bp in length were recovered for the final libraries with the QIAquick gel extraction kit (Qiagen, CA). Mate-pair libraries were prepared with the Nextera Mate-Pair Sample Preparation Kit (Illumina, CA). Briefly, 10ug of high quality genomic DNA was subjected to two tagmentation reactions and run on a 0.6% Megabase agarose gel. Genomic fragments 5-7kb and 8-10kb were size selected, purified on an EluTrap (GE Healthcare Life Sciences, Piscataway, NJ) and circularized. The circles were sonicated with a Covaris M220 and enriched for those fragments containing the biotinylated circularization adaptor. Enriched fragments were end-repaired, A-tailed, adaptored and PCR amplified with the TruSeq DNA Sample Prep kit (Illumina). Four TSLR libraries were constructed with the TruSeq Synthetic Long-Read DNA Library Prep kit (Illumina, CA) following the manufacturer's protocols. The final libraries were run on Agilent bioanalyzer DNA

high-sensitivity chips (Agilent, Santa Clara, CA) to determine the average fragment size and to confirm the presence of DNA of the expected size range. They were also quantitated by qPCR on a BioRad CFX Connect Real-Time System (Bio-Rad Laboratories, Inc. CA) prior to pooling and sequencing. The shotgun and mate-pair libraries were pooled in equimolar concentration based on the qPCR concentration and sequenced on an Illumina HiSeq2500. The DNA fragments were sequenced for 101 cycles from each end using TruSeq SBS sequencing kits v3. The raw .bcl files were converted into demultiplexed compressed fastq files using bcl2fastq v1.8.2 Conversion Software (Illumina). Each TSLR library was sequenced on one lane on an Illumina HiSeq 2500 for 161 cycles from each end of the fragments using Rapid SBS sequencing kits v1. The runs were streamed to BaseSpace and assembled into long reads using the integrated TruSeq Long Read Assembly Software from Illumina.

RNA isolation. Total RNA was isolated using a modification to the method of Verwoerd *et al*. (3) from: (1) *Siraitia* fruits (mix of at least 3 fruit from each stage) harvested during development between 15 DAA to 103 DAA (15, 34, 50, 77, 90, 103 DAA) and, (2) leaves, stems and roots. Briefly, frozen, uniformly ground samples (~3-4 g) were mixed by vortexing in a 50-mL tube with 10 mL hot extraction buffer (80ºC) enclosing equal parts of phenol and RNA isolation buffer contained 0.1 M Tris-HCI (pH 8.0), 0.1 M LiCl, 0.01 M EDTA, 1% (w/v) SDS. After vortexing for 30 second 6 mL mix of chloroform-3-methylbutanol (24:1, v/v) was added, vortexed and centrifuged at 4000*g* for 7 min. The aqueous phase was transferred to a new 50-mL tube and an equal volume of 4 M LiCl was added to the solution. RNAs are allowed to precipitate overnight and collected by centrifugation 12,000 *g* for 10 min at 4º C. The resulting RNA pellet was dissolved in 0.5 mL diethylpyrocarbonate (DEPC) water. After re-precipitation with 1/10 volume of 3 M sodium acetate (pH 5.3) and 2 volumes 95% ethanol, the pellet was dissolved in 100 µL DEPC water. The quality of the RNA was analyzed by ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE) and by electrophoresis on a formaldehyde agarose gel. Total RNA samples were sent to the W.M. Keck Center for Comparative and Functional Genomics (University of Illinois, USA) for the preparation of Illumina RNA-Seq libraries and sequencing.

***De novo* transcriptome assembly and annotation.** Raw reads were subjected to a cleaning procedure using the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html, version 0.0.13.2) including: (1) removing adaptors from reads using fastx_barcode_splitter (2) trimming read-end nucleotides using fastx_trimer; (3) removing sequencing artifacts using fastx artifacts_filter (4) removing reads that had less than 70% base pairs with quality score $\leq$ 22 using fastq_quality_filter, (5) removing poly A-tails from the high quality reads using EMBOSS 6.4: trimmest, (6) removing rRNA, mtDNA and chloroplast sequences. A total of ~119 million clean reads, obtained after processing and cleaning, including 9 single-end libraries of 100bp from 15DAA (days after anthesis), 34DAA, 50DAA, 77DAA, 90DAA,

103DAA fruit, stem, leaves and root) were assembled *de novo* using the CLC-BIO program (http://www.clcbio.com/files/appnotes/CLC_bio_RNA.pdf). The resulting transcripts were annotated using the Basic Local Alignment Search Tool (BLASTX) (4) against the melon protein database (https://melonomics.net/; version 3.5), the Plant Transcription Factors database (http://planttfdb.cbi.pku.edu.cn/), and the SwissProt database, with an E-value cut-off of 10^{-5}. Supplementary Data File 1 includes the transcriptome assembly sequences and annotation, listed as contigs. The resulting transcriptome was mapped to the genome assembly using version 2.1.0 of the bowtie2 software (5).

Protein modelling and localization. SQE was modelled using Phyre2 (6) fold recognition server. Phyre2 was used to model the 3D structure since the closest homologs in the PDB showed only18% sequence identity or less. The server uses advanced remote homology detection methods and through sequential steps, such as profile construction, similarity analysis, and structural properties, selects the best suited templates and generates protein models. All the resulting high quality models were based on Flavin monooxygenase fold. The binding tunnel was calculated using the CAVER (7) program. Potato epoxide hydrolase (PDB entry 2CJP) served as template for modeling SgEPH structure (60% sequence identity). The protein sequences were aligned using HHpred (8) (profile Hidden Markov based alignment). All-atom model of SgEH was then built using the restrained-based modelling approach as implemented in the program MODELLER (9) 9V13. Docking of the epoxycucurbitadienol and epoxysqualene into the SgEPH constructed homology model was carried out using AutoDock Vina (10). The rotatable torsions of the ligands were released during docking calculations as well as the rotatable torsions of several residues in the binding site. Human lanosterol synthase (PDB entry 1W6K) served as template for modeling SgCDS structure (45% sequence identity) performed as for the EPH protein. Cyanobacterial CYP120A1 (PDB entry 2VE3) served as template for modeling SgCYP88L structure (21% sequence identity) performed as for the EPH protein. For modelling of the UGTs, several structural templates were used: *Medicago truncatula* UGT71G1 (PDB entry 2ACW), *Arabidopsis thaliana* UGT72B1 (PDB entry 2VCE), *Vitis vinifera* UFGT (PDB entry 2C1Z), *Medicago truncatula* UGT78G1 (PDB entry 3HBF) and *Medicago truncatula* UGT85H2 (PDB entry 2PQ6). The different SgUGTs share 20-30% sequence identity to the putative templates. The proteins sequences were aligned using multiple sequence alignment tools: HHpred, Promals3D (11) and Expresso (12). All-atom models of the various SgUGTS were then built based on the different sequence alignments, using the restrained-based modelling approach as implemented in the program MODELLER 9V13. The models were evaluated using z-DOPE (13), ProSA (14), ProQ2 (15) and QMEAN (16). For each template, the model showing the best score as judged by consensus prediction carried out by these four evaluation methods was saved for further studies. Docking of mogrol and other mogrosides into SgUGT constructed homology models were carried out using AutoDock Vina. The ligands rotatable torsions were released during docking calculations as well as the rotatable torsions of

several residues in the binding site. Protein localization was performed using the following programs: Bacello (17), Protein Prowler (18), Predotar (19), TargetP (20), Psort (21) and Cello (22).

Supplemental references

1. Chaturvedula, V.S.P **&** Prakash, I. Cucurbitane glycosides from *Siraitia grosvenorii*. *J. Carb. Chem.* 30:16-26 (2011).

2. Geisler K. *et al.* Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants. *Proc Natl Acad Sci USA.* 110:E3360- E3367 (2013).

3. Verwoerd, T.C., Dekker, B.M.M. **&** Hoekema, A. A. Small-scale procedure for the rapid isolation of plant RNAs. *Nucleic Acids Res.* 17:2362 (1989).

4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. **&** Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410 (1990).

5. Langmead, B. **&** Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 9:357-359 (2012).

6. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. **&** Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protoc.* 10:845-858 (2015).

7. Kozlikova, B. *et al.* CAVER Analyst 1.0: graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics* 30:2684-2685 (2014).

8. Söding, J., Biegert, A., **&** Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* 33:W244-W248 (2005).

9. Eswar, N. *et al. Comparative Protein Structure Modeling With MODELLER. In:* Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15:5.6.1-5.6.30 (2006).

10. Trott, O. **&** Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *Journal of Computational Chemistry* 31:455-461 (2010).

11. Pei, J., Kim, B.-H. **&** Grishin, N.V. PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Res*. 36:2295-2300 (2008).

12. Armougom, F. *et al.* Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res*. 34:W604-608 (2006).

13. Shen, M.Y. **&** Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci*. 15:2507-2524 (2006).

14. Wiederstein **&** Sippl. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* 35:W407-W410. (2007).

15. Ray, A., Lindahl, E. **&** Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* 13:224 (2012).

16. Benkert, P., Künzli, M. **&** Schwede, T. QMEAN Server for Protein Model Quality Estimation. *Nucleic Acids Res*. 37:W510-514 (2009).

17. Pierleoni, A. *et al*. BaCelLo: a Balanced subCellular Localization predictor. *Bioinformatics.* 22:E408- E416 (2006).

18. Bodén, M **&** Hawkins, J. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics.* 21:2279-2286 (2005).

19. Small, I., Peeters, N., Legeai, F. **&** Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4:1581-1590 (2004).

20. Emanuelsson, O., Brunak, S., von Heijne, G. **&** Nielsen, H. Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols* 2:953-971 (2007).

21. Nakai, K., **&** Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci*. 24:34-36 (1999).

22. Yu, C.S., Chen, Y.C., Lu, C.H. **&** Hwang, J.K.: Prediction of protein subcellular localization. Proteins: Structure, Function and Bioinformatics 64:643-651 (2006).

Supplementary figures

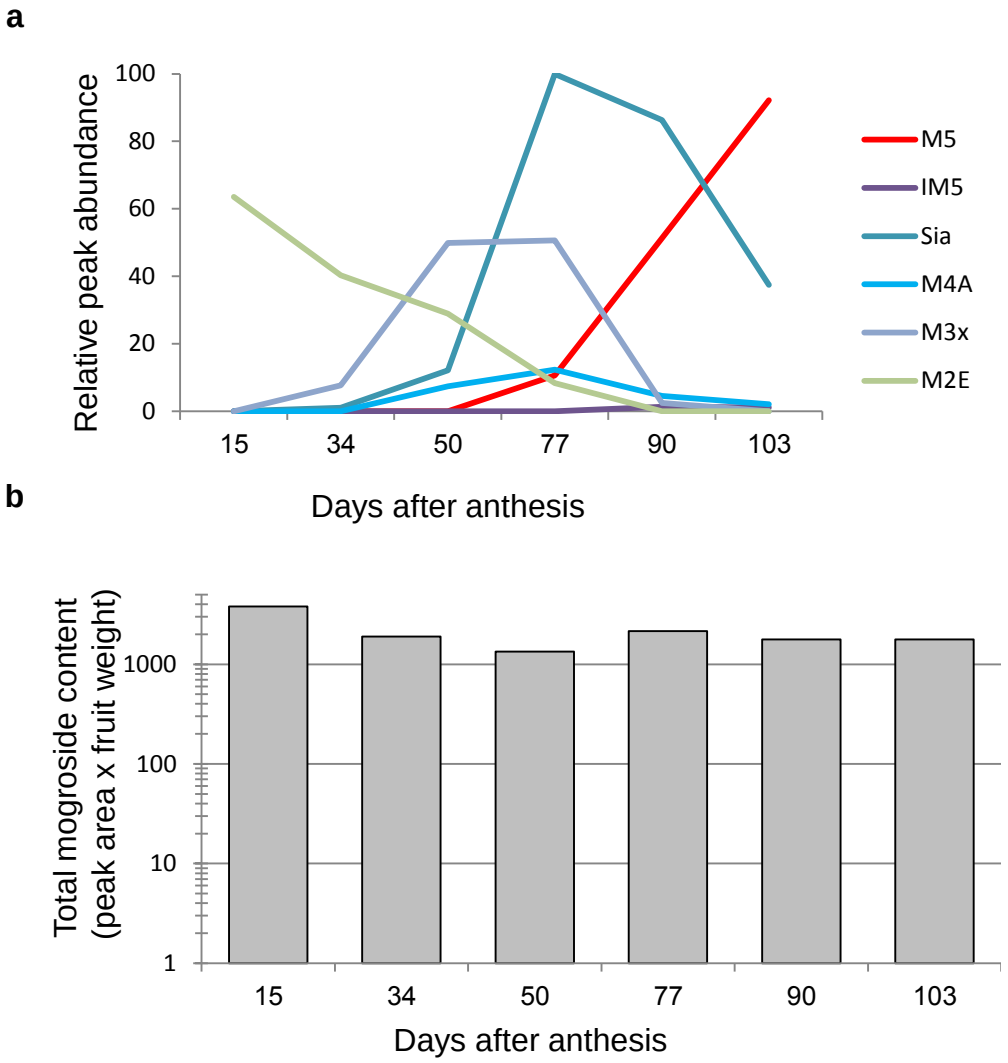


Fig. S1. Mogrosides of *Siraitia grosvenorii* fruit undergo successive glucosylations, but their total content remains stable during development. For each mogroside in developing fruit, the area of assigned peak (m/z) was measured using LC-MS. **a**) Glucosylation pattern of mogrosides alters in the course of *Siraitia* fruit development. The Siamenoside level at 77 DAA was taken as 100% and all other peak areas are related to this. M5, mogroside V; IM5, isomogroside V; Sia, siamenoside; M4A, mogroside IVA; M3x, mogroside 3x; M2E, mogroside IIE. **b**) Total mogroside content of *Siraitia grosvenorii* fruit stays stable during course of development**.** Peak areas of all mogrosides were confirmed for each stage to obtain total mogroside content of the fruit. The combined extract from three fruits of each stage was separated by LC-MS and quantified. DAA, days after anthesis.

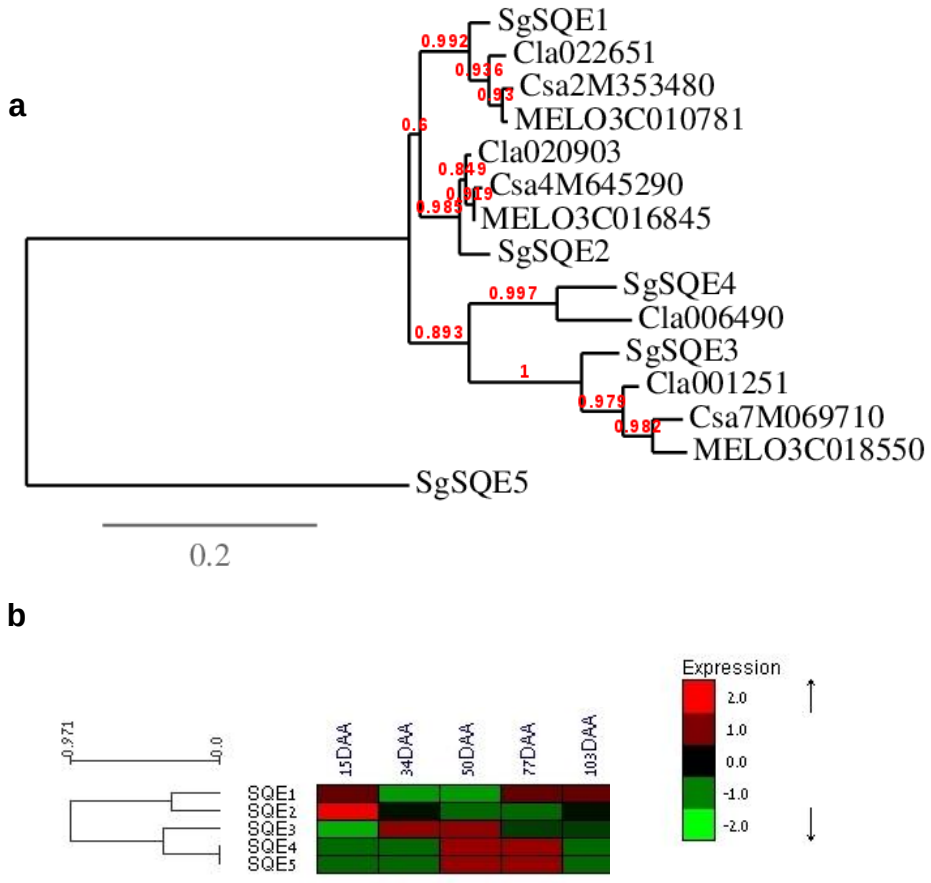


Fig. S2. Squalene epoxidase genes in *Siraitia* and other *Cucurbitaceae*. **a**) Phylogenetic tree of squalene epoxidases in *Siraitia* and additional cucurbits. The Cla and Csa accessions indicate *Citrullus lanatus*, watermelon and *Cucumis sativus*, cucumber, respectively, and are derived from the ICUGI Cucurbit Genomics Database http://www.icugi.org. The MELO accessions are from the C. melo genome and taken from the Melonomics database https://melonomics.net. **b**) Hierarchical clustering of squalene epoxidase gene expression in *Siraitia*. Both SQE1 and SQE2 are significantly expressed in the youngest fruit. Expression RPKM data can be found in Data File S2.

Fig. S3. **Overall structure model of SgSQE1** (contig 16760, residues 60-417). The binding tunnel as calculated by CAVER is shown as green spheres and the bound FAD shown in purple ball & stick representation. The model is based on the structure of the Flavin monooxygenase, Aklavinone 12-hydroxylase RdmE (PDB entry 3IHG), that showed the widest tunnel among the Phyre2 results. More than a dozen Flavin monooxygenase structures were identified in the PDB as good templates for SE modeling with 100% confidence. However, SQE shares very low sequence similarity (at most 18% sequence identity) to any of those Flavin monooxygenases and consequently all the models show high error in the predicted coordinates. Examining the various predicted models reveals a narrow tunnel leading from the SQE surface to the bound FAD. The models differ by the opening and the volume of the tunnel. Nevertheless, most of the predicted channels are wide enough to accommodate the extended squalene conformation and none of them are wide enough to accommodate the cyclicized cucurbitane.

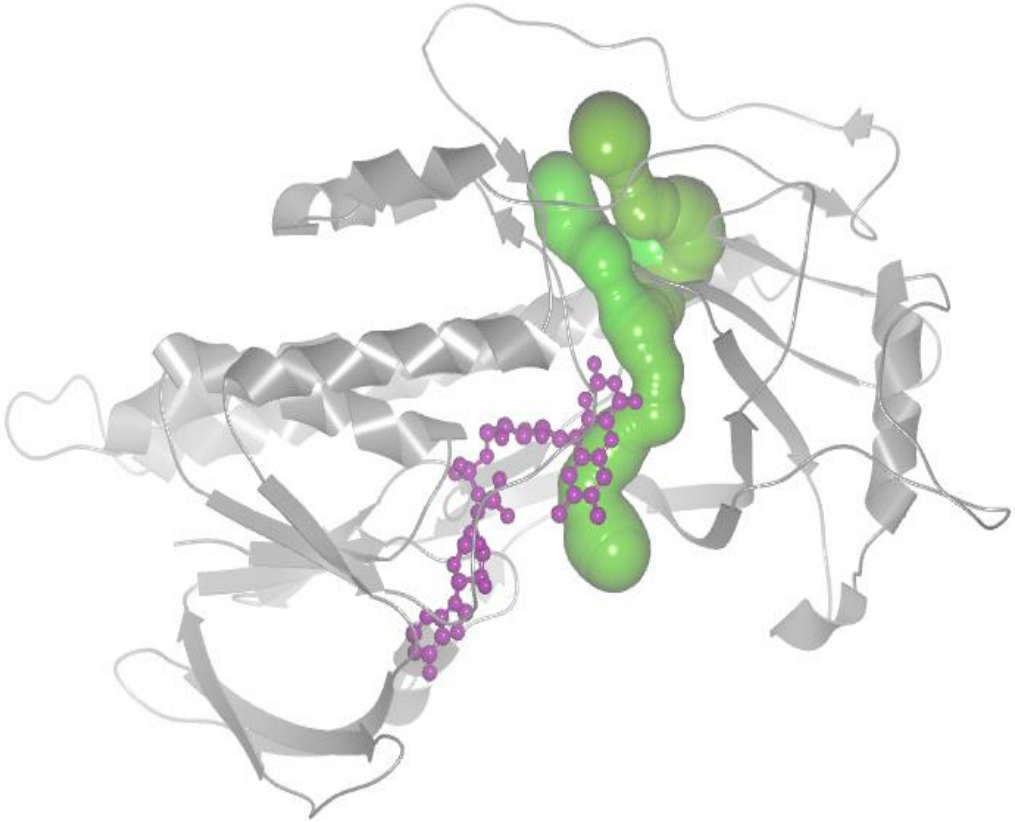
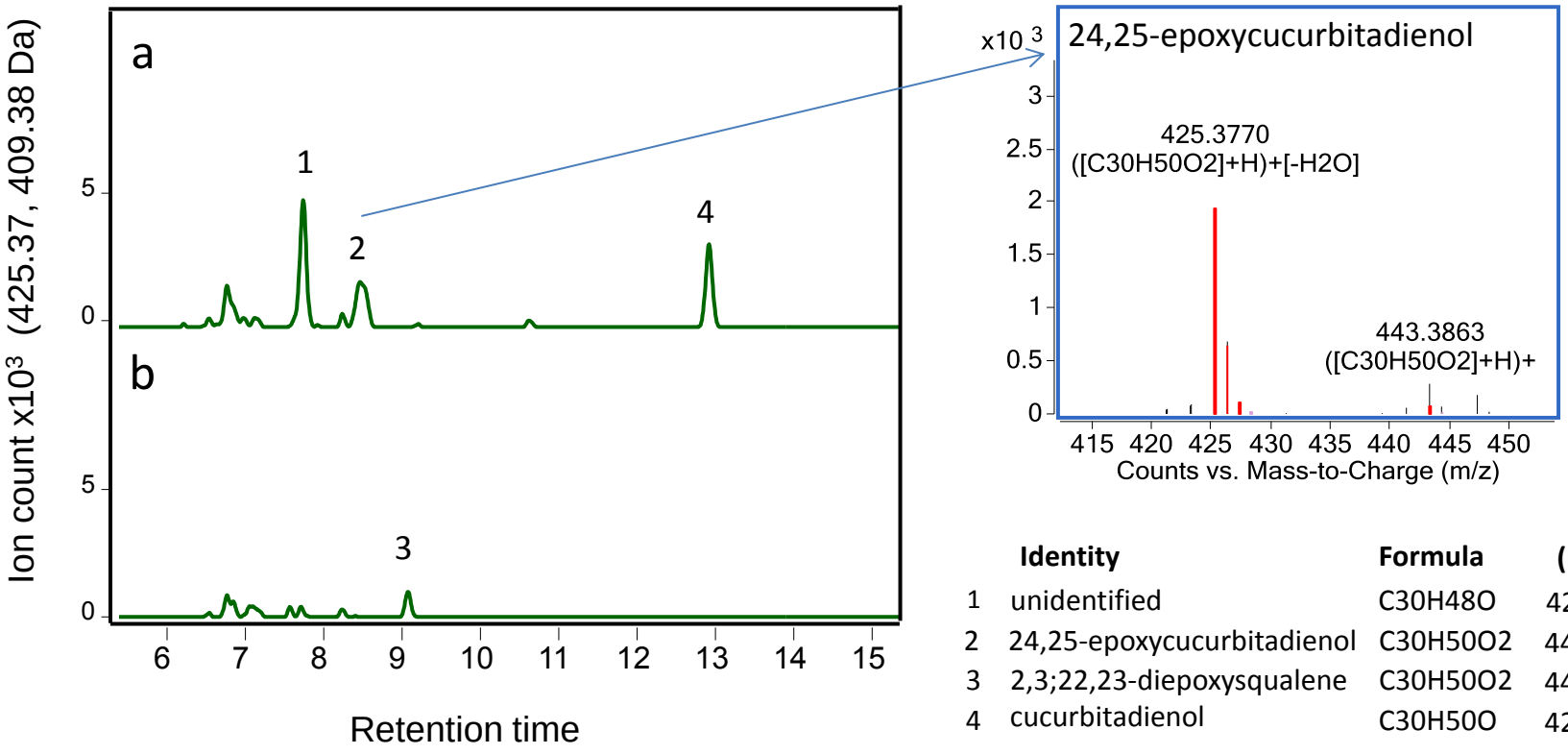


Fig. S4. **Cucurbitadienol and 23,24 epoxycucurbitadienol accumulation in transgenic tobacco plant expressing *Sg*CDS under the control of CaMV 35S promoter.** Extracted ion chromatograms of CDS activity in transgenic tobacco leaves (**a**) and wild type (wt) (**b**) 2,3;22,23-diepoxysqualene (peak #3) is accumulated in wt leaves while cucurbitadienol (peak #4) and 24,25-epoxycucurbitadienol (peak #2) are accumulated in transgenic tobacco leaves expressing SgCDS. Mass spectrum of 24,25 epoxycucurbitadienol from transgenic plants is shown. The mass spectra of other identified compounds are presented in Fig. S5.



	Identity	Formula	$(M+H)^+$
1	unidentified	C30H48O	425.3768
2	24,25-epoxycucurbitadienol	C30H50O2	443.3884
3	2,3;22,23-diepoxysqualene	C30H50O2	443.3884
4	cucurbitadienol	C30H50O	427.3934

Fig. S5. Mass spectra of compounds shown in Fig. 2a. And Fig S7.

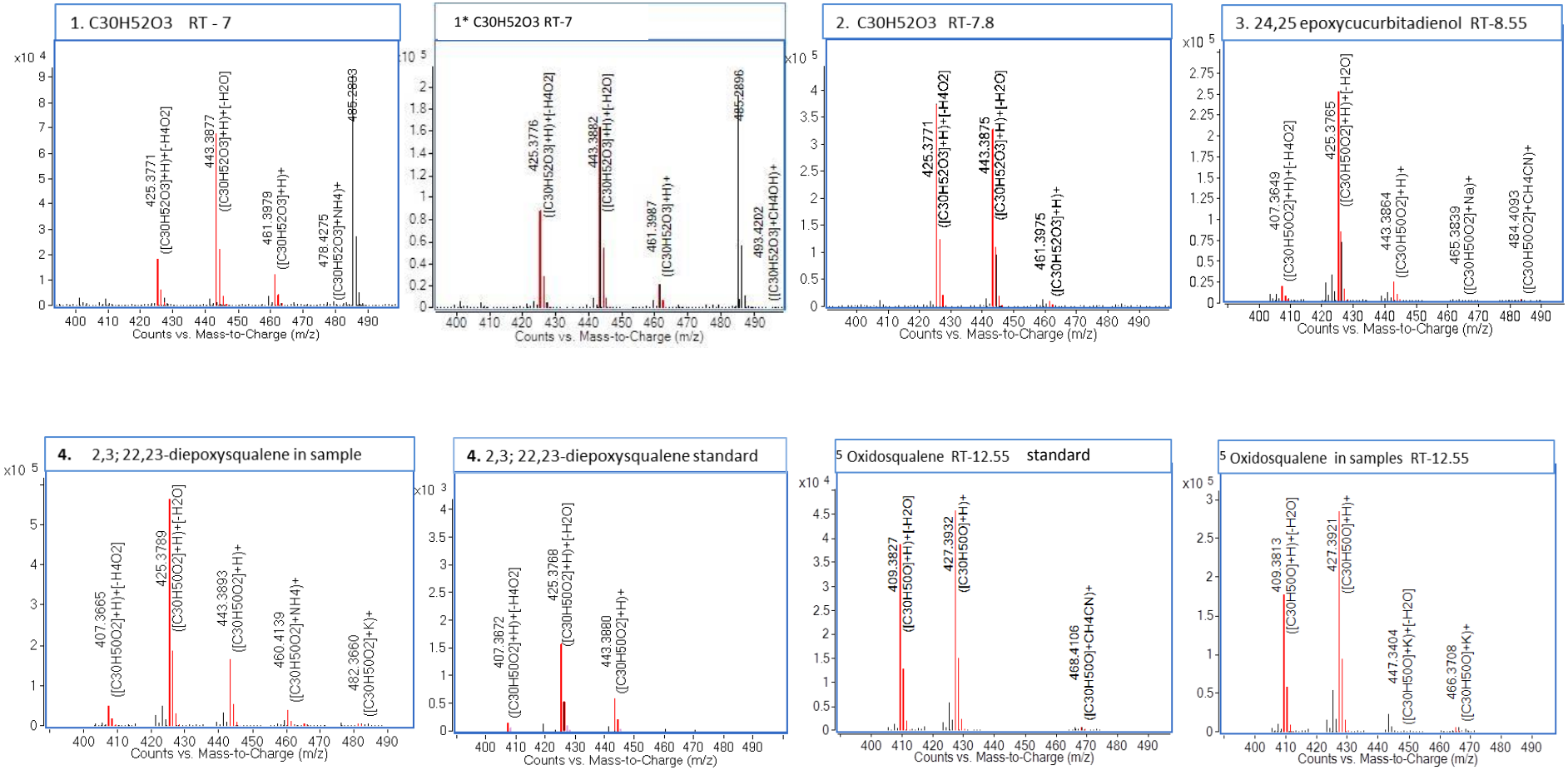
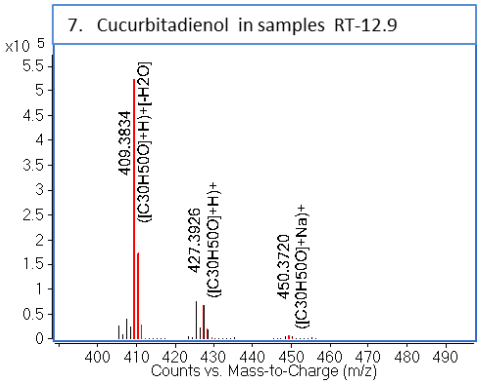
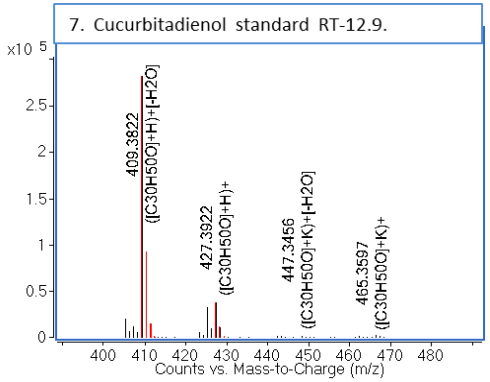
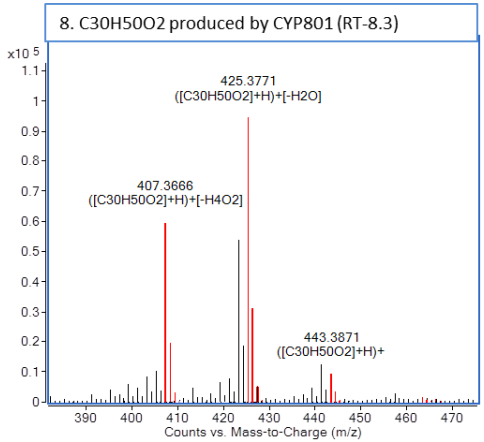
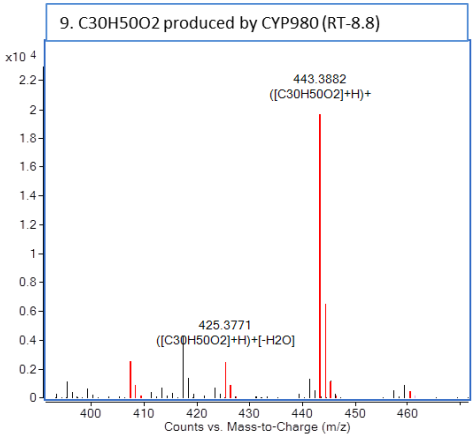


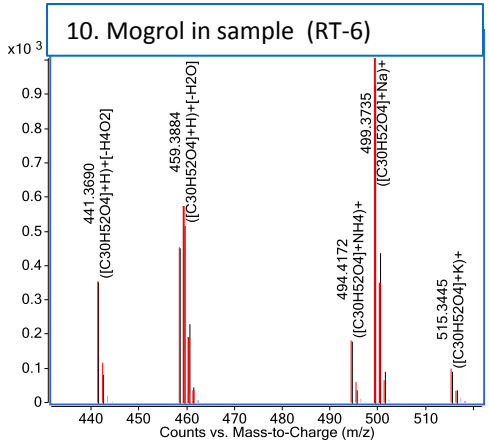
Fig. S5. Continued.

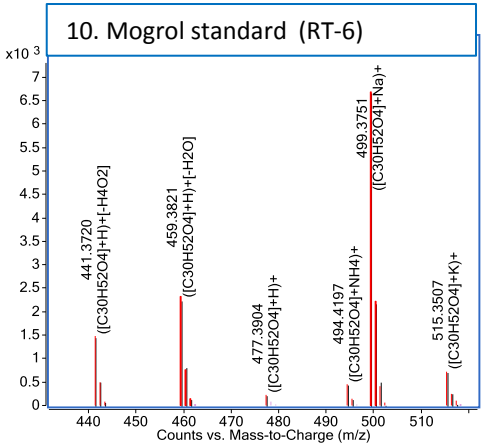












	Identity	Formula	M+H
1-1*	putative dihydroxycucurbitadienol	C30H52O3	461.3989
2	unidentified	C30H52O3	461.3989
3	24, 25 epoxycucurbitadienol	C30H50O2	443.3884
4	2,3; 22,23-diepoxysqualene	C30H50O2	443.3884
5	2,3-oxidosqualene	C30H50O	427.3934
7	cucurbitadienol	C30H50O	427.3934
8	11-hydroxycucurbitadienol	C30H50O2	443.3884
9	19-hydroxycucurbitadienol	C30H50O2	443.3884
10	mogrol	C30H52O4	477.393

Fig. S6. **Detailed docking model of CDS with 24,25 epoxy metabolite.** The anosteryl cation with epoxide moiety in position 24-25 was docked in the CDS model. Purple lines represent hydrophobic interactions and green lines, hydrogen bonding. While the binding pocket is indeed very hydrophobic it accommodates very well the epoxide. The addition of one polar atom (the epoxy oxygen) doesn't effect the binding, likely due to the large amount of interactions and some polarity from nearby main chain atoms.

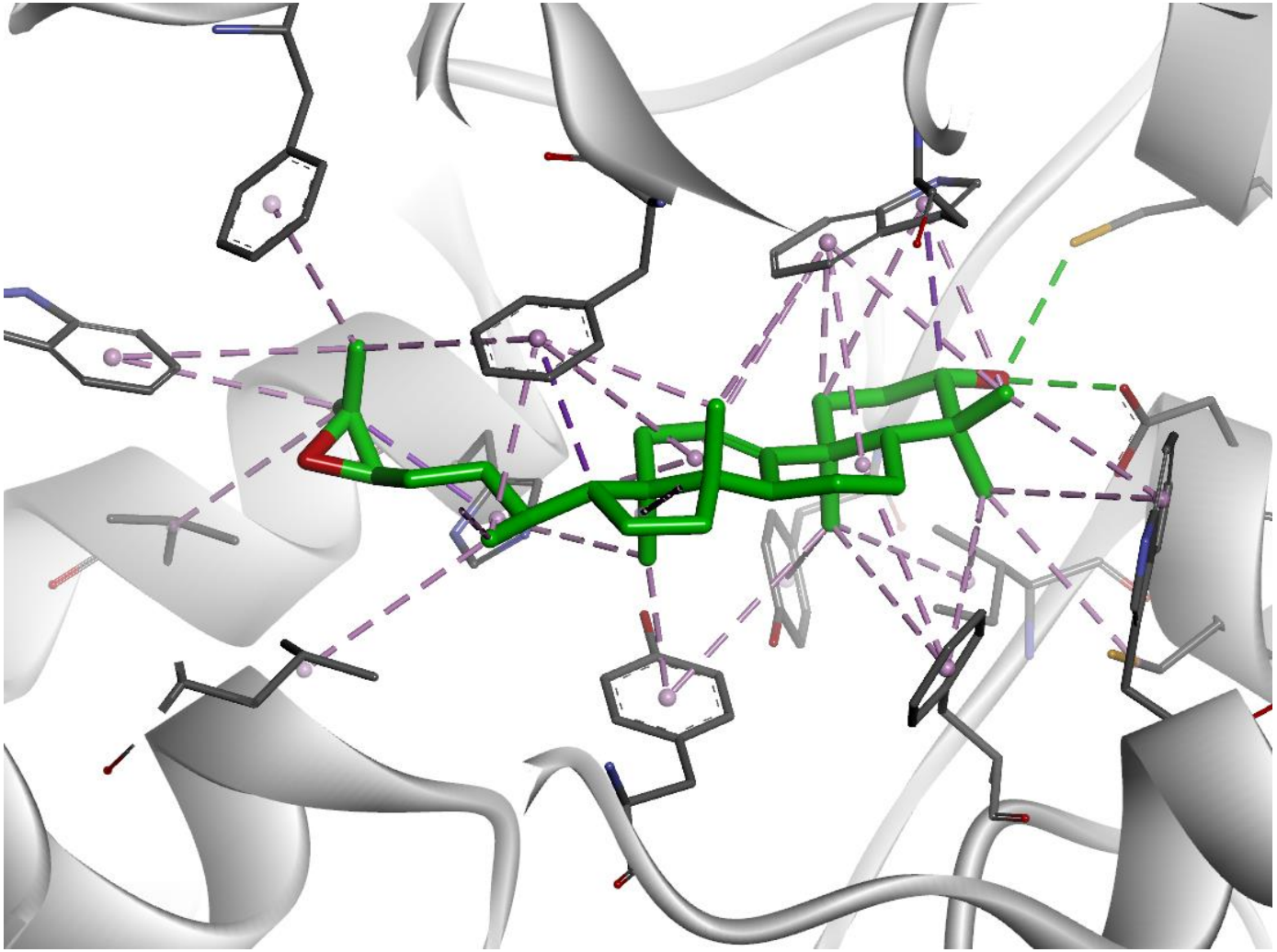
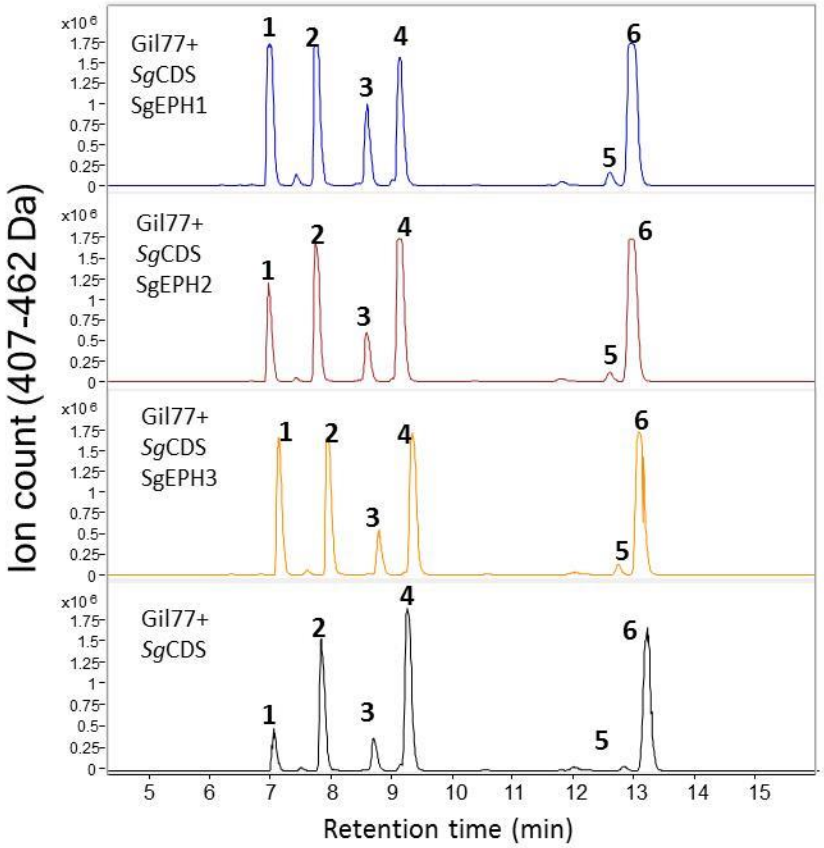


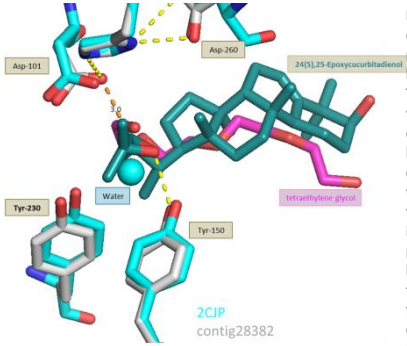
Fig. S7. LC-MS analysis of extracts of yeast coexpressing SgCDS with EPH 1-3. The extracted ion chromatogram of ions m/z=407-444 represents relevant triterpenoid compounds and derivatives accumulated in the yeast. Yeast coexpressing SgCDS with EPH1-3 are represented in the three upper panels and a chromatogram from yeast harboring SgCDS as negative control is presented in the bottom panel. MS spectra and identifications are presented in Fig. S5 .



	Identity	Formula	$(M+H)^+$
1	24, 25 dihydroxycucurbitadienol	C30H52O3	461.3989
2	unidentified	C30H52O3	461.3989
3	24, 25 epoxycucurbitadienol	C30H50O2	443.3884
4	2,3; 22,23-diepoxysqualene	C30H50O2	443.3884
5	2,3-oxidosqualene	C30H50O	427.3934
6	cucurbitadienol	C30H50O	427.3934

Fig. S8. Docking models of SgEPH1-4 and descriptions of their docking characteristics.

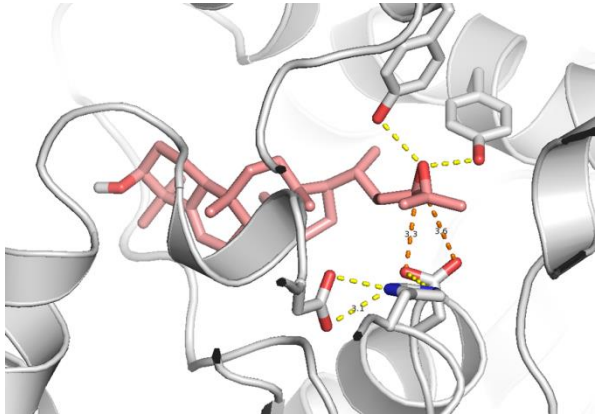
Docking 24(S),25-Epoxycucurbitadienol – Contig28382 Comparison to the Potato EH (2CJP)



In the crystallographic structure of the potato EH, a water molecule was found hydrogen bonded to the two lid tyrosines. This position might indicates the expected position of the oxygen in the epoxide ring of bound substrate. The modeled epoxide oxygen is very close to that position. In addition, tetraethylene glycol was found in the binding pocket of the potato EH (colored hydrophobicnk). HBording to the docking calculation that is the location of the cucurbitadienol. Together this two location overlaps support the docking HBuracy.

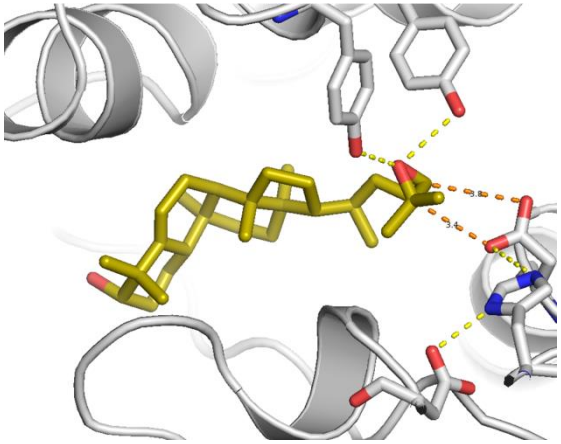
Docking 24(R),25-Epoxycucurbitadienol – Contig28382

Docking the R enantiomer is looking slightly better as the epoxide oxygen found just between the two tyrosines while the nucleophile Asp-101 is in close proximity to both C24 and C25 positions. (Just perfect match). In the next calculations I used both R and S enantiomers, but I will show here only the R enantiomer since it showed in all calculations better match.



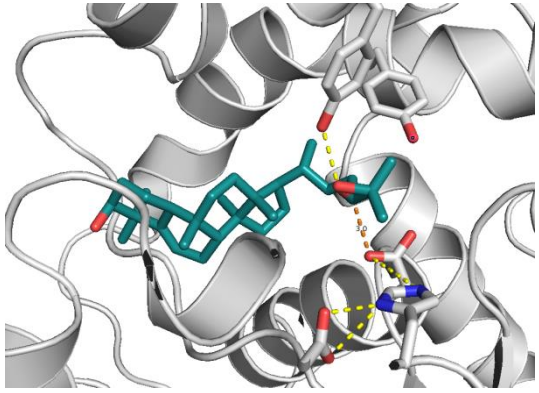
Docking 24(R),25-Epoxycucurbitadienol – Contig73966

The epoxide oxygen found just between the two tyrosines, creating hydrogen bonds to the hydroxyl group, while the nucleophile Asp-101 is in close proximity to both C24 and C25 positions.



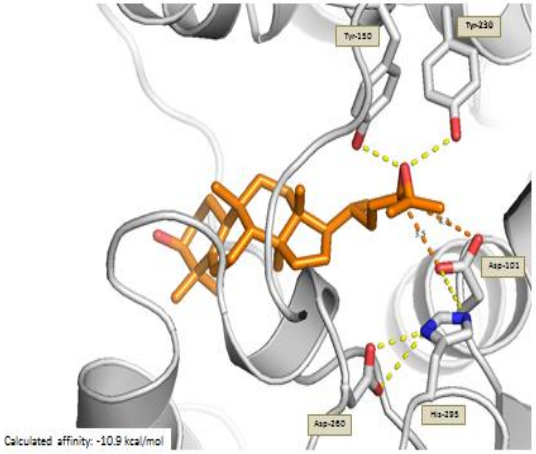
Docking 24(S),25-Epoxycucurbitadienol – Contig28382

Docking of epoxycucurbitadienol was carried out using AutoDock Vina on contig28382 constructed homology model. A low energy model is presented here. This model fit to the known catalytic mechanism of EHs. Asp101 is the catalytic nucleophile, and His295/Asp260 comprise a general base-charge relay pair. The Distance between Asp101 and the substrate C-24 is 3.0Å. Two tyrosine residues from the lid (Tyr150 and Tyr230) are expected to assist ring opening by hydrogen bonding to the oxygen of the substrate's epoxide ring. In the model only Tyr150 creates hydrogen bond to the epoxide.



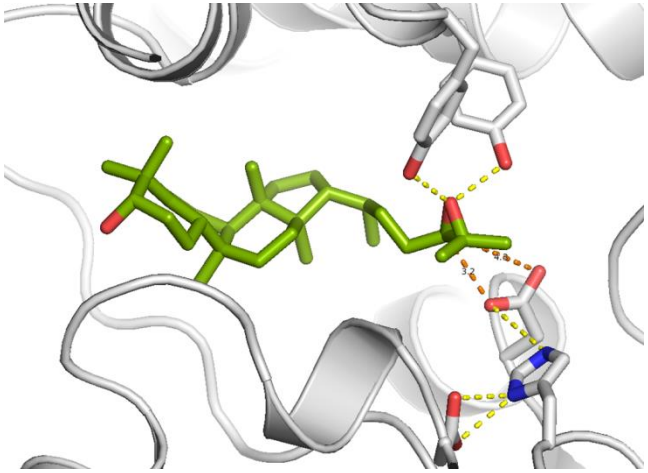
Docking 24(R),25-Epoxycucurbitadienol – Contig86123

The epoxide oxygen found just between the two tyrosines, creating hydrogen bonds to the hydroxyl group, while the nucleophile Asp-101 is in close proximity mainly to both C24 position.



Docking 24(R),25-Epoxycucurbitadienol – Contig102640

The epoxide oxygen found just between the two tyrosines, creating hydrogen bonds to the hydroxyl group, while the nucleophile Asp-101 is in close proximity mainly to C-24 position.



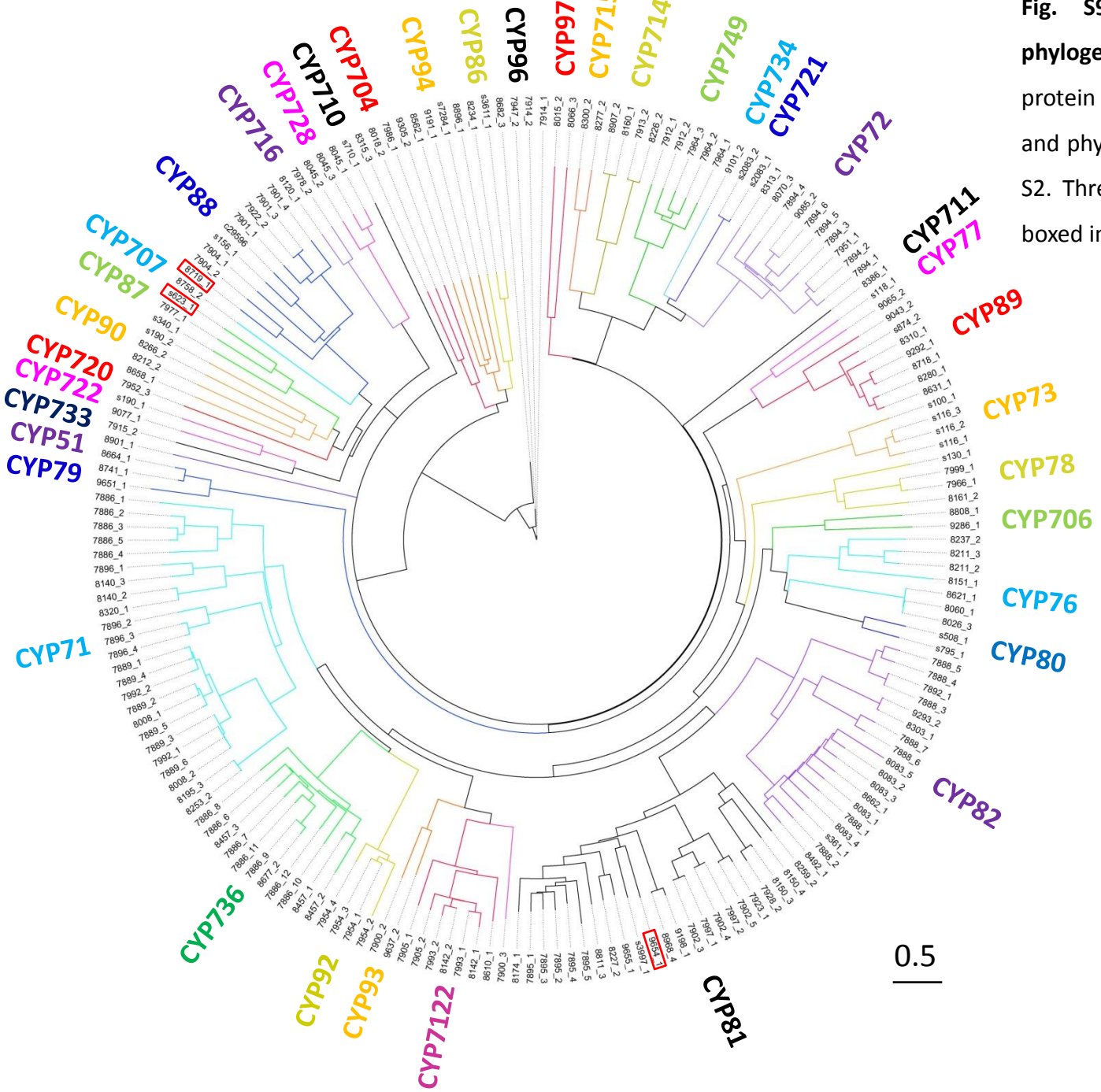


Fig. S9. Expandable version of the phylogenetic tree of *Siraitia* CYP450s. The protein sequences used for the alignment and phylogenetic tree are listed in Data File S2. Three CYPs referred to in the text are boxed in red.

Fig. S10. Expandable version of the hierarchical tree and expression heat map of the expressed Siraitia CYP450s. CYP numbers are according to the CYP scaffolds listed, with RPKM data, in Data File S2. CYP numbers numbered xxxx.x refer to the CYP scaffolds and the last number refers to the number of tandem CYPs in that scaffold. CYPs beginning with the letter S refer to genomic scaffolds and if followed by a decimal point and number refers to the number of tandem CYPs in that genomic scaffold. The two CYPs with activity toward cucurbitadienol are marked with an asterisk *. S623.1 indicates the C11 hydroxylating enzyme.

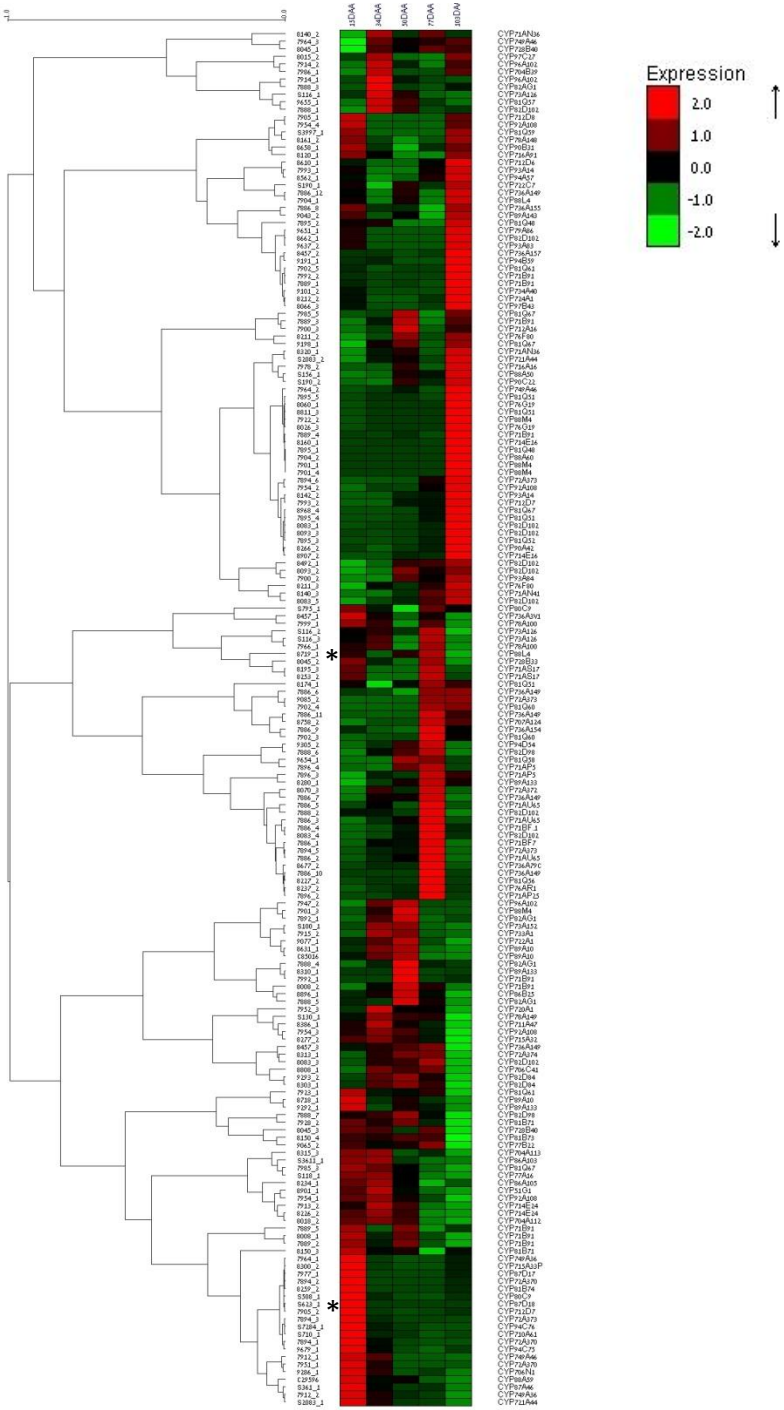
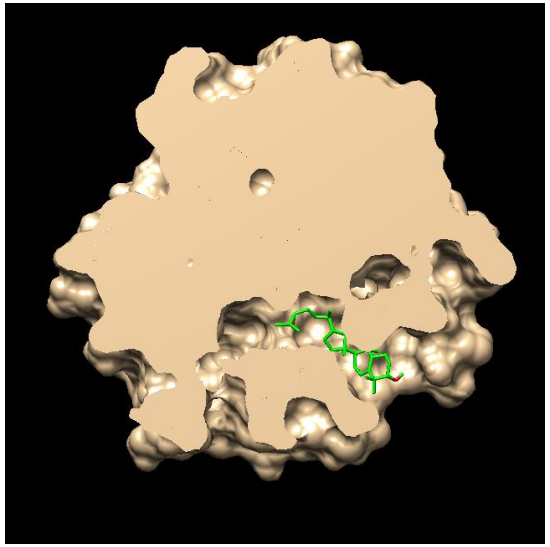
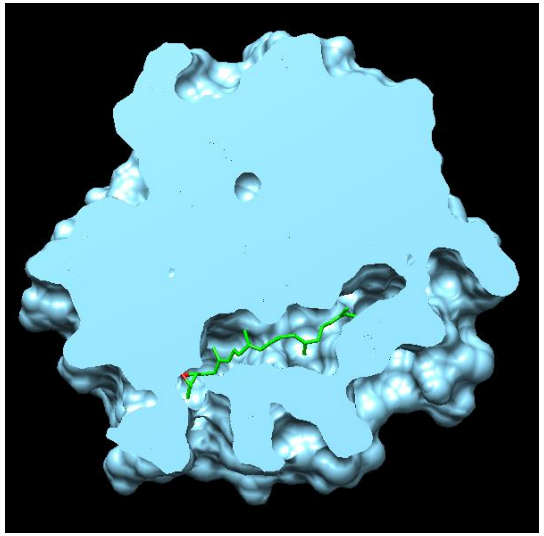


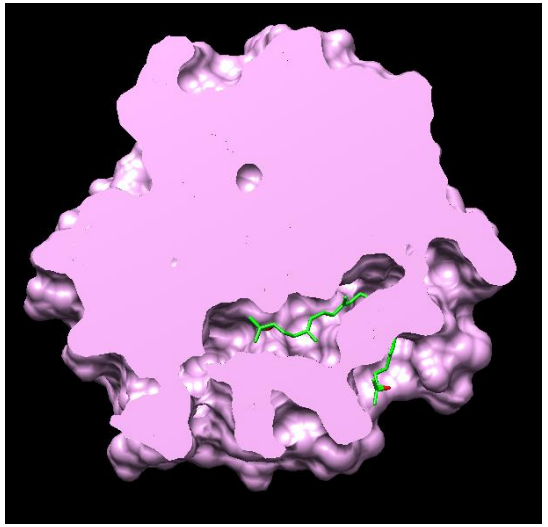
Fig. S11. Preference of EPH for epoxycucurbitadienol substrate. Docking epoxycucurbitadienol and diepoxysqualene on the constructed homology model of EPH (shown in Fig. S8). The results indicate that the reaction with epoxycucurbitadienol is preferred over that with epoxysqualene.



Epoxycucurbitadienol
Calculated affinity -11.8 kcal/mol

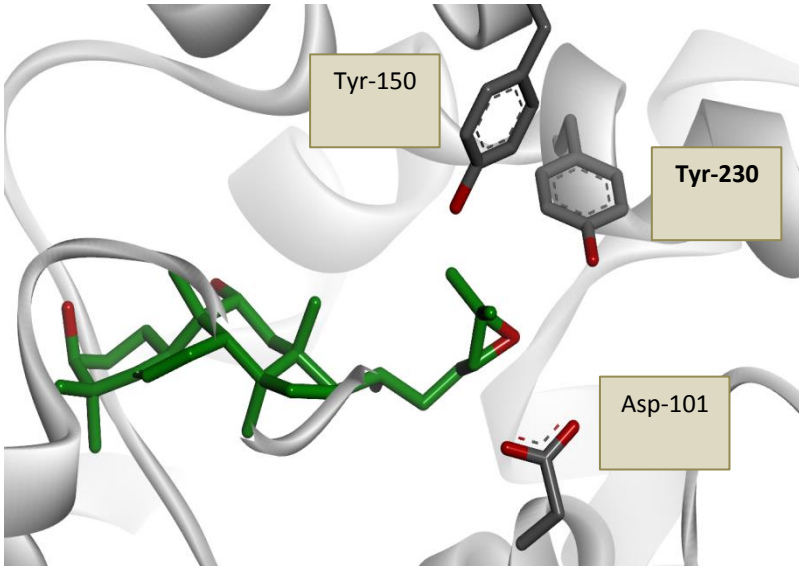


Diepoxysqualene
Calculated affinity: -11.2 kcal/mol
(the epoxide is not in the catalytic site)



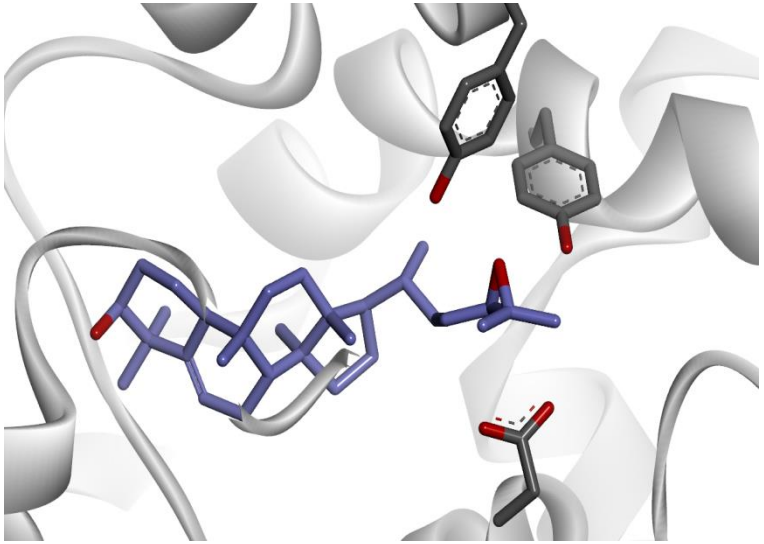
Diepoxysqualene
Calculated affinity: -9.9 kcal/mol
(the epoxide is in the catalytic site)

Fig. S12. Preference of EPH for epoxycucurbitadienol substrate compared to 11-OH epoxycucurbitadienol. Although the epoxycucurbitanedienol with hydroxyl on C11 (R or S configuration) might bind to EPH, all favorably predicted binding modes are not productive as the epoxide is not in the catalytic position. This was due to the highly hydrophobic nature of the pocket.



Hydroxyl on C11 (R configuration):
This conformation cannot lead to hydrolysis of the epoxide by nucleophilic attackof Asp-101

(calculated affinity: -10.8 kcal/mol)



No Hydroxyl on C11:
This conformation can lead to hydrolysis of the epoxide

(calculated affinity: -11.5 kcal/mol)

Fig. S13. Expandable version of the phylogenetic tree of SgUGTs, including functionally identified triterpenoid UGTs from other plants (listed in Table S9). The UGTs identified in this report are boxed in red. The UGT85 family is listed as UGT720 in light of the recent reclassification of the UGT85 family kindly performed by Prof. Michael Court (on behalf of the UGT nomenclature committee).

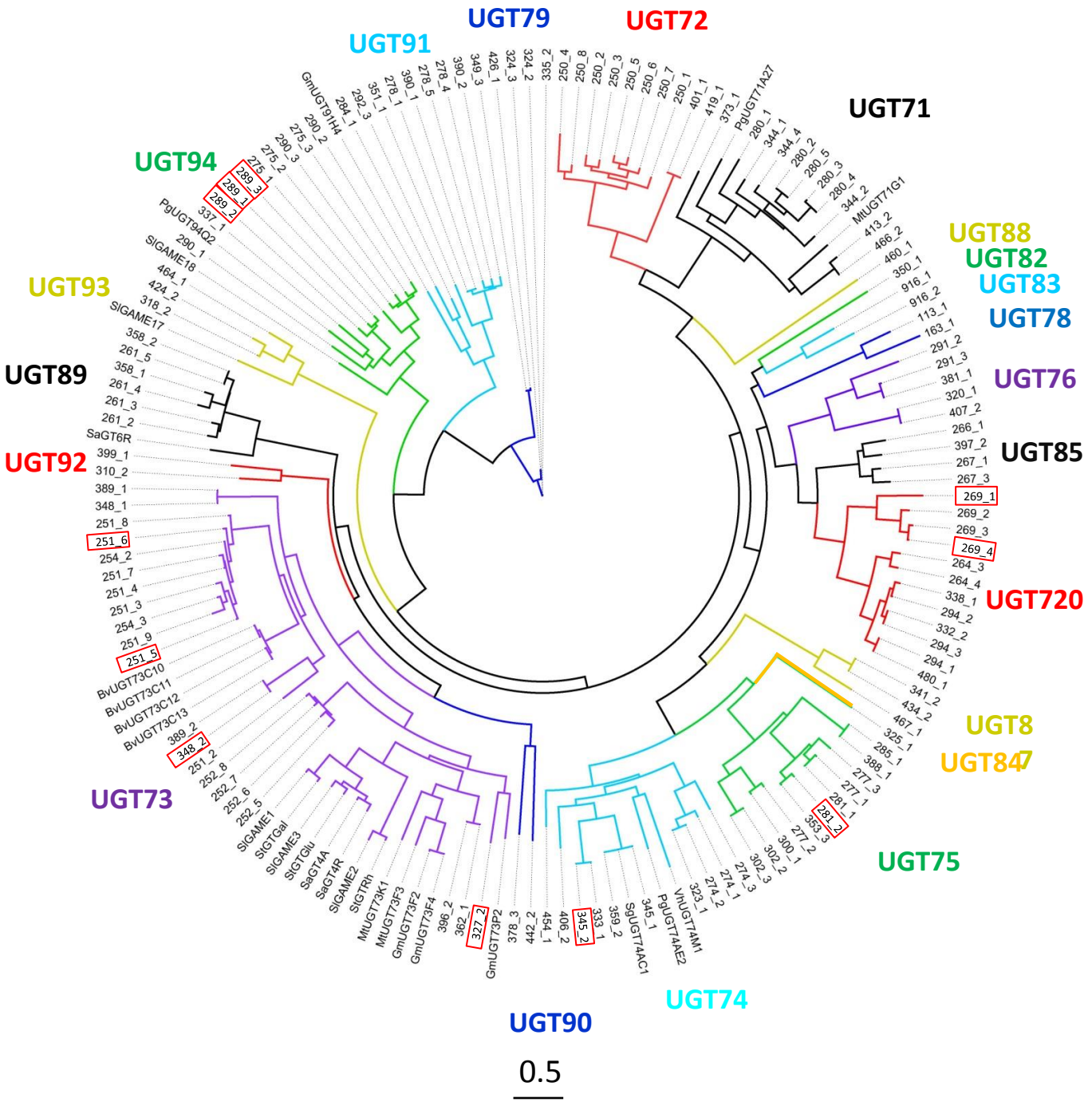


Fig. S14. Expandable version of the hierarchical tree and expression heat map of the *Siraitia* UGTs in developing *Siraitia* fruit. UGT numbers are according to the UGT scaffolds listed in Data File S2. Enzymes identified in this paper are marked by stars. The UGT720 (UGT85) genes 269.1 and 269.4 are highly expressed in the young fruit while the UGT94 genes are highly expressed in the mature fruit. RPKM data can be found in Data File S2.

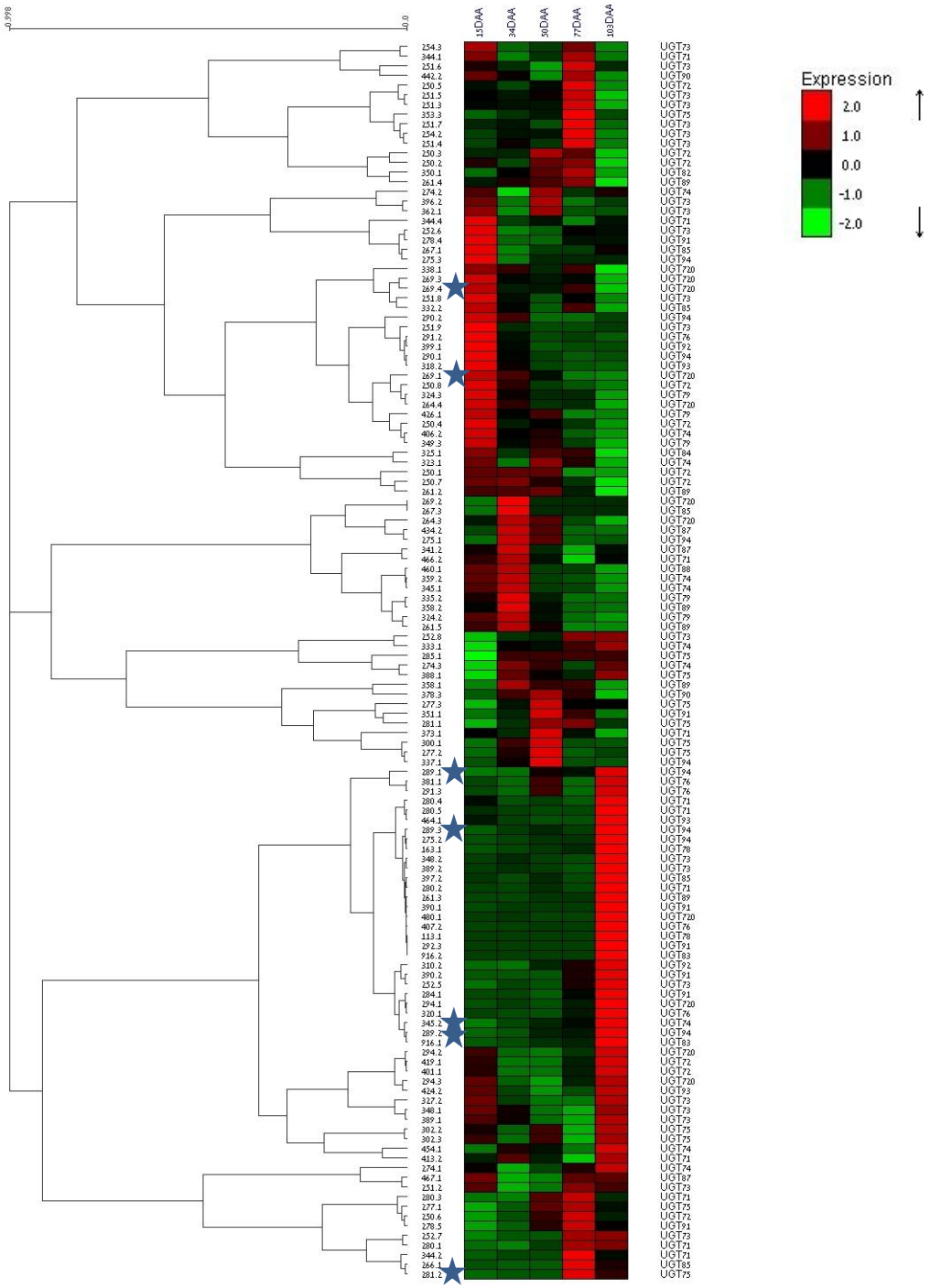


Fig. S15. Reactions, chromatograms and mass spectra of primary C3 and C24 glucosylations schematically presented in Figure 5. EIC (top window) and MS Spectrum (bottom window) results of reaction mixes with active enzymes are shown. Chromatogram window legends: indicated enzyme + substrate and the arrow points to the product shown in the chromatogram. Structures and full names of substrates and products are listed in Table S1. Enzymes participating in reactions are presented in Fig. 1.

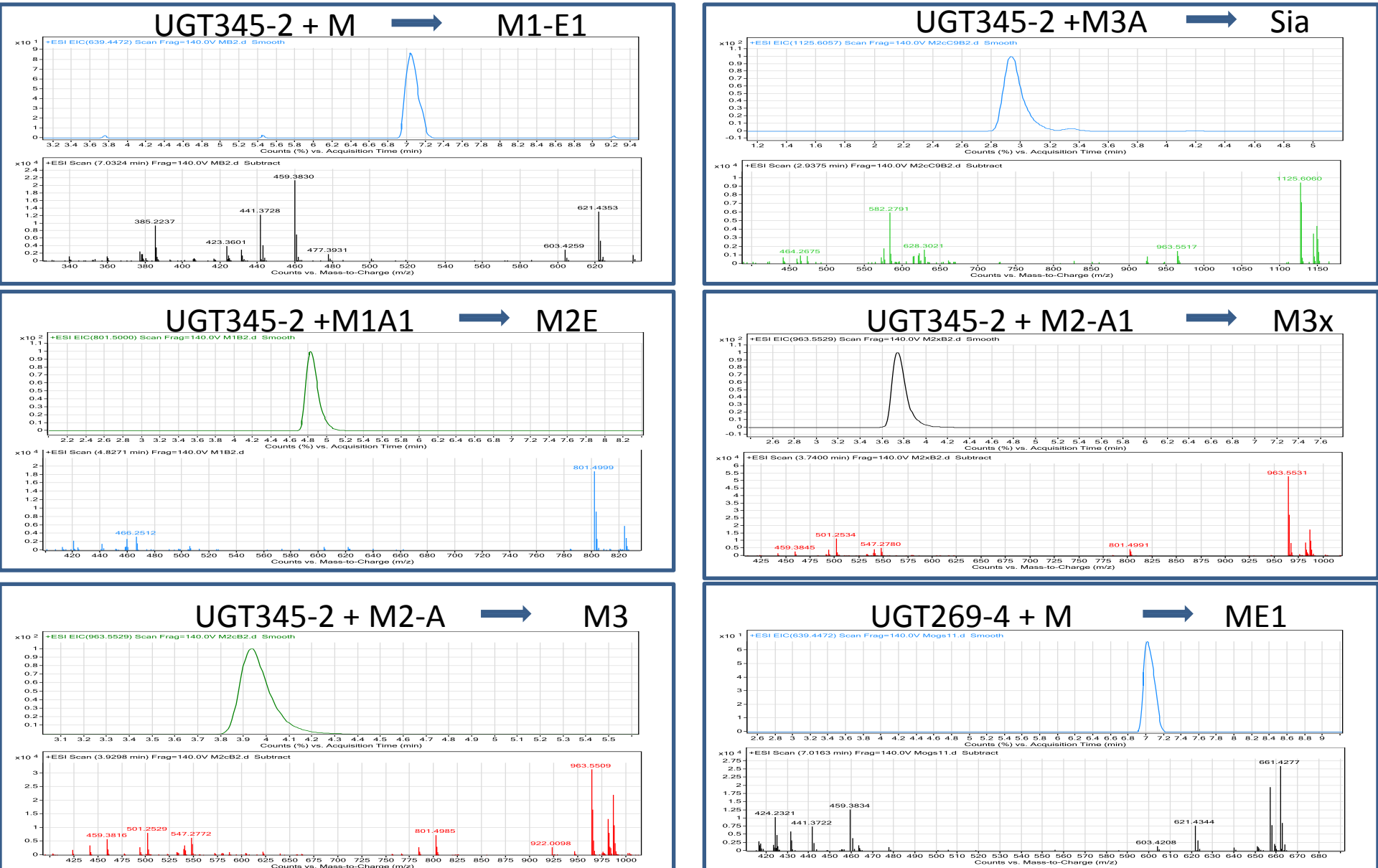
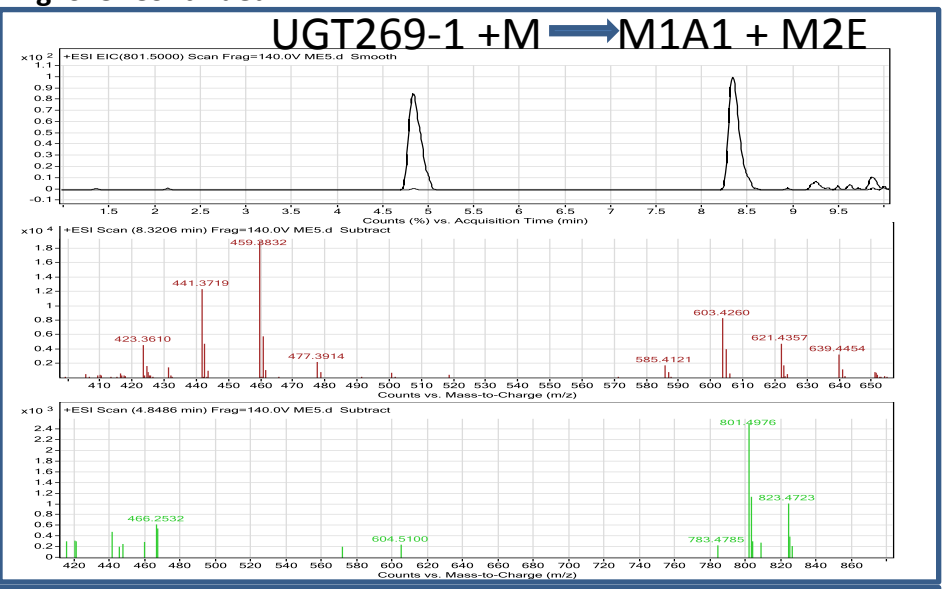
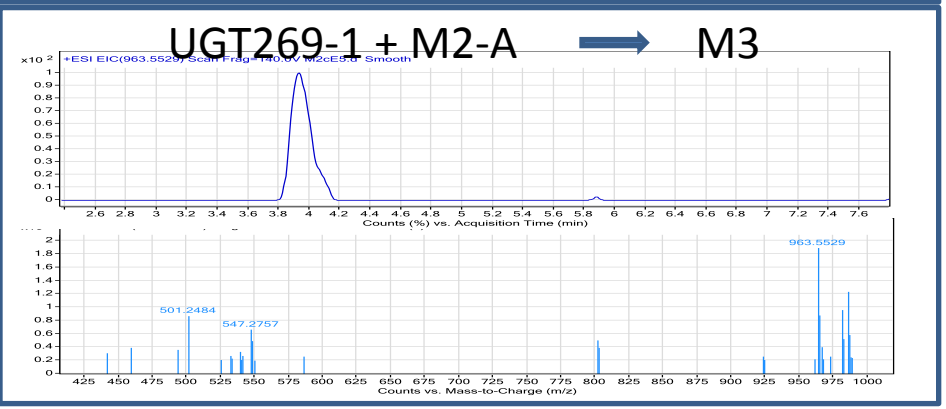
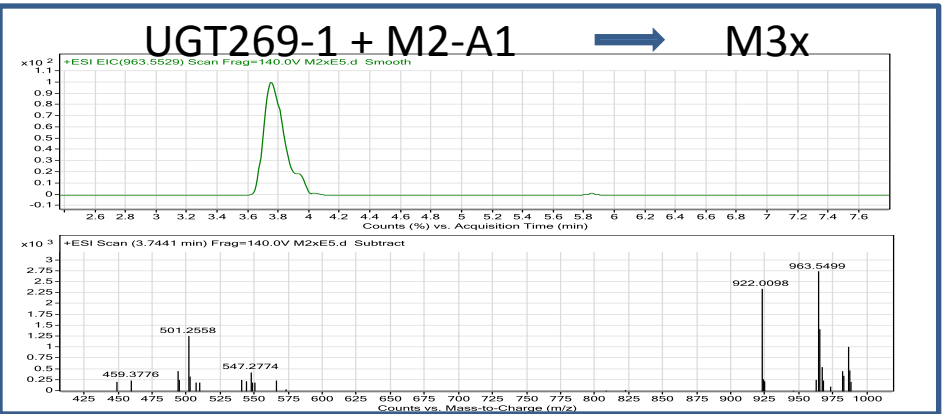
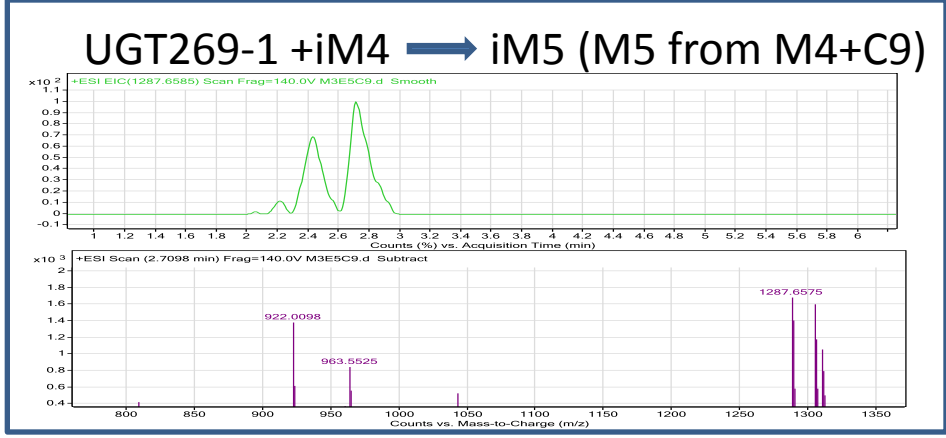


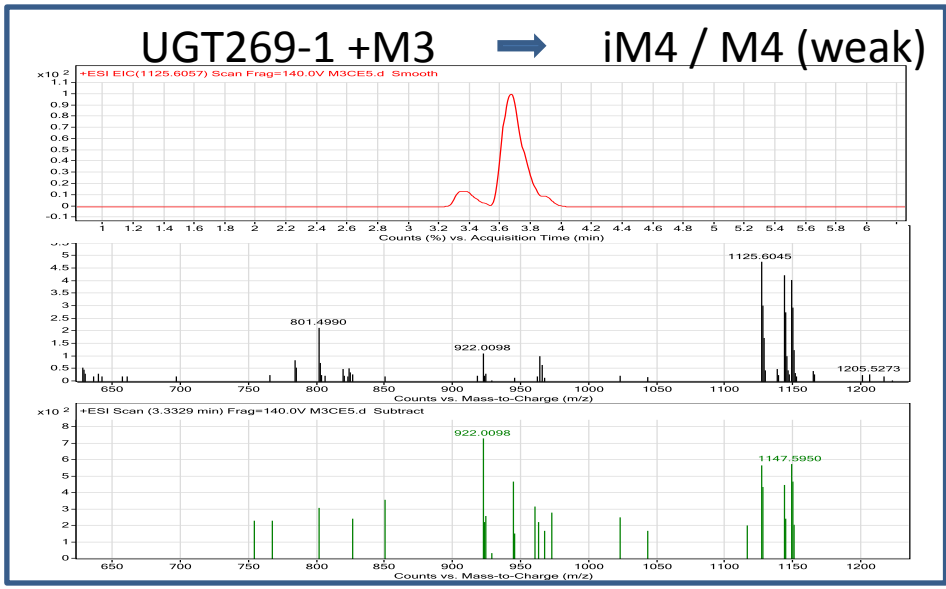
Fig. S15. Continued.











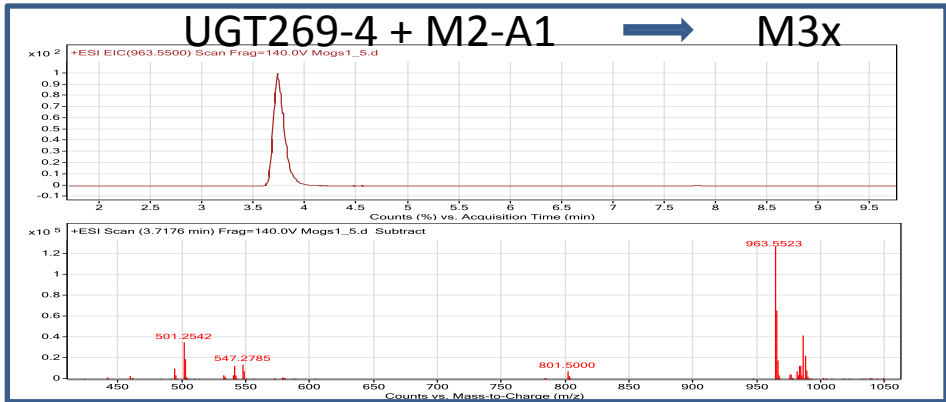
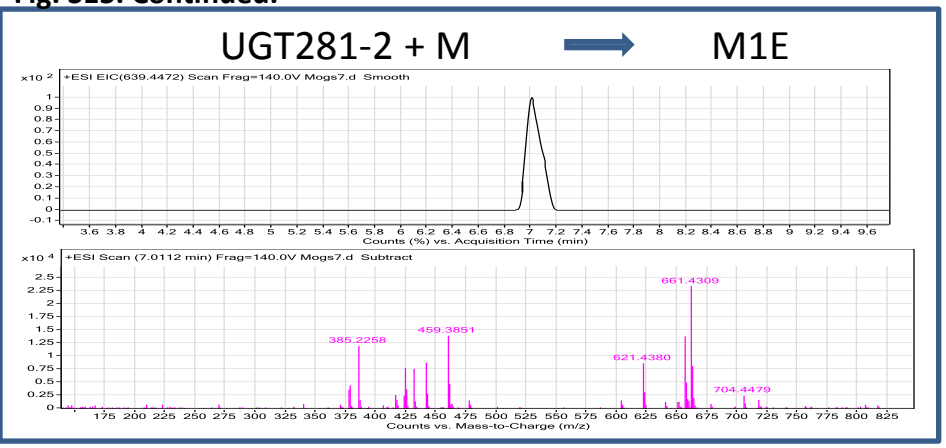
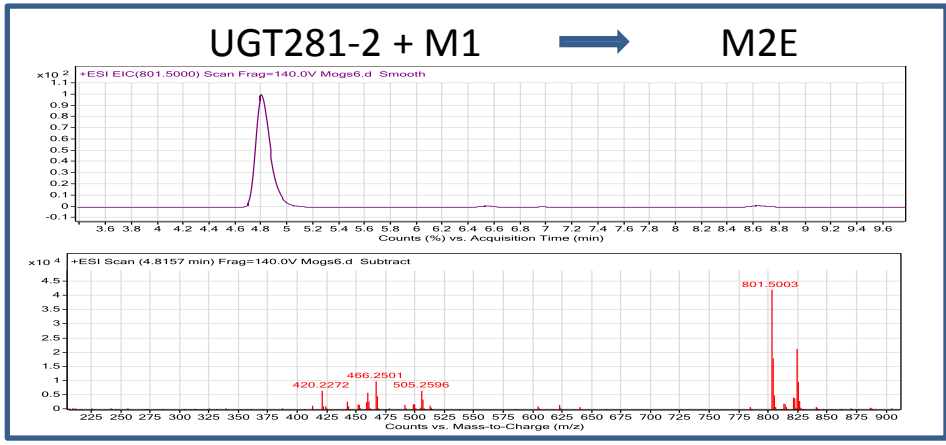
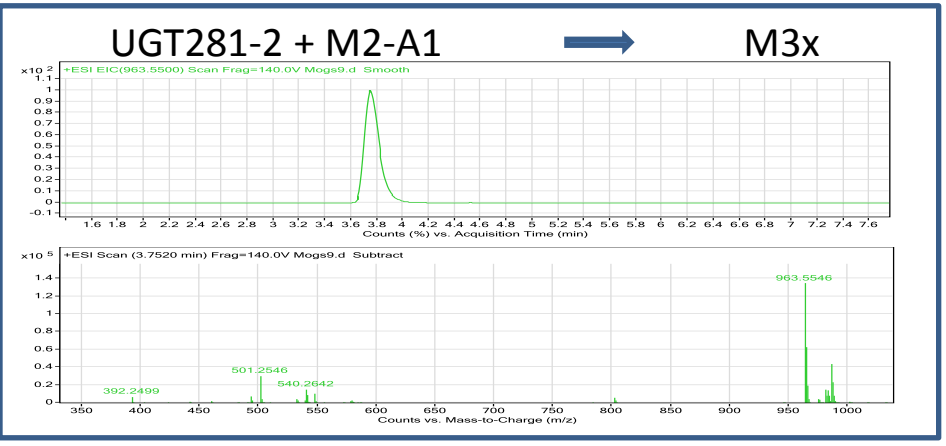
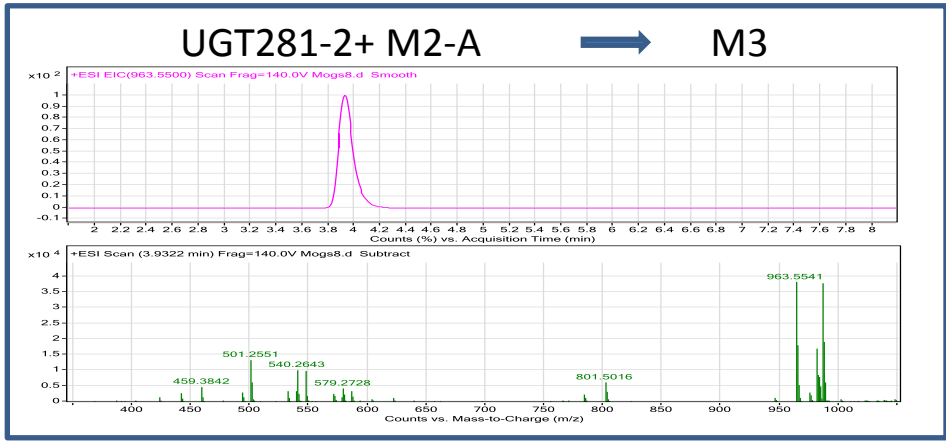


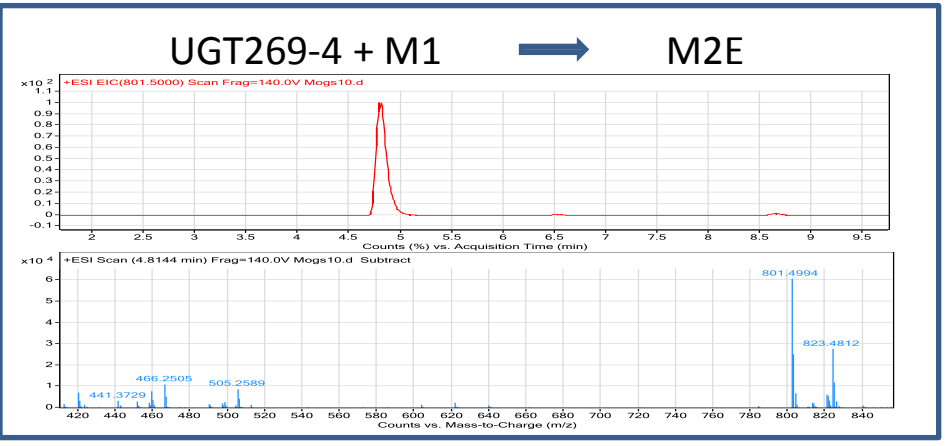
Fig. S15. Continued.











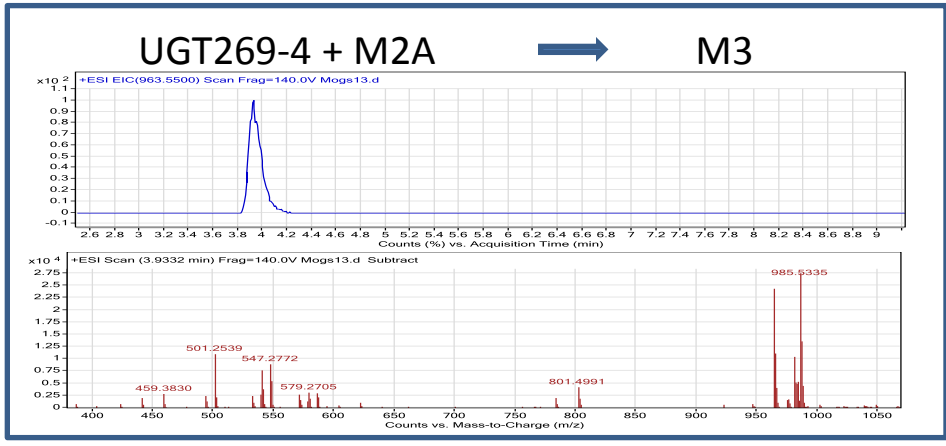


Fig. S16. Reaction, chromatogram and mass spectrum of primary family 73 glucosylation at C25. EIC (top window) and MS Spectrum (bottom window) results of reaction mix with active enzyme UGT73D5 is shown. UGT73D5 is listed in Data file S2 as s63. The C25 position of glucose was confirmed by NMR, presented in Table S8.

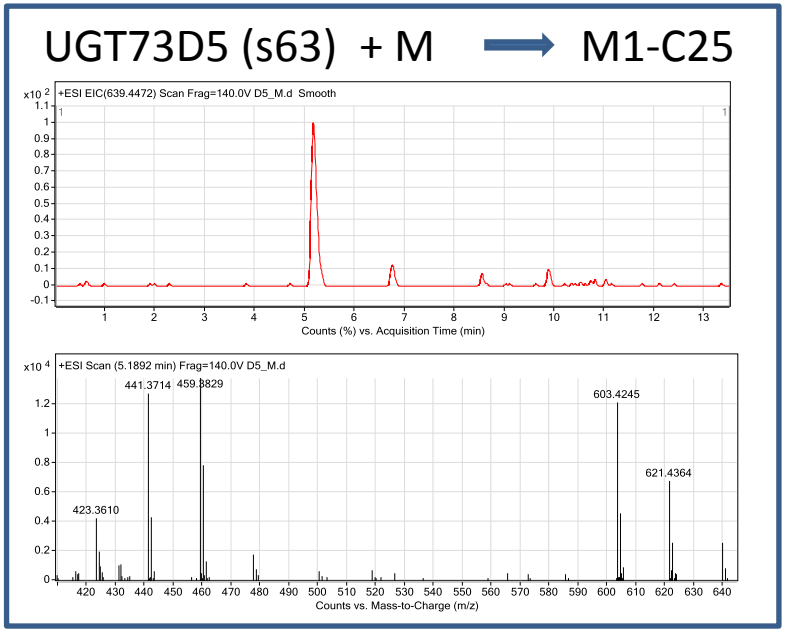


Fig. S17. Models that may explain the regiospecific glycosylations at C24 and C25. In (**a, b**) Mogrol fits into UGT73D5 and UG73-251-6 with only position 25 in close proximity to the catalytic histidine. **c**) Mogrol fits very nicely into UGT720-269-1 with positions 24 and 25 in close proximity both to the catalytic histidine as well as to the UDP-glucose. According to the docking results there is slight preference toward position 24 (shorter distance). However this is almost insignificant due to the expected errors in the model. **d**) The main reason for the difference is the polarity near the catalytic histidine. While the loop that includes the catalytic histine is more hydrophobic in UGT720-269-1, it is polar in UGT73-251-6 and UGT73D5 (s63). As such the hydrophobic rings of mogrol are tilting toward the loop in UGT720-269-1 and tilting away from the loop in UGT73-251-6 and UGT73D5.

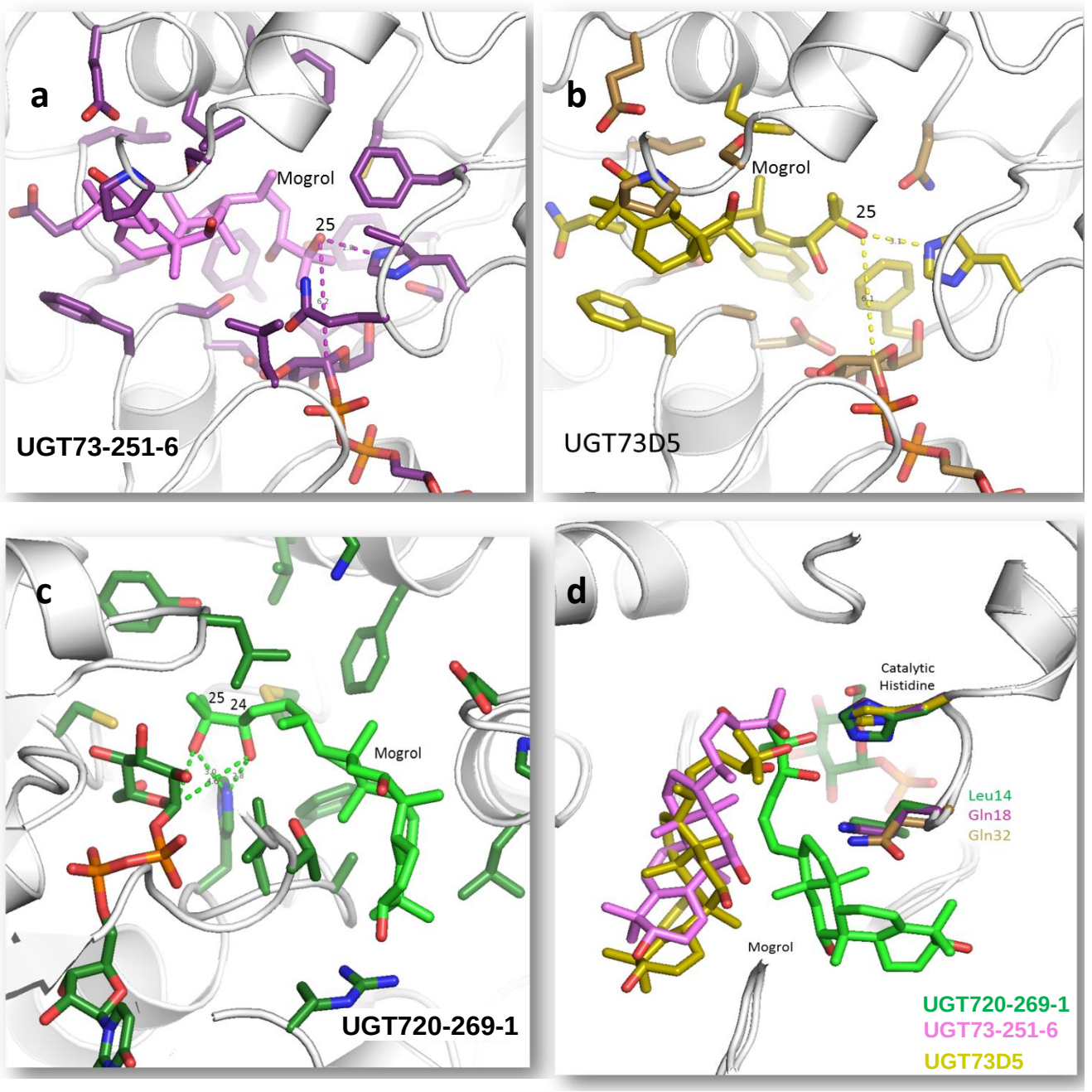
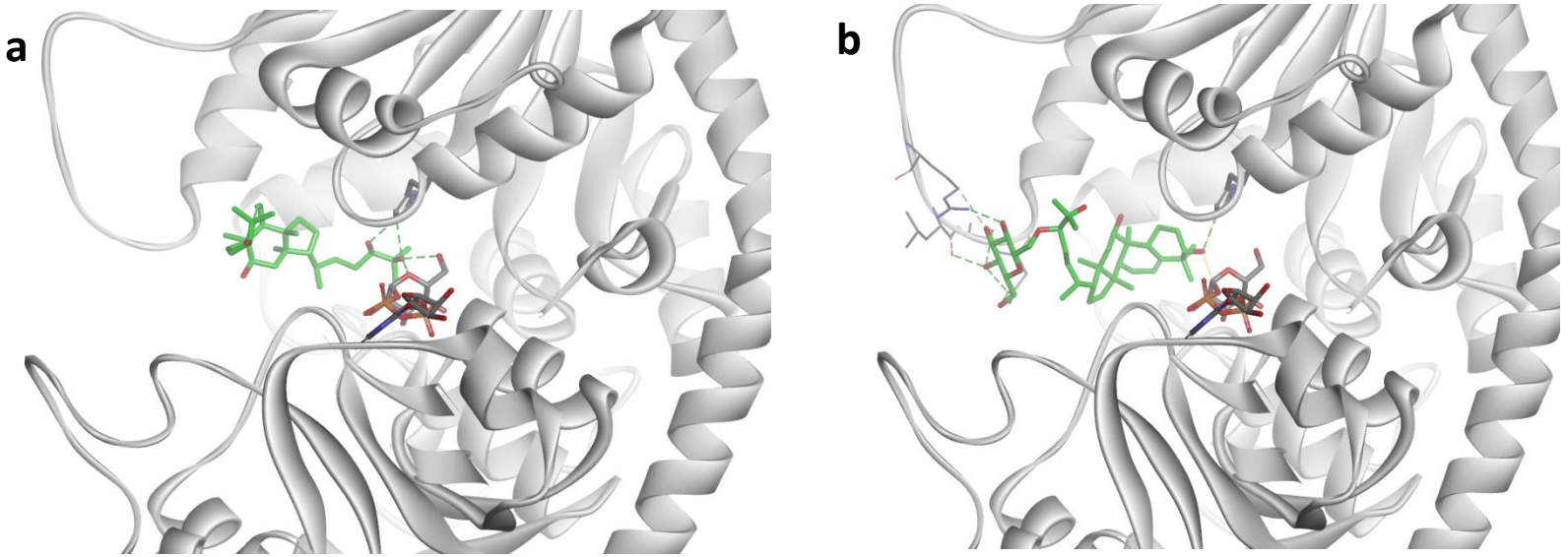
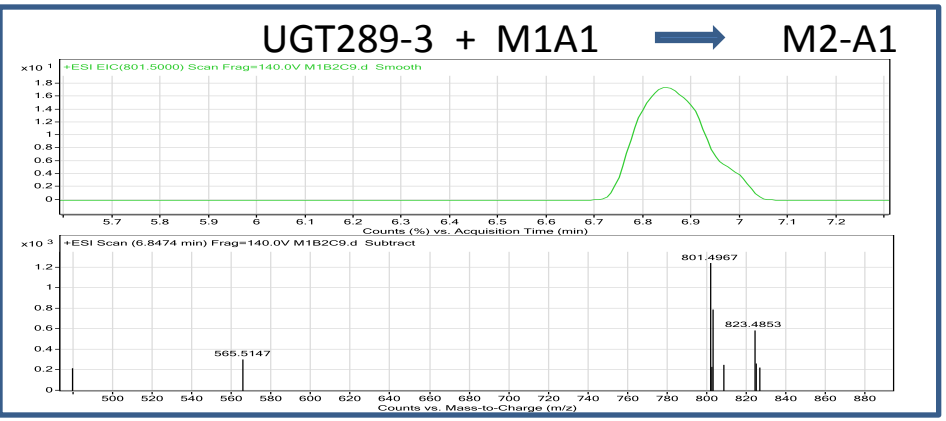
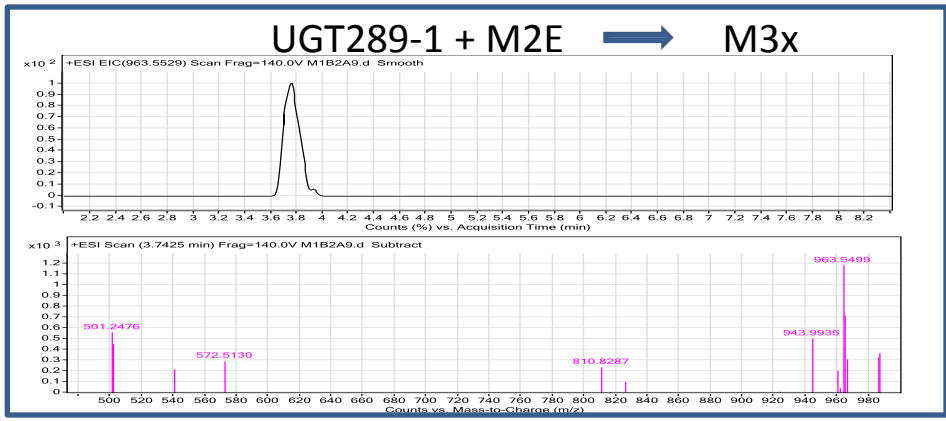
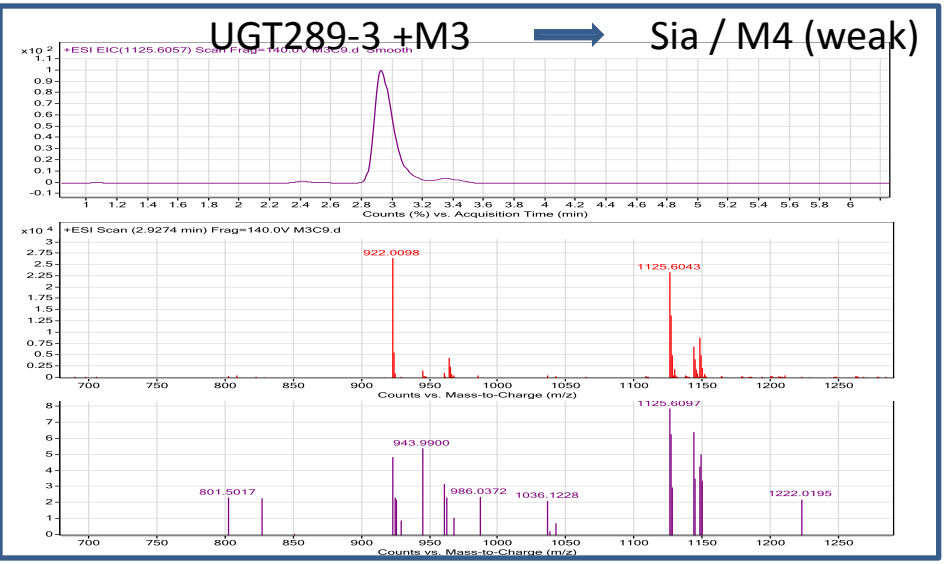


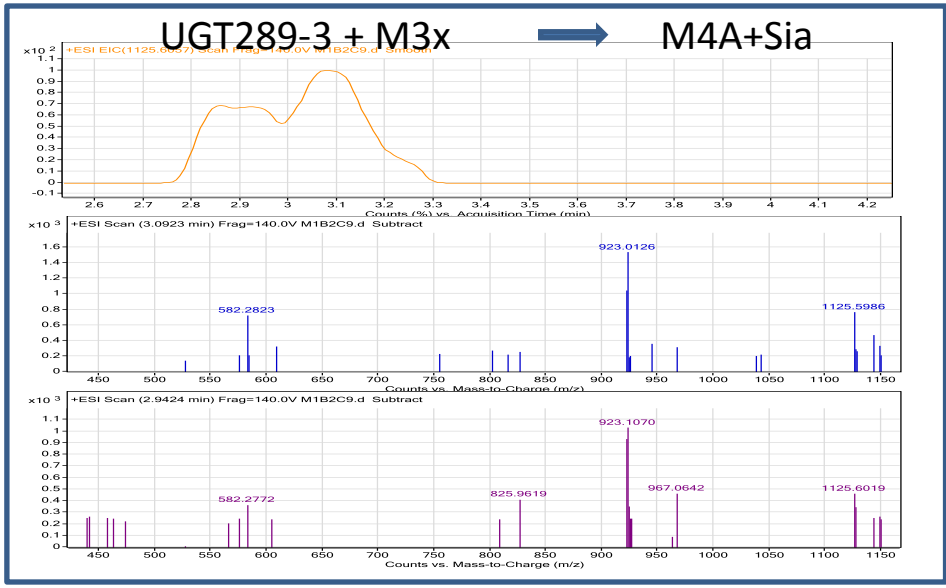
Fig. S18. Protein modelling of UGT720-269-1 showing second glucosylation at C3. **a**) Modelling of glycosylation at C24; **b**) Modelling of second primary glycosylation at C3. UGT720-269-1 structural model reveals highly hydrophobic binding pocket. According to the docking simulation, mogrol preferably binds with C24 in the catalytic site, while C24 glucosylated mogroside will clearly bind with C3 in the catalytic site, maintained by interactions of the C24 glucosyl moiety with amino acids in the loop.











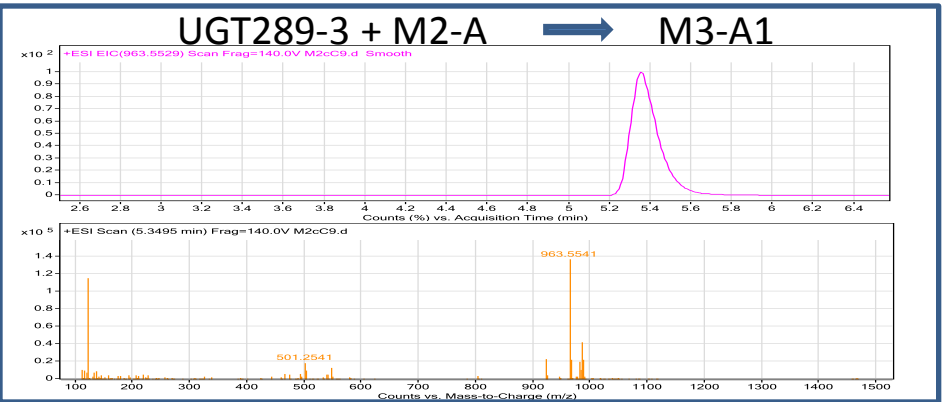
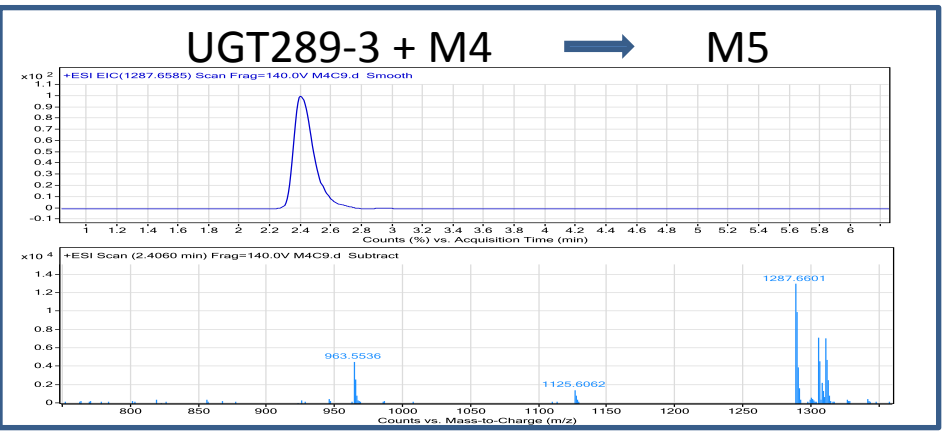
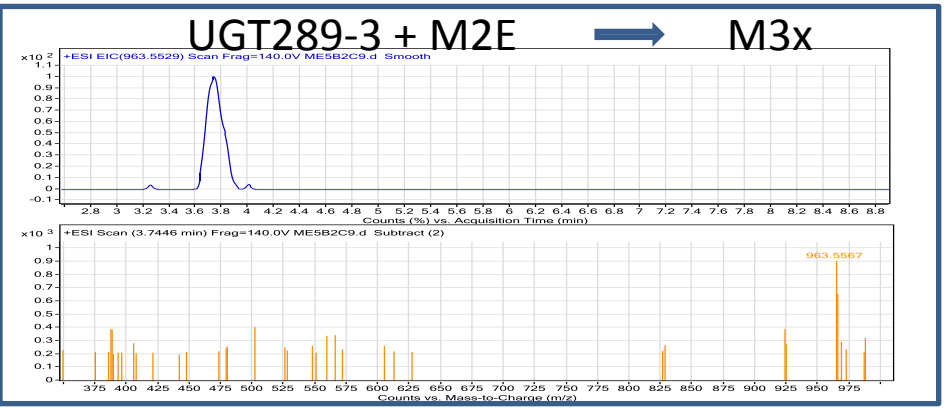
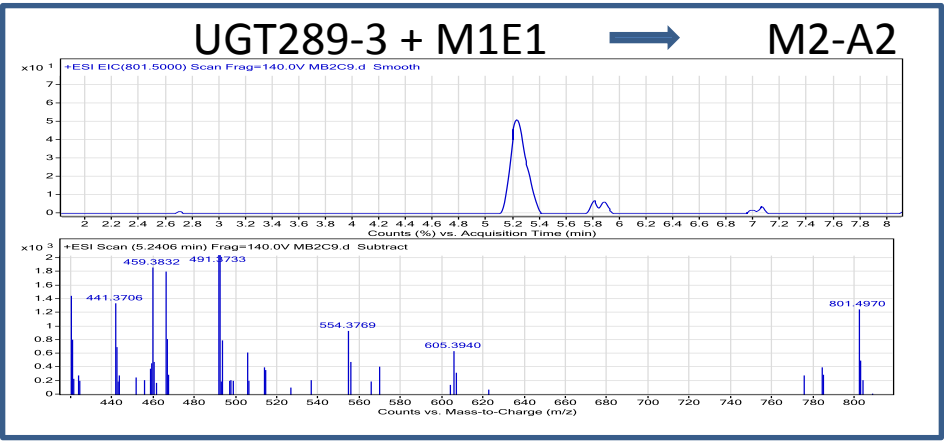


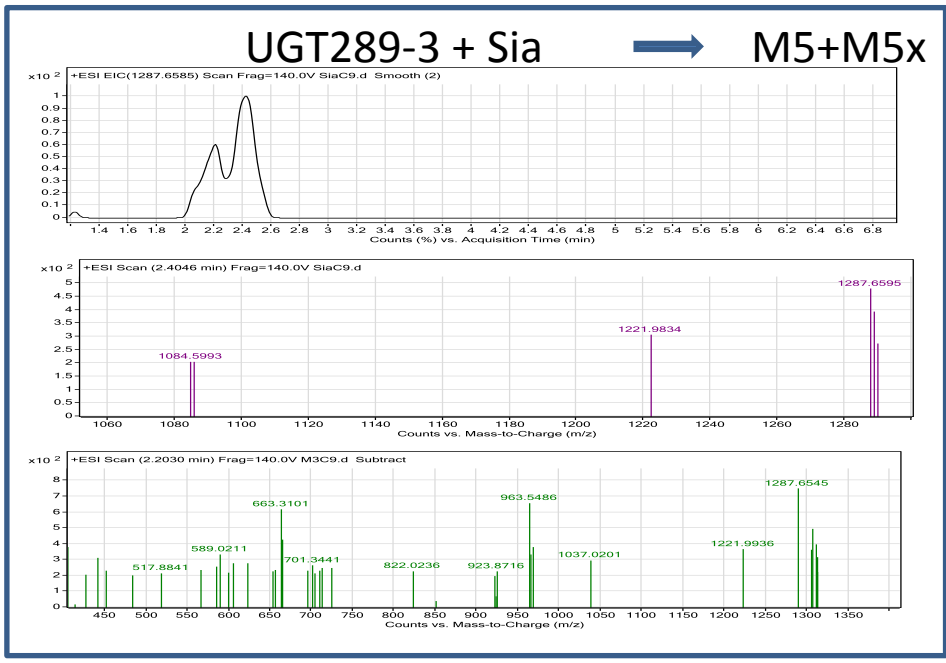
Fig. S19. Reactions, chromatograms and mass spectra of secondary glucosylations. EIC (top window) and MS spectra (bottom window) results of reaction mixes with active enzymes are shown. Structures and full names of substrates and products are listed in Table S1. Enzymes, participating in reactions are listed in Fig. 1.

Fig. S19. Continued









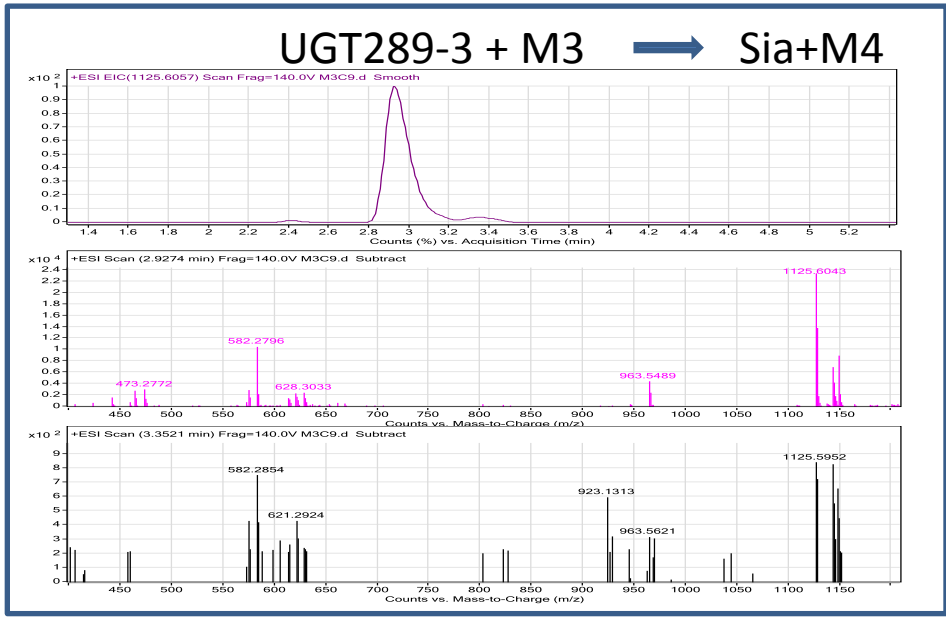
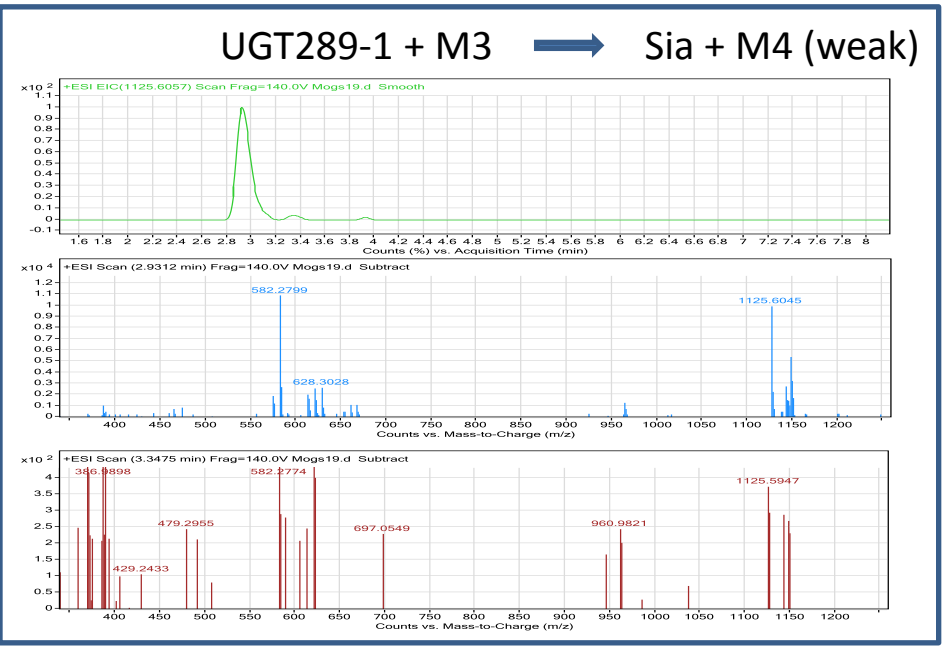
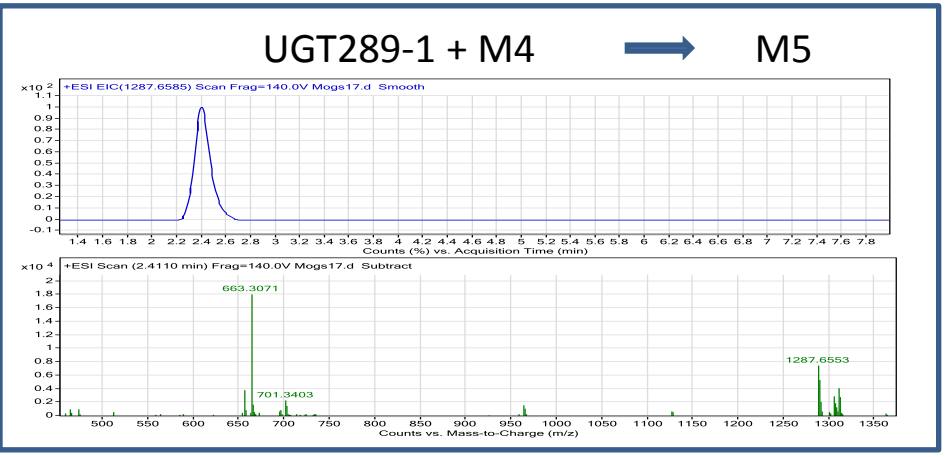
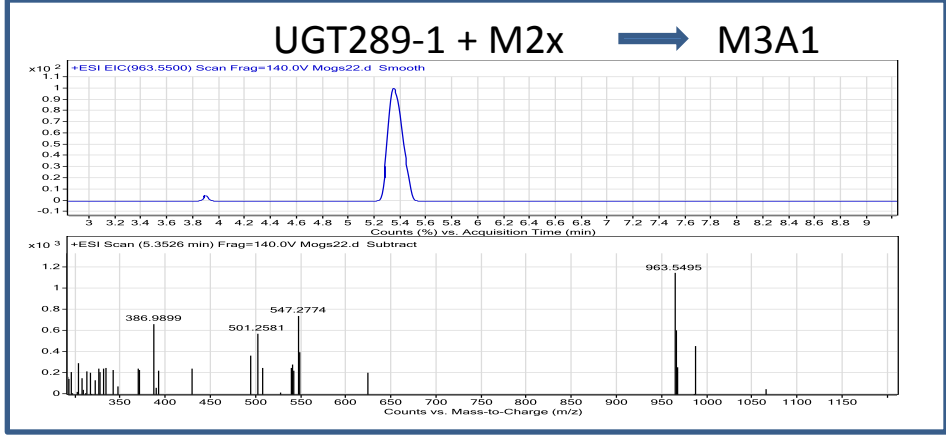
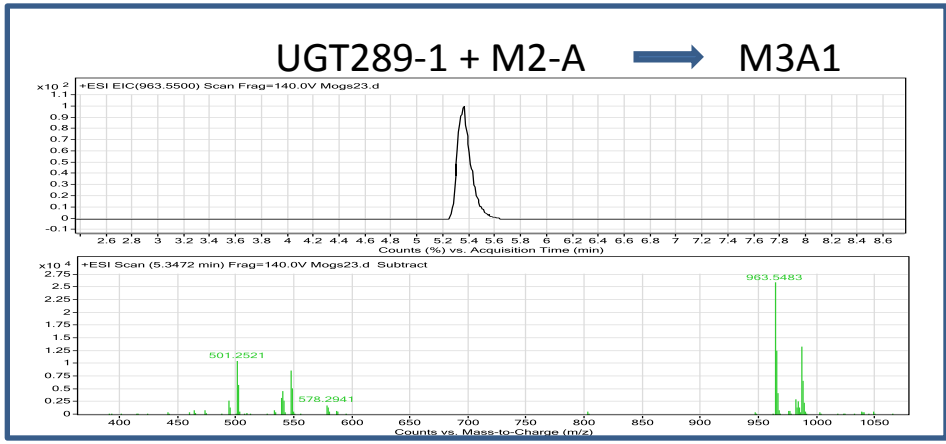


Fig. S19. Continued









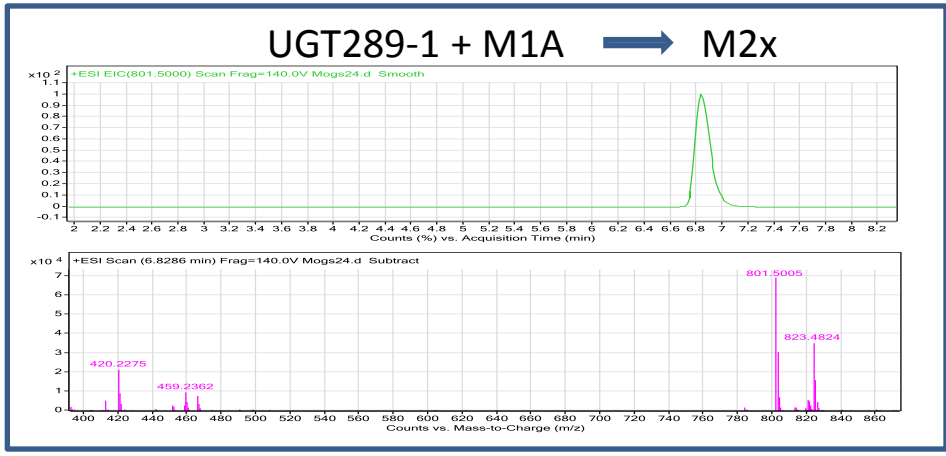
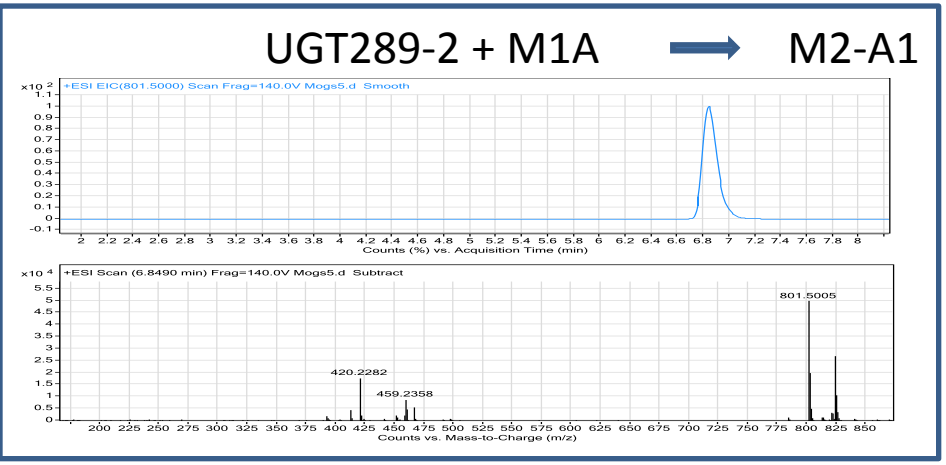
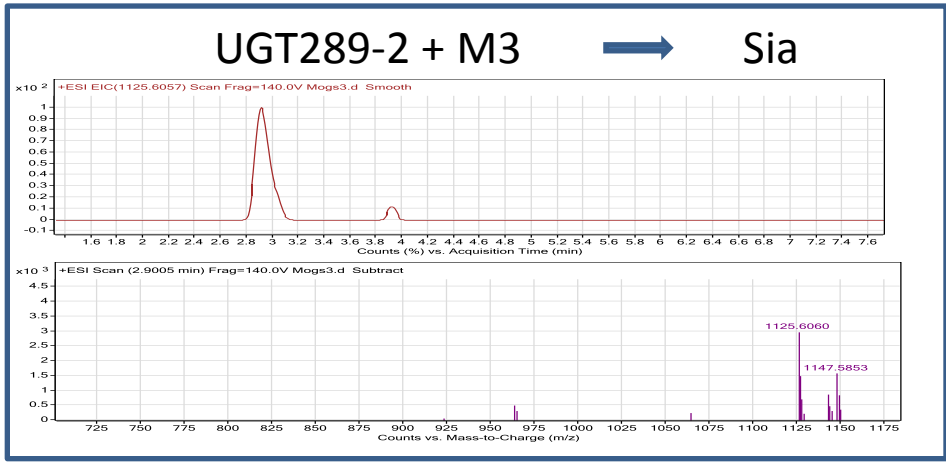
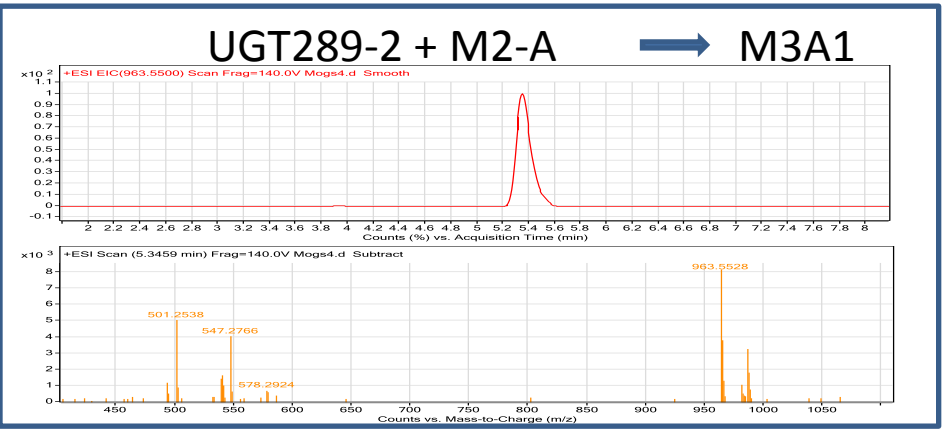


Fig. S19. Continued







a

```
289_2      -MDAQQGHTTTILMLPWVGYGHLLPFLELAKSLSRRKLFHIYFCSTSVSLDAIKPKLPPS
289_1      -MDAQRGHTTTILMFPWLGYGHLSAFLELAKSLSRR-NFHIYFCSTSVNLDAIKPKLPSS
289_3      MDAAQQGDTTTILMLPWLGYGHLSAFLELAKSLSRR-NFHIYFCSTSVNLDAIKPKLPSS
              **:*.******:**:*****  ***********  **********.********* *
289_2      ISSDDSIQLVELRLPSSPE-LPPHLHTTNGLPSHLMPALHQAFVMAAQHFQVILQTLAPH
289_1      S-SSDSIQLVELCLPSSPDQLPPHLHTTNALPPHLMPTLHQAFSMAAQHFAAILHTLAPH
289_3      --FSDSIQFVELHLPSSPE-FPPHLHTTNGLPPTLMPALHQAFSMAAQHFESILQTLAPH
              .****:*** *****: :********.**  ***:***** ******  **:*****
289_2      LLIYDILQPWAPQVASSLNIPAINFSTTGASMLSRTLHPTHYPSSKFPISEFVLHNHWRA
289_1      LLIYDSFQPWAPQLASSLNIPAINFNTTGASVLTRMLHATHYPSSKFPISEFVLHDYWKA
289_3      LLIYDSLQPWAPRVASSLKIPAINFNTTGVFVISQGLHPIHYPHSKFPFSEFVLHNHWKA
           ***** :*****::****:******.***. :::: **  *** ****:******::*:*
289_2      MYTTADGALTEEGHKIEETLANCLHTSCGVVLVNSFRELETKYIDYLSVLLNKKVVPVGP
289_1      MYSAAGGAVTKKDHKIGETLANCLHASCSVILINSFRELEEKYMDYLSVLLNKKVVPVGP
289_3      MYSTADGASTERTRKRGEAFLYCLHASCSVILINSFRELEGKYMDYLSVLLNKKVVPVGP
           **::* ** *:. :*  *::  ***:**.*:*:******* **:****************
289_2      LVYEPNQEGEDEGYSSIKNWLDKKEPSSTVFVSFGTEYFPSKEEMEEIAYGLELSEVNFI
289_1      LVYEPNQDGEDEGYSSIKNWLDKKEPSSTVFVSFGSEYFPSKEEMEEIAHGLEASEVHFI
289_3      LVYEPNQDGEDEGYSSIKNWLDKKEPSSTVFVSFGSEYFPSKEEMEEIAHGLEASEVNFI
           *******:***************************:*************:*** ***.**
289_2      WVLRFPQGDSTSTIEDALPKGFLERAGERAMVVKGWAPQAKILKHWSTGGLVSHCGWNSM
289_1      WVVRFPQGDNTSAIEDALPKGFLERVGERGMVVKGWAPQAKILKHWSTGGFVSHCGWNSV
289_3      WVVRFPQGDNTSGIEDALPKGFLERAGERGMVVKGWAPQAKILKHWSTGGFVSHCGWNSV
           **:******.** ************.***.*******************:********:
289_2      MEGMMFGVPIIAVPMHLDQPFNAGLVEEAGVGVEAKRDSDGKIQREEVAKSIKEVVIEKT
289_1      MESMMFGVPIIGVPMHLDQPFNAGLAEEAGVGVEAKRDPDGKIQRDEVAKLIKEVVVEKT
289_3      MESMMFGVPIIGVPMHVDQPFNAGLVEEAGVGVEAKRDPDGKIQRDEVAKLIKEVVVEKT
           **.********.****:********.************ ******:**** *****:***
289_2      REDVRKKAREMGEILRSKGDEKIDELVAEISLLRKKAPCSI
289_1      REDVRKKAREMSEILRSKGEEKMDEMVAAISLFLKI-----
289_3      REDVRKKAREMSEILRSKGEEKFDEMVAEISLLLKI-----
           ***********.*******:**:**:** ***: *
```

b

	289_1	289_2	289_3
289_1		83.3	88.1
289_2	91.1		83
289_3	92.9	89.3	

Fig. S20. Three tandem functionally active UGT94 family Siraitia genes from scaffold 289 share high level of identity. a) Multiple alignment of the three UGT94 family proteins from contig 289. **b**) Identity and similarity matrix between family 94 UGT scaffold 289 member genes. Similarity and identity scores between three family 94 genes (showing enzymatic activity) from *Siraitia* were determined using MatGAT 2.02 (http://bitincka.com/ledion/matgat/) run with BLOSUM62. In the lower left side of the figure - the percentage of similarity is presented, whereas in the upper right side of the figure there are values of the identity percentage between three proteins. 289_1 is UGT94-289-1, 289_2 is UGT94-289-2 and 289_3 is UGT94-289-3.

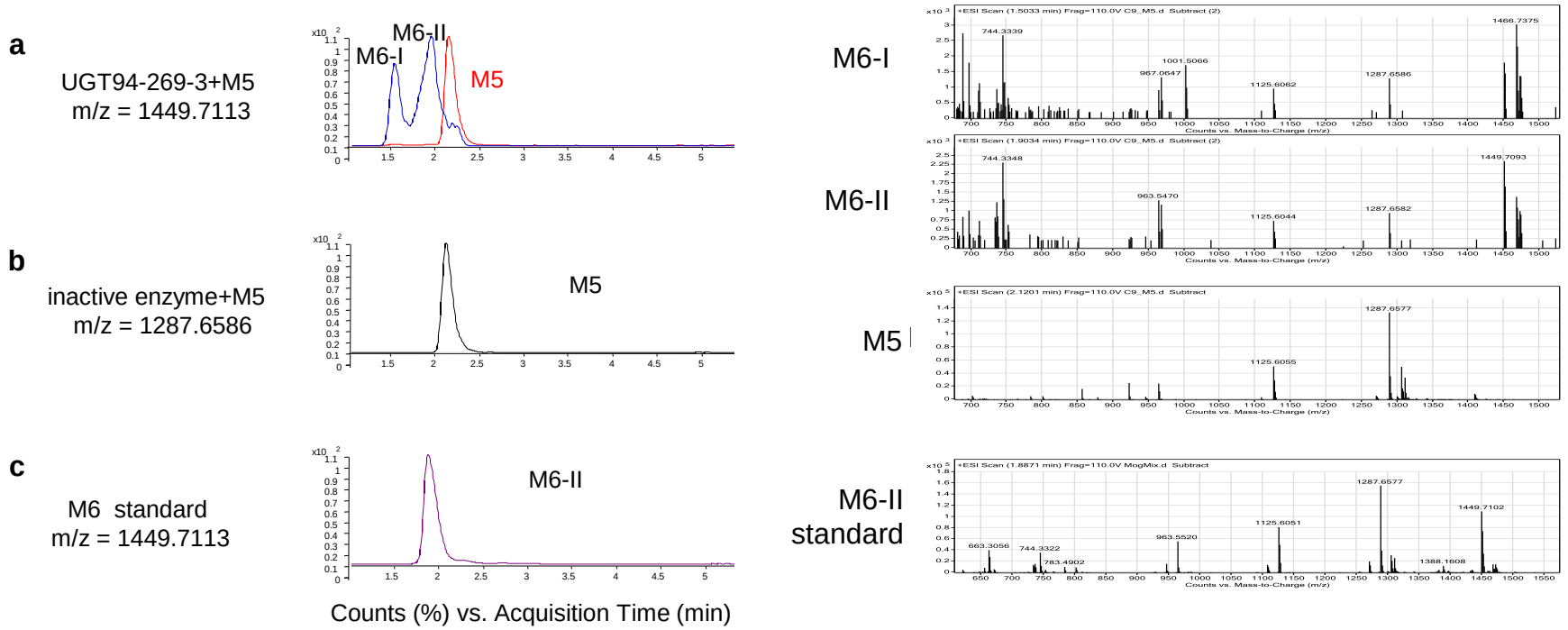
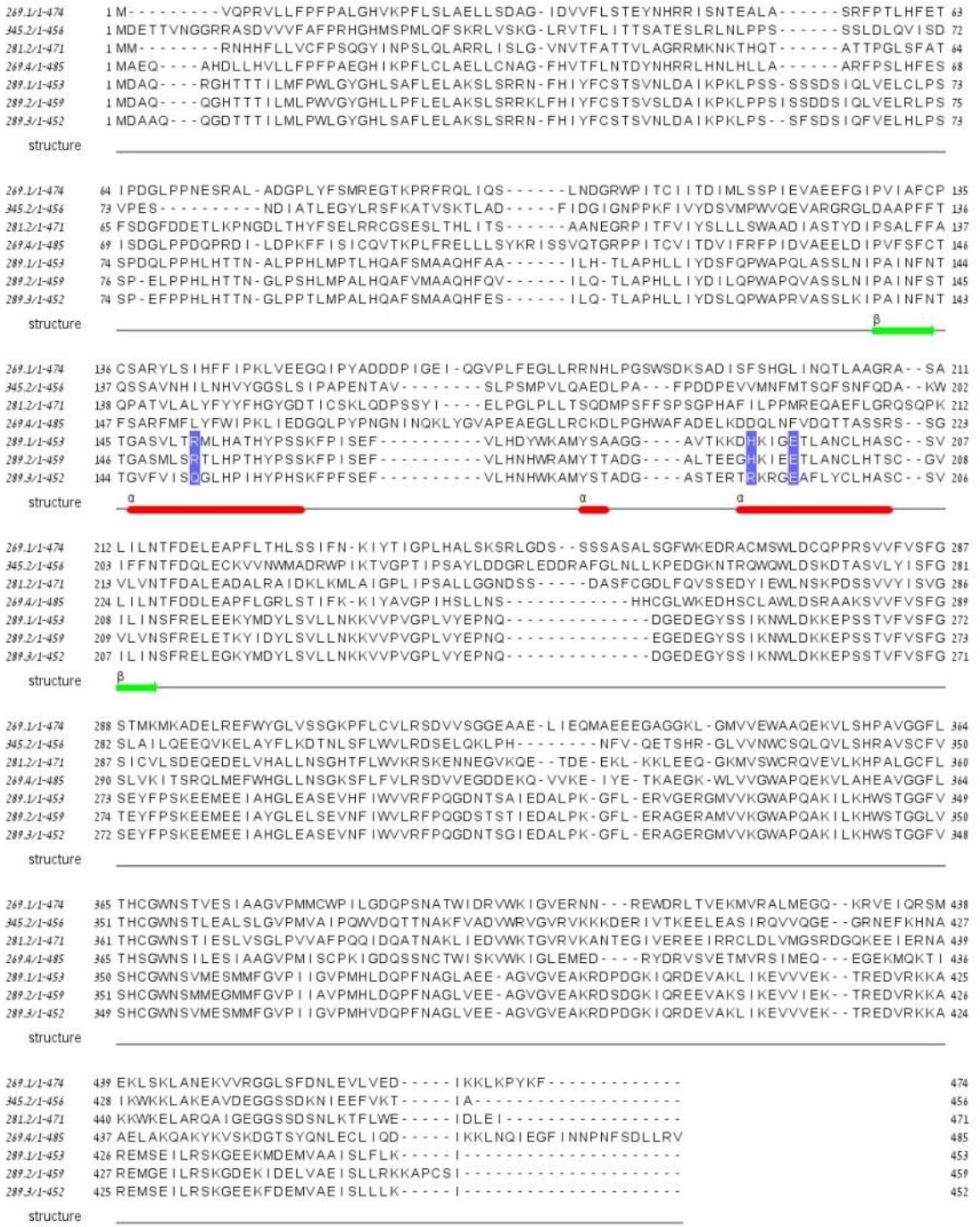


Fig. S21. UGT94-289-3 was occasionally shown to catalyze production of Mogroside VI using Mogroside V as a substrate. Peak "M6-II" , eluting at 1.9 min in (**a**) (m/z=1449.7113) coelutes with the M6 standard in (**c**) indicates accumulation of Mogroside VI in the reaction mix, compared to inactive enzyme control (**b**). Residual Mogroside V that was not completely converted to Mogroside VI in reaction mix, elutes at 2.1 min in (**a**) and (**b**). **c**) Standard of Mogroside VI. The reaction products were checked using LC-MS, as described in SI Methods section. Spectrum is shown for two Mogroside VI products according to the m/z values but the structural differences were not deciphered. To discriminate between the two Mogrosides VI they were marked I (eluting at 1.5min) and II (eluting at 1.9 min).

Fig. S22. Complete alignment for branching UGTs. The complete sequences of the seven UGTs partially presented in Fig. 6b showing the conserved polar amino acids characteristic of the branching UGTs, in blue. Sequences are presented in Data File S2.



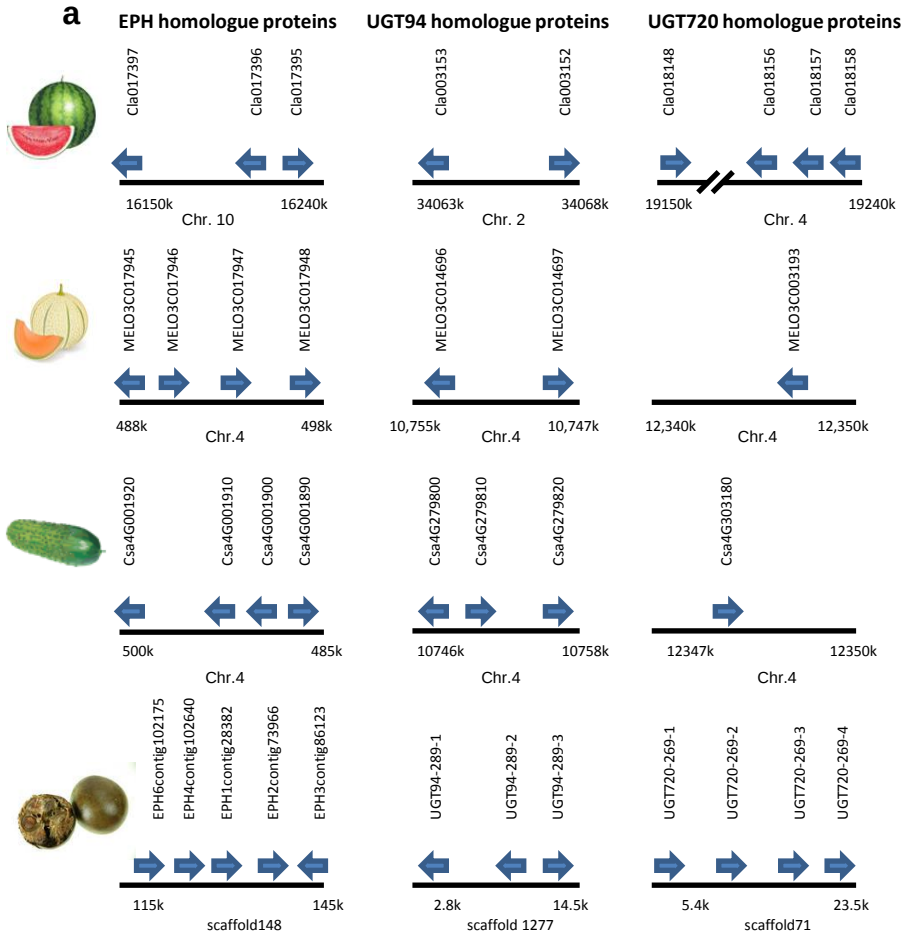


Fig. S23. Arrangement of tandem *EPH* and *UGT* (families 94 and 720) genes in watermelon, melon, cucumber and *Siraitia* genomes/scaffolds (b-d). The length of the three *Siraitia* scaffolds presented (scaffolds 148, 1277 and 71) are 192.4, 51.1 and 452.5 kb, respectively. **a**) Syntenous tandem arrangement of *EPH* and *UGT* genes in watermelon, melon, cucumber and *Siraitia* genomes/contigs. Identity matrixes between syntenous tandem genes from EPH (**b**), UGT94 (**c**) and UGT720 (**d**) families in watermelon, melon, cucumber ad *Siraitia*, indicating the onthologous nature of the genes.

b EPH

	1. MELO3C017948	2. MELO3C017947	3. MELO3C017946	4. MELO3C017945	5. Cla017397	6. Cla017396	7. Cla017395	8. Csa4G001920	9. Csa4G001890	10. Csa4G001900	11. Csa4G001910	12. EPH3contig86123	13. EPH6contig102175	14. EPH1contig28382	15. EPH2contig73966	16. EPH4contig102640
1. MELO3C017948		73.2	60.7	65.9	86.7	61.9	65.6	66.2	82.2	74.6	60.4	67.8	71.6	80.8	64.9	60.6
2. MELO3C017947	86.1		59.2	63.6	74.4	61.3	63	64.8	69.5	91.7	59.8	65.4	65.4	78.4	65.4	57.1
3. MELO3C017946	76.4	76.5		65.2	61.3	87.7	64.3	64.6	60.1	59.9	92.1	67.4	64.3	63.9	81.1	55
4. MELO3C017945	82	81.2	80.5		65.2	68	95.9	99.8	63.4	61.8	64.9	88	70.9	69.3	72.6	59.4
5. Cla017397	92.4	84.9	76.7	80.4		60.6	64.9	65.5	83.2	74.8	61	66.1	69.6	78.8	65.5	57.2
6. Cla017396	77.4	78.4	93.7	82.1	76.4		67.4	67.7	60.1	61.7	88.7	71.8	64.6	64.6	85.2	56.2
7. Cla017395	82.6	81.5	80.2	98.4	80.7	81.4		94.3	63.4	60.9	64.6	87	69.9	68.4	72.6	58.2
8. Csa4G001920	82.3	81.5	80.2	98.8	80.7	81.4	97.8		63.4	62.7	64.6	87	70.6	69.3	71.9	59.1
9. Csa4G001890	91.4	84	74.8	78.2	91.1	75.2	78.8	78.2		71.5	59.4	65	68.8	73.5	64.2	56
10. Csa4G001900	87	96.9	77.6	82	85.7	78.9	82	82.3	84.8		59.6	64.9	66.1	79.5	64.6	56.5
11. Csa4G001910	76.4	77.2	95	80.5	76.7	94	80.2	80.2	75.2	78.3		68.7	63.6	63.9	81.4	55.9
12. EPH3contig86123	82.9	82.1	81.1	93	81.3	83	93	93	79.7	83.2	82.1		72.8	72.2	75.4	59.1
13. EPH6contig102175	84.8	80.9	78.9	85.4	82.9	79.9	85.1	85.1	82.3	82	78.6	85.8		75	68.8	63.7
14. EPH1contig28382	90.5	89.2	80.2	85.8	89.2	81.1	86.1	85.8	86.7	90.1	80.2	86.7	87		70.3	61.2
15. EPH2contig73966	79.7	79.6	89.3	83.2	78.8	92.8	83.2	82.9	77.5	80.1	89.6	84.5	81.3	82.6		59.3
16. EPH4contig102640	74.5	76	70.2	74.8	72	71.4	74.8	75.4	69.8	73.8	71.4	73.8	75.4	75.7	71.7	

c UGT94

	1. 289-1	2. 289-2	3. 289-3	4. Csa4G279820	5. Csa4G279810	6. Csa4G279800	7. Melo3c014696	8. Melo3c014697	9. Cla003153	10. Cla003152
1. 289-1		83.3	88.1	62.6	64.1	68.9	63.6	68.3	66.6	69.5
2. 289-2	91.1		83	59.4	62	66.6	62.1	65.4	62.7	65.9
3. 289-3	92.9	89.3		63.6	66.6	73	66.2	72.4	69.3	73.5
4. Csa4G279820	78.4	76.9	79.4		87.8	67	85.6	66.1	74.4	68.2
5. Csa4G279810	79.7	78	81.2	93.4		68.8	90	67.8	77.2	69.5
6. Csa4G279800	82.8	80.4	85	83.4	83.6		68.9	91.2	68.4	81.3
7. Melo3c014696	78.8	77.3	79	92.5	93.4	83.4		67.5	76	69.8
8. Melo3c014697	83.3	80.4	85.1	82.2	82.9	95.4	82.4		68.2	81.8
9. Cla003153	81	77.6	82.7	85.6	87.6	83.2	85.6	82.6		69.4
10. Cla003152	83.5	81.8	84.6	83.5	84.4	90	83.5	91.5	83.5	

d UGT720

	1. 269-1	2. 269-2	3. 269-3	4. 269-4	5. Cla018148	6. Cla018156	7. Cla018157	8. Cla018158	9. Csa4G303180	10. Melo3c003193
1. 269-1		54.2	51.9	50.3	50.3	44.1	48.8	72.1	50.1	50.3
2. 269-2	72.8		80.6	80.5	67.9	60.5	64.9	51.1	66	66.5
3. 269-3	73.4	90		91.3	66.5	55.7	62.1	49.9	64.7	65
4. 269-4	72.2	88.5	94.6		65.5	54.7	61.1	48.6	63.6	62.9
5. Cla018148	71.5	83.7	82.5	81.4		70.9	79.3	48.5	79.3	78.5
6. Cla018156	61.6	74.2	70.8	70.1	79.1		67.6	43.3	67.2	68.1
7. Cla018157	69	80	77.8	76.5	87.5	80.1		47.3	76.9	77.3
8. Cla018158	85.8	70.1	69	68.7	70.1	59.8	67.8		47.8	47.6
9. Csa4G303180	70.9	83	81.5	80.2	89.9	78.3	87.3	69.9		93
10. Melo3c003193	71.7	82.6	80.1	79.2	89	78.2	86.9	69.5	96.6	

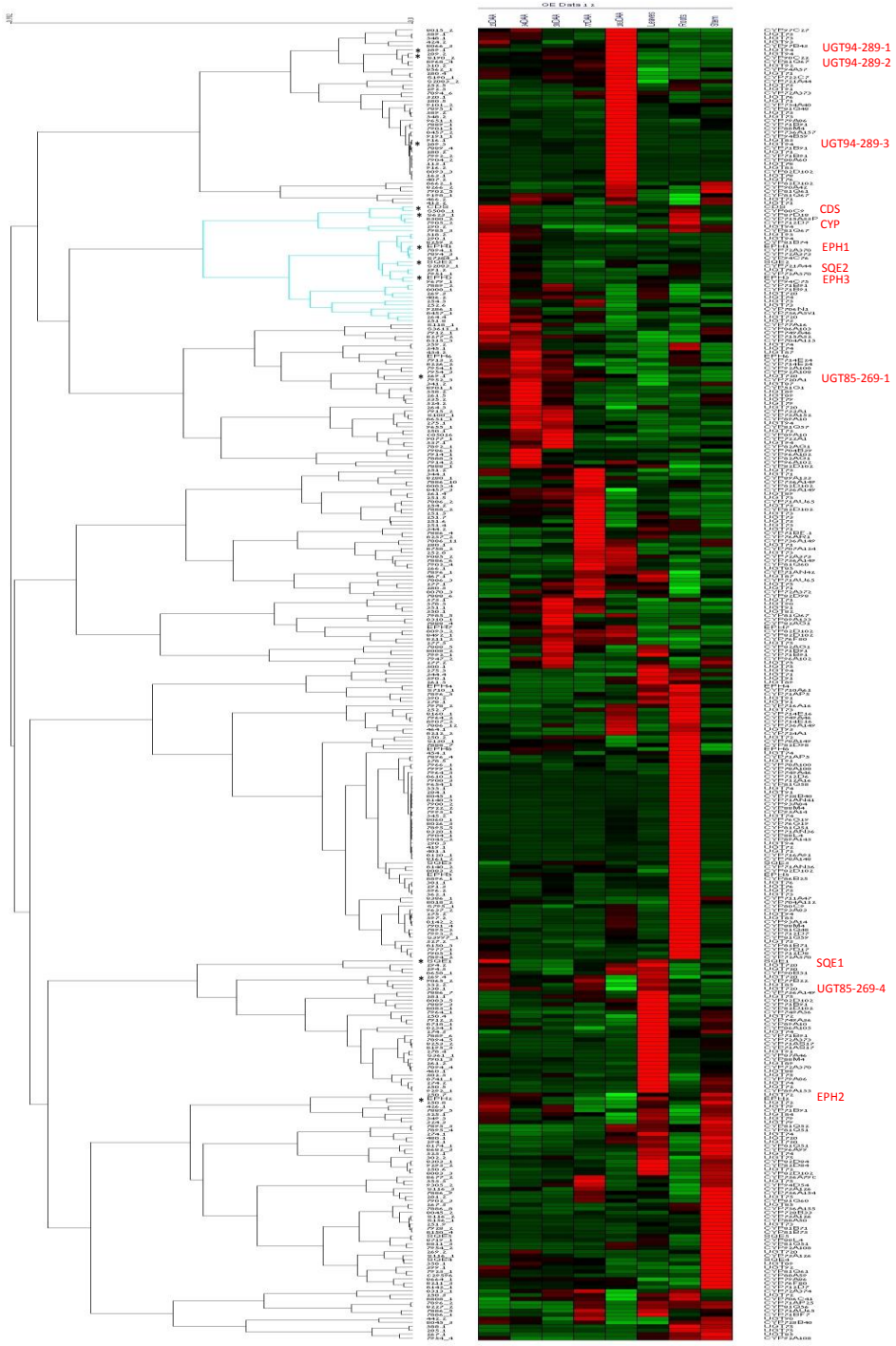


Fig. S24. Expandable version of the normalized hierarchical tree and expression heat map of the combined members of the five enzyme families (SQE, CDS, EPH, CYP and UGT) reported in this paper and shown in Figure 7. Enzymes identified in this paper are marked by asterisks*. Expression data for the non-fruit tissues stems, roots and leaves are included. The young fruit expression of SQE1 and EPH2 is marked by asterisks in the lower third of the figure. The mature-fruit specific expression of the UGT94 branching family can be seen in the upper portion. The early-fruit expression pattern of UGT85-269-1 is also presented, clustering close to the mogrol genes.

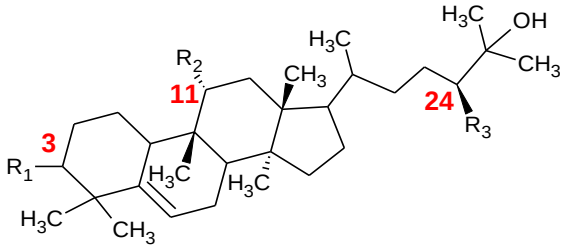


Table S1. Structures and additional data regarding the mogrosides referred to in this study. Mogrosides were identified using LCMS (m/z) and NMR (Table S8), and by comparing the eluting peak m/z and retention time to that of a known standard, when available.

R_1(C3)	R_2(C11)	R_3 (C24)	Name	m/z used for measurement	Verification	Source
-OH	-OH	-OH	Mogrol (M)	477.3944	NMR	Mild acid hydrolysis
-OH	-OH	-Glu	Mogroside I-A1 (M1A1)	639.4472	NMR	Mild acid hydrolysis
-Glu	-OH	-OH	Mogroside I-E1 (M1E1)	639.4472	MS	Enzymatic reaction
-OH	-OH	-Glu(1-6) Glu	Mogroside II-A1 (M2A1) (M2x)	801.5000	NMR	Mild acid hydrolysis
-OH	-OH	-Glu(1-2) Glu	Mogroside II-A (M2A) (M2c)	801.5000	NMR	Enzymatic hydrolysis
-Glu	-OH	-Glu	Mogroside II-E1 (M2E1)	801.5000	MS	Enzymatic reaction
-OH	-OH	-Glu(1-6) Glu / -Glu(1-2) Glu	Mogroside III-A1 (M3A1)	963.5500	MS	Enzymatic reaction
-OH	-OH	-Glu(1-2) Glu	Mogroside III -1-2 (M3-our)	963.5500	NMR	Mild acid / Enzymatic hydrolysis
-OH	-OH	-Glu(1-6) Glu	Mogroside III (M3x)	963.5500	MS	Enzymatic reaction
-Glu(1-4) Glu	-OH	-Glu(1-2) Glu	Isomogroside IV (iM4)	1125.6057	MS	Enzymatic reaction
-Glu(1-6) Glu	-OH	-Glu(1-6) Glu	Mogroside IV-A (M4A)	1125.6057	MS	Enzymatic reaction
-Glu(1-6) Glu	-OH	-Glu(1-2) Glu	Mogroside IV (M4)	1125.6057	NMR	Chaturvdula Prakash and Prakash, 2011
-Glu	-OH	-Glu(1-6) Glu / -Glu(1-2) Glu	Siamenoside I (Sia)	1125.6057	NMR	Chaturvdula Prakash and Prakash, 2011
-Glu(1-4) Glu	-OH	-Glu(1-6) Glu / -Glu(1-2) Glu	Isomogroside V (iM5)	1287.6585	NMR	Chaturvdula Prakash and Prakash, 2011
-Glu(1-6) Glu	-OH	-Glu(1-6) Glu / -Glu(1-2) Glu	Mogroside V (M5)	1287.6585	NMR	Chaturvdula Prakash and Prakash, 2011
-Glu(1-6) Glu	=O	-Glu(1-6) Glu / -Glu(1-2) Glu	11-oxo-Mogroside V (11OM5)	1285.6429	NMR	Chaturvdula Prakash and Prakash, 2011
-Glu(1-6) Glu / -Glu(1-2) Glu	-OH	-Glu(1-6) Glu / -Glu(1-2) Glu	Mogroside VI (M6)	1449.7113	NMR	Chaturvdula Prakash and Prakash, 2011

Table S2. Description of DNA libraries used for genome assembly. In light of the large read lengths of the Moleculo reads only a small percentage of the mate-paired reads were necessary for the scaffolding. Data were deposited in the NCBI Sequence Read Archive (SRA) database as Bioproject XXX (to be deposited upon acceptance).

Raw data

Source	reads	length (bp)
moleculo long (1.5kb to 10 kb)	408,881	2.2 GB
moleculo short(0.5kb to 1.49kb)	166,746	147.5 MB
pair-end (100bp)	30,089,629	3.0GB
mate-paired (100bp)	101,219,634	10.1GB

Assembly input

Source	reads	length(bp)
Moleculo (0.5kb to 10kb)	575,627	2,322,119,138
SuperReads (100bp to 2000bp)	22,151,273	3,066,902,982
Mate-paired (100bp)	2,338,717	233,871,700

Siraitia hybrid assembly RunCA parameters (spec file settings)

Parameter	Setting	Notes
doOBT	1	Overlap Based Trimming
doFragmentCorrection	0	
merSize	22	
overlapper	ovl	
merylMemory	12800	Calculates K-mer seeds
merylThreads	15	
ovlMerThreshold	75	Calculates overlaps
ovlHashBits	25	
ovlHashBlockLength	1000000000	
ovlRefBlockSize	100000000	
ovlThreads	1	
ovlConcurrency	15	
frgMinLen	64	
ovlMinLen	40	
ovlStoreMemory	32768	Mbp
frgCorrThreads	1	Error correction
frgCorrConcurrency	30	
ovlCorrBatchSize	1000000	
ovlCorrConcurrency	15	
unitigger	bogart	
utgGenomeSize	441559616	
cnsConcurrency	20	consensus

Table S3. DNA genomic assembly statistics. Reads were filtered to remove chloroplast and mitochondrial genome sequences, as determined by blast analysis compared to the melon chloroplast and mitochondrial genome (https:/melonomics.net/genome/).

Estimate of genome size	~420M
Number of scaffolds (≥100 bp)	12,772
Total size of assembled scaffolds	420,148,549
N50 (scaffolds)	101,068
Longest scaffold	802,427
Number of contigs (≥100 bp)	25,166
Total size of assembled contigs	411,093,625
N50 (contigs)	34,151
Longest contig	395,621
GC content	33.39%

Table S4. Number of RNA-Seq reads in *Siraitia* fruits, leaves, stem and root. Following cleaning as described in the Methods section, the remaining reads were assembled into transcript contigs, described in Table S5.

	Library tissue									
	15DAA	**34DAA**	**50DAA**	**77DAA**	**90DAA**	**103DAA**	**Stem**	**Leaves**	**Root**	**Total**
#Raw reads	22,174,072	31,407,983	15,416,546	15,936,051	28,689,192	18,081,587	11,657,397	18,349,281	14,800,462	176,512,571
#Clean reads	15,140,655	28,574,943	9,559,931	12,934,972	16,982,487	4,979,298	6,558,456	10,982,862	13,230,409	118,944,013

Table S5. Statistics of the *de novo* transcriptome assembly. Methods for assembly are described in the methods section. Results are presented in Data File S1.

	Filtered transcriptome
Number of contigs	111,084
N50	780
GC content	40.66
Average contig size	561
Total assembled bases	62,407,413

Table S6. Squalene synthase genes in other *Cucurbitaceae* (a) and the expression of the single gene in *Siraitia* (b). Gene names for the three published genomes are derived by blast analysis from the databases ICUGI Cucurbit Genomics Database http://www.icugi.org for cucumber and watermelon and the Melonomics database https://melonomics.net for melon.

a) Single squalene synthase genes in *Cucurbitaceae* genomes

Plant species	Squalene synthase gene
Cucumis melo	MELO3C023346
Cucumis sativus	Csa2M251460
Citrullus vulgaris	Cla016602
Siraitia grosvenorii	scaffold938|60,000..67,000

b) Squalene synthase gene expression in *Siraitia*

	fruit					leaves	root	stem
	15D	34D	50D	77D	103D			
RPKM	59.5	51.1	45.8	65.6	1.9	23.2	90.7	22.4

Table S7. List of functionally identified triterpenoid CYPs and their families, derived from published studies.

Gene name	Species	Accession number	Sugar acceptor	Reference
Cyp51H10	*Avena strigosa*	ABG88965.1	β-amyrin	Qi et al., 2006
CYP705A5	*Arabidopsis thaliana*	Q9FI39	thalian-diol	Field and Osbourn, 2008
Cyp710A1	*Arabidopsis thaliana*	O64697	beta-sitosterol	Morikawa et al., 2006
CYP716A12	*Medicago truncatula*	ABC59076	β-amyrin	Carelli et al., 2011
CYP716A47	*Panax ginseng*	ABB84472	dammarenediol-II	Han et al., 2011
CYP716A53v2	*Panax ginseng*	I7CT85	Protopanaxadiol	Han et al., 2012
CYP72A154	*Glycyrrhiza uralensis*	H1A988	11-oxo- β-amyrin	Seki et al., 2011
CYP72A63	*Medicago truncatula*	H1A981	11-oxo- β-amyrin	Seki et al., 2011
CYP734A7	*Solanum lycopersicum*	NP_001233940	castasterone	Ohnishi et al., 2006
CYP81	*Cucumis sativus*	AIT72037	19-hydroxycucurbitadienol	Shang et al., 2014
CYP85A2	*Arabidopsis thaliana*	NP_566852	brassinosteroid	Castle et al., 2005
CYP88	*Cucumis sativus*	AIY67847	cucurbitadienol	Shang et al., 2014
CYP88D6	*Glycyrrhiza uralensis*	B5BSX1	β-amyrin	Seki et al., 2008
Cyp90A1	*Arabidopsis thaliana*	AED90909	22-hydroxycampesterol	Ohnishi et al., 2005
CYP93E1	*Glycine max*	NP_001236154	β-amyrin	Shibuya et al., 2006
CYP93E3	*Glycyrrhiza uralensis*	B5BT05	β-amyrin	Seki et al., 2008

Table S8. NMR data of identified compounds presented in this paper.

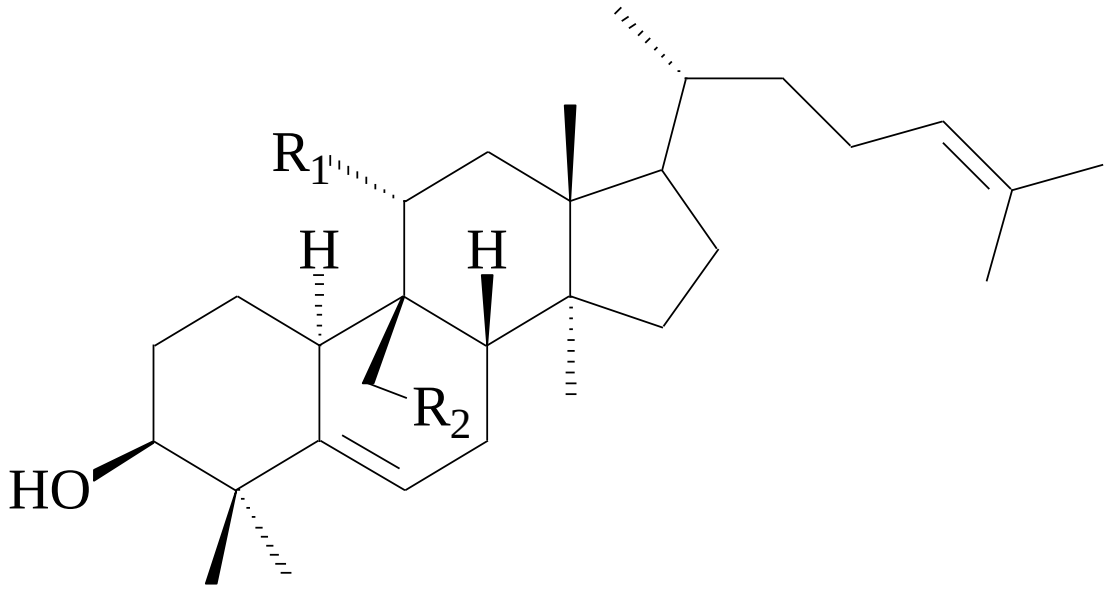
a) Triterpene (aglycone) chemical shifts (all in CD_3OD)

	1		**2**		**3**		**4**		**5**	
	^{1}H	**^{13}C**	**^{1}H**	**^{13}C**	**^{1}H**	**^{13}C**	**^{1}H**	**^{13}C**	**^{1}H**	**^{13}C**
1	1.50, 2.23	26.47	1.54, 1.42	21.58	1.50, 2.23	26.47	1.50, 2.23	26.47	1.50, 2.23	26.47
2	1.54, 2.00	30.53	1.67, 1.88	30.68	1.55, 2.00	30.52	1.55, 2.00	30.52	1.55, 2.00	30.53
3	3.41	77.58	3.43	77.45	3.41	77.57	3.42	77.57	3.42	77.59
4	—	42.66	—	42.07	—	42.65	—	42.66	—	42.67
5	—	144.06	—	142.89	—	144.05	—	144.05	—	144.07
6	5.50	120.60	5.58	121.91	5.50	120.62	5.50	120.63	5.50	120.64
7	1.80, 2.42	25.13	1.78, 2.35	25.03	1.80, 2.42	25.13	1.80, 2.43	25.14	1.80, 2.43	25.14
8	1.67	44.78	2.35	35.39	1.68	44.78	1.68	44.78	1.68	44.80
9	—	40.95	—	40.28	—	40.95	—	40.95	—	40.97
10	2.49	37.21	2.41	39.40	2.49	37.10	2.48	37.20	2.49	37.14
11	3.85	79.36	1.44, 1.88	27.10	3.85	79.37	3.85	79.37	3.85	79.39
12	1.81	41.09	1.57, 1.73	31.61	1.83	41.10	1.83	41.11	1.80, 1.85	41.13
13	—	48.28	—	47.20	—	48.29	—	48.29	—	48.31
14	—	50.61	—	50.20	—	50.62	—	50.59	—	50.63
15	1.13, 1.18	35.35	1.25	36.21	1.13, 1.20	35.35	1.14, 1.20	35.39	1.12, 1.20	35.37
16	1.29, 1.91	29.13	?	29.11	1.34, 1.96	28.83	1.38, 1.95	29.14	1.34, 1.95	28.89
17	1.61	51.61	1.53	51.91	1.61	51.84	1.63	51.72	1.61	51.88
18	0.90	17.10	0.93	15.35	0.92	17.15	0.92	17.16	0.92	17.17
19	1.15	26.27	3.28, 3.50	66.71	1.15	26.27	1.15	26.26	1.15	26.28
20	1.47	37.01	1.46	37.07	1.51	37.20	1.47	37.53	1.50	37.22
21	0.96	19.18	0.93	19.24	0.97	19.16	0.97	19.11	0.96	19.16
22	1.04, 1.43	37.49	?	37.61	1.28, 1.50	34.50	1.47, 1.52	34.29	1.30, 1.47	34.64
23	1.89, 2.03	25.83	?	25.85	1.35	29.12	1.47, 1.61	29.72	1.34	29.11
24	5.09	126.20	5.09	126.29	3.21	79.76	3.45	89.71	3.35	77.94
25	—	131.84	—	131.78	—	73.91	—	73.67	—	81.52

26	1.66	25.93	1.67	25.94	1.12	24.96	1.16	24.67	1.20	22.65
27	1.60	17.74	1.60	17.73	1.16	25.75	1.15	26.62	1.24	23.03
28	1.10	26.45	1.11	26.16	1.10	26.47	1.10	26.47	1.10	26.47
29	1.05	27.44	1.01	28.24	1.05	27.43	1.05	27.43	1.05	27.45
30	0.86	19.86	0.88	18.98	0.87	19.86	0.87	19.89	0.87	19.88

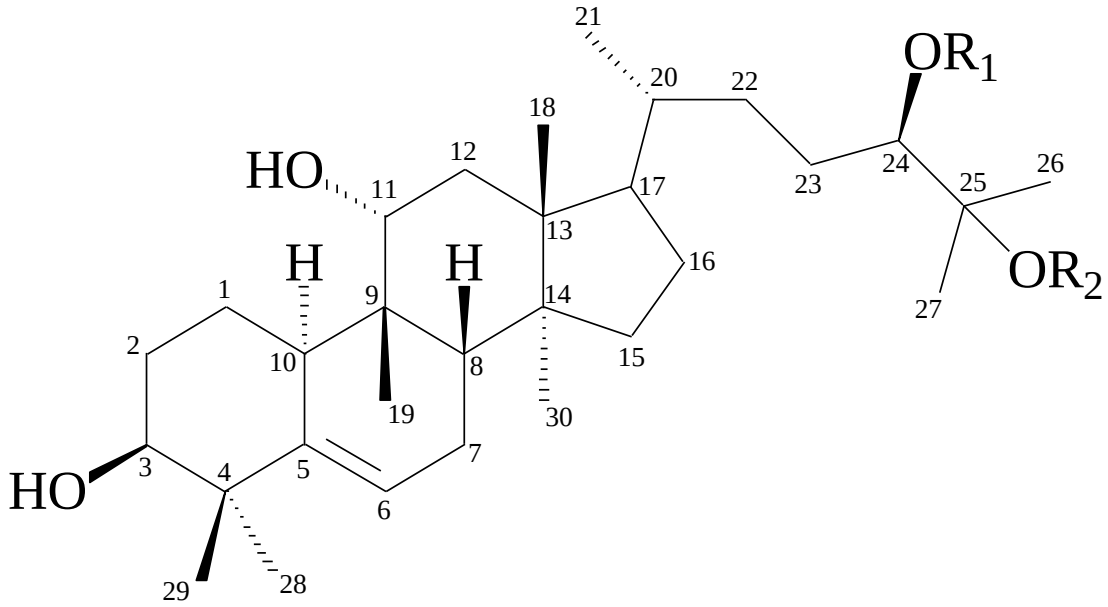
b) Glucose moiety chemical shifts (all in CD_3OD)

	4		**5**	
	^{1}H	**^{13}C**	**^{1}H**	**^{13}C**
1'	4.33	105.96	4.51	98.06
2'	3.22	75.37	3.15	75.40
3'	3.35	78.16	3.36	78.30
4'	3.30	71.63	3.28	71.71
5'	3.27	78.05	3.27	77.83
6'	3.64, 3.85	62.66	3.64, 3.82	62.81



1. R_1 = OH, R_2 = H

2. R_1 = H, R_2 = OH



3. R_1 = R_2 = H (mogrol)

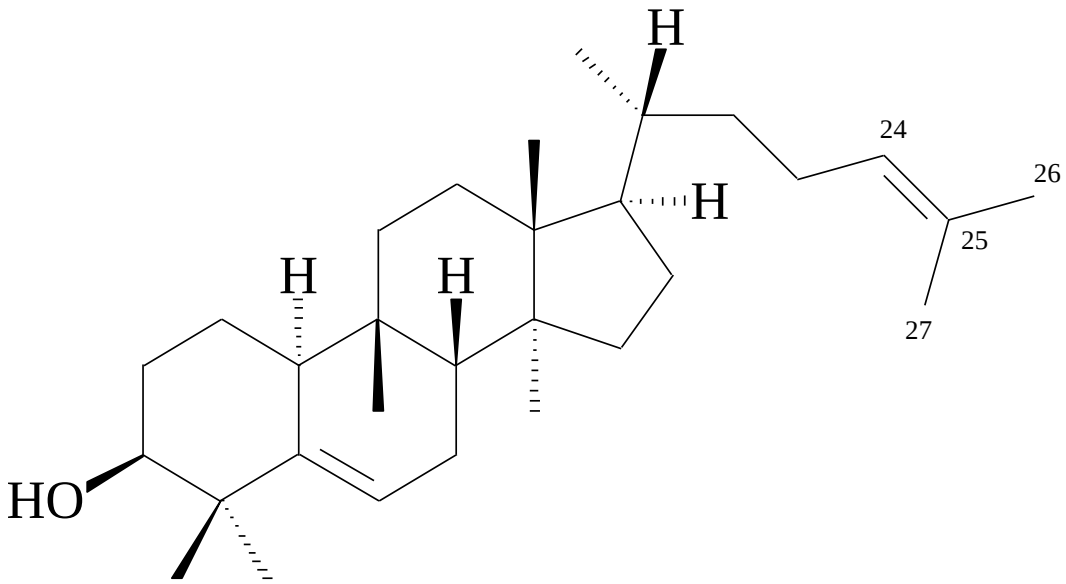
4. R_1 = 1-β-glucose, R_2 = H (M1)

5. R_1 = H, R_2 = 1-β-glucose (M1-C25)

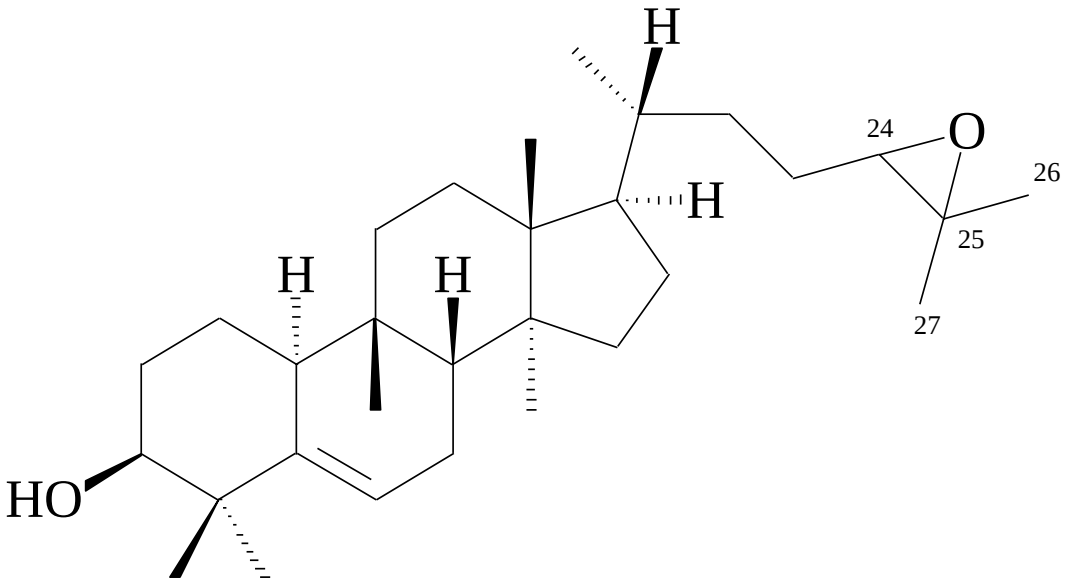
c) Cucurbitadienol (6) and its 24,25-epoxide (7), in $CDCl_3$

	6		7	
	^{1}H	^{13}C	^{1}H	^{13}C
1	1.46, 1.57	21.14	1.46, 1.58	21.14
2	1.71, 1.88	28.90	1.72, 1.88	28.93
3	3.47	76.66	3.48	76.66
4	—	41.45	—	41.45
5	—	141.23	—	141.24
6	5.59	121.51	5.59	121.49
7	1.80, 2.38	24.39	1.80, 2.38	24.38
8	1.76	43.67	1.76	43.65
9	—	34.49	—	34.48
10	2.27	37.85	2.27	37.84
11	1.44, 1.66	32.34	1.44, 1.65	32.32
12	1.51, 1.68	30.46	1.50, 1.68	30.46
13	—	46.27	—	46.28
14	—	49.17	—	49.19
15	1.12, 1.20	34.77	1.13, 1.22	34.75
16	1.27, 1.88	27.94	1.31, 1.88	27.93
17	1.50	50.47	1.50	50.43
18	0.85	15.37	0.86	15.40
19	0.92	28.06	0.92	28.05
20	1.43	35.82	1.50	35.88
21	0.90	18.67	0.91	18.66
22	1.03, 1.42	36.46	1.29, 1.61	32.91
23	1.85, 2.02	24.87	1.45, 1.55	25.86
24	5.09	125.25	2.68	64.99
25	—	130.94	—	58.13

26	1.68	25.73	1.30	24.96
27	1.60	17.63	1.26	18.63
28	1.02	27.26	1.02	27.25
29	1.14	25.44	1.14	25.47
30	0.80	17.83	0.81	17.84



6. Cucurbitadienol



7. 24,25-epoxycucurbitadienol

Table S9. List of functionally identified triterpenoid UGTs, based on published literature.

Gene name	Species	Accession number	Sugar acceptor	Reference
SlGAME1	*Solanum lycopersicum*	HQ293016	Tomatidine	Itkin et al., 2011
SlGAME2	*Solanum lycopersicum*	HQ293018	Tomatidine glucoside	Itkin et al., 2014
SlGAME3	*Solanum lycopersicum*	HQ293017	Tomatidine	Itkin et al., 2011
SlGAME18	*Solanum lycopersicum*	Solyc07g043500	*γ*-tomatine	Itkin et al., 2014
SlGAME17	*Solanum lycopersicum*	Solyc07g043480	Tomatidine galactoside	Itkin et al., 2014
*St*SGT3 (Rha)	*Solanum tuberosum*	ABB84472	*β*-solanine, *β*-chaconine	McCue et al., 2007
*St*SGT1 (Gal)	*Solanum tuberosum*	AAB48444	Solanidine	McCue et al., 2005
*St*SGT2 (Glu)	*Solanum tuberosum*	ABB29873	Solanidine	McCue et al., 2006
Sa UGT73L1 (GT4A)	*Solanum aculeatissimum*	BAD89042	Diosgenin, nuatigenin, tigogenin, solanidine, solasodine, tomatidine	Kohara et al., 2005
MtUGT73F3	*Medicago truncatula*	FJ477891	Hederagenin	Naoumkina et al., 2010
MtUGT73K1	*Medicago truncatula*	AY747626	Hederagenin	Achnine et al., 2005
MtUGT71G1	*Medicago truncatula*	AY747627	Medicagenic acid	Achnine et al., 2005
SvUGT74M1	*Saponaria vaccaria*	DQ915168	Gypsogenic /16*α*-hydroxygypsogenic acids and Gypsogenin	Meesapyodsuk et al., 2007
GmUGT73F2	*Glycine max*	BAM29362	Saponins from A group	Sayama et al., 2012
GmUGT73F4	*Glycine max*	BAM29363	Saponins from A group	Sayama et al., 2012
GmUGT73P2	*Glycine max*	FJ433879	Soyasapogenol B	Shibuya et al., 2010
GmUGT91H4	*Glycine max*		Soyasaponin III	Shibuya et al., 2010
BvUGT73C10	*Barbarea vulgaris*	JQ291613	Hederagenin, Oleanolic acid	Augustin et al., 2012
BvUGT73C11	*Barbarea vulgaris*	AFN26667	Hederagenin, Oleanolic acid	Augustin et al., 2012
BvUGT73C12	*Barbarea vulgaris*	AFN26668	Hederagenin, Oleanolic acid	Augustin et al., 2012
BvUGT73C13	*Barbarea vulgaris*	AFN26669	Hederagenin, Oleanolic acid	Augustin et al., 2012
PgUGT74AE2	*Panax ginseng*	JX898529	Protopanaxadiol, Compound K	Jung et al., 2014
PgUGT94Q2	*Panax ginseng*	JX898530	Ginsenoside F2	Jung et al., 2014
PgUGT71A27	*Panax ginseng*	KF377585	Protopanaxadiol	Jung et al., 2014
SgUGT74AC1	*Siraitia grosvenorii*	AEM42999	Mogrol	Dai et al., 2015

Table S10. **Gene locations of mogroside gene orthologs in other cucurbits.** The syntenous nature of the tandem gene families (UGT94 and EPH) can be seen among all the species. Gene names and positions for the three published genomes are derived from the databases ICUGI Cucurbit Genomics Database http://www.icugi.org for cucumber and watermelon and the Melonomics database https://melonomics.net for melon.

Gene	***Siraitia grosvenorii***		***Cucumis melo***			***Cucumis sativus***			***Citrullus lanatus***		
	scaffold	**position**	**LG**	**scaffold**	**gene**	**chr**	**position**	**gene**	**chr**	**position**	**gene**
SQE1	Read_65550-Barcode=BC266:length=4362		3	14	MELO3c010781	Chr2	16352391..16356007	Csa2G353480	Chr8	25404900..25408164	Cla022651
SQE2	scaffold60|size415656	212374..216712	7	29	MELO3c016845	Chr4	21766118..21772333	Csa4G645290	Chr5	25744891..25750880	Cla020903
EPH1	scaffold148|size192353	118500..120200	7	31	MELO3c017948	Chr4	488292..489884	Csa4G001900	Chr10	16232403..16234720	Cla017397
EPH2	scaffold148|size192353	125500-128000	7	31	MELO3c017946	Chr4	491904..492259	Csa4G001910	Chr10	16206533..16210293	Cla017396
EPH3	scaffold148|size192353	142000-144500	7	31	MELO3c017945	Chr4	497804..498023	Csa4G001920	Chr10	16153806..16154994	Cla017395
CDS	scaffold1407|size46898	30150..38365	11	52	MELO3c022374	Chr6	4857000..4862947	Csa08595	Chr6	1545386..1554594	Cla007080
CYP87D18	scaffold623|size327546	274290..276990	12	1	MELO3c002192	Chr1	4911784..4911897	Csa1G044890	Chr1	10099457..10102279	Cla008354
UGT720 269-1	scaffold71|size452539	145000..146150	8	2	MELO3c003193	Chr4	12347724..12347757	Csa4g303180	Chr4	19233380..19237092	Cla18158
UGT94 289-1	ctg7180289Length = 21327	2800-4200	5	22	MELO3c014696	Chr4	10755907..10757265	Csa4G279820	Chr2	34067241..34068626	Cla003152
UGT94 289-2	ctg7180289Length = 21327	9500-11000	5	22	MELO3c014697	Chr4	10748480..10748928	Csa4G279810	Chr2	34064202..34065530	Cla003153
UGT94 289-3	ctg7180289Length = 21327	13000-14500				Chr4	10746090..10747448	Csa4G279800			

Table S11. **Syntenous organization of CDS clusters in *Siraitia* and other cucurbits.** The *Siraitia* cluster is presented in 2 scaffolds, 1407 and 2217 of *Siraitia* genome, which were not combined by the assembly program due to the large intron within the CDS gene. The CDS coding sequence was manually identified in the two scaffolds and the total scaffold size encompasses in total about 75 kbp. The genes are aligned according to the respective CDS genes. ACT, acyltransferase; CDS, cucurbitadienol synthase; CYP, cytochrome P450.

Siraitia grosvenorii			***Cucumis melo***				***Cucumis sativus***					***Citrullus lanatus***			
Scaffold	**Position**	**Annotation**	**LG**	**Scaffold**	**Gene**	**Annotation**	**Chr**	**Position**	**Gene**	**Annotation**		**Chr**	**Position**	**Gene**	**Annotation**
			11	52	MELO3c022377	CYP81	Chr6	6,065,000	Csag6088160	CYP81		Chr6	1526791..1528422	Cla007077	CYP81
Scaffold1407|size46898	7334…5752	CYP81 (c20848)	11	52	MELO3c022376	CYP89	Chr6	6,068,000	Csag6088170	CYP89		Chr6	1531545..1533098	Cla007078	CYP89
Scaffold1407|size46898	18676…17360	ACT	11	52	MELO3c022375	CYP81	Chr6	6,071,000	Csag6088180	CYP81		Chr6	1538623..1540306	Cla007079	CYP81
Scaffold1407|size46898 and Scaffold2217|size27566	33740…31667 23000-24000	**CDS**	11	52	MELO3c022374	**CDS**	Chr6		Csa08595	**CDS**		Chr6	1545386..1554594	Cla007080	**CDS**
Scaffold2217|size27566	9000..10000	CYP87 (c82338)	11	52	MELO3c022373	ACT	Chr6	6,065,000	Csag6088700	ACT		Chr6	1570224..1572546	Cla007081	ACT
			11	52	MELO3c022372	CYP87	Chr6	6,095,000	Csag6088710	CYP87		Chr6	1579830..1583830	Cla007082	CYP87
			Based on *C. melo* genome browser https://melonomics.net/genome/				Based on *C. sativus* genome browser http://icugi.org/cgi-bin/gb2/gbrowse/cucumber_v2/					Based on *C. lanatus* genome browser http://icugi.org/cgi-bin/gb2/gbrowse/watermelon_v1/			

Table S12. **Synteny of CYP450 C-11 hydroxylase cluster in *Siraitia* and other cucurbits.** The complete *Siraitia* cluster is presented in a single scaffold of 327kb. The gene arrangements indicate inversions in the gene order. BAHD, BAHD acyltransferase; Adh, alcohol dehydrogenase.

Siraitia			***Cucumis melo***				***Cucumis sativus***			***Citrullus lanatus***		
scaffold	**position**	**annotation**	**LG**	**scaffold**	**gene**	**annotation**	**chr**	**gene**	**annotation**	**chr**	**gene**	**annotation**
scaffold623; 327kbp	281992-284018	BAHD	9	51	MELO3c022188	transporter	Chr 1	Csa1G044900	BAHD	Chr 1	Cla008353	BAHD
scaffold623; 327kbp	279100-281100	BAHD	9	51	MELO3c022189	Adh	Chr 1	Csa1G044870	transport testa	Chr 1	Cla008354	**CYP87A3**
scaffold623; 327kbp	274078-276024	**CYP87D18 (c102801)**	9	51	MELO3c022190	transport testa	Chr 1	Csa1G044820	transporter	Chr 1	Cla008355	**CYP87A3**
scaffold623; 327kbp	266000-267000	momilactone synthase	9	51	MELO3c022191	momilactone synthase	Chr 1	Csa1G044860	Adh	Chr 1	Cla008356	momilactone synthase
scaffold623; 327kbp	263000-265000	Adh	9	51	MELO3c022192	**CYP87A3**	Chr 1	Csa1G044880	momilactone synthase	Chr 1	Cla008357	transport testa
scaffold623; 327kbp	262000-263000	transport testa	9	51	MELO3c022193	BAHD	Chr 1	Csa1G044890	**CYP87A3**	Chr 1	Cla008358	Adh
			based on *C. melo* genome browser https://melonomics.net/genome/				based on *C. sativus* genome browser http://icugi.org/cgi-bin/gb2/gbrowse/cucumber_v2/			based on *C. lanatus* genome browser http://icugi.org/cgi-bin/gb2/gbrowse/watermelon_v1/		

Table S13. Expression of the mogroside pathway orthologs in developing melon and watermelon fruit. Data represent expression data (RPKM) in developing fruit of 3 varieties of watermelon and 3 varieties of melon, of the respective orthologs of the 6 mogrol biosynthesis genes coordinately expressed in *Siraitia* fruit. All varieties were sampled at 10, 20, 30 and ripe (about 40) days after pollination. Watermelon varieties are Orangeglo (OG), Yellow Crimson (YC) and Crimson Sweet (CS). Melon varieties are Doya (a flexuosus type), Noy Yizre'el (NY), a cantaloupensis type, and Faqus (FAQ), a flexuosus type. Data are the average of the results from three individual RNA-seq libraries, each.

gene name	watermelon gene number	OG				YC				CS			
		10	20	30	ripe	10	20	30	ripe	10	20	30	ripe
SQE1	Cla022651	33.74	7.08	15.80	10.53	17.28	6.67	9.59	7.63	13.71	8.97	9.19	3.64
SQE2	Cla020903	76.12	52.78	54.49	73.53	58.80	27.91	21.43	20.95	37.82	22.72	19.01	21.21
EPH1	Cla017397	71.06	66.40	38.82	45.34	66.77	72.14	29.73	17.58	58.36	38.16	13.78	8.68
EPH3	Cla017395	5.39	1.84	0.50	0.59	4.32	1.35	0.36	0.30	3.59	1.44	0.58	0.49
CDS	Cla007080	0	0	0	0	0	0.03	0	0	0	0	0	0
CYP	Cla008354	0	0	0	0	0.04	0	0	0	0.03	0	0	0

gene name	melon gene number	Doya				NY				FAQ			
		10	20	30	ripe	10	20	30	ripe	10	20	30	ripe
SQE1	MELO3C010781	79.68	44.46	77.08	208.40	43.05	42.92	29.74	321.22	68.26	36.96	45.89	131.32
SQE2	MELO3C016845	32.18	36.11	53.26	47.92	26.79	17.74	29.68	13.44	36.22	29.81	33.58	56.32
EPH1	MELO3C017947	0.19	0	0	0	0.81	0.38	0.23	0	0	0.12	0	0
EPH3	MELO3C017945	0.40	0.56	0.09	0.30	1.03	1.16	1.32	0	0.82	2.38	2.12	0
CDS	MELO3C022374	0	0	0	0	0	0	0.03	0	0	0	0	0
CYP	MELO3C002192	0.12	0	0	0	0.08	0	0	0	0	0	0	0

Table S14. Expression of alternative triterpene synthases during fruit development and in vegetative tissues of *Siraitia*. bAM, beta-amyrin synthase; CAS, cycloartenol synthase; CDS, cucurbitadienol synthase. RPKM, reads per kilobase of transcript per million mapped reads. DAA, days after anthesis, indicating fruit age. SgCDS is the most highly expressed terpene synthase in young *Siraitia* fruit.

		RPKM							
	***Siraitia* contig**	**DAA 15**	**DAA 34**	**DAA 50**	**DAA 77**	**DAA 103**	**Leaves**	**Root**	**Stem**
bAM1	c74269	0.4	0.8	2.8	6.3	0.0	0.8	0.0	0.0
bAM2	c31969	11.5	9.1	44.0	40.6	0.1	1.8	0.0	4.9
CAS	c83509	21.8	75.5	45.2	34.8	1.1	1.7	43.8	2.6
CDS	c102303	103.3	1.0	0.5	0.3	0.3	14.3	140.9	6.0

Table S15. Subcellular localization predictions of the mogroside enzymes based on six localization prediction algorithms. The best hit from each program is presented. The classifiers used by each program are listed below. The references for the programs are listed as supplemental references 17-22.

Protein	Prediction Program					
	BaCello	**ProteinProwler**	**Predotar**	**TargetP**	**Psort**	**Cello**
Squalene synthase	Chloroplast	Other	none	other	plasma membrane	Plasma Membrane
SQE1	Secretory	Secretory Pathway	ER	Secretory Pathway	ER (membrane)	Plasma Membrane
SQE2	Chloroplast	Secretory Pathway	ER	Secretory Pathway	ER (membrane)	Plasma Membrane
CDS	Nucleus	Other	none	other	microbody (peroxi)	Lysosomal
EPH1	Cytoplasm	Other	none	other	microbody (peroxi)	Cytoplasmic
EPH2	Cytoplasm	Other	none	other	microbody (peroxi)	Cytoplasmic
EPH3	Chloroplast	Other	none	other	microbody (peroxi)	Plasma Membrane
EPH4	Cytoplasm	Other	none	other	ER (membrane)	Cytoplasmic
CYP87D18	Cytoplasm	Secretory Pathway	ER	Secretory Pathway	ER (membrane)	Mitochondrial
UGT269.1	Secretory	Other	ER	Secretory Pathway	microbody (peroxi)	Chloroplast
UGT289.1	Chloroplast	mTP	ER	other	microbody (peroxi)	Chloroplast

Available classifiers:

BaCello: secretory pathway (SP), cytoplasm, nucleus, mitochondrion (mTP) and chloroplast (cTP).

ProteinProwler: SP, mTP, cTP, other

Predotar: Mito, Plastid, ER, Elsewhere

TargetP: cTP, mTP, SP, other

Psort: plasma membrane, ER (membrane), ER (lumen), microbody (peroxi), Chloroplast thylakoid membrane, Golgi, mitochondrial inner membrane, mitochondrial matrix space, lysozome (lumen), cytoplasm

Cello: PlasmaMembrane, Lysosomal, Cytoplasmic, Chloroplast, Mitochondrial, Peroxisomal, ER, Extracellular, Vacuole, Golgi, Nuclear, Cytoskeletal