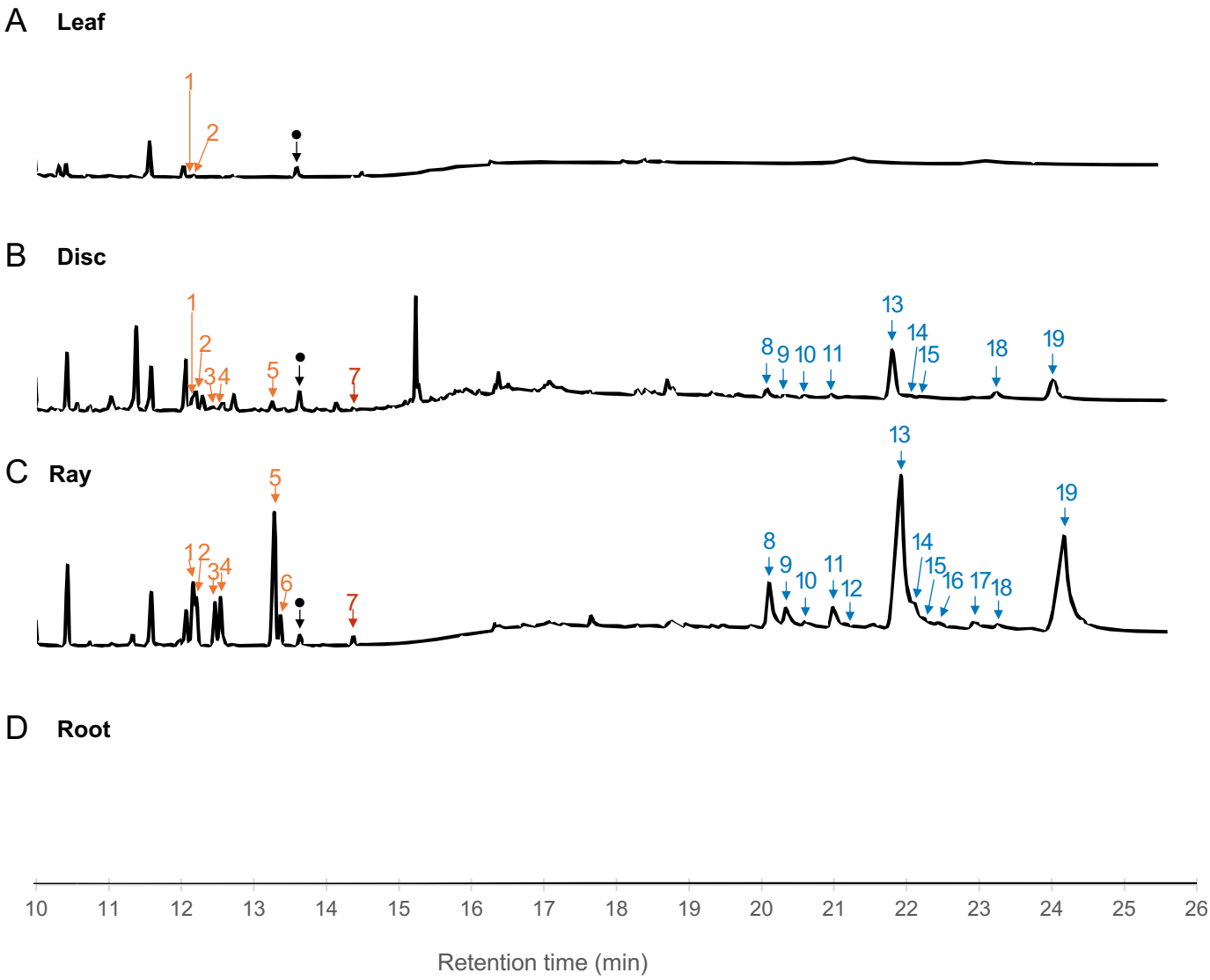


Triterpene monols

1. β-amyrin
2. isofucosterol
3. α-amyrin
4. lupeol
5. ψ-taraxasterol
6. taraxasterol

- friedelin (internal standard)

Triterpene diols

7. faradiol

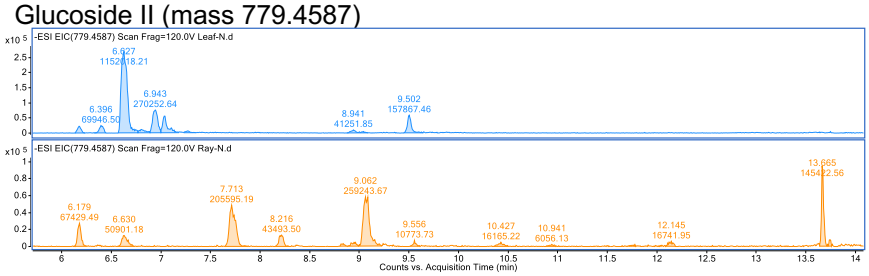
Triterpene fatty acid esters

8. faradiol/arnidiol laurate
9. β-amyrin myristate
10. maniladiol myristate
11. calenduladiol myristate
12. ψ-taraxasterol/taraxasterol myristate
13. faradiol/arnidiol myristate
14. β-amyrin palmitate
15. lupeol palmitate
16. manaladiol palmitate
17. calenduladiol palmitate
18. ψ-taraxasterol/taraxasterol palmitate
19. faradiol/arnidiol palmitate

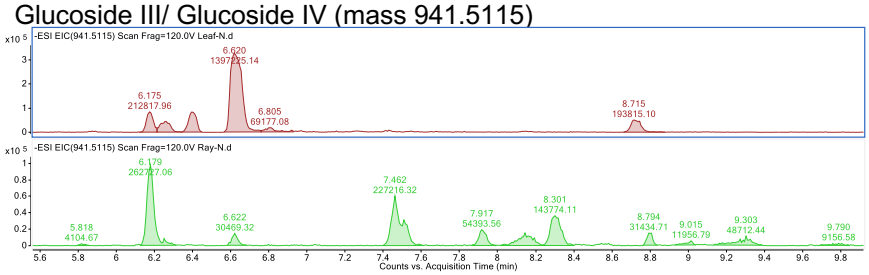
Supplementary Figure 1. Metabolite analysis of *Calendula officinalis* by GC-MS. A. Total Ion Chromatogram (TIC) of derivatised leaf extract. **B.** TIC of derivatised disc floret extract. **C.** TIC of derivatised ray floret extract. **D.** TIC of derivatised root extract. Labelled peaks = triterpene monols (orange), triterpene diols (red), triterpene fatty acid esters (blue) and internal standard (grey).

A **Glucosides** (numbered according to Szakiel et al. (2005) DOI 10.1007/s11101-005-4053-9)

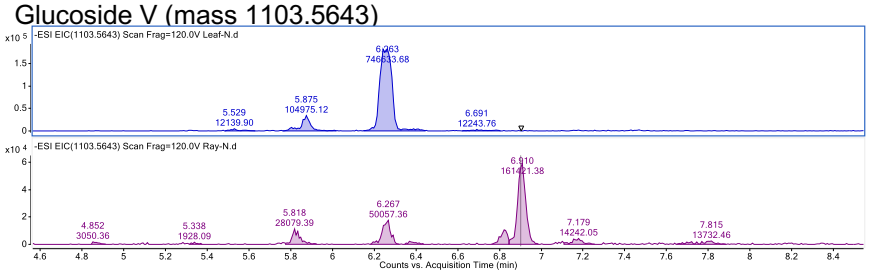
Glucoside II (mass 779.4587)



Glucoside III/ Glucoside IV (mass 941.5115)

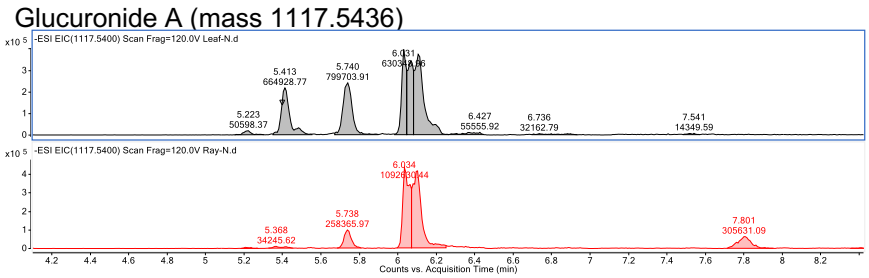


Glucoside V (mass 1103.5643)

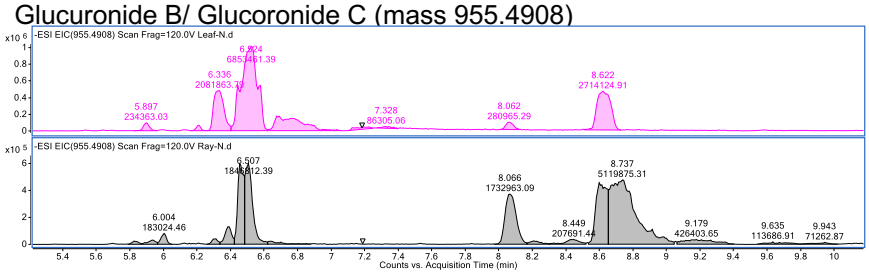


B **Glucoronides** (numbered according to Szakiel et al. (2005) DOI 10.1007/s11101-005-4053-9)

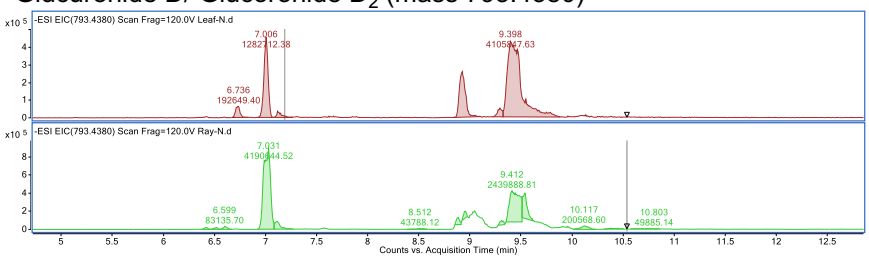
Glucuronide A (mass 1117.5436)



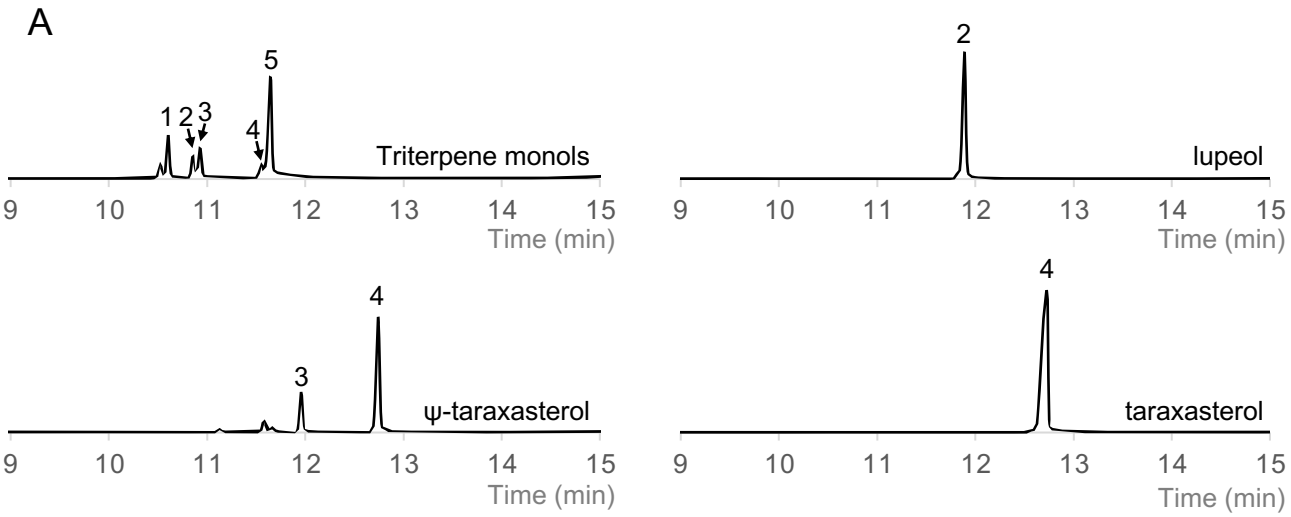
Glucuronide B/ Glucoronide C (mass 955.4908)

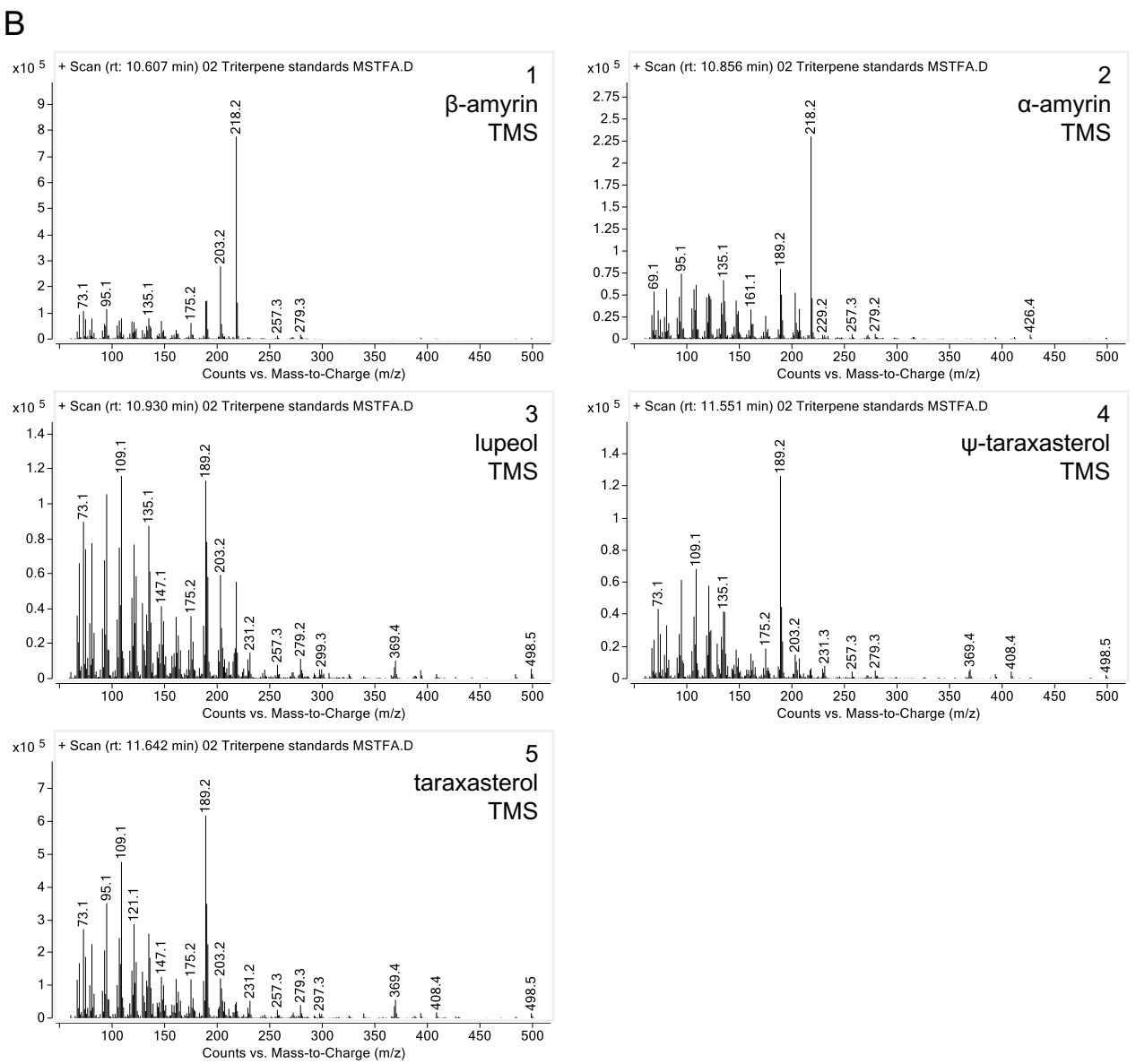


Glucuronide D/ Glucoronide D_2 (mass 793.4380)

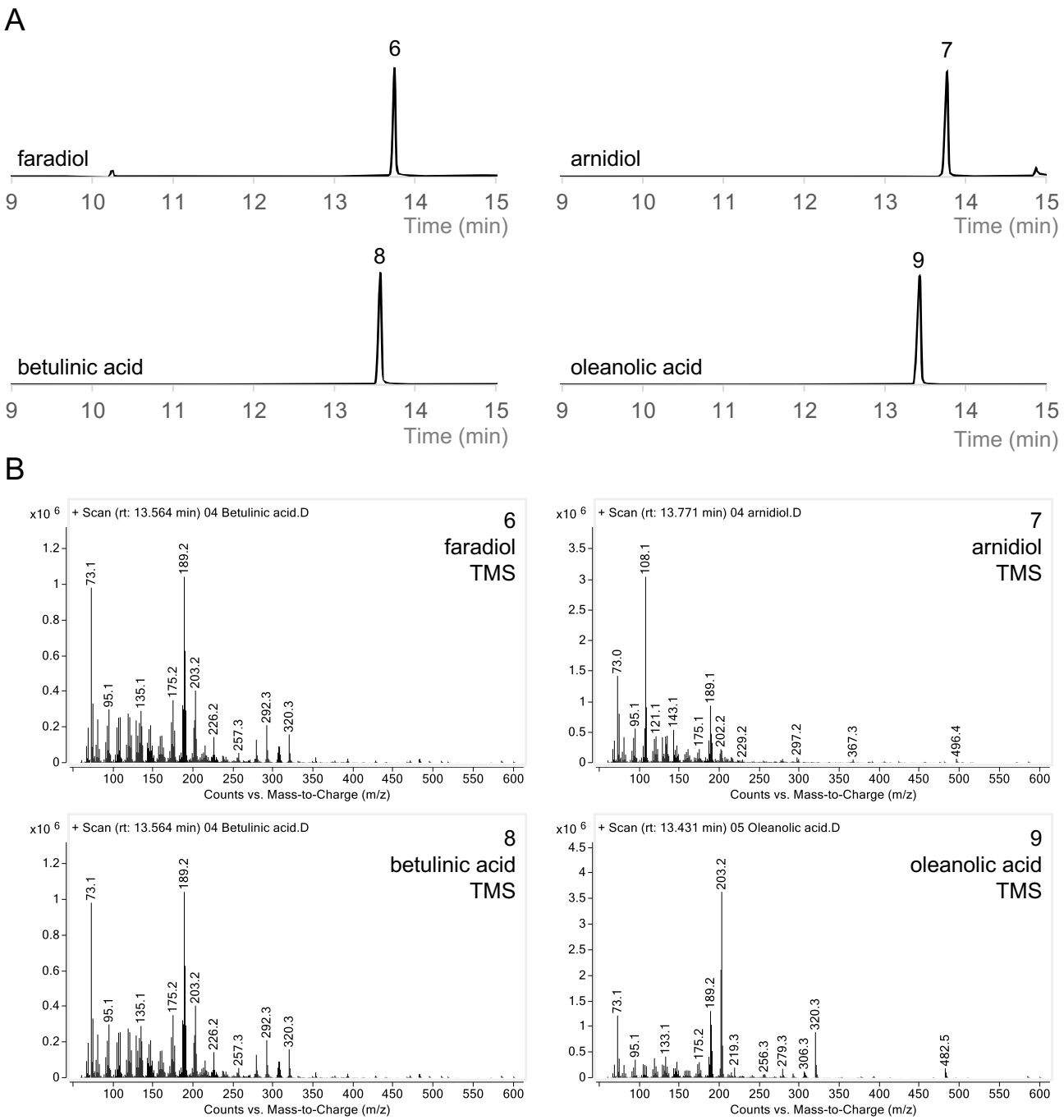


Supplementary Figure 2. Metabolite analysis of *Calendula officinalis* by LC-MS
A. Extracted ion chromatograms (EIC) of *Calendula officinalis* leaf and ray extracts corresponding to masses of oleanolic acid glucosides. **B.** Extracted ion chromatograms (EIC) of *Calendula officinalis* leaf and ray extracts corresponding to masses of oleanolic acid glucuronides. Compounds are numbered according to Szakiel et al. (2005) DOI 10.1007/s11101-005-4053-9.

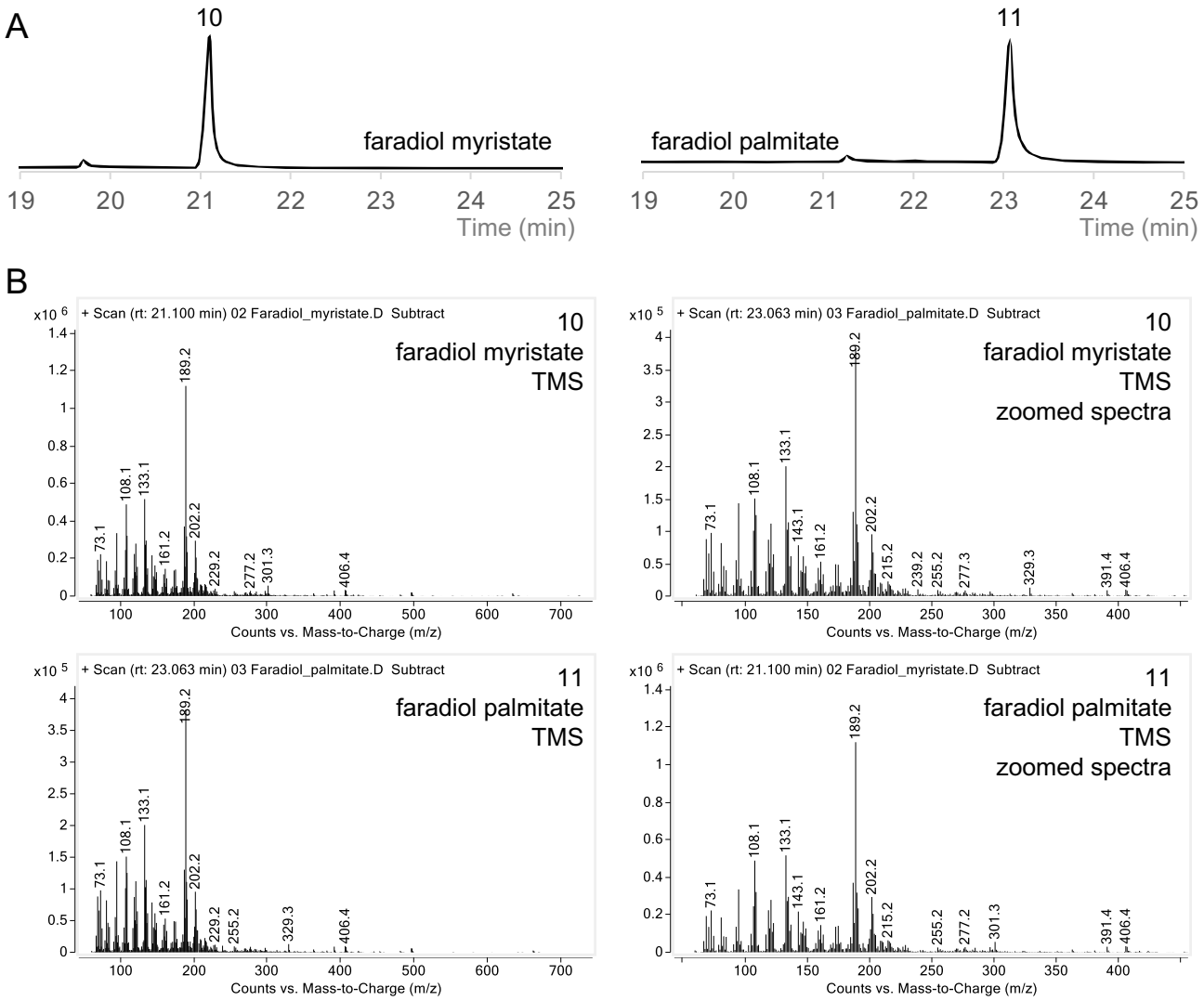




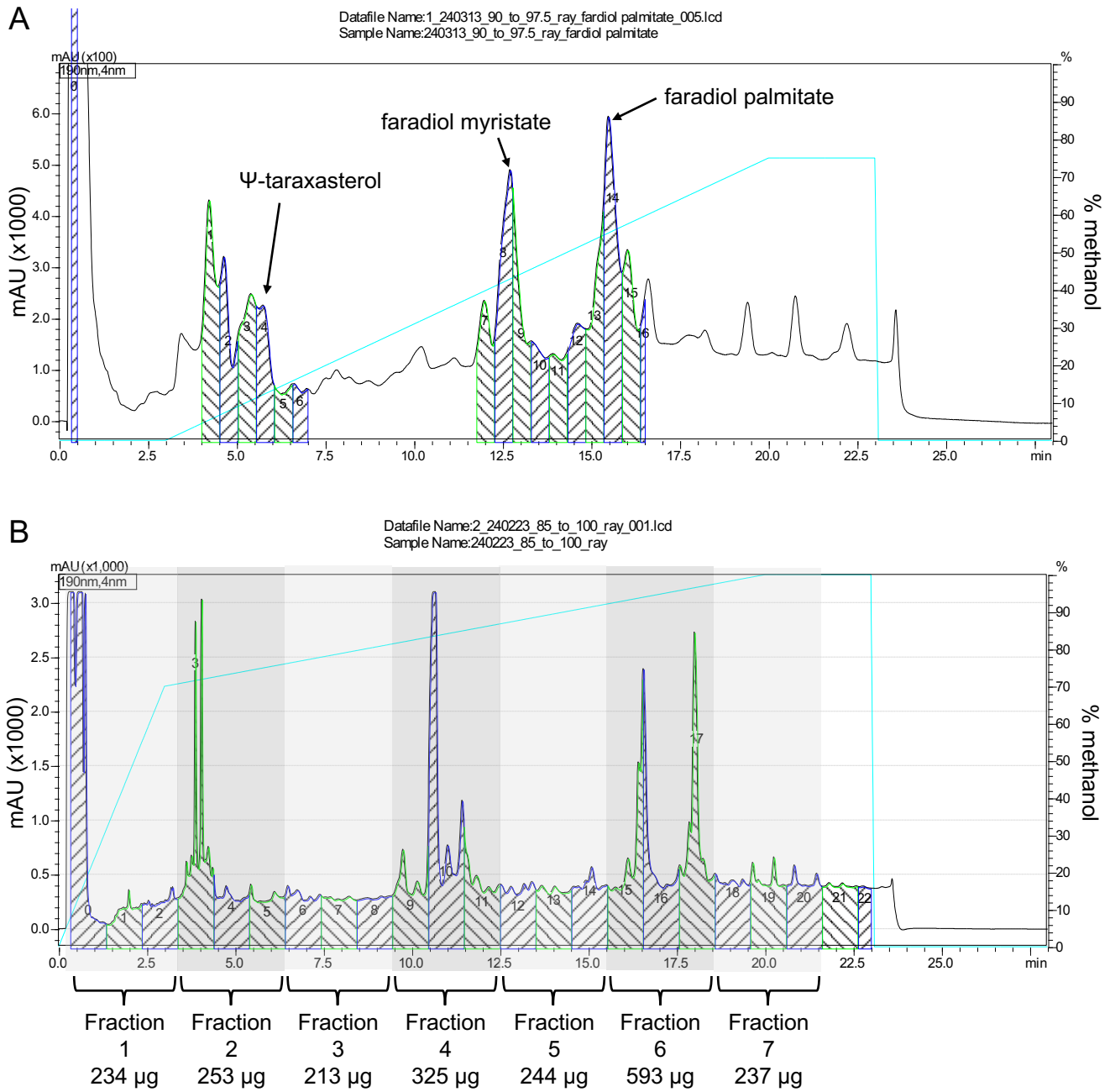
Supplementary Figure 3. GC-MS standards and mass spectra for triterpene monols. A. GC-MS chromatograms of trimethylsilyl (TMS) derivatised triterpene monols. Triterpene monols is a mixture of 1. β-amyrin, 2. α-amyrin, 3. lupeol, 4. ψ-taraxasterol, 5. taraxasterol. β-amyrin, α-amyrin and lupeol were kindly provided by Anne Osbourn, John Innes Centre, Norwich, UK; taraxasterol were purchased from Merck, Darmstadt, Germany (PHL84272-10MG). ψ-taraxasterol was purified from *Calendula officinalis* (Supplementary Figures 6-8) **B.** Mass spectra for each of the TMS derivatised triterpene monols.



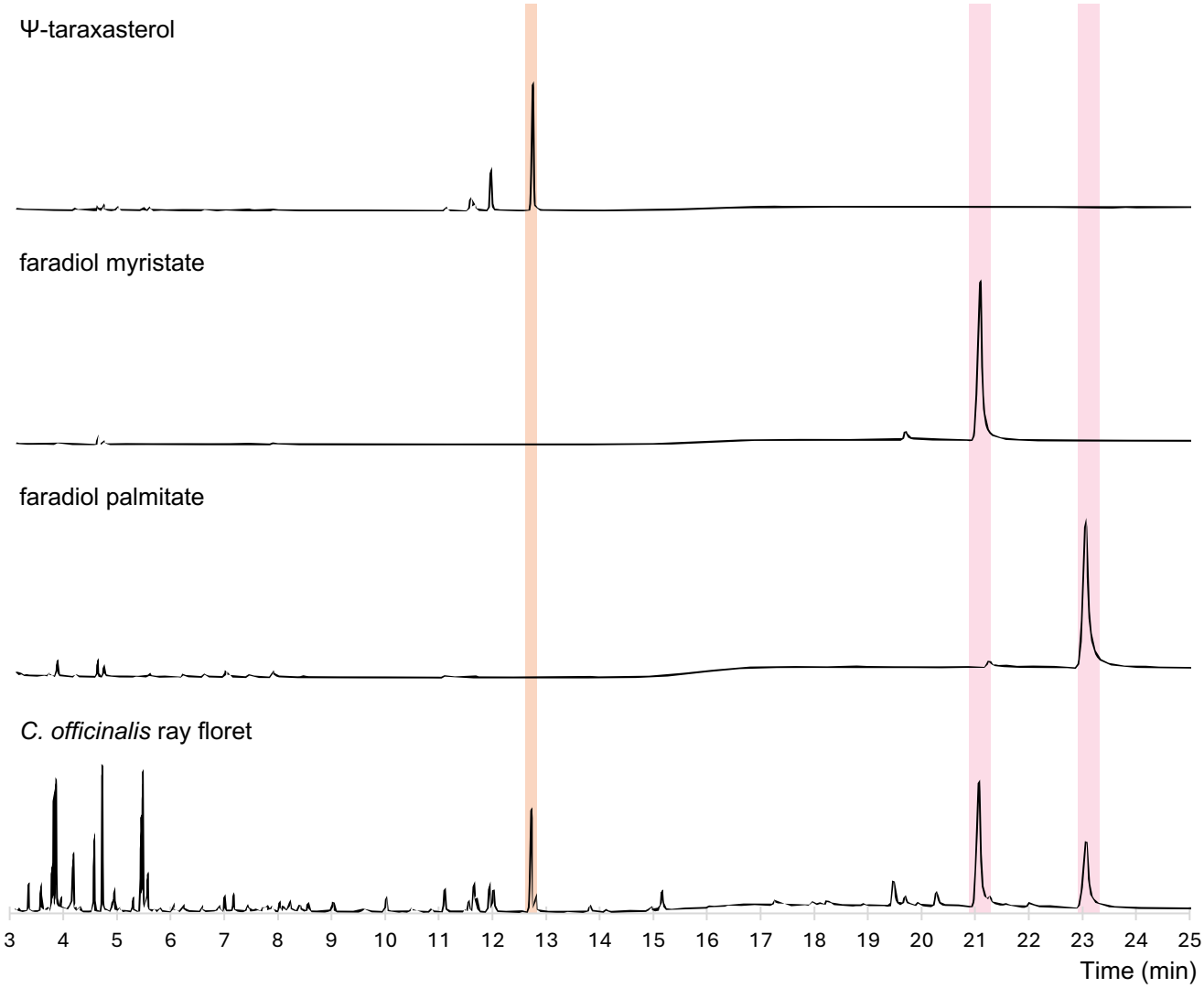
Supplementary Figure 4. GC-MS standards and mass spectra for triterpene diols and acids. A. GC-MS chromatograms of TMS derivatised triterpene diol and acids. Faradiol, betulinic acid and oleanolic acid were purchased from Merck, Darmstadt, Germany (PHL82536-10MG, 91466-10MG and 42515-10MG respectively). Arnidiol was purchased from MedChem Express, New Jersey, USA (HY-N4165-1mg) **B.** Mass spectra for each of the TMS derivatised triterpene diols and acids.



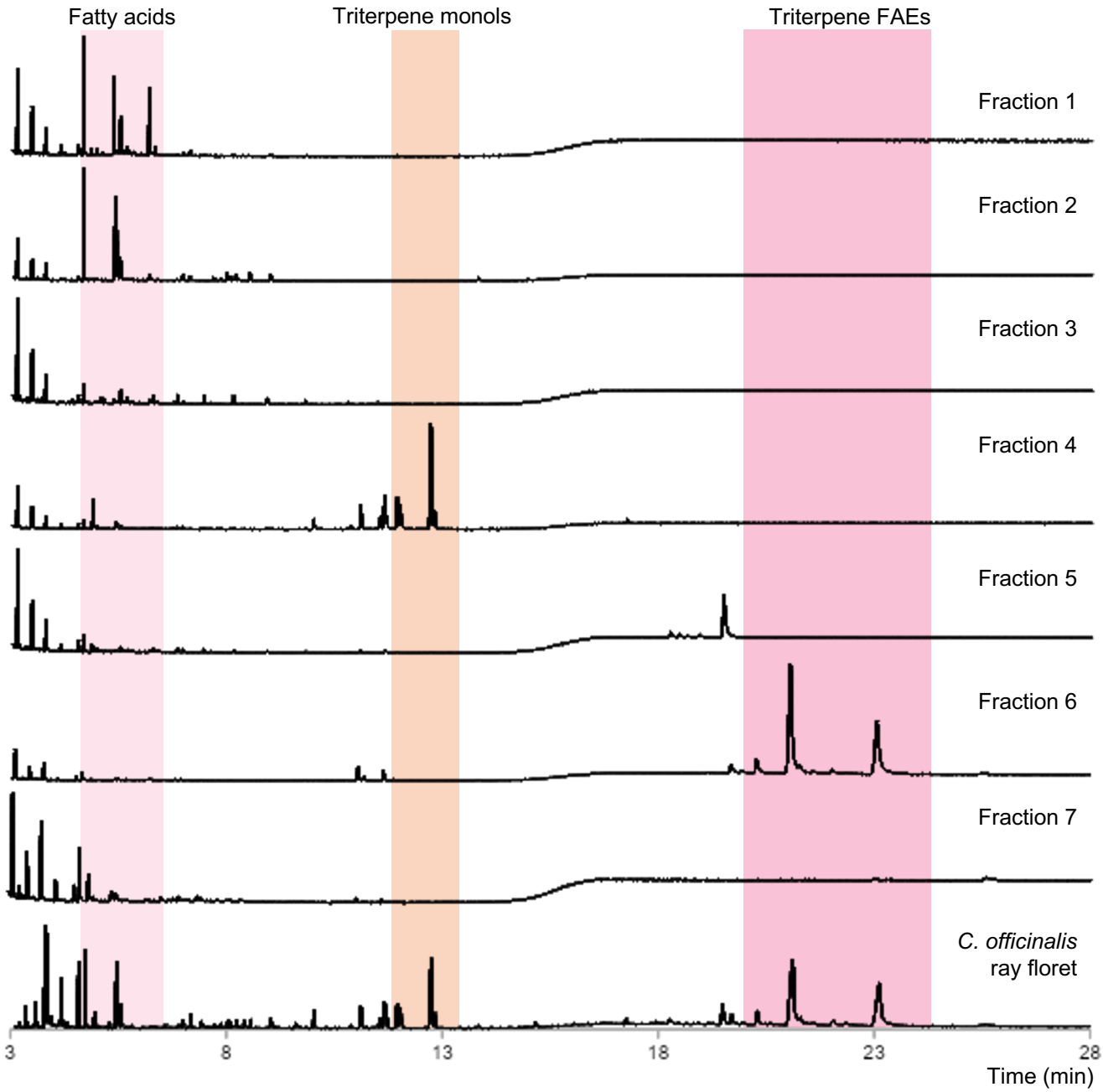
Supplementary Figure 5. GC-MS standards and mass spectra for faradiol fatty acid esters. A. GC-MS chromatograms of TMS derivatised faradiol fatty acid esters. Faradiol palmitate and faradiol myristate were purified from *Calendula officinalis* flowers and verified by NMR (Supplementary Figures 6-14; Supplementary Data 1-2) **B.** Mass spectra for each of the TMS derivatised faradiol fatty acid esters.



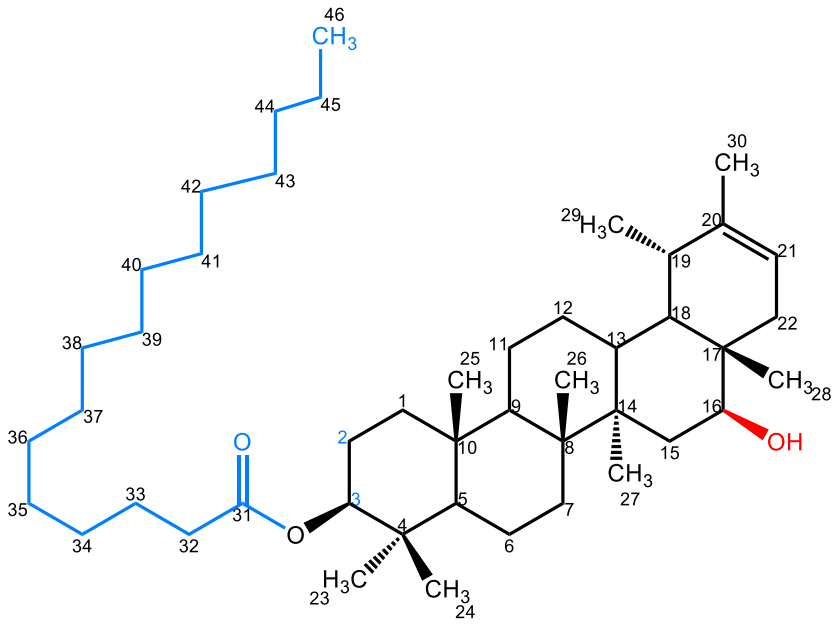
Supplementary Figure 6. Semi-preparative uHPLC chromatograms of methanol extracts of *Calendula officinalis* ray florets. The methanol gradient used for separation is shown as a cyan line. **A - Fractionation for purification.** Selected regions were collected as fractions (as indicated in green and blue shading). Multiple runs were performed and fractions corresponding to compounds of interest were pooled, verified by GC-MS analysis and dried down for NMR confirmation. In this run, the following fractions were taken for the compounds of interest: Ψ-taraxasterol (fraction 4); faradiol myristate (fractions 8 and 9); and faradiol palmitate (fraction 14). **B - Fractionation for bioassays.** Selected regions were collected as fractions (as indicated in green and blue shading). A total of 8 runs were performed and fractions were pooled into seven groups of three fractions and dried down for use in bioassays. The yield of each pooled fraction is also shown.



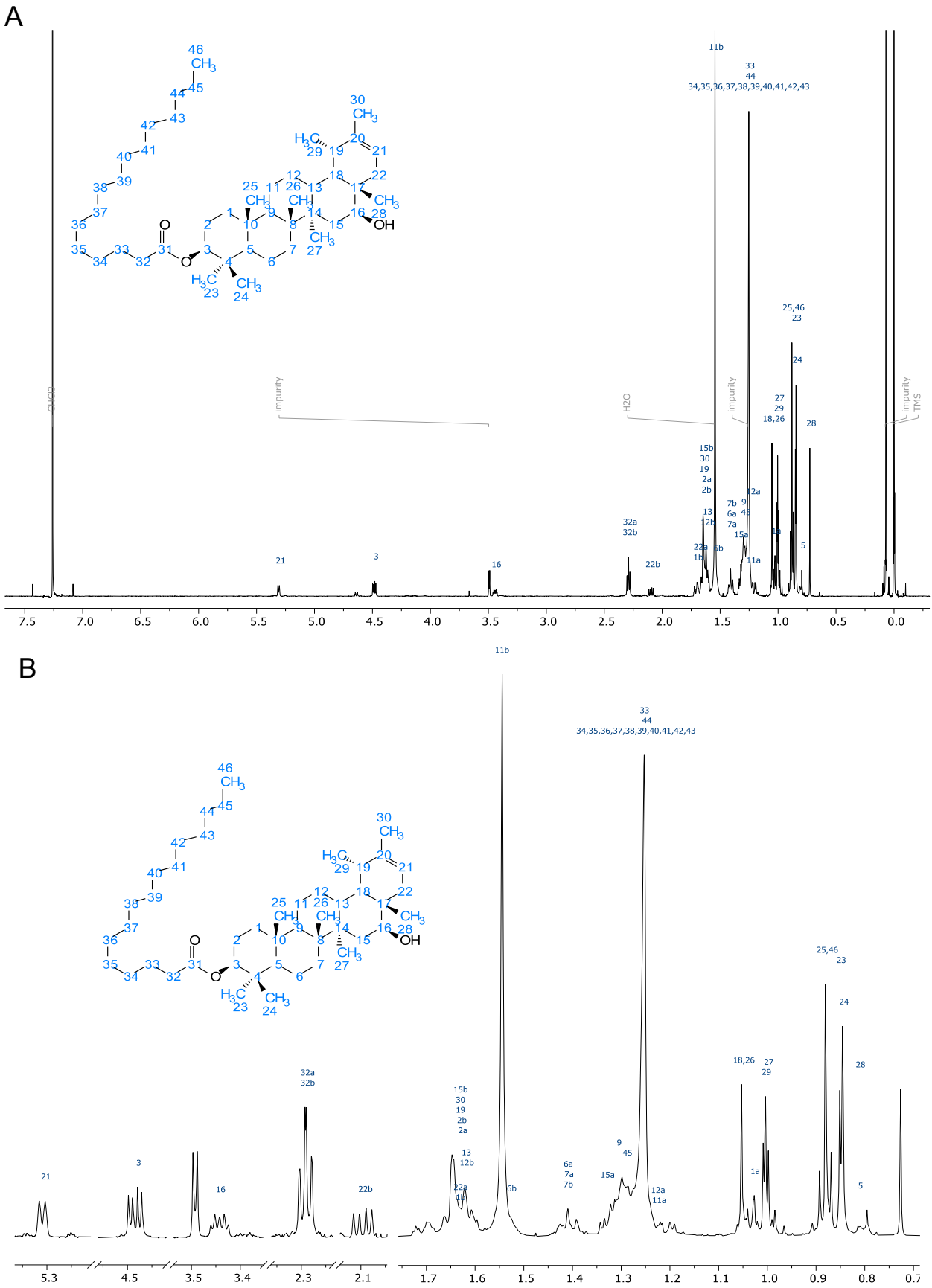
Supplementary Figure 7. GC-MS analysis of compounds purified from *Calendula officinalis* ray floret extracts. GC-MS chromatograms of TMS derivatised pooled fractions from purification of triterpene compounds from *C. officinalis* ray floret extract. The ψ-taraxasterol fraction contains a mixture of ψ-taraxasterol (62%) and α-amyrin (19%). The purity of the faradiol myristate and faradiol palmitate fractions is 96% and 98% respectively. Triterpene monols were identified by comparing retention times and mass fragmentation patterns with those of authentic standards (Supplementary Figures 3-5). Faradiol palmitate was confirmed by NMR analysis (Supplementary Figures 9-14; Supplementary Data 1-2).



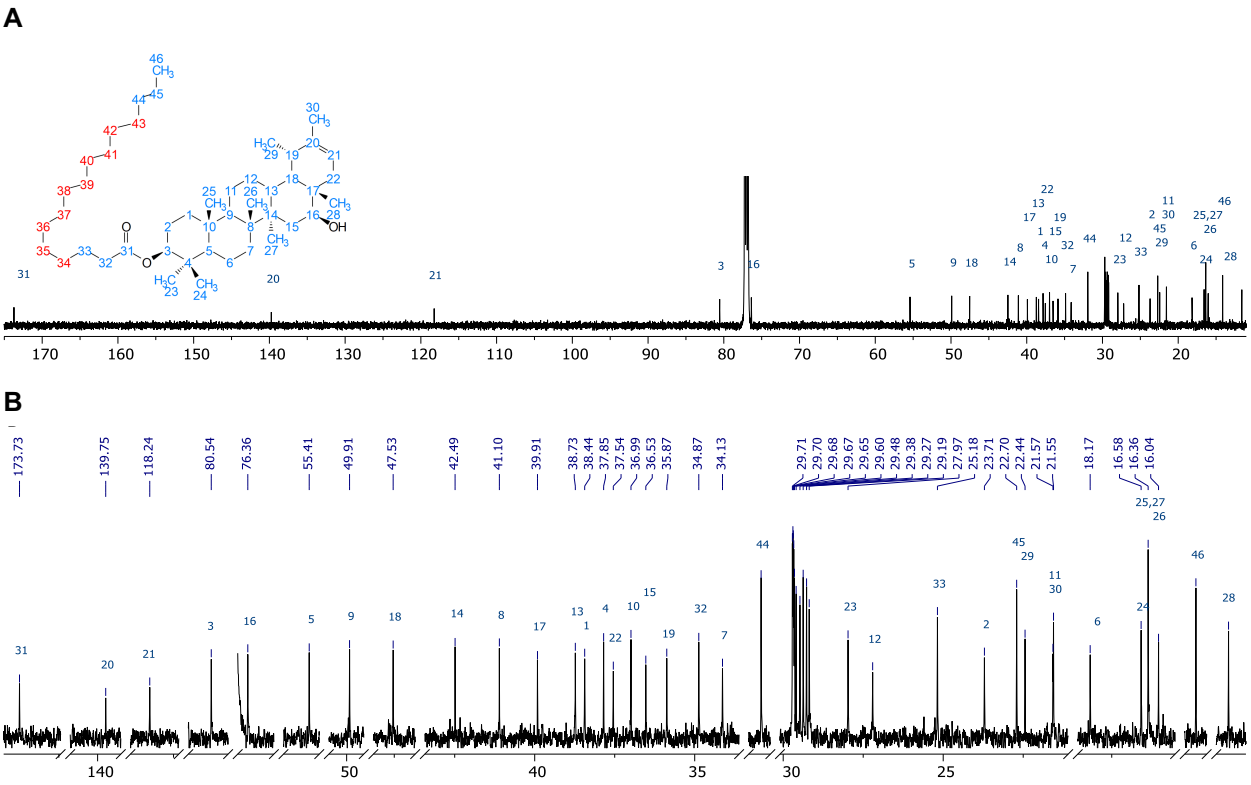
Supplementary Figure 8. GC-MS analysis of *Calendula officinalis* ray floret fractions. GC-MS chromatograms of TMS derivatised pooled fractions from fractionation of the *C. officinalis* ray floret extract. Fatty acids were found in fractions 1 and 2; triterpene monols in fraction 4 and triterpene fatty acid esters in fractions 5 and 6. Triterpene monols and diols were identified by comparing retention times and mass fragmentation patterns with those of authentic standards (Supplementary Figures 3-5).



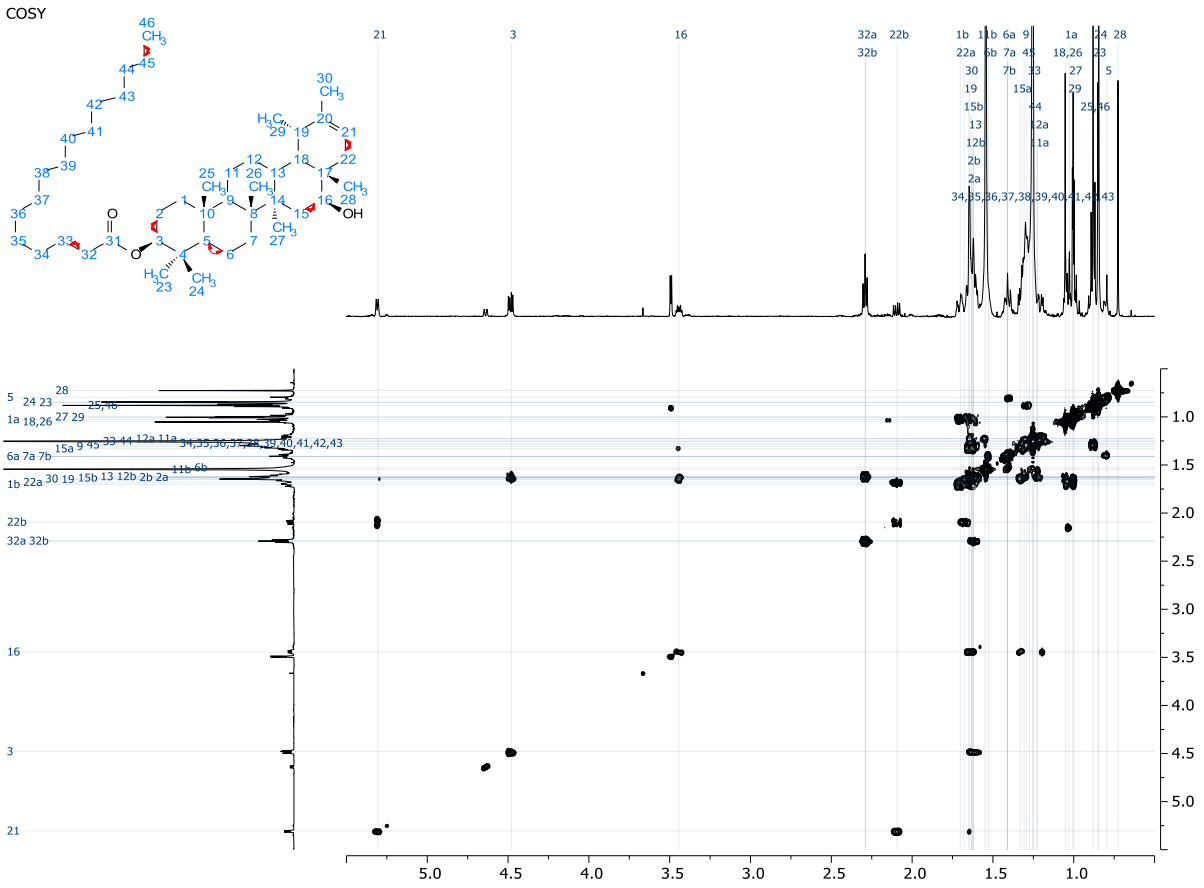
Supplementary Figure 9. Structure of 3-*O*-palmitoyl faradiol (faradiol palmitate)
The structure of purified 3-O-palmitoyl faradiol was determined by 1D and 2D NMR -1H, 13 C, COSY, HMQC, and HMBC. The 1H and 13C NMR spectra of 3-O-palmitoyl faradiol were in a good agreement with spectra published by Ezzat et al (2017) Nat Prod Res. 31(6):676-680. Using HMBC spectra it was possible to identify the positions of C31, C32, C45 and C46 of palmitoyl residue that were not previously assigned. Furthermore, cross peaks of 4.48 ppm (in proton spectra dimension) with 173.7 (C31), 37.8 (C-4), 28.0 (C23) and 16.6 (C24) in HMBC spectra indicate that the signal at 4.48 ppm belongs to the resonance of H3 thus establishing the position 3 as a position of esterification of faradiol. In 1H, the NMR spectra resonance of methine proton H3 (4.48 ppm) appears 1 ppm downfield compared to the resonance of H16 (3.44 ppm) of hydromethine group. This characteristic downfield shift due to esterification confirms the position of the palmitoyl attachment to O-3.



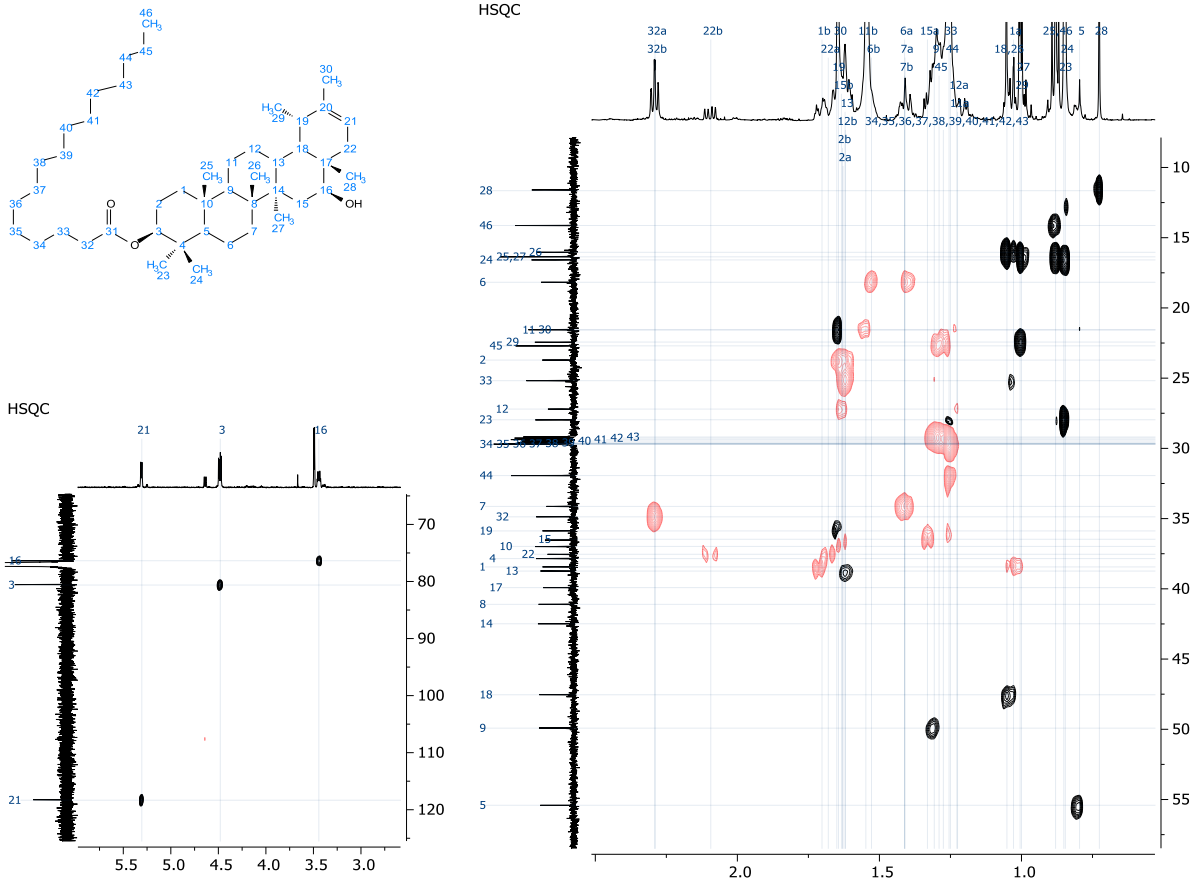
Supplementary Figure 10. ^{1}H NMR spectra of faradiol palmitate. A. Overview of ^{1}H NMR (600 MHz, CDCl3, 298 K) of 3-O-palmitoyl faradiol. **B.** Expansion of 3-O-palmitoyl faradiol resonances in ^{1}H NMR (600 MHz, CDCl3, 298 K).



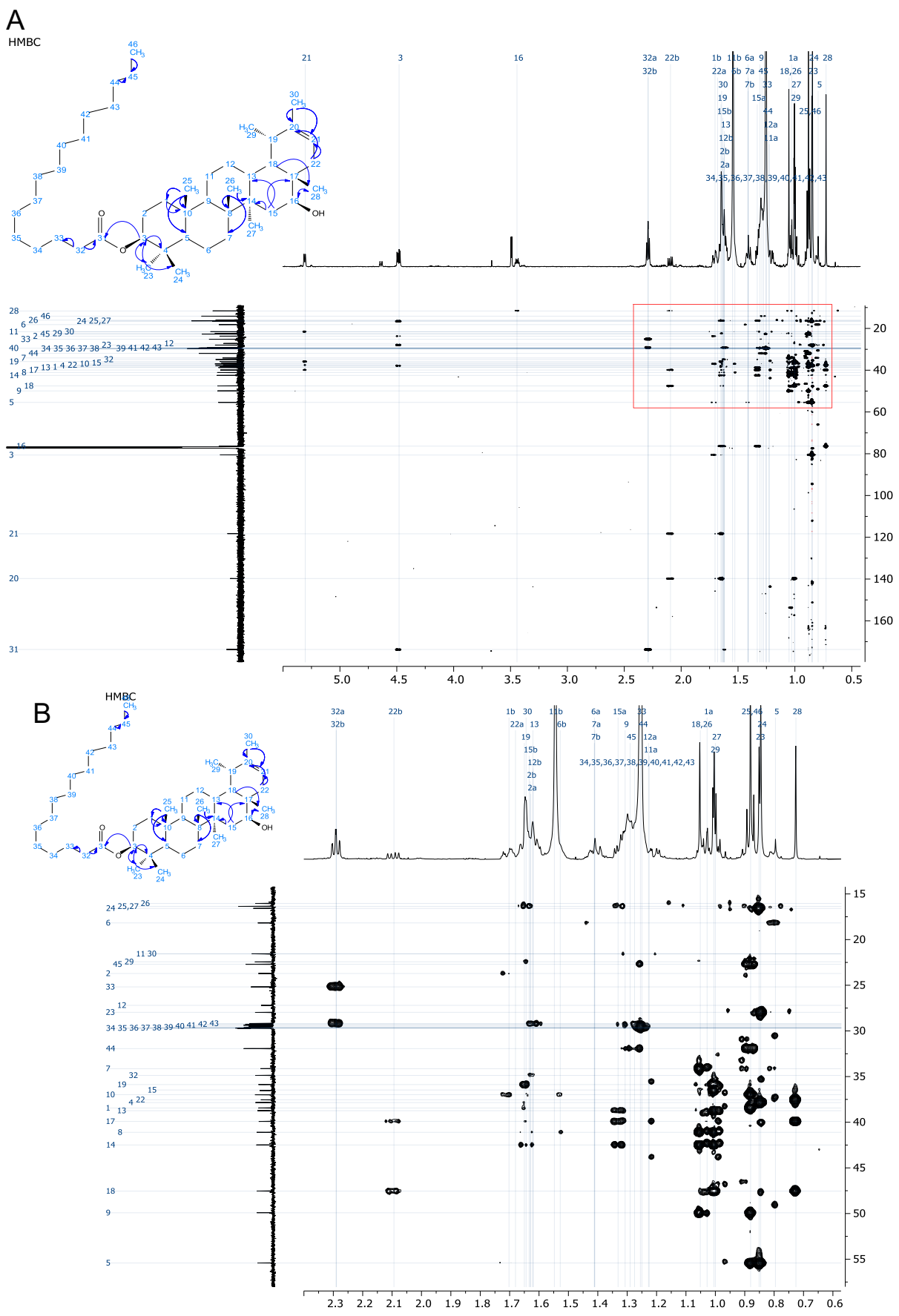
Supplementary Figure 11. ^{13}C NMR spectra of faradiol palmitate. A. Overview; **B.** Expansion.



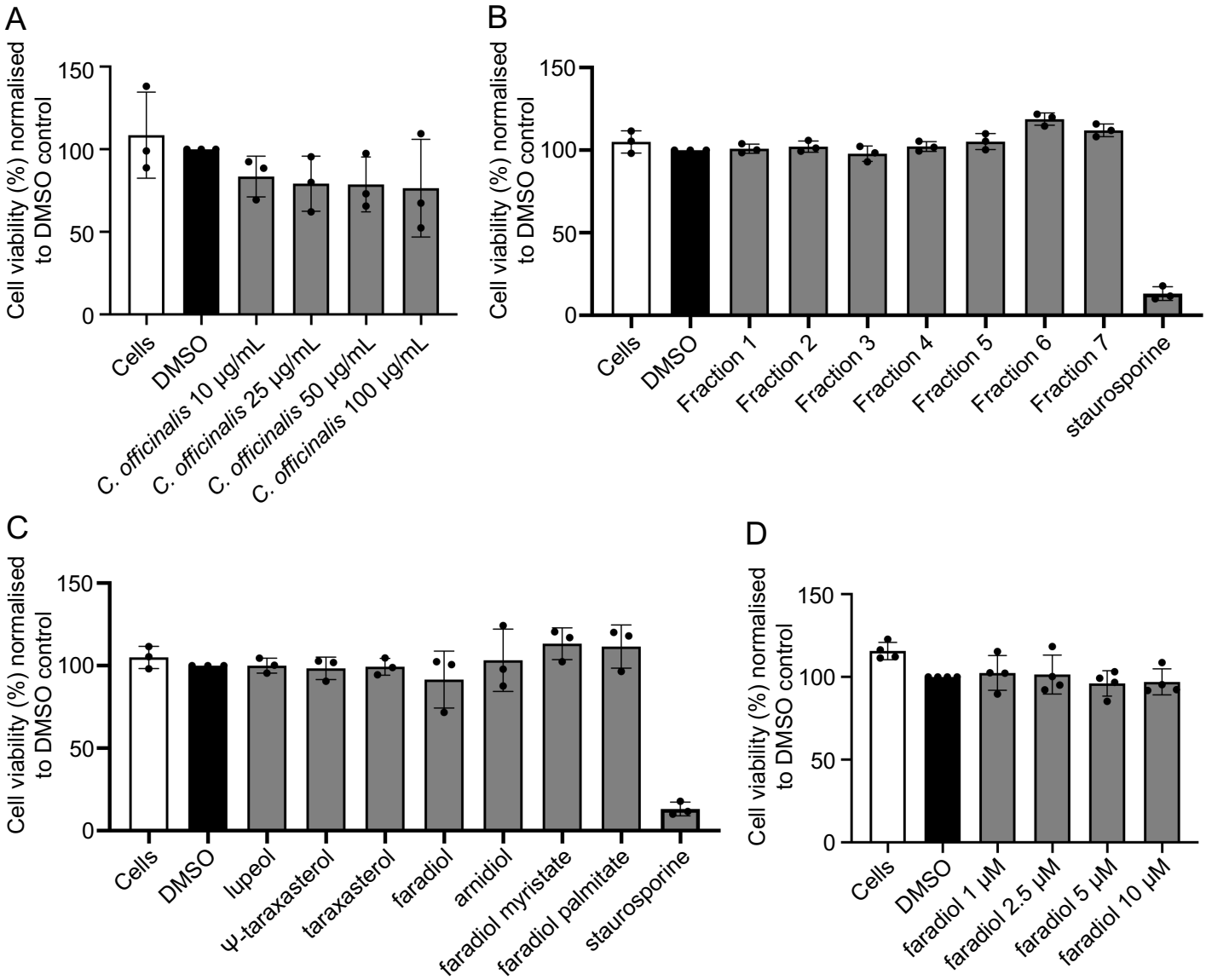
Supplementary Figure 12. ^{1}H-^{1}H COSY NMR (600 MHz, CDCl3, 298 K) of faradiol palmitate.
Only limited number of correlations can be found due to signals overlap.



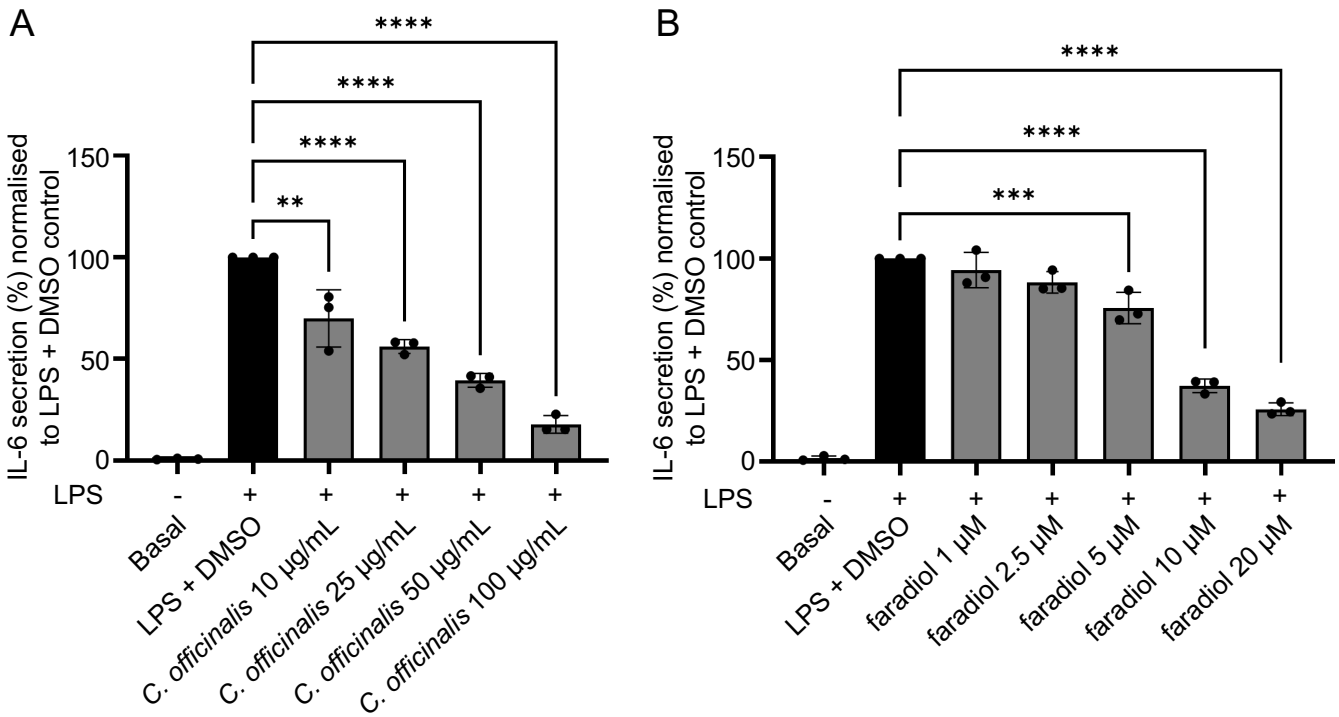
Supplementary Figure 13. ^{1}H-^{13}C-HSQC-edited NMR (600 MHz, CDCl3, 298 K) of faradiol palmitate.



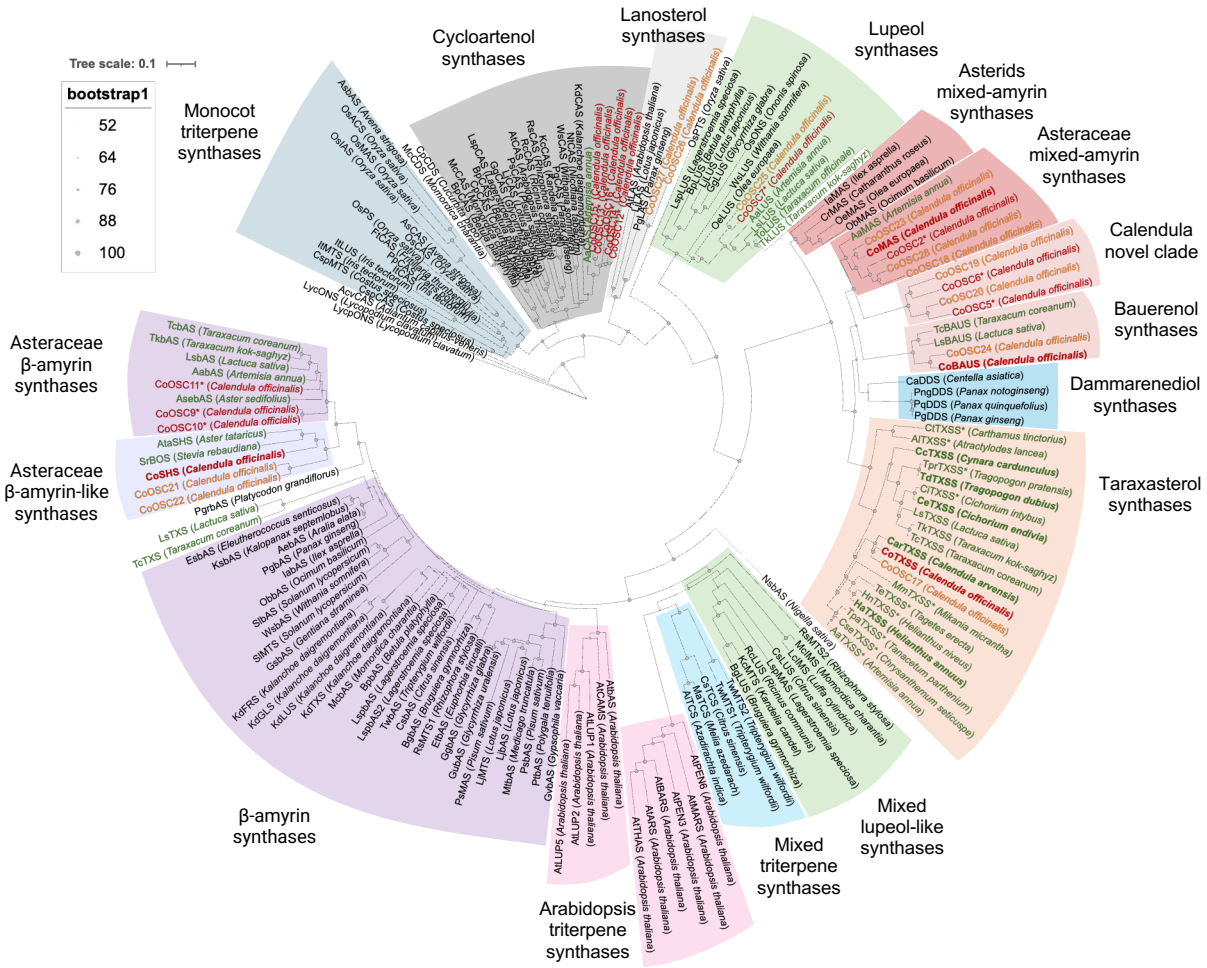
Supplementary Figure 14. ^{1}H-^{13}C HMBC NMR (600 MHz, CDCl3, 298 K) of faradiol palmitate. A. Overview **B.** Expansion of area highlighted in A. Curly arrows indicate observed HMBC interactions.



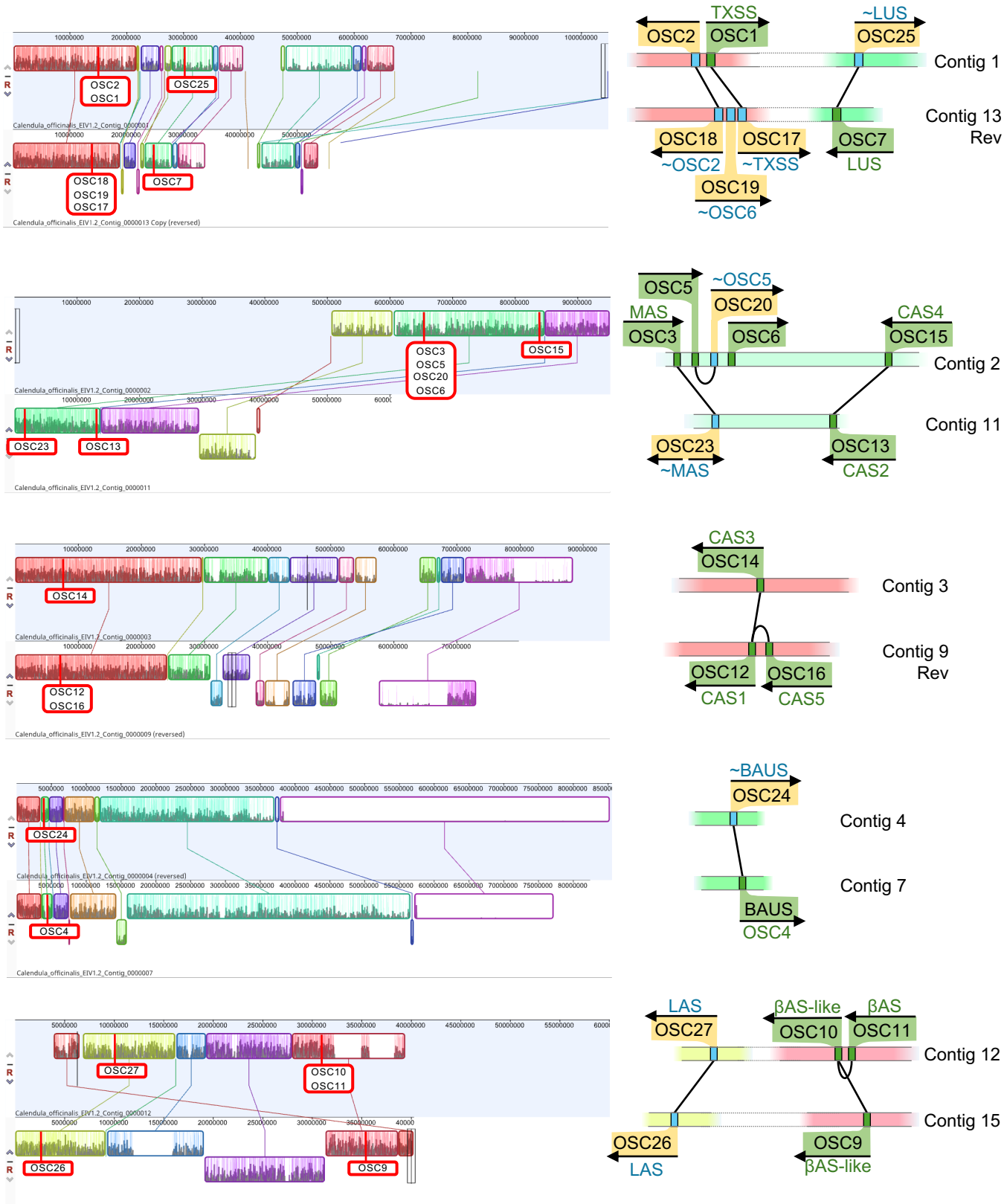
Supplementary Figure 15. Effects of Calendula officinalis (pot marigold) extracts and triterpenoids on the viability of human monocytic (THP-1) cells. A Cell viability following application of pot marigold floral extracts at different concentrations. **B** Cell viability following application of extract fractions. C Cell viability following application of triterpenoid compounds (20 µM). D Cell viability following application of faradiol at different concentrations.



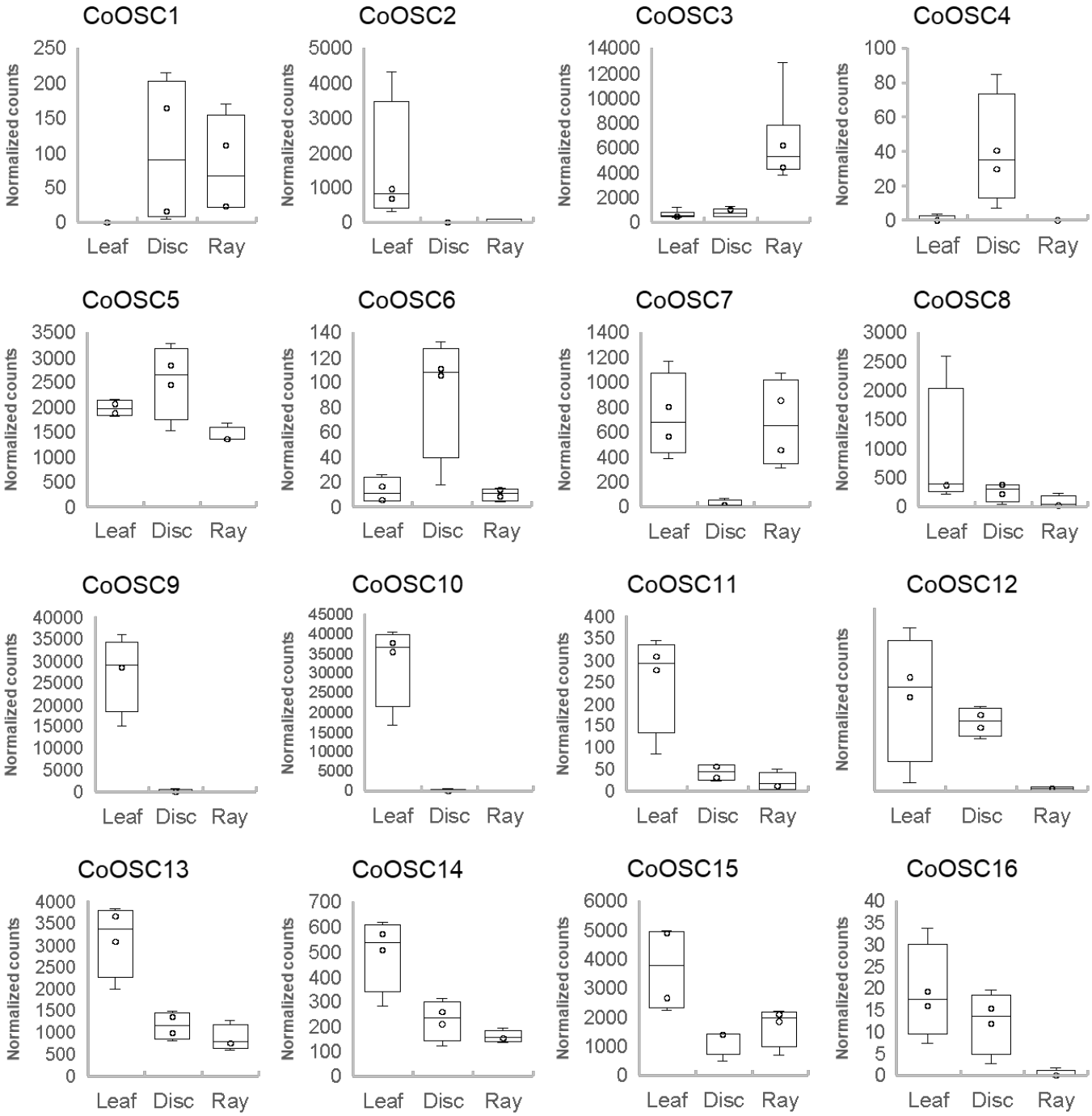
Supplementary Figure 16. Concentration-dependent responses to floral extracts and faradiol. Concentration-dependent anti-inflammatory effect of: **A.** Calendula officinalis (pot marigold) floral extracts and **B.** Faradiol , on the release of IL-6 in (LPS)-activated human monocytic (THP-1) cells.



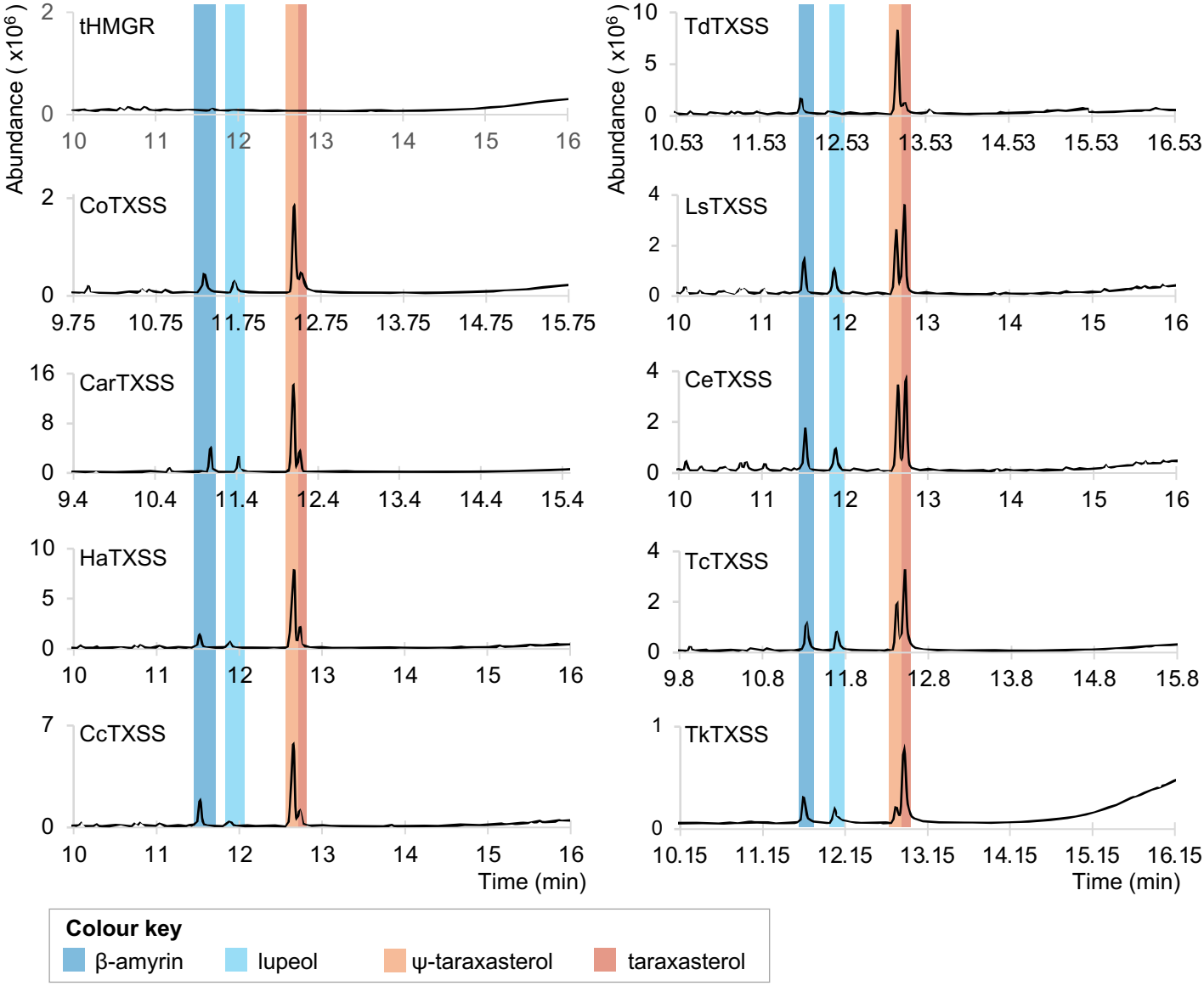
Supplementary Figure 17. Maximum-likelihood tree of plant oxidosqualene cyclases (OSCs).
Functionally characterised non-Asteraceae and Asteraceae plant OSCs are shown in black and green respectively. *Calendula officinalis* plant OSCs that are expressed are shown in red with likely pseudogenes shown in orange. Genes that were functionally characterised in this paper are shown in bold and any uncharacterised genes are indicated with an asterisk (*). The maximum likelihood tree was constructed in IQ-Tree using the JTT matrix-based model and visualised using Interactive Tree of Life (iTOL) v6. Filled grey circles indicate bootstrap supports for each node. Scale bar represents the number of substitutions per site. Clades are highlighted and labelled according to product specificity.



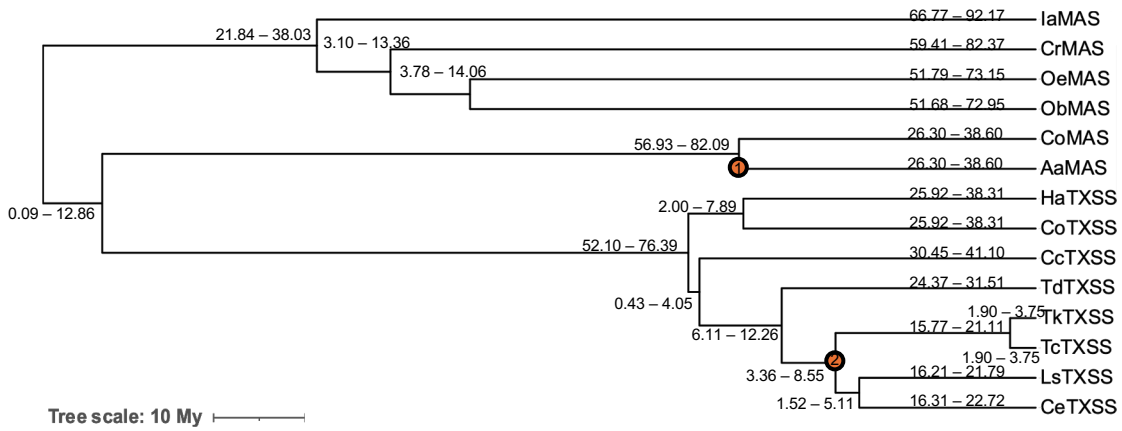
Supplementary Figure 18. Genomic locations of *Calendula officinalis* (pot marigold) *OSC* genes. A graphical representation of the relative position, orientation and similarity of *CoOSC* genes. Contigs are paired to show synteny between genes found on homeologous contigs. Genes that were expressed in the pot marigold transcriptome are shown in green. Those for which expression was not detected are shown in yellow and those not expressed and with mutations indicating pseudogenisation are prefixed "~". Functional annotations based on phylogenetic relationships to previously characterised OSCs: TXSS = taraxasterol synthase; βAS = β-amyrin synthase; LAS = Lanosterol synthase; BAUS = baurenol synthase; CAS = cycloartenol synthase; MAS = mixed amyrin synthase. LUS = lupeol synthase.



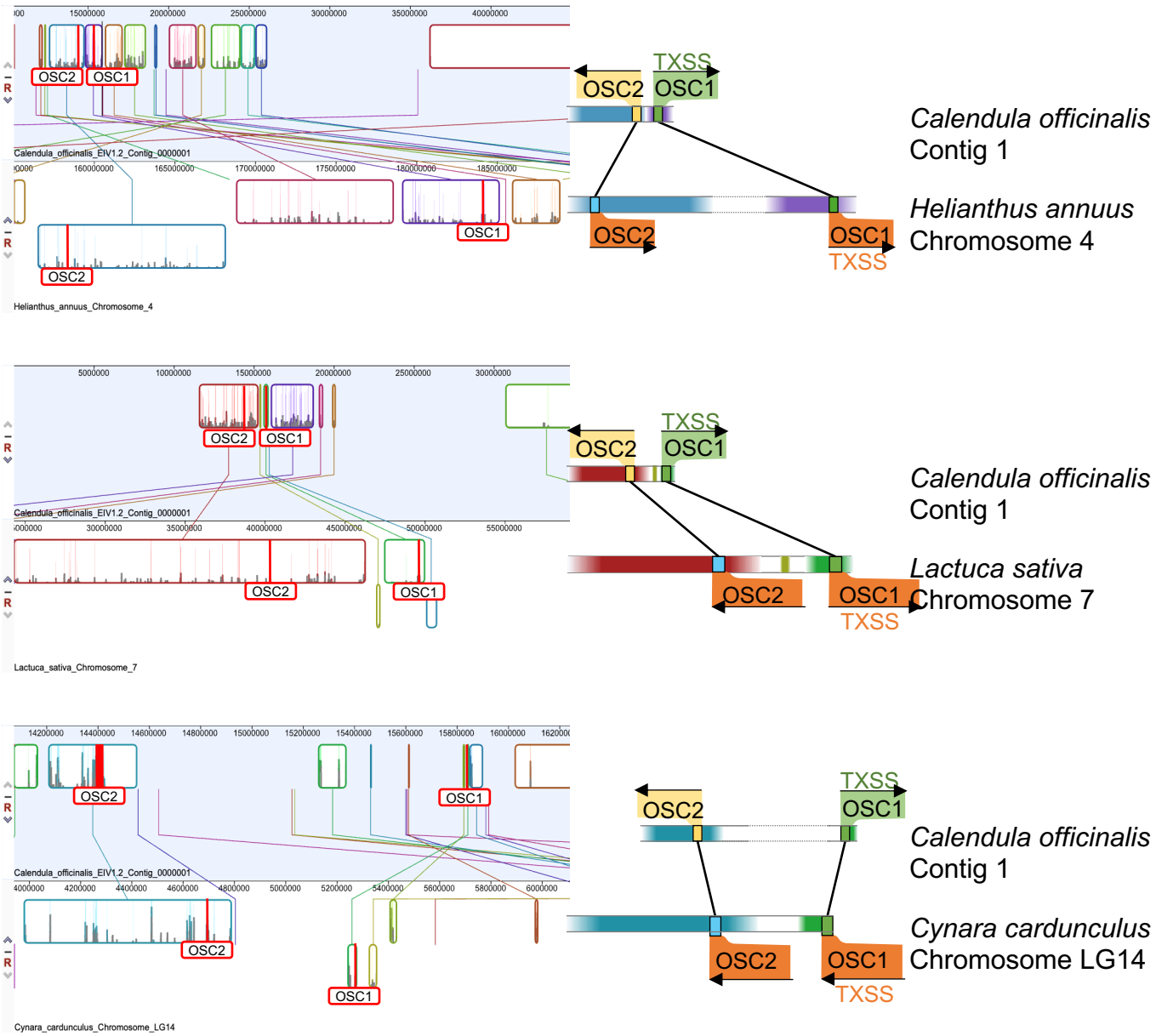
Supplementary Figure 19. Differential expression of *Calendula officinalis OSC* genes. Gene expression of 16 *OSC* genes expressed in leaf, disc and ray floret tissues. Gene expression for each tissue is displayed in a box plot as normalised counts following transcript quantification using Salmon v1.2.0 and differential gene expression analysis using DESeq2. The centre line of the boxplot denotes the median, the box denotes the 25th to 75th percentiles of the data and the whiskers show the minimum and maximum of the data.



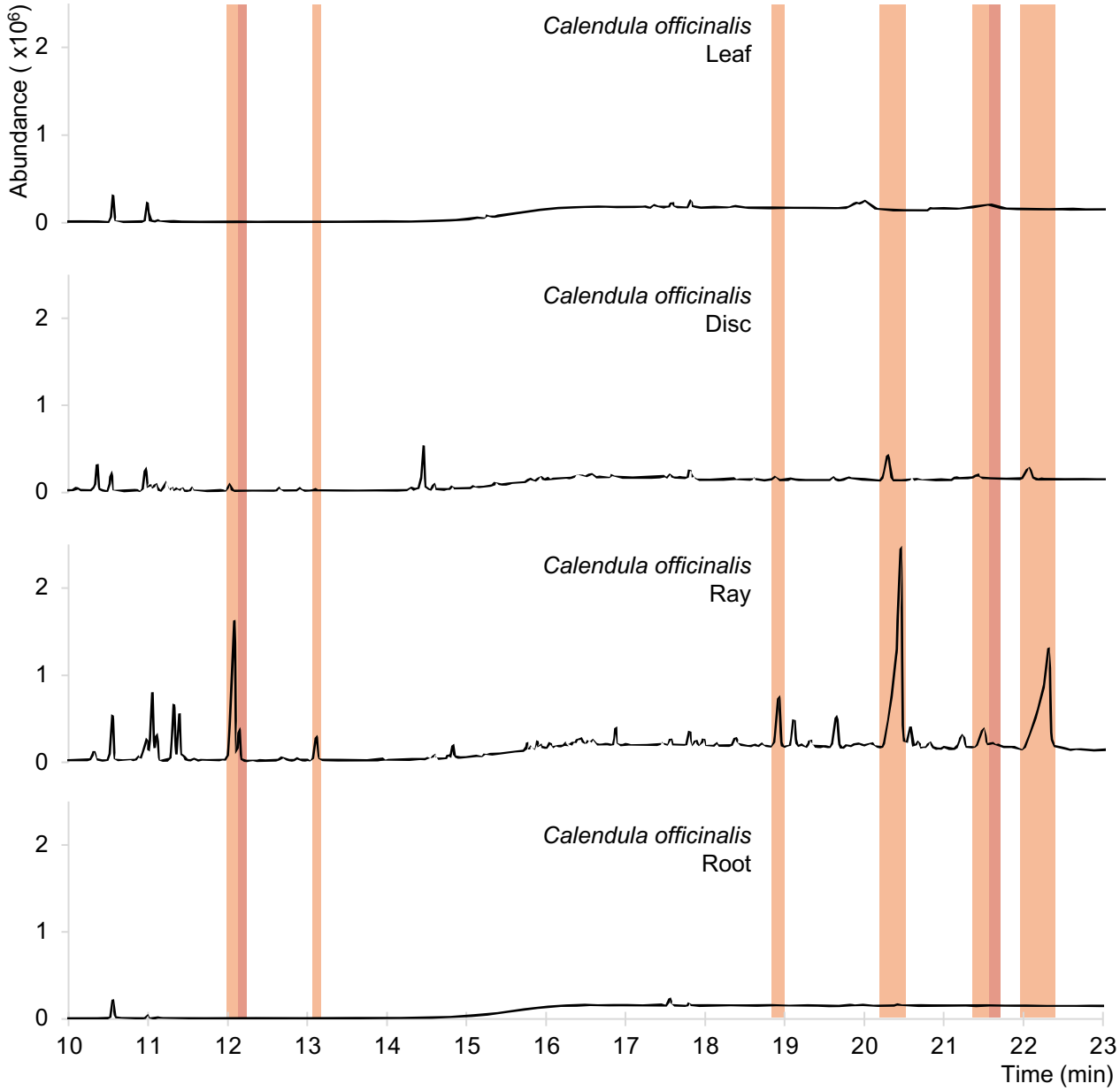
Supplementary Figure 20. GC-MS analysis of *Nicotiana benthamiana* expressing taraxasterol synthases (TXSS). Representative total ion chromatograms of ethyl acetate extracts of *Nicotiana benthamiana* leaves transiently expressing *TXSS* genes from Asteraceae species. *CoTXSS* and *CarTXSS* were mined from data generated for this manuscript; *HaTXSS, CcTXSS, TdTXSS* and *CeTXSS* were mined from publicly available datasets; *LsTXSS*, *TcTXSS* and *TkTXSS* were previously characterised. Key metabolites are highlighted and detailed in the colour key. *Co = Calendula officinalis; Car = Calendula arvensis; Ha = Helianthus annuus; Cc = Cynara cardunculus; Td = Tragopogon dubius; Ls = Lactuca sativa; Ce = Cichorium endivia; Tc = Taraxacum coreanum; TkTXSS = Taraxacum kok-saghyz*.



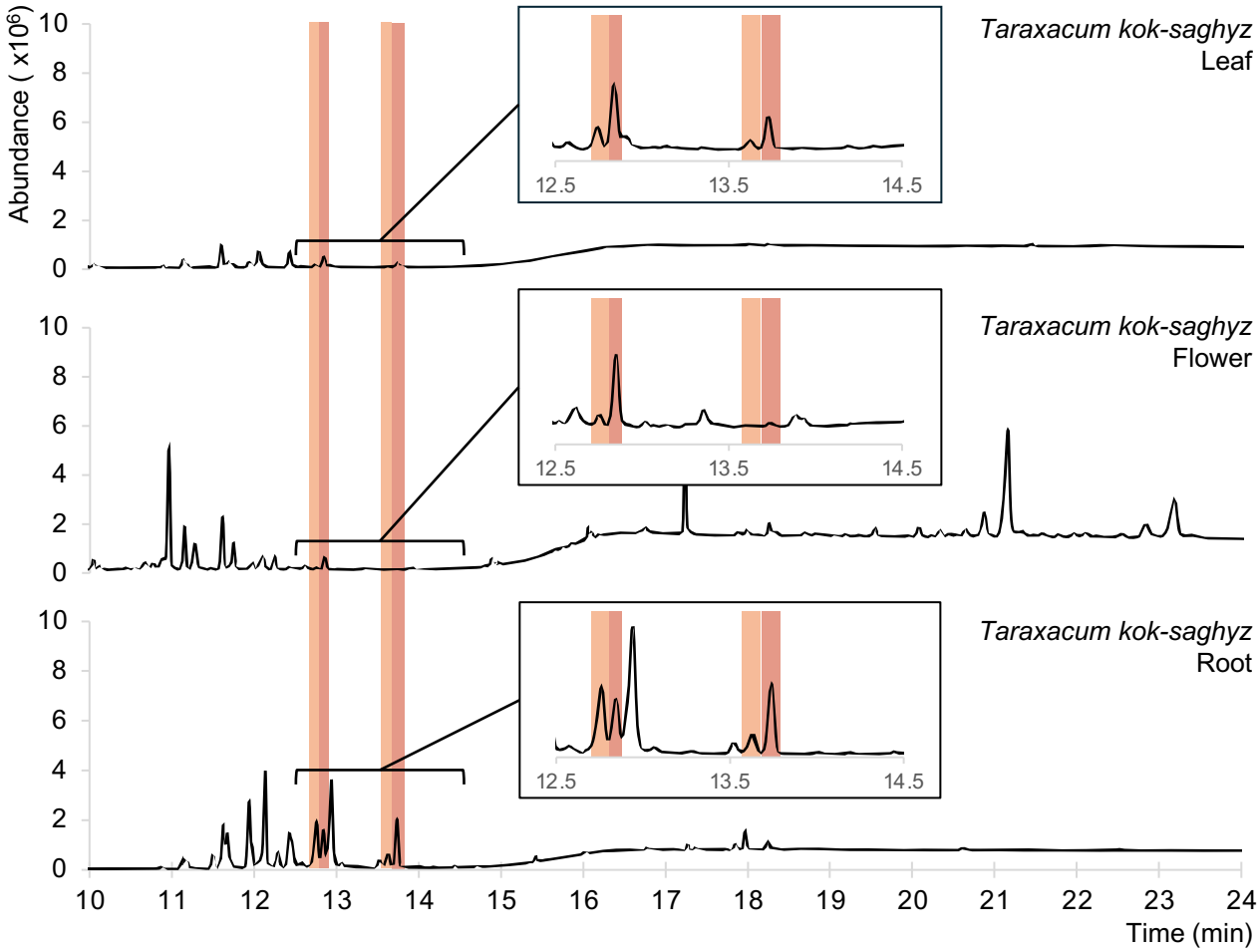
Supplementary Figure 21. Bayesian phylogenetic tree of mixed amyrin synthases (MAS) and taraxasterol synthases (TXSS). Bayesian phylogenetic analysis was performed using BEAST2 with a GTR model and an *Artemisia* fossil (orange circle; 1.31 mya) and a *Cichorium intybus* type fossil (orange circle; 2.22 mya) as calibrators. Numbers show branch length range. Ia=*Ilex asprella ;* Cr=*Catharanthus roseus;* Oe=*Olea europaea;* Ob=*Ocimum basilicum;* Co=*Calendula officinalis;* Aa=*Artemisia annua;* Ha=*Helianthus annuus;* Cc=*Cynara cardunculus;* Td=*Tragopogon dubius;* Tk=*Taraxacum kok-saghyz;* Tc=*Taraxacum coreanum;* Ls=*Lactuca sativa;* Ce=*Cichorium endivia*.



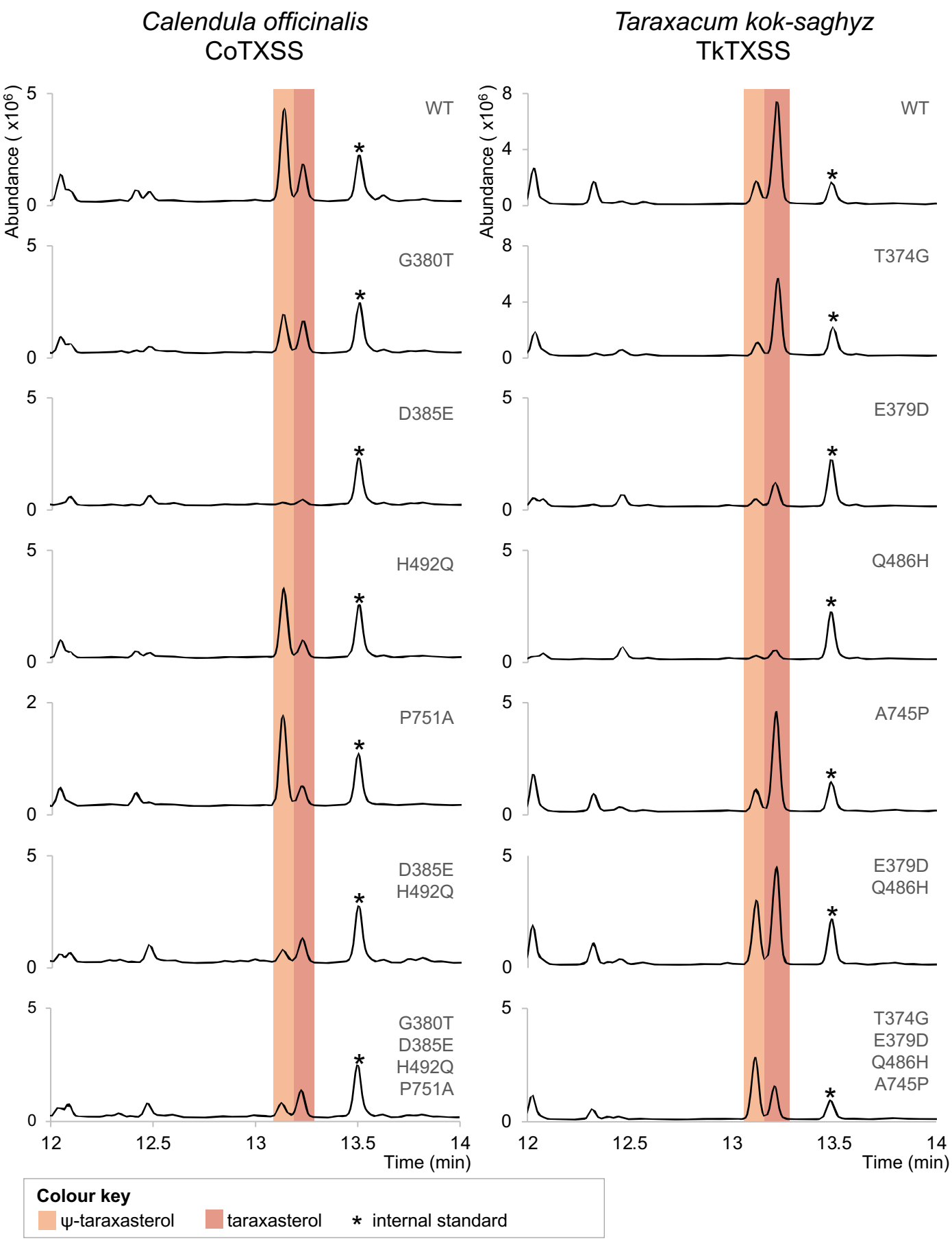
Supplementary Figure 22. Genomic locations of *Calendula officinalis* (pot marigold), *Helianthus annuus* (sunflower), *Lactuca sativa* (lettuce) and *Cynara cardunculus* (globe artichoke) *TXSS* genes. A graphical representation of the relative position, orientation and similarity of genomic regions encoding *TXSS* genes. Contigs and chromosomes are paired to show synteny between genes found on homologous regions. Genes that were expressed in the pot marigold transcriptome are shown in green. Those for which expression was not detected in pot marigold leaf and floral tissues are shown in yellow. Homologues in sunflower, lettuce and globe artichoke are shown in orange.



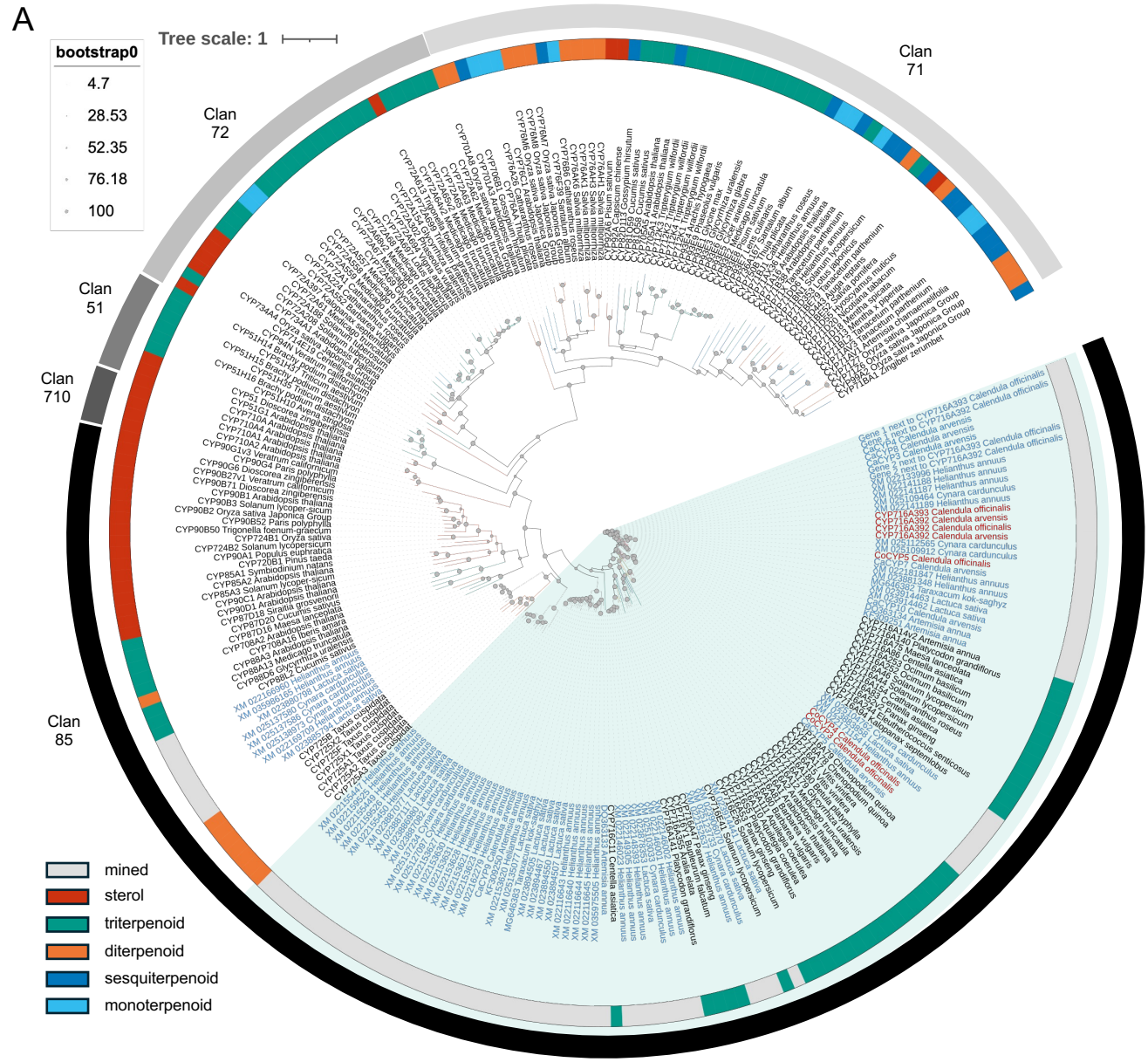
Supplementary Figure 23. GC-MS profile of *Calendula officinalis* (pot marigold) tissues. Representative total ion chromatograms of ethyl acetate extracts of pot marigold tissues to identify ψ-taraxasterol and taraxasterol metabolites and their derivatives. ψ-taraxasterol and its derivatives are highlighted in orange and taraxasterol and its derivatives are highlighted in red.

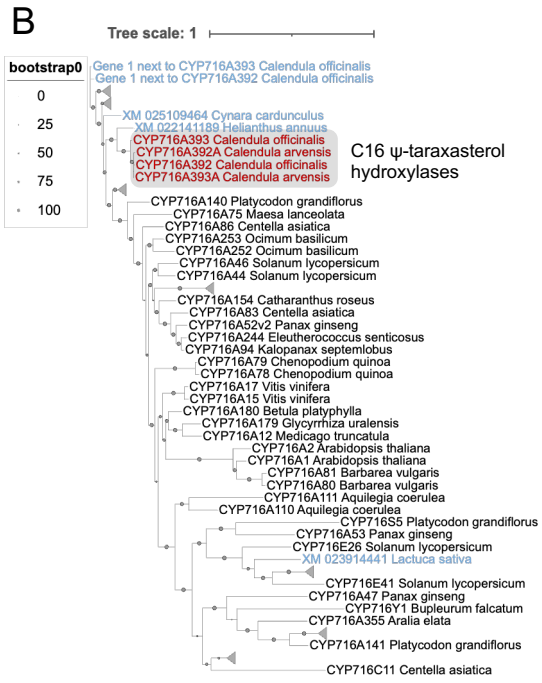


Supplementary Figure 24. GC-MS profile of *Taraxacum kok-saghyz* (Russian dandelion) tissues. Representative total ion chromatograms of ethyl acetate extracts of Russian dandelion tissues to identify ψ-taraxasterol and taraxasterol metabolites and their derivatives. ψ-taraxasterol and its derivatives are highlighted in orange and taraxasterol and its derivatives are highlighted in red.



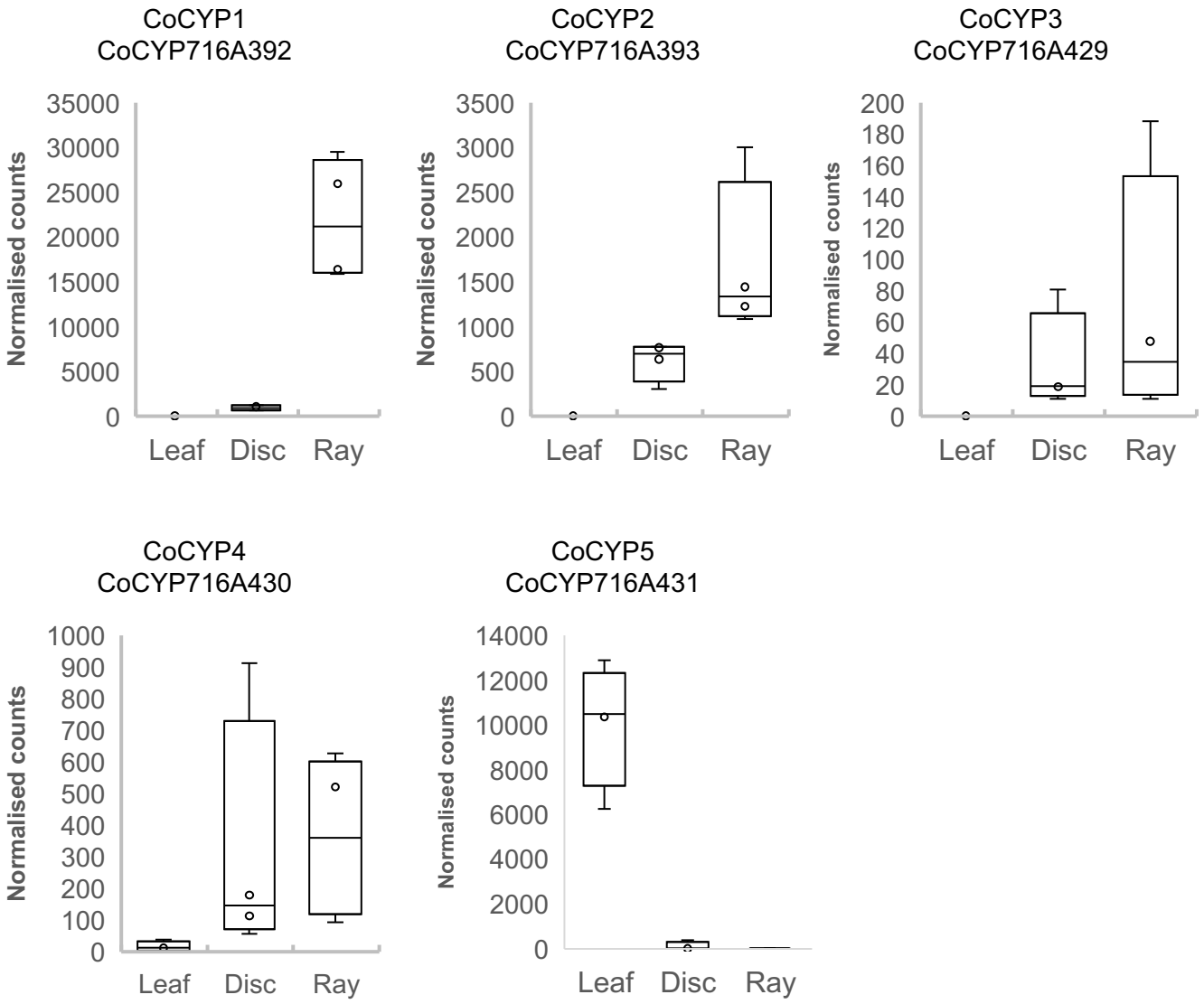
Supplementary Figure 25. GC-MS analysis of *Nicotiana benthamiana* expressing taraxasterol synthase (TXSS) mutants. Total ion chromatograms of extracts of *N. benthamiana* leaves transiently expressing wild type and mutated CoTXSS and TkTXSS. Key metabolites are highlighted and detailed in the colour key.



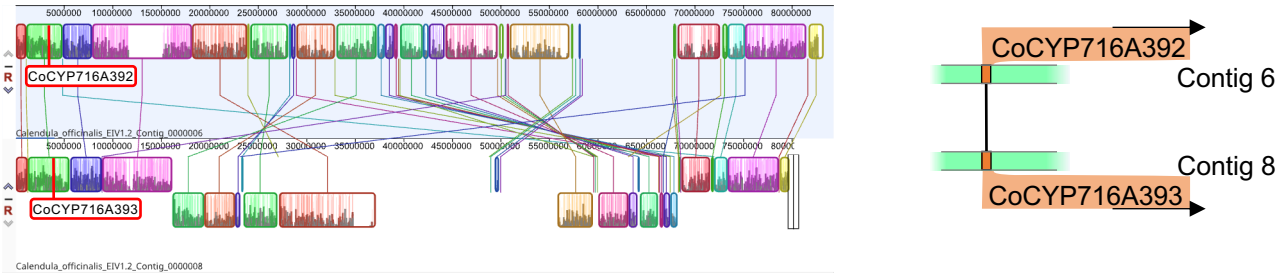


Supplementary Figure 26. Maximum-likelihood tree of plant cytochrome p450s (CYPs).

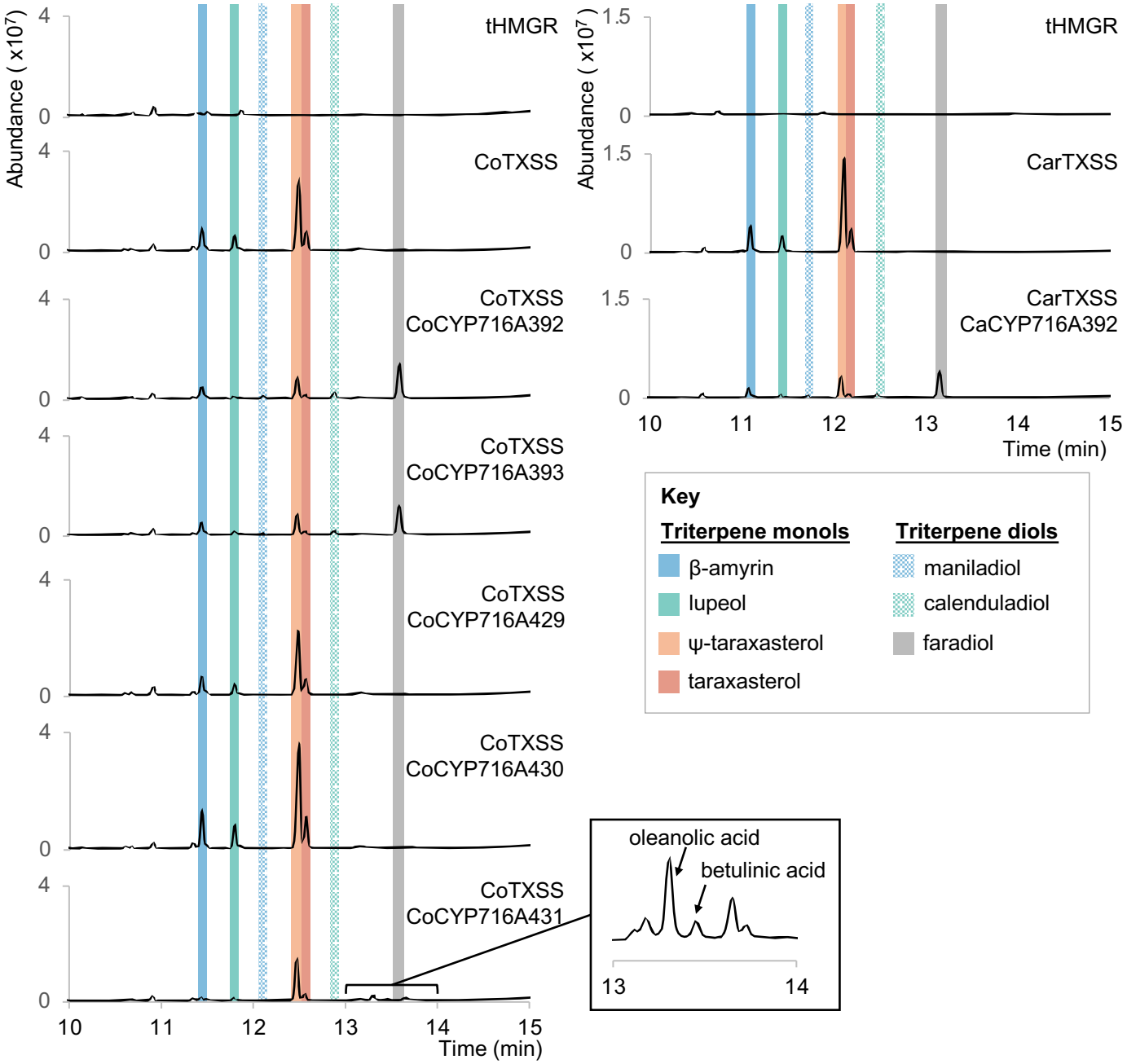
A. Maximum likelihood tree was constructed in IQ-Tree using the LG+F+I+G4 matrix-based model and visualised using Interactive Tree of Life (iTOL) v6. Substrate specificity is shown in the inner ring; CYPs clans are shown in the outer ring, CYP716 family triterpenoid modifying clade is highlighted in light teal**.** Filled grey circles indicate bootstrap supports for each node. The scale bar represents the number of substitutions per site. Functionally characterised taxa are shown in black. CYPs sequences mined from the publicly available genomes of Asteraceae, *Calendula officinalis* genome and transcriptome datasets, and *Calendula arvensis* transcriptome, are shown in blue. *C. officinalis* and *C. arvensis* C16 ψ-taraxasterol hydroxylases are shown in red. **B.** Extracted CYP716 family showing characterised p450s (black text), uncharacterised genes (blue text) and C16 ψ-taraxasterol hydroxylases (red text, grey shading)



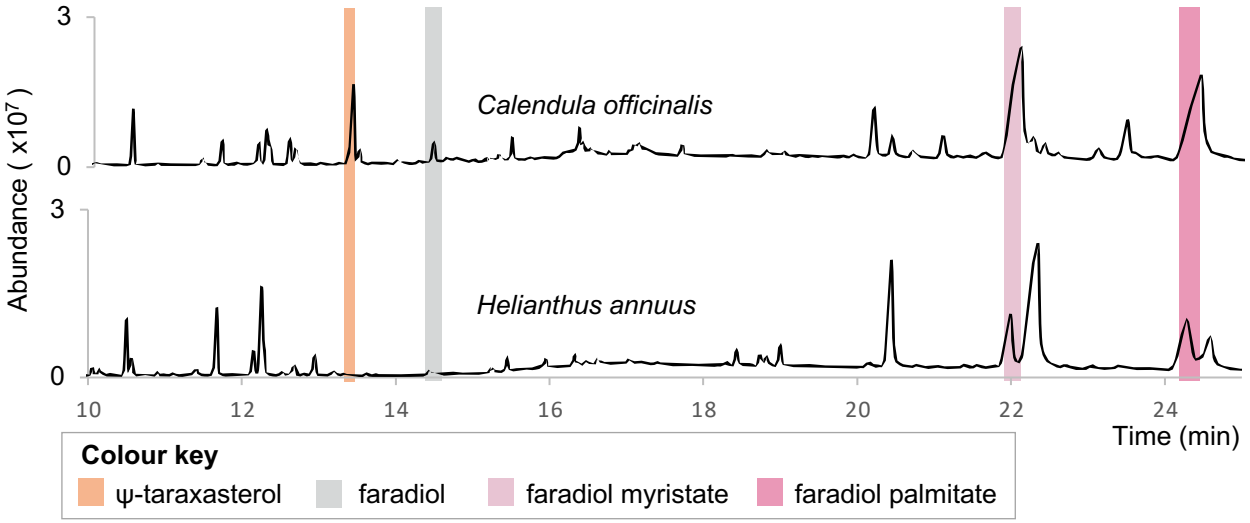
Supplementary Figure 27. Differential expression of *Calendula officinalis* (pot marigold) CYP genes. Gene expression of five cytochrome P450 (CYP) genes expressed in leaf, disc and ray floret tissues. Gene expression for each tissue is displayed in a box plot as normalised counts following transcript quantification using Salmon v1.2.0 and differential gene expression analysis using DESeq2. The centre line of the boxplot denotes the median, the box denotes the 25th to 75th percentiles of the data and the whiskers show the minimum and maximum of the data.



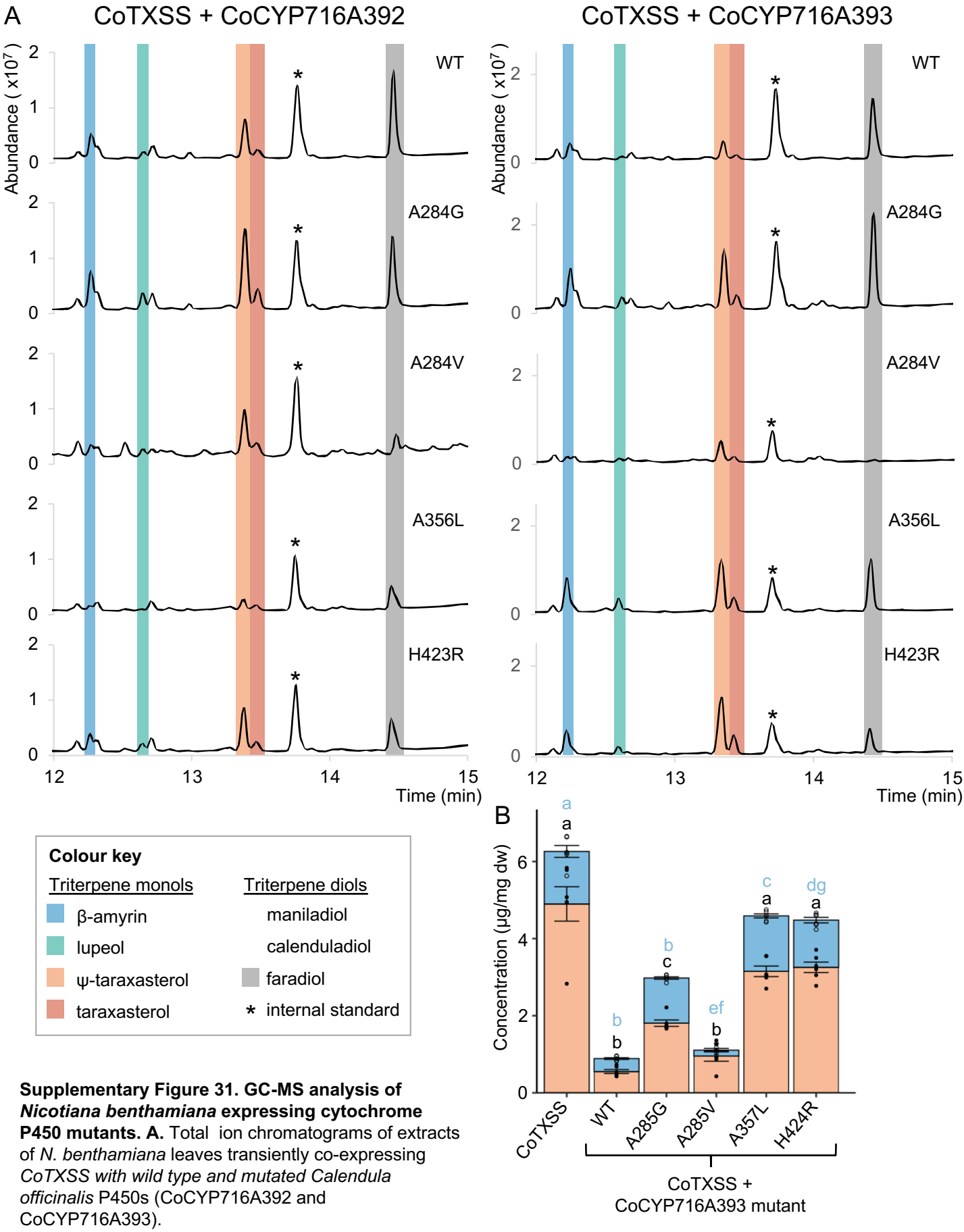
Supplementary Figure 28. Genomic location and synteny of *Calendula officinalis* (pot marigold) genes encoding CoCYP716A392 and CoCYP716A393. Chromosome synteny and graphical representation of the contig positions of cytochrome P450 and acyltransferase genes involved in the triterpene tatty acid ester (TFAE) pathway. Contigs are paired to show synteny between genes found on homeologous contigs. A graphical representation of the relative position, orientation and similarity of genes within each contig pair are shown to the right of the contig synteny plot.



Supplementary Figure 29. GC-MS analysis of *Nicotiana benthamiana* expressing TXSS and CYPs. Representative total ion chromatograms of ethyl acetate extracts of *N. benthamiana* leaves transiently expressing *TARAXASTEROL SYNTHASE* (*TXSS)* genes and *CYP* candidate genes mined from *Calendula officinalis* and *Calendula arvensis* transcriptomes.

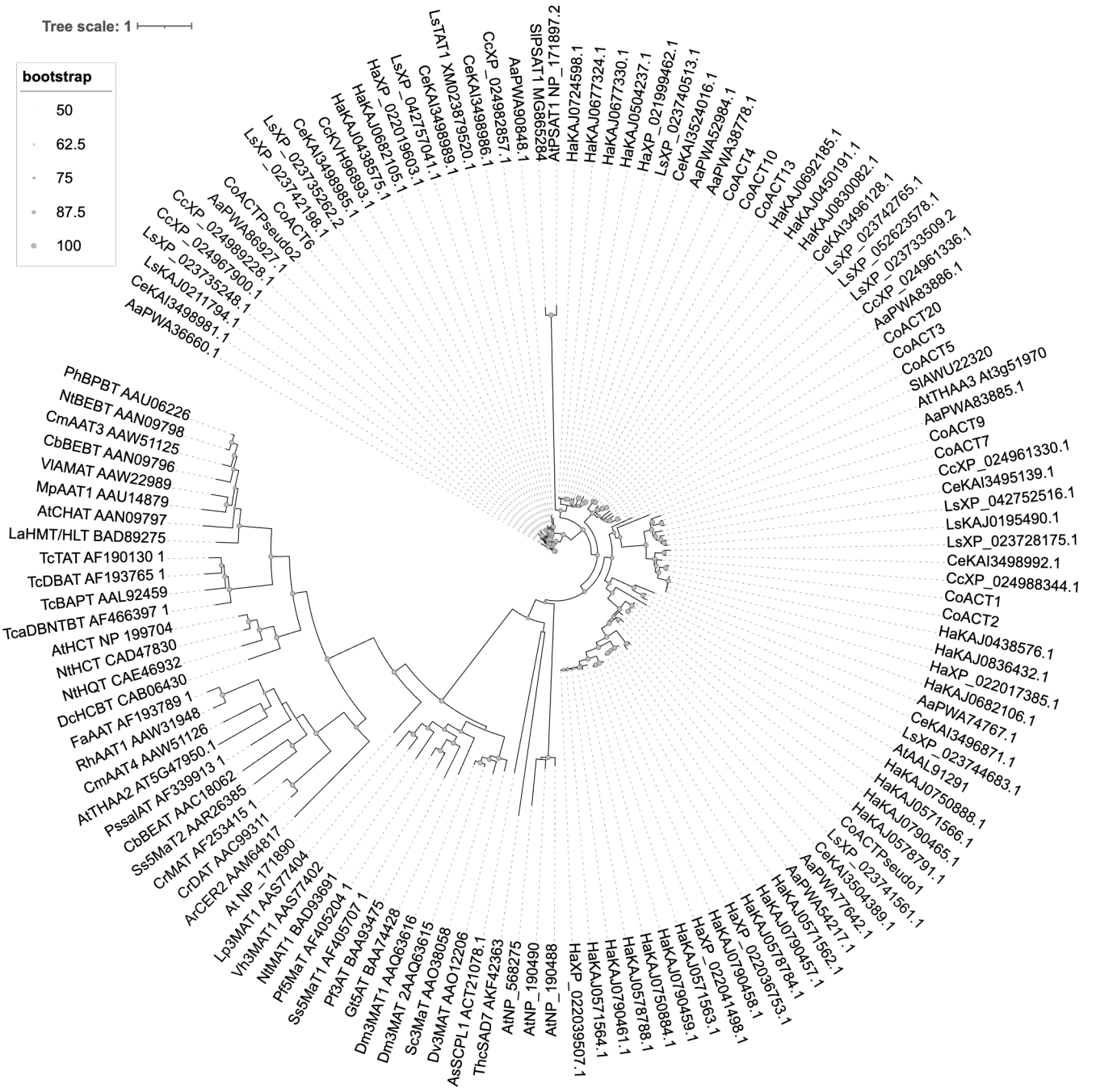


Supplementary Figure 30. GC-MS profiling of Asteraceae floral extracts. Total ion chromatograms of ethyl acetate extracts of floral tissues from five Asteraceae species identified to encode a protein with high similarity to CoCYP716A392/CoCYP716A393. Faradiol is highlighted in grey; its fatty acid derivatives, faradiol myristate and faradiol palmitate, are highlighted in light pink and pink respectively; its precursor ψ-taraxasterol is highlighted in orange.

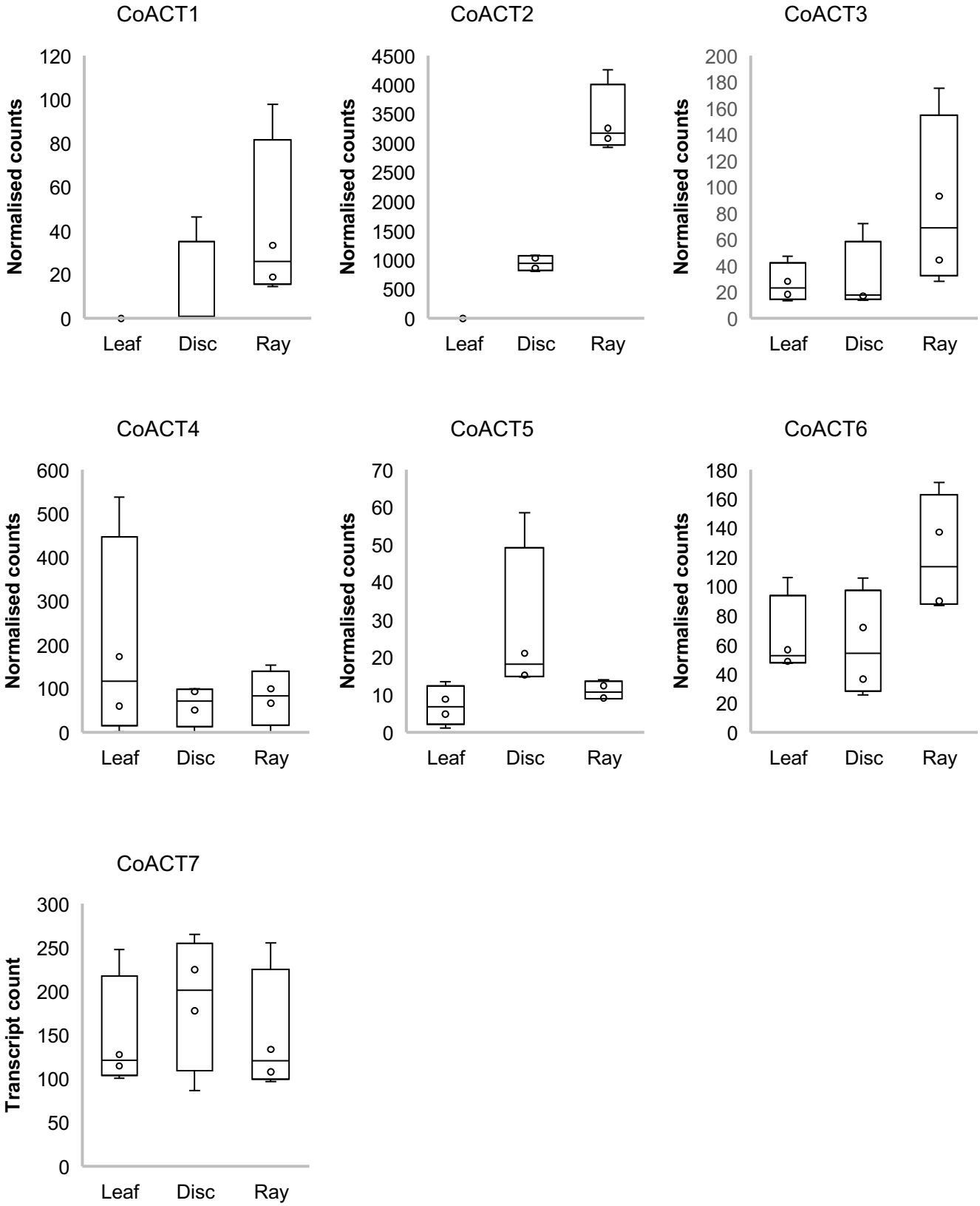


Supplementary Figure 31. GC-MS analysis of *Nicotiana benthamiana* expressing cytochrome P450 mutants. A. Total ion chromatograms of extracts of *N. benthamiana* leaves transiently co-expressing *CoTXSS with wild type and mutated Calendula officinalis* P450s (CoCYP716A392 and CoCYP716A393).

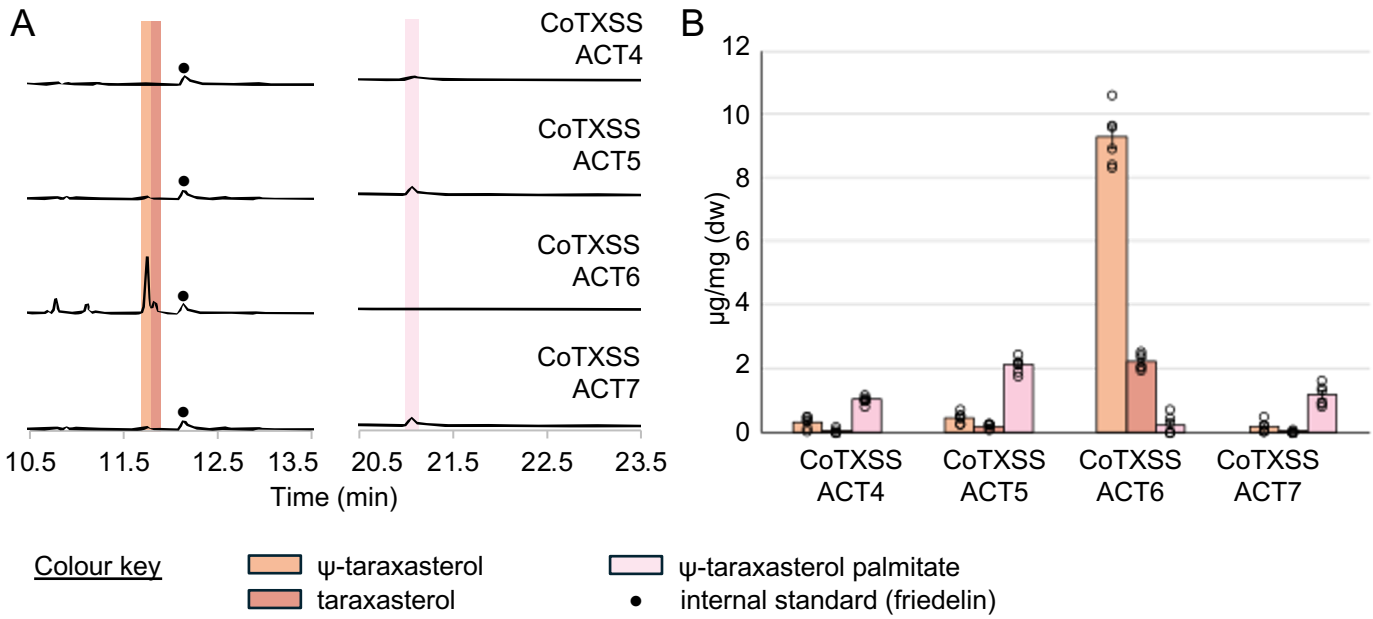
B. Quantification of triterpenes depleted by wild type and mutated CoCYP716A392; n=6; error bars indicate standard error. Significant differences in total Ψ-taraxasterol/β-amyrin content compared to wild type CoCYP716A393 (black lowercase letters), and significant differences in taraxasterol/β-amyrin ratio compared to wild type CoCYP716A393 (blue lowercase letters) were analysed using a Kruskal-Wallis test followed by post-hoc Wilcoxon rank sum test with a Benjamini-Hochberg correction (Supplementary Data 9). Samples that do not share the same lower-case letter are significantly different from each other ($p<0.05$).



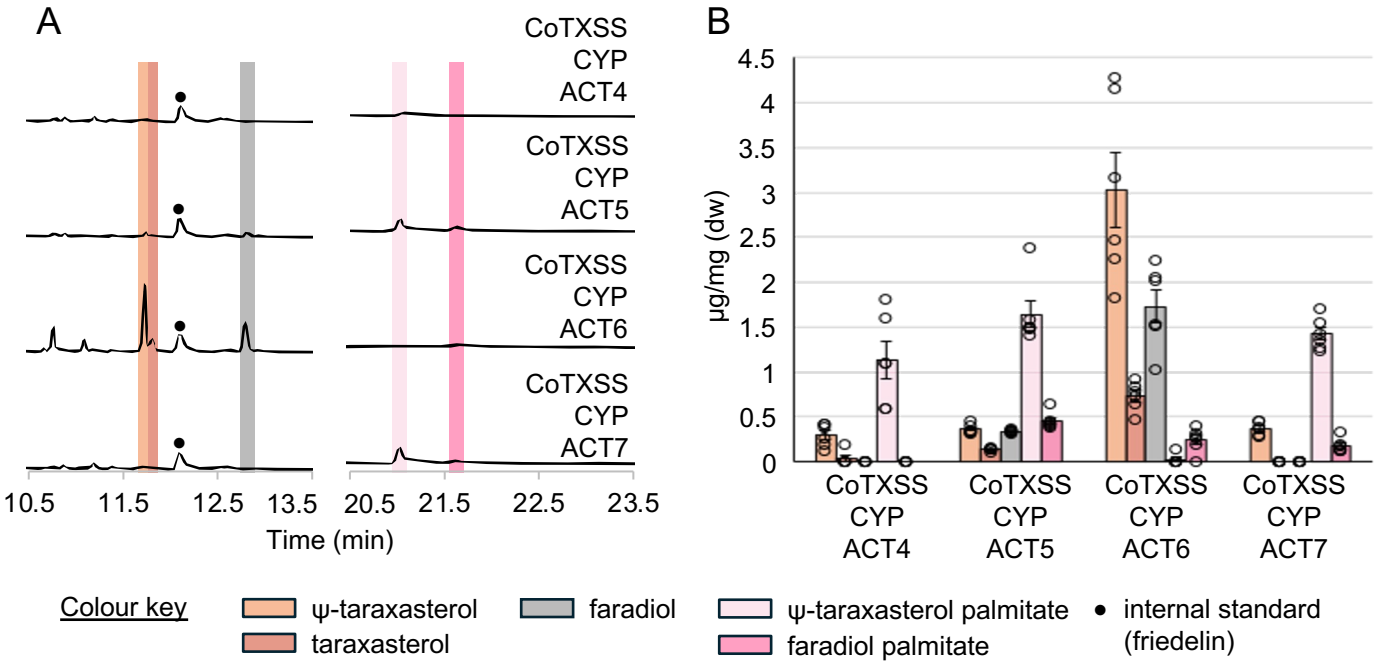
Supplementary Figure 32. Maximum-likelihood tree of plant acyltransferases. Maximum-likelihood tree of characterised plant ACTs constructed in IQTree using the JTT matrix-based model. Taxa annotated with an asterisk are likely pseudogenes. Filled grey circles indicate bootstrap supports for each node. The scale bar represents the number of substitutions per site. Accession numbers are indicated by prefixe indicating the species: Aa, *Artemisia annua*; As, *Avena strigosa*; At, *Arabidopsis thaliana*; Cb, *Clarkia breweri;* Cc, *Cynara cardunculus* var. scolymus; Ce, *Cichorium endivia*; Cm, *Cucumis melo*; Co, *Calendula officinalis*; Cr, *Catharanthus roseus*; Dc, *Dianthus caryophyllus*; Dm, *Dendranthema x morifolium*; Dv, *Dahlia variabilis;* Fa, *Fragaria x ananassa*; Gt, *Gentiana triflora*; Ha, *Helianthus annuus*; La, *Lupinus albus*; Lp, *Lamium purpureum*; Ls, *Lactuca sativa*; Mp, *Malus domestica*; Nt, *Nicotiana tabacum*; Pf, *Perilla frutescens*; Ph, *Petunia x hybrida*; Ps, *Papaver somniferum*; Rh, Rosa hybrid cultivar; Sc, *Senecio cruentus*; Sl, *Solanum lycopersicum*; Ss, *Salvia splendens*; Tc, *Taxus cuspidata*; Tca, *Taxus canadensis*; Thc, *Theobroma cacao*; Vh, *Verbena hybrida*; Vl, *Vitis labrusca*.



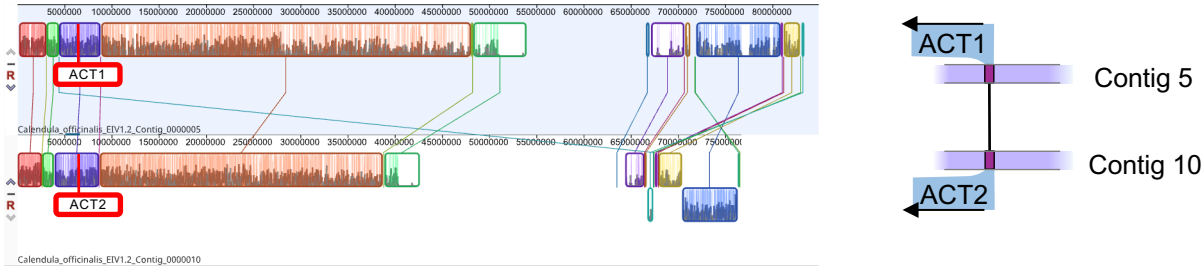
Supplementary Figure 33. Differential expression of *Calendula officinalis ACYLTRANSFERASE (ACT)* genes. Gene expression of seven *ACT* genes expressed in leaf, disc and ray floret tissues. Gene expression for each tissue is displayed in a box plot as normalised counts following transcript quantification using Salmon v1.2.0 and differential gene expression analysis using DESeq2. The centre line of the boxplot denotes the median, the box denotes the 25th to 75th percentiles of the data and the whiskers show the minimum and maximum of the data.



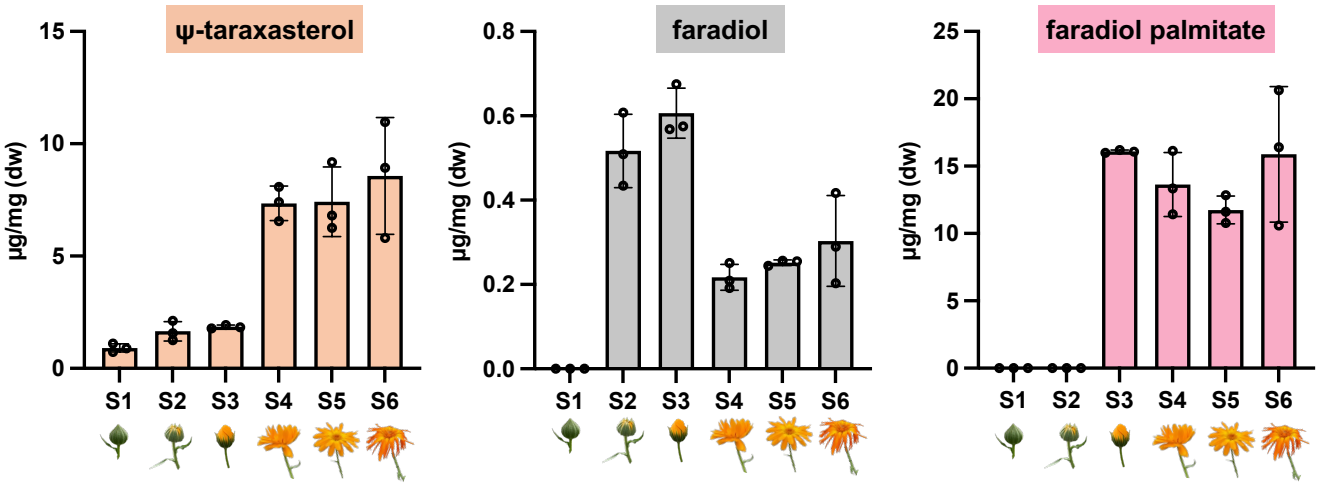
Supplementary Figure 34. GC-MS analysis of *Nicotiana benthamiana* expressing TXSS and ACTs.
A. Total ion chromatograms of extracts of *N. benthamiana* leaves transiently co-expressing CoTXSS with CoACT4-7. **B.** Peak area analysis showing products of pathway reconstruction in *N. benthamiana* through agroinfiltration of indicated genes, quantified compared to friedelin internal standard, n=6, error bars represent standard error.



Supplementary Figure 35. GC-MS analysis of *Nicotiana benthamiana* expressing TXSS, CYPs and ACTs. **A.** Total ion chromatograms of extracts of *N. benthamiana* leaves transiently co-expressing CoTXSS and CoCYP716A392 with CoACT4-7. **B.** Peak area analysis showing products of pathway reconstruction in *N. benthamiana* through agroinfiltration of indicated genes, quantified compared to friedelin internal standard, n=6, error bars represent standard error.



Supplementary Figure 36. Genomic location and synteny of *Calendula officinalis* (pot marigold) *ACYLTRANSFERASES* (*ACTs*). A graphical representation of the relative genomic position, orientation and similarity of *CoACT1* and *CoACT2,* which add fatty acid groups to faradiol C3.



Supplementary Figure 37. Peak area analysis showing accumulation of ψ-taraxasterol, faradiol and faradiol palmitate through *Calendula officinalis* pot marigold floral development. Metabolites were quantified compared to friedelin internal standard; n=3; error bars indicate standard deviation.

Supplementary Table 1. Triterpenes detected in *Calendula officinalis.*
Content and distribution of triterpenes detected in extracts of leaf, disc floret, ray floret and root tissue. Metabolite concentrations (µg/mg dry weight) were determined by GC-MS peak area analysis using friedelin as an internal standard. Data are means ± SD of 4 independent plants. n.d. = not detected.

Metabolite	Leaf	Disc floret	Ray floret	Root
µg/mg dry weight ± S.D.				
Triterpene scaffold				
β-amyrin	0.02 ± 0.001	0.246 ± 0.069	1.965 ± 0.443	n.d.
isofucosterol	0.052 ± 0.015	0.33 ± 0.02	1.068 ± 0.303	n.d.
α-amyrin	0.018 ± 0.003	0.082 ± 0.029	1.126 ± 0.213	n.d.
lupeol	0.02 ± 0.002	0.056 ± 0.021	1.46 ± 0.513	n.d.
ψ-taraxasterol	n.d.	0.209 ± 0.065	4.645 ± 1.104	n.d.
taraxasterol	n.d.	n.d.	0.848 ± 0.244	n.d.
Hydroxylated triterpenes				
faradiol	n.d.	0.07 ± 0.006	0.294 ± 0.071	n.d.
Triterpene fatty acid esters				
faradiol/arnidiol laurate	n.d.	0.255 ± 0.231	2.486 ± 0.186	n.d.
β-amyrin myristate	n.d.	0.062 ± 0.06	1.311 ± 0.161	n.d.
manaladiol myristate	n.d.	0.046 ± 0.049	0.401 ± 0.098	n.d.
calenduladiol myristate	n.d.	0.062 ± 0.072	1.282 ± 0.197	n.d.
ψ-taraxasterol/taraxasterol palmitate	n.d.	n.d.	0.298 ± 0.143	n.d.
faradiol/arnidiol myristate	n.d.	2.257 ± 2.201	13.178 ± 1.171	n.d.
β-amyrin palmitate	n.d.	0.059 ± 0.064	1.769 ± 0.307	n.d.
lupeol palmitate	n.d.	0.094 ± 0.076	0.735 ± 0.225	n.d.
manaladiol palmitate	n.d.	n.d.	0.505 ± 0.132	n.d.
calenduladiol palmitate	n.d.	n.d.	0.782 ± 0.168	n.d.
ψ-taraxasterol/taraxasterol palmitate	n.d.	0.595 ± 0.565	1.125 ± 0.983	n.d.
faradiol/arnidiol palmitate	n.d.	1.082 ± 1.163	12.463 ± 0.569	n.d.

Supplementary Table 2. Table of genome assembly statistics

Statistics	**Final Calendula officinalis genome** Calendula_officinalis_EIV1.2
Genome size	1,301,655,687
GC content	36.5%
Contigs	2,811
Largest contig	104,728,146
N50	80.2 Mb
Complete BUSCOs	2256 (96.9%)
Single Copy BUSCOs	586 (25.2%)
Duplicated BUSCOs	1670 (71.8%)
Fragmented BUSCOs	10 (0.4%)
Missing BUSCOs	60 (2.6%)

Supplementary Table 3. Table of *Calendula officinalis oxidosqualene cycle* (*OSC*) transcripts. *OSC* genes identified in the *Calendula officinalis* transcriptome (Calendula_officinalis_EIV1.1.mikado.loci.cds.fa) and the *Calendula arvensis* transcriptome (CombineCarv.combined.cdhit98.fa.transdecoder.cds.cds). Genes that were functionally characterised have their functions shown in bold, predicted functions are in normal text.

Species	**Gene name**	**Transcript number**	**Characterised**/predicted function
Calendula officinalis	CoOSC1	0435770.1	**Taraxasterol synthase**
Calendula officinalis	CoOSC2	0275220.1	Mixed-amyrin synthase
Calendula officinalis	CoOSC3	0109510.1	**Mixed-amyrin synthase**
Calendula officinalis	CoOSC4	0055930.1	**Bauerenol synthase**
Calendula officinalis	CoOSC5	0279660.1	Novel/multifunctional triterpene synthase
Calendula officinalis	CoOSC6	0762650.1	Novel/multifunctional triterpene synthase
Calendula officinalis	CoOSC7	0553310.1	Lupeol synthase
Calendula officinalis	CoOSC8	0142220.1	**Shionone synthase**
Calendula officinalis	CoOSC9	0358550.1	β-amyrin-like synthase
Calendula officinalis	CoOSC10	0553470.1	β-amyrin-like synthase
Calendula officinalis	CoOSC11	0264370.1	β-amyrin synthase
Calendula officinalis	CoOSC12	0025710.1	Cycloartenol synthase
Calendula officinalis	CoOSC13	0135220.1	Cycloartenol synthase
Calendula officinalis	CoOSC14	0433710.1	Cycloartenol synthase
Calendula officinalis	CoOSC15	0546580.1	Cycloartenol synthase
Calendula officinalis	CoOSC16	0553310.1	Cycloartenol synthase
Calendula arvensis	CarOSC1	DN42097_c0_g2_i3	**Taraxasterol synthase**

Supplementary Table 4. Taxa and accession numbers of sequences used for positive selection analysis in Figure 5B.

Abbreviation	Enzyme name	Species	Accession no
AaMAS	Multifunctional triterpene synthase	*Artemisia annua*	AHF22084
CoMAS	Mixed amyrin synthase	*Calendula officinalis*	This study
CrMAS	Mixed amyrin synthase	*Catharanthus roseus*	AEX99665
IaMAS	Mixed amyrin synthase	*Ilex asprella*	AIS39793
ObMAS	Mixed amyrin synthase	*Ocimum basilicum*	AFH53506
OeMAS	Mixed amyrin synthase	*Olea europaea*	BAF63702
CcTXSS	Taraxasterol synthase	*Cynara cardunculus*	XP_024993293
CoTXSS	Taraxasterol synthase	*Calendula officinalis*	This study
HaTXSS	Taraxasterol synthase	*Helianthus annuus*	XP_022035366
LsTXSS	Taraxasterol synthase	*Lactuca sativa*	QIB85400
TcTXSS	Taraxasterol synthase	*Taraxacum coreanum*	QBO24611
TkTXSS	Taraxasterol synthase	*Taraxacum kok-saghyz*	AXU93516
TdTXSS	Taraxasterol synthase	*Tragopogon dubius*	DDRL-2016809 (1KP)
CeTXSS	Taraxasterol synthase	*Cichorium endivia*	GGQM01030486.1

Supplementary Methods

1. **Library construction and sequencing**
2. **Genome Assembly**
3. **Functional annotation**

1. Library construction and sequencing

PacBio HiFi

A low input library was prepared from a starting input of 984 ng gDNA. The sample was sheared using the Megaruptor 3 instrument (Diagenode, P/N B06010003), 5 ng/µl at a speed setting of 30. The sample underwent AMPure® PB bead (PacBio®, P/N 100-265-900) purification and concentration at a ratio of 1.8X before library construction using the SMRTbell template prep kit 2.0 (PacBio®, P/N 100-983-900).

The HiFi library was prepared from 767 ng of sheared gDNA, according to the instructions in the low input protocol version 05 (PacBio®, P/N 101-730-400). A barcoded adapter (PacBio ®, P/N 101-628-500) was ligated to create the SMRTbell library. The final SMRTbell library was purified with AMPure beads. The final library size was estimated from a smear analysis performed on the FEMTO Pulse® System (Agilent, P/N M5330AA) and quantified by fluorescence (Invitrogen Qubit™ 3.0, P/N Q33216).

Loading calculations for sequencing were made using the PacBio® SMRT®Link Binding Calculator v9.0.0.92017. Sequencing primer v4 was annealed to the adapter sequence of the library. Binding of the library to the sequencing polymerase was completed using Sequel® II Binding Kit v2.0 (PacBio®, P/N 101-842-900). The Sequel® II DNA internal control was spiked into the library complex at the standard concentration prior to sequencing. The sequencing chemistry used was Sequel® II Sequencing Plate 2.0 (PacBio®, P/N 101-820-200) and the Instrument Control Software v9.0.0.92233.

The library was sequenced on four Sequel II SMRT®cells 8M on the Sequel II instrument. The parameters for sequencing per cell were as follows: diffusion loading, CLR sequencing mode, 30-hour movie, 2-hour immobilisation time, 2-hour pre-extension time, 20 pM on plate loading concentration.

A total of 1284 Gb of PacBio data was produced from four SMRT cells. Circular consensus sequencing (CCS) analysis (SMRT Link 9.0.0.92188) yielded 6,098,205 CCS reads with a median length of 13.0 kb at an estimated accuracy greater than Q20 (or 99.9% accuracy). The Q20 CCS reads comprised a total 81Gb of Hifi data (or 58X coverage based on an estimated genome size of 1.4Gb). HiFI reads were processed with CutAdapt v3.2 (http://code.google.com/p/cutadapt/) [--action lowercase -O 20 --errors 8 -b ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT -rc] to remove reads containing SMRT adapter sequences.

10X Genomics linked reads

Genomic DNA was quantified using the Qubit dsDNA HS (High Sensitivity) Assay Kit (Q32854, Thermo Fisher Scientific). DNA integrity was checked using the FEMTO Pulse System Genomic DNA 165 kb Kit (FP-1002-0275, Agilent Technologies), which determined that 64.4% of the DNA was greater than 50 kb.The Chromium 10x platform (10x Genomics) micro-fluidic Genome Chip (PN-120216) was used to produce three barcoded linked read libraries, with 1.14ng DNA input, using the Chromium™ Genome Library Kit & Gel bead Kit v2 (120258) following the Chromium Genome Reagent Kits Version 2 User Guide (CG00043). The micro-fluidic chip produces a library of genome gel beads by combining the DNA with the gel beads, reaction mastermix and partitioning oil to create Gel Bead in Emulsions (GEMs). Isothermal incubation of the GEMs was followed by recovery of the barcoded fragments which were then used to create libraries for Illumina sequencing. Library

yields were quantified using the Qubit dsDNA HS Assay; insert size was determined with Agilent 2100 Bioanalyzer High Sensitivity DNA chip (5067-4627, Agilent Technologies), and finally checked by qPCR (07960204001, Roche Diagnostics Ltd.) prior to equimolar pooling and sequencing.

Omni-C

A sample of fresh leaf was provided to Dovetail Genomics (Scotts Valley, CA, USA).

Illumina RNA-Seq

Libraries were constructed using the NEBNext Ultra II RNA Library prep for Illumina kit (NEB#E7760L), NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB#E7490L) and NEBNext Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs) (E6440S/L) at a concentration of 10 µM. Library preparation was performed on the Perkin Elmer (formerly Caliper LS) Sciclone G3 (PerkinElmer PN: CLS145321). mRNA was purified from 1 µg of total RNA using a Poly(A) mRNA Magnetic Isolation Module. Isolated mRNA was fragmented for 12 minutes at 94°C, and first strand cDNA was synthesised. Directionality was retained by adding dUTP during second strand synthesis and subsequent cleavage of the uridine containing strand using USER Enzyme. NEBNext Adaptors were ligated to end-repaired, dA-tailed DNA. The ligated products were purified using Beckman Coulter AMPure XP beads (A63882). Adaptor-ligated DNA was enriched in ten cycles of PCR (30 secs at 98°C, 10 cycles of: 10 secs at 98°C _75 secs at 65°C _5 mins at 65°C). Barcodes (NEBNext Multiplex Oligos for Illumina®) were incorporated during PCR. Library quality was determined using a Perkin Elmer DNA High Sensitivity Reagent Kit (CLS760672) with DNA 1K / 12K / HiSensitivity Assay LabChip (760517) and the concentration measured with a Quant-iT™ dsDNA Assay Kit, high sensitivity (Plate Reader) assay from ThermoFisher (Q-33120). The final libraries were equimolar pooled, verified by qPCR and sequenced on two lanes of a NovaSeq 6000 SP flow cell with 150 bp PE reads, yielding ~100 Gb.

PacBio IsoSeq

Libraries were constructed from 300 ng of total RNA. Reverse transcription cDNA synthesis was performed using NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module (NEB, E6421). cDNA samples were amplified with barcoded primers for 12 cycles and pooled at equimolar concentration before SMRTbell library construction. The library pool was prepared according to the Iso-Seq protocol version 02 (PacBio, 101-763-800), using SMRTbell express template prep kit 2.0 (PacBio, 102-088-900). The library pool was quantified using a Qubit Fluorometer 3.0 (Invitrogen) and sized using the Bioanalyzer HS DNA chip (Agilent Technologies, Inc.).

The loading calculations were performed using the PacBio SMRTlink Binding Calculator v8.0.0.78409. Sequencing primer v2 was annealed to the Iso-Seq library pool and complexed to the sequencing polymerase with the Sequel II binding kit V2.1 (PacBio, 101-843-000). The sequencing internal control complex 1.0 (PacBio, 101-717-600) was spiked at a standard concentration before sequencing. The Iso-Seq pool was sequenced on the Sequel II instrument with a single Sequel II SMRT®cell 8M cell using Sequel® II Sequencing Plate 2.0 (PacBio®, 101-820-200) and the Instrument Control Software v8.0.0.78867. The parameters for sequencing were diffusion loading, 30 hr movie, 2 hr immobilisation time, 2 hr pre-extension time, 40 pM loading concentration.

PacBio subreads (single SMRT cell) were analysed via the smrtlink-8.0.0.80529 pb_demux_isoseq3 application (default parameters). Processing generated Circular Consensus Sequences (CCS), which were demultiplexed, primers removed, concatemers identified, polyA tails removed and clustered into high quality (predicted accuracy ≥ 0.99) and low quality (predicted accuracy < 0.99) transcripts. In total, 108,882, 79,287 and 94,014

high- and low-quality transcripts were output for three libraries, CoDisc (*Calendula officinalis* Disc), CoLeaf (*C. officinalis* Leaf) and CoRay (*C. officinalis* Ray) and 105,790 high- and low-quality transcripts from *Calendula arvensis.*

2. Genome Assembly

Six assemblies were generated from the HiFI data using four alternative assembly tools Hicanu v2.0 (Nurk et al. 2020), Hifiasm v0.12 (Cheng et al. 2021), WTDBG v2.5 (Ruan and Li 2020) and IPA 1.3.1 (https://github.com/PacificBiosciences/pbipa). The “best” assembly was selected based on contiguity, kmer completeness / coherence against the content of the HiFi data and an assessment of potential misassembles using 10X linked-read information (Jackman et al, 2018). The hifiasm L2 assembly was provided to Dovetail genomics for scaffolding with Omni-C data and the HiRise scaffolding algorithm (see below).

Repeat identification

RepeatModeler (v1.0.11 - http://www.repeatmasker.org/RepeatModeler/) was used for *de novo* identification of repetitive elements from the assembled *C. officinalis* genome. High copy protein-coding genes potentially included in the RepeatModeler library were identified and effectively removed by running RepeatMasker v4.0.72 using the *Lactuca sativa* coding genes to hard mask the RepeatModeler library; transposable element genes were first excluded from the *Lactuca sativa* coding-geneset by running TransposonPSI (r08222010). Unclassified repeats were searched in a custom BLAST database of organellar genomes (mitochondrial and chloroplast sequences from eudicotyledons in the NCBI nucleotide division). Any repeat families matching organellar DNA were also hard-masked.

Repeat identification was completed by running RepeatMasker v4.0.72 with a RepBase Viridiplantae library and with the customized RepeatModeler library (i.e. after masking out protein coding genes), both using the -nolow setting.

Genome quality control and evaluation

An inspection of the kmer spectrum of HiFi reads was performed using the Kmer Anlaysis Toolkit (KAT) v2.4.2 (Mapleson et al. 2017). A kmer frequency histogram of the full set of Hifi data was produced at a kmer size of 22 leaving all other parameters at default settings. The kmer spectrum shows three clear peaks, consistent with *C. officinalis* having an allotetraploid background with the third peak representing content that is shared by both subgenomes (**Figure 1**). The first peak, representing heterozygous kmers, is small in comparison to the main peak, indicating that Marigold displays a moderate level of heterozygosity.

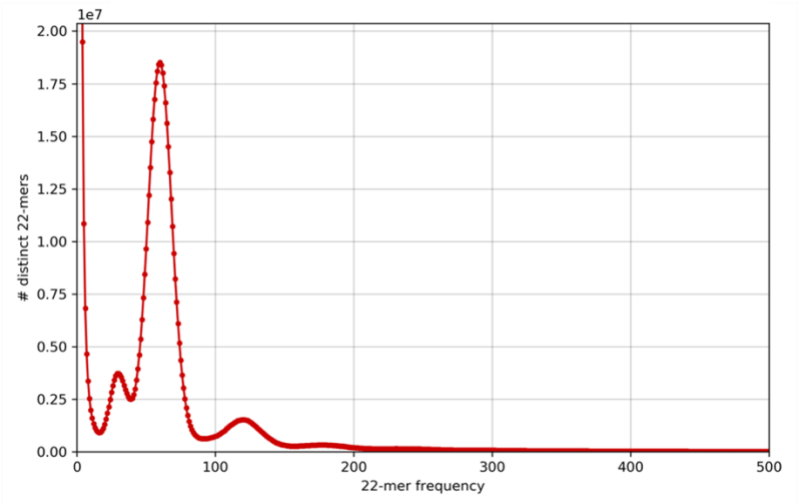


Figure 1. Kmer frequency histogram of Hifi data (k = 22).

Alternative draft genome assemblies were generated using 4 tools, Hicanu v2.0 (Nurk et al. 2020), Hifiasm v0.12 (Cheng et al. 2021), WTDBG v2.5 (Ruan and Li 2020) and IPA 1.3.1 (https://github.com/PacificBiosciences/pbipa). For Hifiasm we evaluated alternative parameter options for the merge level (-l parameter) controlling the aggressiveness of haplotype collapsing by hifiasm's internal implementation of Purge Dupes, and the –high-het setting (in beta in version 0.12). Hifiasm and IPA both produce a haplotype collapsed 'primary' assembly representing a single allele per genomic locus and an additional set of 'alternate' haplotigs. Hicanu and WTDBG produce a single haplotype expanded assembly. The Hicanu and WTDBG assemblies were therefore further processed with the Purge Dupes software v0.0.3 (Guan et al. 2020) to produce a primary assembly and associated set of 'alternate' contigs. Purge Dupes was run with default settings to auto detect the coverage threshold at which to collapse haplotigs.

Assemblies produced by the above tools were assessed using a range of assembly QC metrics. The full set of Hifi reads were aligned back to all assemblies using minimap2 (version 2.11) with parameters [-ax asm20]. Read alignments were passed to the software package Asset v1.0.0, to compute read alignment breakpoints, following the protocol described on the Asset github page (https://github.com/dfguan/asset). The count of breakpoints was used as a QC metric. Consensus quality of assemblies was summarised by QV value, computed by comparison of the kmer content of the assemblies to the kmer content of the Hifi reads using the software YAK v0.1 (https://github.com/lh3/yak) (**Table 1**). BUSCO scores were produced for each assembly using the BUSCO pipeline v3.0 (Simão et al. 2015) run with default settings.

Table 1. Genome assembly QC summary

Assembly	Size	N50	QV	Breakpoints	BUSCO Complete
Hifiasm L=1	1.3 Gb	36.6 Mb	51.5	7,245	91.7%
Hifiasm L=2	1.3 Gb	36.6 Mb	51.5	7,096	91.9%
Hifiasm L=2, Hihet = True	1.3 Gb	27.8 Mb	52.3	6,948	91.8%
Hicanu	1.1 Gb	18.75 Mb	40.4	12,657	91.2%
IPA	1.1 Gb	6.9 Mb	42.6	1,818	91.9%
WTDBG2	928 Mb	258 kb	38.7	14,876	89.8%

To examine assembly coherence against the content of the reads, we used the Kmer Analysis Toolkit (KAT) v2.4.2 (Mapleson et al. 2017) comp function to generate spectra-cn plots (**Figure 2**).

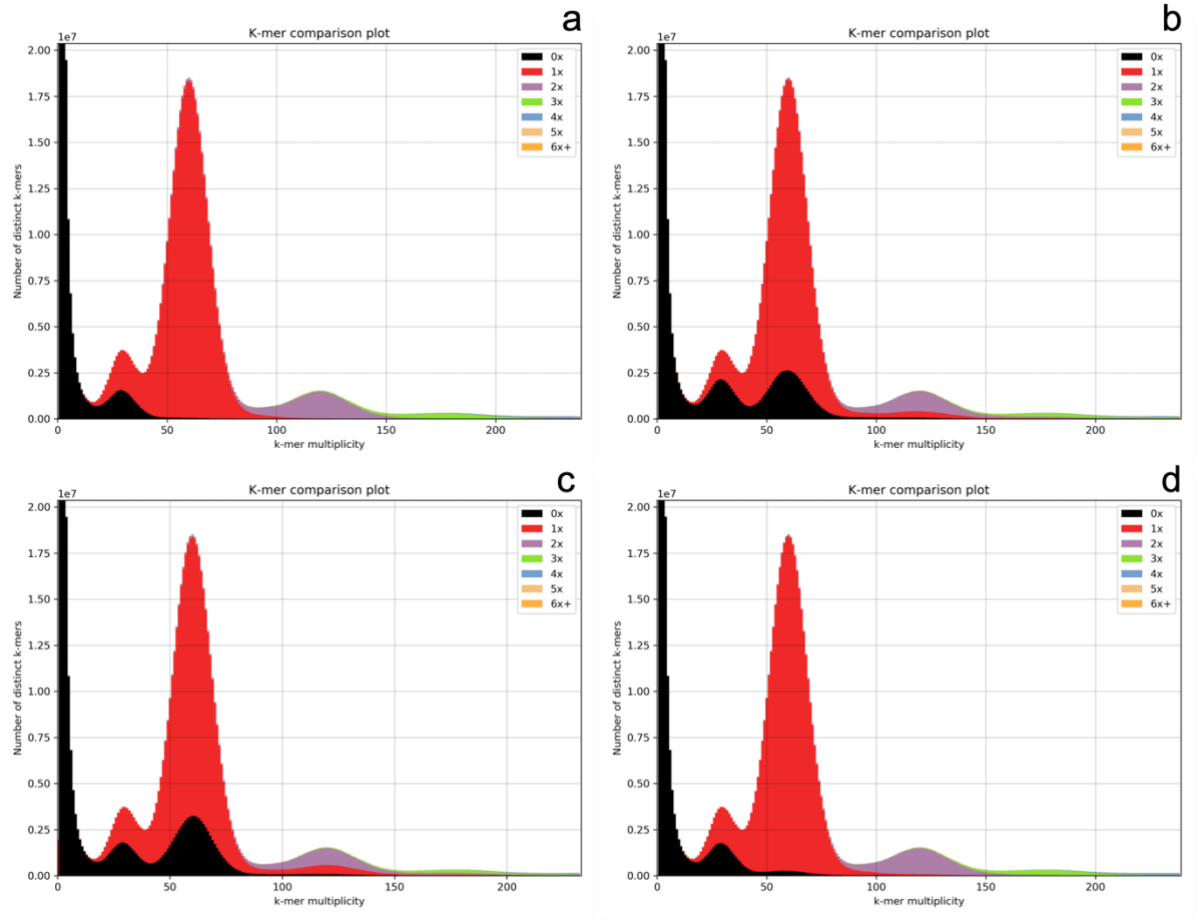


Figure 2. Spectra-cn plots of alternative HiFi assemblies, a) Hifiasm L=2, b) Hicanu, c) WTDBG2 and d) IPA

The spectra-cn plots for the Hifiasm assemblies (only L=2 shown) fit with expectations. The first peak of heterozygous kmers is coloured roughly half red and half black, indicating half of the heterozygous kmers present in the reads are present in the assembly (**Figure 2**). This is as expected for a haplotype mosaic assembly, where one allele is represented at each heterozygous locus. The second peak of single copy genomic content is coloured red, indicating that the content is represented once in the assembly. Lastly the third peak of content shared between both sub genomes is coloured purple indicating that all this content is present twice in the assembly. Once again, this is concomitant with a veridical assembly of the genome. Kmer spectrum analysis shows a significant amount of missing content from the single copy portion of the kmer spectrum in both the WTDBG and Hicanu assemblies, indicating that these assemblies do not completely represent the unique content in the genome. No further analysis was conducted on the WTDBG and Hicanu assemblies.

The hifiasm assemblies are more contiguous than the IPA assembly but have more read alignment breakpoints. 10X genomics linked reads were used to assess Hifiasm and IPA assemblies. Linked read data was pre-processed using the Longranger Basic pipeline (https://github.com/10XGenomics/longranger) to trim reads and prepare fastq headers with barcode information. Pre-processed linked read data was passed to the Tigmint pipeline (Jackman et al., 2018) with default settings. Tigmint aligns reads to the assembly using longranger, infers molecule extents, and identifies misassembles using linked-read information. The number of breaks to contigs in different size fractions was used as a QC metric (**Table 2**).

Table 2. Summary of 10X contig breaking

	IPA	Hifiasm L=1	Hifiasm L=2	Hifiasm L=2 Hihet = True
Total Breaks	342	2099	1971	2016
Broken Contigs	132	474	463	477
Total Length Broken Contigs	104.8 Mb	110.3 Mb	70.9 Mb	200.9 Mb
Broken contigs > 1Mb	20	3	2	6
Broken contigs 100kb – 1Mb	9	30	27	29
Broken contigs 10-100kb	94	441	434	442
Broken contigs 1-10kb	9	0	0	0

Analysis of the Tigmint assembly breaking indicates that the IPA assembly contains more potential misassemblies in contigs over 1Mb based on alignment of 10X data (**Table 2**). Based on this powerful orthogonal data set and the contiguity and kmer analysis, it was decided to accept the Tigmint broken, hifiasm L2 assembly as the final contig assembly.

Breaks were introduced at positions where barcode coverage (i.e. the number of inferred DNA molecules covering each position in the assembly) dropped below 20X. These breaks were mostly introduced at the ends of long contigs.

The final contig assembly was provided to Dovetail Genomics for scaffolding with Omni-C data and the HiRise scaffolding algorithm (Putnam et al. 2016). The Hirise software made a total of 5 breaks and 33 joins to the input contigs. Dovetail Genomics made a total of 7 manual joins to the assembly. The linked read data again provides an orthogonal data set with which to validate the scaffolding process. The broken assembly was independently scaffolding using the 10X data and the ARCS algorithm (Yeo et al. 2018) and the results were compared with the results obtained from Hirise. Of the 33 joins made by Hirise, 20 were supported by 10X data. The ARCS software makes a total of 23 joins based on 10X data, i.e., an additional three which were not made by Hirise. Of these three additional 10X based joins, two supported manual edits made by Dovetail genomics while the final one would have linked two chromosome scale scaffolds. Since this last join is biologically implausible it was not incorporated into the final assembly.

The final contig set was screened for contaminants using Kraken2 v2.0.7 (Wood, Lu, and Langmead 2019) and an index built from the Refseq genomes database. Kraken2 was. Run using default settings. In addition, proteins from *Helianthus annuus* (sunflower) were aligned to the assembly with diamond (version 0.9.18) and parameters [--id 70 --subject-cover 80]. Diamond hits were filtered to retain only those with greater than 80% identity and alignment length within +/- 5% of the Sunflower protein length. Contigs with at least one filtered Sunflower gene hit, or which have a Kraken2 annotation of Viridiplantae were considered plant contigs. A total of 33 small contigs remained unclassified but no evidence was found of any contaminant contigs. The assembly was screened for organelle contigs via BLAST and with manual review of the alignments, a total of 697 contigs were removed, 622 (9136324 bp) designated chloroplast and 75 (1625390) designated mitochondria.

3. Genome Annotation

Gene models were annotated using the Robust and Extendable eukaryotic Annotation Toolkit (REAT, https://github.com/EI-CoreBioinformatics/reat) and Minos (https://github.com/EI-CoreBioinformatics/minos) using a workflow incorporating repeat identification, RNA-Seq mapping / assembly, alignment of protein sequences from related species and evidence guided gene prediction with AUGUSTUS (see below). The REAT transcriptome workflow was run with RNA-Seq reads and IsoSeq transcripts from 3 tissues (disc, leaf and ray). Illumina RNA-Seq reads were mapped to the genome with HISAT2 v2.1.0 (Kim et al. 2019) and high-confidence splice junctions identified by Portcullis (Kim et al. 2019; Mapleson et al. 2018). The aligned reads were assembled for each tissue with

StringTie2 v1.3.3 (Kovaka et al. 2019) and Scallop v0.10.2 (Shao and Kingsford 2017), IsoSeq transcripts were aligned with minimap2 (Li H., 2018). From the combined set of RNA-Seq assemblies and IsoSeq alignments a filtered set of non-redundant gene-models were derived using Mikado (https://github.com/EI-CoreBioinformatics/mikado; (Venturini et al. 2018)). The REAT homology workflow was used to generate gene models based on alignment of protein sequences from 11 Asteraceae species. These together with the transcriptome derived models were used to train the AUGUSTUS (Stanke and Morgenstern 2005), gene predictor, with transcript and protein alignments plus repeat annotation provided as hints in evidence guided gene prediction using the REAT prediction workflow. Three alternative AUGUSTUS gene builds were generated using different evidence inputs or weightings. These together with the gene models derived from the transcriptome and homology workflow were consolidated into a single set of gene models using Minos (https://github.com/EI-CoreBioinformatics/minos). The Minos pipeline scores alternative models based on the level of supporting evidence (protein homology, transcriptome data) and gene structure characteristics (e.g., CDS, UTR features) to select a representative gene model and alternative splice variants.

Reference guided transcriptome reconstruction

Gene models were derived from the RNA-Seq and IsoSeq transcripts using the REAT transcriptome workflow (https://github.com/EI-CoreBioinformatics/reat). HISAT2 (Kim et al. 2019) was selected as the short read aligner with IsoSeq transcripts aligned with minimap2 (Kim et al. 2019; Li 2018)), maximum intron length was set as 50,000 bp and minimum intron length to 20bp. IsoSeq alignments were required to meet 90% coverage and 95% identity. High-confidence splice junctions were identified by Portcullis. RNA-Seq reads were assembled for each tissue with StringTie2 (Kovaka et al. 2019) and Scallop (Shao and Kingsford 2017). Gene models were derived from the RNA-Seq assemblies and IsoSeq alignments with Mikado (https://github.com/EI-CoreBioinformatics/mikado; (Venturini et al. 2018)). Mikado was run with all Scallop, StringTie and IsoSeq alignments and a second run with only IsoSeq alignments.

Table 3. Reference guided transcriptome assembly statistics.

Stat	CoDisc (StringTie)	CoDisc (scallop)	CoLeaf (StringTie)	CoDisc (scallop)	CoRay (StringTie)	CoRay (scallop)
Number of genes	96657	79669	88418	79669	84618	68308
Number of Transcripts	162235	200517	141734	200517	138904	171285
Transcripts per gene	1.68	2.52	1.6	2.52	1.64	2.51
Number of monoexonic genes	21899	10074	23357	10074	20140	8672
Monoexonic transcripts	23268	14285	24490	14285	21229	12217
Transcript mean size cDNA (bp)	1669.59	1852.64	1743.52	1852.64	1705.18	1836.14
Transcript median size cDNA (bp)	1375	1593	1420	1593	1395	1586
Min cDNA	200	200	200	200	200	200
Max cDNA	29526	32610	25392	32610	36617	38844
Total exons	871337	1087218	763945	1087218	764548	958715
Exons per transcript	5.37	5.42	5.39	5.42	5.5	5.6
Exon mean size (bp)	310.86	341.69	323.47	341.69	309.8	328.05
Intron mean size (bp)	429.22	399.35	403.29	399.35	410.56	383.84

Table 4. Mikado consolidated gene sets, gene model statistics

Stat	mikado (only IsoSeq)	mikado (IsoSeq + StringTie + scallop)
Number of genes	40650	85954
Number of Transcripts	49415	144680
Transcripts per gene	1.22	1.68
Number of monoexonic genes	4664	10169
Monoexonic transcripts	4683	10256
Transcript mean size cDNA (bp)	2004.98	1753.3
Transcript median size cDNA (bp)	1759	1539
Min cDNA	305	200
Max cDNA	12107	21329
Total exons	356703	821530
Exons per transcript	7.22	5.68

Exon mean size (bp)	277.76	308.77
CDS mean size (bp)	219	214.49
Transcript mean size CDS (bp)	1434.73	982.76
Transcript median size CDS (bp)	1236	801
Min CDS	0	0
Max CDS	11646	16149
Intron mean size (bp)	275.22	363.17
5'UTR mean size (bp)	249.73	313.01
3'UTR mean size (bp)	285.07	356.74

Cross-species protein alignment

Protein sequences from 5 Asteraceae species *Cynara cardunculus* (cardoon), *Mikania micrantha,* (bitter vine), *Lactuca sativa* (lettuce), *Helianthus annuus* (sunflower) and *Artemisia annua* (sweet wormwood) were aligned to the *C. officinalis* assembly using the reat homology workflow (https://github.com/EI-CoreBioinformatics/reat) with options --annotation_filters aa_len --alignment_species Eudicoty --filter_max_intron 20000 --filter_min_exon 10 --alignment_filters aa_len internal_stop intron_len exon_len splicing --alignment_min_coverage 90 --junction_f1_filter 40. The reat homology workflow aligns proteins with spaln v2.4.6 (Gotoh 2021) and filters and generates metrics to remove misaligned proteins. The aligned proteins are clustered into loci and a consolidated set of gene models are derived via Mikado. A second reat homology run was also performed with predicted protein sequences derived from *de novo* transcriptome assemblies and IsoSeq transcripts of *C. arvensis*. Illumina RNA-Seq reads were assembled with trinity v2.8.5, combined with IsoSeq transcripts and filtered and clustered with the tr2aacds pipeline (tr2aacds4.pl, -MINAA=90) of the EvidentialGene package. The filtered set of transcripts (tr2aacds.pl okayset) were further filtered to retain only transcripts with greater than 0.3 TPM as determined via Salmon quantification against the illumina RNA-Seq reads for each species. The CDS sequences of the retained transcripts were translated, and the protein sequences used as input to the reat homology workflow, aligning the proteins to the *C.officinalis* assembly and generating a consolidated set of protein derived gene models.

Evidence guided gene prediction

The evidence guided annotation of protein coding genes based on repeats, RNA-Seq mappings, transcript assembly and alignment of protein sequences was created using the reat prediction workflow. The pipeline has four main steps: (1) the reat transcriptome and homology Mikado models are categorised based on alignments to uniprot proteins to identify models with likely full-length CDS and which meet basic structural checks i.e. having complete but not excessively long UTR's and not exceeding a minimum CDS/cDNA ratio. A subset of gene models were then selected from the classified models and used to train the AUGUSTUS gene predictor (Stanke et al., 2005); (2) Augustus is run in both *ab initio* mode and with extrinsic evidence generated in the reat transcriptome and homology runs (repeats, protein alignments, RNA-Seq alignments, splice junctions, categorised Mikado models). Three evidence guided AUGUSTUS predictions were created using alternative bonus scores and priority based on evidence type. (3). AUGUSTUS models, reat transcriptome / homology models, protein and transcriptome alignments were provided to EVidenceModeler (EVM) (Haas et al. 2008) to generate consensus gene structures. (4) EVM models are processed through Mikado to add UTR features and splice variants.

Gene model consolidation

The final set of gene models was selected using Minos (https://github.com/EI-CoreBioinformatics/minos)**.** Minos is a pipeline that generates and utilises metrics derived from protein, transcript, and expression data sets to create a consolidated set of gene models. In this annotation, the three alternative evidence guided Augustus gene builds described earlier, and the gene models derived from the reat transcriptome and homology runs were filtered and consolidated into a single set of gene models.

Gene models were classified as biotypes protein_coding_gene, predicted_gene and transposable_element_gene, and assigned as high or low confidence based on the criteria below:

a) **High confidence protein_coding_gene:** Any protein coding gene where any of its associated gene models have a BUSCO v4.0.6 (Seppey, Manni, and Zdobnov 2019)protein status of Complete/Duplicated OR have blastp (v2.9.0+) coverage (average across query and target coverage) >= 80% against the list protein datasets mentioned in Section 4. Or alternatively have average blastp coverage (across query and target coverage) >= 60% against the list protein datasets AND have transcript alignment F1 score (average across nucleotide, exon and junction F1 scores based on RNA-Seq transcript assemblies) >= 40%.

b) **Low confidence protein_coding_gene**: Any protein coding gene where all of its associated transcript models do not meet the criteria to be considered as high confidence protein coding transcripts.

c) **High confidence transposable_element_gene**: Any protein coding gene where any of its associated gene models have coverage >= 40% against the combined interspersed repeats (see methods).

d) **Low confidence transposable_element_gene**: Any protein coding gene where all of its associated transcript models do not meet the criteria to be considered as high confidence and assigned as a transposable_element_gene (see c).

e) **Low confidence predicted_gene**: Any protein coding gene where all of its associated transcript models do not meet the criteria to be considered as high confidence protein coding transcripts. And in addition where any of the associated gene models have average blastp coverage (across query and target coverage) < 30% against the list protein datasets mentioned in Section 4 AND having a protein-coding potential score < 0.25 calculated using CPC2 0.1 (Kong et al. 2007).

f) **Discarded models**: Any models having no BUSCO protein hit AND no protein alignment score (average across nucleotide, exon and junction F1 scores based on protein alignments) AND no transcript alignment F1 score (average across nucleotide, exon and junction F1 scores based on RNA-Seq transcript assemblies) AND no blastp coverage (average across query and target coverage) AND Kallisto v0.44 (Bray et al. 2016) expression score <0.3 from across RNA-Seq reads OR having short CDS <30bps.

Table 5. Minos classified gene models

Gene Model Classification			
Biotype	**Confidence**	**Gene**	**Transcript**
protein_coding_gene	High	72324	110138
transposable_element_gene	High	14620	14900
protein_coding_gene	Low	9858	10221
transposable_element_gene	Low	5301	5318
predicted_gene	Low	2352	2383
Total		104455	142960

Functional annotation

All proteins were annotated using EI-FunAnnot pipeline v1.2 (https://github.com/EI-CoreBioinformatics/eifunannot) utilising AHRD v.3.3.3 https://hdl.handle.net/20.500.11811/6420; https://github.com/groupschoof/AHRD/blob/master/README.textile). Sequences were blasted against the reference proteins (Arabidopsis thaliana TAIR10, TAIR10_pep_20101214_updated.fasta.gz - https://www.araport.org) and the UniProt viridiplantae sequences (data download 20Dec2021), both Swiss-Prot and TrEMBL datasets (The UniProt Consortium, 2014). Proteins were BLASTed (v2.6.0; blastp) with an e-value of

1e-5. We have also provided interproscan (v5.22.61) (Jones et al. 2014) results to AHRD. We adapted the standard AHRD example configuration file path test/resources/ahrd_example_input_go_prediction.yml, distributed with the AHRD tool, changing the following apart from the location of input and output files:

1. we included the GOA mapping from uniprot (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gaf.gz) as parameter 'gene_ontology_result',

2. we also included the interpro database (ftp://ftp.ebi.ac.uk/pub/databases/interpro/61.0/interpro.xml.gz) and provided as parameter 'interpro_database',

3. we changed the parameter 'prefer_reference_with_go_annos' to 'false' and did not use the parameter 'gene_ontology_result'

4. The regular expression used to analyse the reference protein fasta header was amended to the custom header format:

fasta_header_regex: "^>(?<accession>[aA][tT][0-9mMcC][gG]\\d+(\\.\\d+)?)\\s+\\|[^\\|]+\\|\\s+(?<description>[^\\|]+)(\\s*\\|.*)?$"

References:

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. “Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm.” *Nature Methods* 18 (2): 170–75.

Gotoh, Osamu. 2021. “Cooperation of Spaln and Prrn5 for Construction of Gene-Structure-Aware Multiple Sequence Alignment.” *Methods in Molecular Biology (Clifton, N.J.)* 2231:71–88.

Guan, Dengfeng, Shane A. McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. 2020. “Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies.” *Bioinformatics (Oxford, England)* 36 (9): 2896–98.

Haas, Brian J., Steven L. Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E. Allen, Joshua Orvis, Owen White, C. Robin Buell, and Jennifer R. Wortman. 2008. “Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments.” *Genome Biology* 9 (1): R7.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics (Oxford, England)* 30 (9): 1236–40.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype.” *Nature Biotechnology* 37 (8): 907–15.

Kong, Lei, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. 2007. “CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine.” *Nucleic Acids Research* 35 (Web Server issue): W345–49.

Kovaka, Sam, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. 2019. “Transcriptome Assembly from Long-Read RNA-Seq Alignments with StringTie2.” *Genome Biology* 20 (1): 278.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Mapleson, Daniel, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J. Clavijo. 2017. “KAT: A K-Mer Analysis Toolkit to Quality Control NGS

Datasets and Genome Assemblies." *Bioinformatics (Oxford, England)* 33 (4): 574–76.
Mapleson, Daniel, Luca Venturini, Gemy Kaithakottil, and David Swarbreck. 2018. "Efficient and Accurate Detection of Splice Junctions from RNA-Seq with Portcullis." *GigaScience* 7 (12). https://doi.org/10.1093/gigascience/giy131.
Nurk, Sergey, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. 2020. "HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads." *Genome Research* 30 (9): 1291–1305.
Putnam, Nicholas H., Brendan L. O'Connell, Jonathan C. Stites, Brandon J. Rice, Marco Blanchette, Robert Calef, Christopher J. Troll, et al. 2016. "Chromosome-Scale Shotgun Assembly Using an in Vitro Method for Long-Range Linkage." *Genome Research* 26 (3): 342–50.
Ruan, Jue, and Heng Li. 2020. "Fast and Accurate Long-Read Assembly with wtdbg2." *Nature Methods* 17 (2): 155–58.
Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. "BUSCO: Assessing Genome Assembly and Annotation Completeness." *Methods in Molecular Biology (Clifton, N.J.)* 1962:227–45.
Shao, Mingfu, and Carl Kingsford. 2017. "Accurate Assembly of Transcripts through Phase-Preserving Graph Decomposition." *Nature Biotechnology* 35 (12): 1167–69.
Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics (Oxford, England)* 31 (19): 3210–12.
Stanke, Mario, and Burkhard Morgenstern. 2005. "AUGUSTUS: A Web Server for Gene Prediction in Eukaryotes That Allows User-Defined Constraints." *Nucleic Acids Research* 33 (Web Server issue): W465–67.
Venturini, Luca, Shabhonam Caim, Gemy George Kaithakottil, Daniel Lee Mapleson, and David Swarbreck. 2018. "Leveraging Multiple Transcriptome Assembly Methods for Improved Gene Structure Annotation." *GigaScience* 7 (8). https://doi.org/10.1093/gigascience/giy093.
Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20 (1): 257.
Yeo, Sarah, Lauren Coombe, René L. Warren, Justin Chu, and Inanç Birol. 2018. "ARCS: Scaffolding Genome Drafts with Linked Reads." *Bioinformatics (Oxford, England)* 34 (5): 725–31.