Article

Parallel evolution of cannabinoid biosynthesis

Paula Berman [1,7] ✉, Luis Alejandro de Haro [1,7], Adam Jozwiak [1,7], Sayantan Panda [1], Zoe Pinkas[1], Younghui Dong[1], Jelena Cveticanin[1], Ranjit Barbole [2,3], Rotem Livne[1], Tali Scherf [4], Eyal Shimoni[4], Smadar Levin-Zaidman[4], Nili Dezorella[4], Ekaterina Petrovich-Kopitman[5], Sagit Meir [1], Ilana Rogachev[1], Prashant D. Sonawane [1,6] ✉ & Asaph Aharoni [1] ✉

Modulation of the endocannabinoid system is projected to have therapeutic potential in almost all human diseases. Accordingly, the high demand for novel cannabinoids stimulates the discovery of untapped sources and efficient manufacturing technologies. Here we explored *Helichrysum umbraculigerum*, an Asteraceae species unrelated to *Cannabis sativa* that produces *Cannabis*-type cannabinoids (for example, 4.3% cannabigerolic acid). In contrast to *Cannabis*, cannabinoids in *H. umbraculigerum* accumulate in leaves' glandular trichomes rather than in flowers. The integration of de novo whole-genome sequencing data with unambiguous chemical structure annotation, enzymatic assays and pathway reconstitution in *Nicotiana benthamiana* and in *Saccharomyces cerevisiae* has uncovered the molecular and chemical features of this plant. Apart from core biosynthetic enzymes, we reveal tailoring ones producing previously unknown cannabinoid metabolites. Orthology analyses demonstrate that cannabinoid synthesis evolved in parallel in *H. umbraculigerum* and *Cannabis*. Our discovery provides a currently unexploited source of cannabinoids and tools for engineering in heterologous hosts.

For many decades, *Cannabis sativa* L. (*Cannabis*) has been the predominant plant species associated with cannabinoid production. These molecules exert diverse biological and pharmacological effects via the modulation of the endocannabinoid system[1–4]. In *Cannabis*, major cannabinoids such as cannabigerolic acid (CBGA), (−)-Δ[9]-*trans*-tetrahydrocannabinolic acid (Δ[9]-THCA), cannabichromenic acid (CBCA) and cannabidiolic acid (CBDA) are synthesized and accumulate in the glandular trichomes of female inflorescences (Fig. 1a)[5–8]. Structurally related molecules have also been identified in several species[9–11], although some claims still require additional proof[11].

One of these plants is *Helichrysum umbraculigerum* Less, an Asteraceae plant that has been used in fumigation rituals and traditional medicine by native South African tribes (Fig. 1b). More than 40 years ago, Bohlmann and Hoffmann[12] identified CBGA in this plant, making it the only known species other than *Cannabis* from which a *Cannabis*-like cannabinoid has been isolated and characterized[11]. However, a more recent study could not confirm the presence of CBGA in this species[13].

In our work, we found that CBGA accumulates in all the aerial parts of *H. umbraculigerum*, reaching up to 4.3% of dry weight in leaves. Its accumulation, along with additional new cannabinoids, was

[1]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot, Israel. [2]Biochemical Sciences Division, CSIR-National Chemical Laboratory, Pune, India. [3]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India. [4]Department of Chemical Research Support, Weizmann Institute of Science, Rehovot, Israel. [5]Life Science Core Facilities, Weizmann Institute of Science, Rehovot, Israel. [6]Present address: Department of Natural Product Biosynthesis, Max-Planck Institute for Chemical Ecology, Jena, Germany. [7]These authors contributed equally: Paula Berman, Luis Alejandro de Haro, Adam Jozwiak. ✉e-mail: bermansh@gmail.com; psonawane@ice.mpg.de; asaph.aharoni@weizmann.ac.il

accompanied by a diversity of other classes of metabolites, including amorfrutins, prenyl-acyl-phloroglucinoids, prenylchalcones and prenyl-flavanones. We located this assortment in glandular trichomes of leaves and flowers and suggest that their biosynthesis arises via five parallel pathways. Furthermore, we discovered three core cannabinoid biosynthetic enzymes and two previously unknown types absent in *Cannabis* that glycosylate and *O*-acylate cannabinoids, generating previously unreported cannabinoids. Our findings exemplify a case of parallel evolution in cannabinoid formation in two genetically distant plant families, providing new enzymes for metabolic engineering strategies. Moreover, we demonstrate the potential for another perennial, fast-growing, commercially viable plant source for bioactive cannabinoids.

Results

H. umbraculigerum produces CBGA

As two earlier reports regarding the presence of cannabinoids (specifically CBGA (**1**)) in *H. umbraculigerum* were contradictory[12,13], we decided to carry out comprehensive chemical profiling of cannabinoids in various *H. umbraculigerum* tissues. We confirmed that **1** is a major component of *H. umbraculigerum*, accumulating up to 4.3% on a dry-weight basis in leaves (Fig. 1c,d), comparable to the maximum typically measured concentrations in inflorescences of medical *Cannabis* chemotypes (Fig. 1d)[14]. CBGA, its phenethyl analogue heliCBGA (**2**) and pre-amorphastilbol (APHA) (**3**), the stilbene form of **2**, represent three of the major peaks in the total ion chromatogram of a fresh-leaf ethanolic extract (Fig. 1c,e, Extended Data Fig. 1a and Supplementary NMR Data 1 and 2).

We predicted that **1** and **2** biosynthesis originates from hexanoic acid and phenylalanine, respectively (Fig. 1a). We therefore fed *H. umbraculigerum* leaves with unlabelled and stable isotopically labelled hexanoic acid (hexanoic-D_{11} acid) or phenylalanine (phenylalanine-D_5 or phenylalanine-$^{13}C_9$) and compared the labelled versus non-labelled masses and their respective tandem mass spectrometry (MS/MS) spectra (Extended Data Fig. 1b). Consequently, newly derived isotopologues were detected as co-eluting chromatographic peaks (unlabelled and labelled forms) with mass shifts and MS/MS fragmentation patterns corresponding to the isotopically labelled parts of the molecule. These findings validate the existence of the alkyl and aralkyl cannabinoids in *H. umbraculigerum* and confirm that their biosynthesis derives from the polyketide and phenylpropanoid pathways, respectively. Feeding experiments revealed the presence of some additional major prenyl-acyl-phloroglucinoids, prenylchalcones and prenylflavanones with similar chemical formulas as **1**–**3**. On the basis of previously identified core structures[12,13] and each metabolite MS/MS fragmentation spectrum, we assigned these peaks to the structures shown in Fig. 1e (Supplementary Figs. 1 and 2; geranylphlorocaprophenone (**4**) and geranyl-pinocembrin chalcone (**6**) structures were confirmed by nuclear magnetic resonance (NMR) spectroscopy, Supplementary NMR Data 3 and 4).

Cannabinoids accumulate in glandular trichomes. We employed various high-resolution imaging technologies to examine whether,

like *Cannabis*, *H. umbraculigerum* develops and accumulates cannabinoids in glandular trichomes. We found that in flowers, the involucral bracts of the capitula had numerous non-glandular and glandular trichomes. In individual florets, glandular trichomes were particularly abundant on the tips of the corolla lobe (Extended Data Fig. 2a,b). In leaves, both the adaxial and abaxial surfaces were densely covered with both non-glandular and glandular trichomes (Fig. 1f). The glandular trichomes were slightly elevated from the epidermis and consisted of a biseriate stalk and a globose head (Extended Data Fig. 2c). Two disk cells (DCs) were observed in the subcuticular space of the globose head (Fig. 1g). In *Cannabis*, cannabinoid biosynthesis takes place in these cells[10]. The multicellular biseriate structure of the trichomes further consisted of two basal cells (not always observed), stalk cells, neck cells and a secretory cavity (SCv) (Fig. 1h). DCs of trichomes at the secretion stage showed exudation of electron-transparent secretions from plastids into vesicles, followed by exocytosis of their contents into the periplasmic space (PSP), where they accumulated prior to secretion into the SCv (Fig. 1i and Extended Data Fig. 2d–f).

We next applied matrix-assisted laser desorption/ionization–MS imaging (MALDI–MSI) to spatially localize cannabinoids in *H. umbraculigerum*. We first analysed the abaxial and adaxial leaf surfaces following partial removal of trichomes (Extended Data Fig. 2g,h; imaging m/z (M + H)$^+$ = 361.237 Da corresponding to **1** and **4**). As shown, metabolites were detected in the intact parts, while areas with partially or fully removed trichomes showed less or no signals, respectively. We further analysed cross-sections of *H. umbraculigerum* leaves and flowers. We sectioned leaves crosswise so that trichomes on the adaxial and abaxial parts were exposed on each side (Fig. 2a). In flowers, we sectioned the receptacle, exposing trichomes on the outer surface of the involucral bracts (Extended Data Fig. 2i). As shown in Fig. 2b and Extended Data Fig. 2j for the leaf and flower samples, respectively, **1** was found exclusively in glandular trichomes.

H. umbraculigerum produces classical and previously unknown cannabinoids. *Cannabis* produces various CBGA-type analogues with aliphatic chains of different lengths (one to seven carbons), derived from different linear short-chain and medium-chain fatty acids (FAs)[9,15,16]. We observed several of these analogues in leaves of *H. umbraculigerum*, including cannabigerovarinic acid (**9**), cannabigerol butyric acid (**10**), cannabigerohexolic acid (**11**) and cannabigerophorolic acid (**12**), corresponding to three-, four-, six- and seven-carbon-atom chains, respectively (Supplementary Fig. 3). We also observed two metabolites with similar masses and fragmentation patterns as **1** and **11**, which we assigned as cannabinoids derived from branched FAs (**13** and **14**, respectively; Supplementary Fig. 3). These branched cannabinoids have not been identified in *Cannabis*. We also found small amounts of CBCA (**15**) and its aromatic analogue helicannabichromenic acid (**16**) and their hydroxylated forms (**17** and **18**, respectively; Supplementary Fig. 4), and the isoprenyl forms of **1** and **2** according to MS/MS fragmentation (CBPA (**19**) and heliCBPA (**20**), respectively, Supplementary Fig. 5). We did not detect Δ^9-THCA-type or CBDA-type cannabinoids in any of the tissues.

Fig. 1 | *H. umbraculigerum* biosynthesizes CBGA and other terpenophenols in all aerial plant parts. a, Proposed biosynthetic pathways of **1** and **2**. **b**, Photographs of the *H. umbraculigerum* plant inflorescence (top) and shoot (bottom). **c**, Total ion chromatogram of an ethanolic extract of fresh *H. umbraculigerum* leaves. The most abundant peaks of identified metabolites are marked and colour-coded according to the class of terpenophenol. **1** and **2** are highlighted in red and blue, respectively. **d**, Absolute quantification of **1** in different plant tissues (%w/w per fresh weight, mean ± s.d.; n = 3; for lyophilized leaves, %w/w per dry weight (DW), mean ± s.d.; n = 5). Reported *Cannabis* values[14] are provided for comparison. ND, not detected. **e**, Chemical structures and names of selected terpenophenols with similar chemical formulas as **1**–**3**. **f,g**, Representative cryo scanning electron microscopy (cryo-SEM) (**f**) and confocal micrographs (**g**)

of the adaxial top-view domain of leaves showing stalked glandular trichomes (marked by arrows). The micrographs are representative of multiple (n > 3) leaf areas sampled on the same day. **h**, Transmission electron microscopy (TEM) micrograph showing the multicellular structure of the different cell types in a stalked glandular trichome at the secretory stage. BC, basal cell; SC, stalk cell; NC, neck cell; V, vacuole; SCv, secretory cavity. The dashed line marks the surface of the SCv. **i**, High-magnification image showing the ultrastructure of DCs. CW, cell wall; M, mitochondria; P, plastid; Vs, vesicle; PSP, periplasmic space. The arrows mark active secretions from vesicles to the PSP by exocytosis. The micrographs are representative of multiple trichomes (n > 3) from different sections of young leaves.

Some additional peaks had MS/MS fragments and chemical formulas corresponding to one or two hydroxylations of the metabolites with five-carbon-atom chains, which were labelled following feeding with hexanoic-D_{11} acid (**21–33**; Supplementary Fig. 6). Interestingly, hydroxylated amorfrutins were observed with similar fragmentation patterns as the cannabinoids (with an m/z difference of 33.984 Da), suggesting similar chemical structures and enzymes associated with their metabolism (**34–46**; Supplementary Fig. 6). We purified metabolite **26** from this group and identified a new tetrahydroxanthane-type cannabinoid by NMR spectroscopy (12-OH-cyclocannabigerolic acid (**26**); Supplementary NMR Data 5). According to its MS/MS fragmentation pattern, we also putatively identified cyclocannabigerolic acid (**47**) and analogous amorfrutin types (12-OH-heli-cyclocannabigerolic acid (**39**) and heli-cyclocannabigerolic acid (**48**); Supplementary Fig. 7).
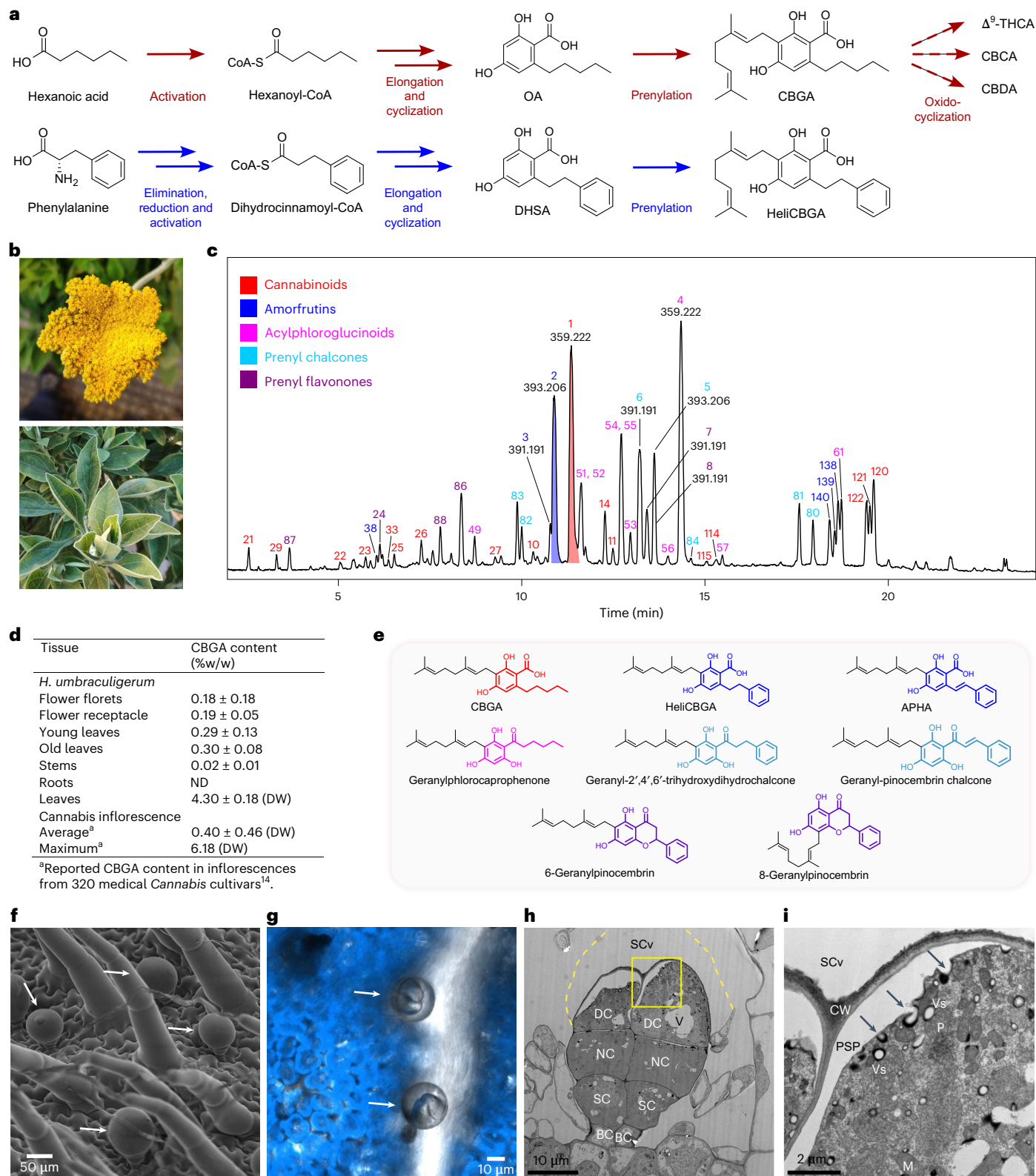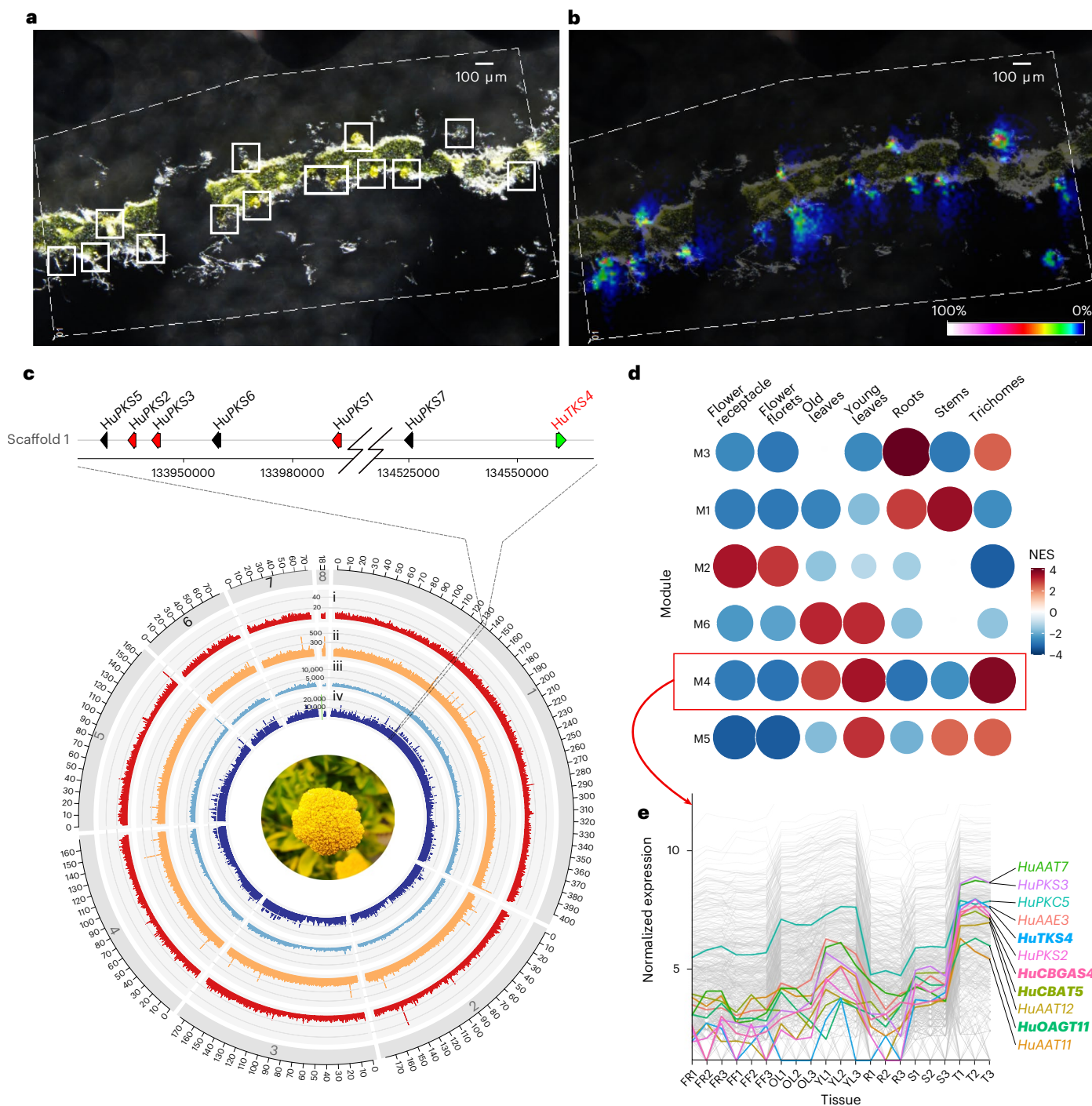
Fig. 2 | Cannabinoid-associated gene expression is correlated with cannabinoid metabolite accumulation in *H. umbraculigerum* glandular trichomes. a,b, Optical image (**a**) and MALDI–MSI of *m/z* 361.23 ± 0.01 Da (**b**) of a cross-sectioned leaf showing that **1** accumulates in stalked glandular trichomes of leaves. Glandular trichomes in **a** are marked to improve interpretation. The signals in **b** correspond to the protonated *m/z* of **1** and **4**. The analyses are representative of two individual samples. **c**, Genomic landscape of the eight longest scaffolds of *H. umbraculigerum* assembly. Track i represents the gene density, ii represents repeat element density, iii represents 3′ Tran-Seq coverage and iv represents TruSeq coverage. These metrics are calculated in 0.1 Mb non-overlapping windows. Magnification of the marked area in scaffold 1 reveals a tandem gene cluster containing seven *PKS*s. The enzymes HuPKS1–3 and HuTKS4 were cloned and functionally characterized in this study. **d**, Normalized enrichment score (NES) of each co-expressed module in each tissue. Module M4 is highlighted because it is highly expressed in trichomes and leaves. **e**, The expression profile of module M4. The expression levels of individual genes are shown in grey lines. The coloured lines highlight the expression of candidate genes from the pathway. Active enzymes identified and functionally characterized in this study are marked in bold. FR, flower receptacles; FF, flower florets; OL, old leaves; YL, young leaves; R, roots; S, stems; T, trichomes.

According to our feeding experiments, prenyl-acyl-phloroglucinoids, prenylchalcones and prenylflavanones were derived from similar precursors as the cannabinoids and amorfrutins (**49–91**; Supplementary Figs. 8–12 and Supplementary NMR Data 6–10). The identified metabolites **1–91** are summarized in Supplementary Tables 1–5.

Proposed biosynthetic pathways in *H. umbraculigerum*. We postulated that the core cannabinoid pathway leading to **1** in *H. umbraculigerum* consists of similar types of enzymes and reactions as in *Cannabis* (Extended Data Fig. 3). These include an acyl-activating enzyme (AAE) for the activation of hexanoic acid into hexanoyl-CoA, a type III polyketide synthase (PKS) and a polyketide cyclase (PKC) to produce olivetolic acid (OA) (**92**), and a membrane-bound aromatic prenyl transferase (PT) for the geranylation of **92** to **1**. In addition to **1** and other cannabinoids, we propose that all the identified terpenophenols are produced via five parallel pathways (Extended Data Fig. 3). According to this scheme, cannabinoids and phloroglucinoids derive from a common linear or branched FA precursor activated via the same AAE. Amorfrutins and chalcones derive from cinnamic or coumaric acids, which originate from phenylalanine, and are also activated via an AAE (similar or different from the polyketide one). These activated intermediates can be further reduced by a double bond reductase to form dihydro intermediates. The activated precursors are elongated using three malonyl CoAs by one or more PKS-type enzymes and are further cyclized by the PKS in a Claisen reaction to form the phloroglucinoid or chalcone backbone, or in an aldol reaction assisted by a PKC to form the cannabinoids and amorfrutins. The fifth pathway employs a chalcone isomerase enzyme that cyclizes chalcones to flavanones. All these intermediates are further prenylated by one or more PTs to form the different types of terpenophenols. Although most of the molecules enclosed monoprenyls, other prenyl types were also observed. The terpenophenols can be further cyclized by berberine-bridge-like enzymes to produce cyclized metabolites such as **15**, cyclocannabinoids and cycloamorfrutins (**26**, **47**, **39** and **48**), and cyclophloroglucinoids previously identified by Pollastro et al.[13]. Additional functional groups and rearrangements include hydroxylation, double bond isomerization and reduction, among others. In support of these five pathways, we identified in *H. umbraculigerum* the primary intermediates (before prenylation) from all the corresponding metabolic routes (**92**–**101**; Supplementary Fig. 13 and Supplementary Table 6).

Elucidation of the core cannabinoid pathway. To identify the enzymes responsible for cannabinoid biosynthesis in *H. umbraculigerum*, we obtained a haplotype resolved dual genome assembly using 44× Pacbio HiFi reads and 200 M reads of Illumina HiC chromatin interaction data (haploid size of ~1.3 Gb; Supplementary Figs. 14 and 15 and Supplementary Table 7). After scaffolding, the N50 of the primary assembly was 174 Mb with eight scaffolds >10 Mb (Fig. 2c, Supplementary Fig. 16 and Supplementary Table 7). We also obtained RNAseq data using PacBio Iso-Seq, Illumina TruSeq and Illumina UMI-aware 3′ Transeq of different tissues (Supplementary Tables 8–10). The genome was soft masked (Supplementary Table 11), and gene models were obtained reaching BUSCO completeness values of 98.7% for the primary assembly and 99.3% for all transcripts, including those missing in the genome (Supplementary Fig. 17). On the basis of Fig. 2b, we expected that the biosynthetic genes would be highly expressed in trichomes. Weighted gene co-expression network analysis of *H. umbraculigerum* tissue transcriptomic data (Supplementary Table 10) revealed a transcriptional module enriched in FA and terpenoid biosynthetic genes induced in trichomes and leaves (Fig. 2d,e and Supplementary Fig. 18). This module included two AAEs, three PKSs, one stress-related protein (a potential PKC) and one PT (Fig. 2e and Supplementary Table 12). Notably, these three PKSs were also located in a tandem gene cluster consisting of seven enzymes of the same type (Fig. 2c). This region exhibited strong footprints of long terminal repeat transposition activity, which might explain the observed patterns of gene duplication (Supplementary Table 13). Overall, we selected six HuAAEs, four HuPKSs, five HuPKCs and four HuPTs for further characterization (Fig. 3a and Supplementary Table 14; Hu signifies *H. umbraculigerum*). The four selected PKSs

showed subtle amino acid differences that would have been overlooked without the genomic sequence, and the inability to amplify the different variants from complementary DNA led us to produce the genes synthetically.

The first step in cannabinoid biosynthesis involves the formation of acyl-CoA thioesters by members of the AAE superfamily. As different acyl moieties are substrates for these enzymes, we tested acetic, butyric, hexanoic, octanoic, cinnamic and coumaric acids. In vitro assays with purified recombinant proteins showed that HuAAE2 and HuAAE4 efficiently produced butyryl-CoA and that HuAAE2 presented higher activity against acetic acid and formed acetyl-CoA (Fig. 3b and Extended Data Fig. 4a). HuAAE6 (HuCoAT6) was the only enzyme with activities towards both medium-chain alkyl (for example, hexanoic and octanoic acids) and aralkyl (for example, cinnamic and coumaric acids) precursors required for the five types of terpenophenols observed in *H. umbraculigerum*. Interestingly, while HuAAE4 belongs to the same clade as the most active *Cannabis* enzyme[5], HuCoAT6 is located within the clade of long-chain acyl-CoA synthetases (Extended Data Fig. 5a and Supplementary Table 15).

In *Cannabis*, the next step is performed by a coupled enzymatic reaction involving olivetol synthase (CsOLS)[6] and the accessory protein OA cyclase (CsOAC)[7], resulting in the condensation of hexanoyl-CoA with three molecules of malonyl-CoA to yield **92**. In in vitro assays, derailment of the unstable intermediates occurs, producing additional by-products not naturally identified in plant extracts (olivetol, pentyl acyl diacetic acid lactone (PDAL) and hexanoyl acyl triacetic acid lactone (HTAL); Extended Data Fig. 4b). PDAL and HTAL are produced by spontaneous lactonization of the tri- and tetra-ketide unstable intermediates, whereas olivetol is produced by CsOLS in the absence of CsOAC in an aldol decarboxylation cyclization reaction resembling the production of resveratrol by a stilbene synthase (STS)[17]. When CsOAC is also present in the reaction, **92** is produced at the expense of olivetol. Here we cloned and expressed in *Escherichia coli* HuPKS1-4, HuPKC1-5, CsOLS and CsOAC enzymes, and using hexanoyl-CoA and malonyl-CoA, we tested their ability to form **92** in coupled in vitro assays with all the possible combinations (Supplementary Fig. 19a). In the absence of PKCs, all the HuPKSs produced the PDAL and HTAL by-products, while HuPKS1, HuPKS2 and HuPKS4 also produced olivetol (Fig. 3c and Extended Data Fig. 4c). When the reactions were performed coupled to CsOAC, olivetol decreased and **92** increased, especially for HuPKS4 (HuTKS4) (Fig. 3c and Extended Data Fig. 4c). However, in all the reactions, we observed considerably smaller amounts of olivetol and **92** than HTAL and PDAL (Fig. 3c). Interestingly, regardless of CsOAC, all HuPKSs produced the phloroglucinoid precursor phlorocaprophenone (PCP) (**95**), present in *H. umbraculigerum* (Fig. 3c and Supplementary Fig. 13). This suggests that the same HuPKS enzyme can carry out both the aldol and Claisen cyclization reactions. This phenomenon has been observed previously for chalcone synthase (CHS) and STS enzymes producing different amounts of both naringenin and resveratrol[18] and for PKSs producing different ratios of both resorcinolic-acid and phloroglucinoid products[19,20]. Interestingly, the HuPKS protein sequences did not cluster with known resorcinolic-acid- or phloroglucinoid-producing PKSs such as CsOLS, *Rhododendron dauricum* orcinol synthase (RdOS)[20] or *Humulus lupulus* valerophenone synthase (HlVPS)[21] (Extended Data Fig. 5b). None of the combinations including HuPKS1 through HuPKS4 or CsOLS with the HuPKC enzymes (selected on the basis of their expression profile and sequence homology to CsOAC) resulted in the formation of **92** (Supplementary Fig. 19a,b). This suggests that the cyclization and possibly stabilization of the tetraketide intermediate is mediated by a different type of enzyme than in *Cannabis*. This was previously suggested to occur in *R. dauricum* in the production of orselinic acid by RdOS and a yet-to-be-identified PKC enzyme[20]. Alternatively, *H. umbraculigerum* may contain another CsOAC homologue that we did not characterize in this study.
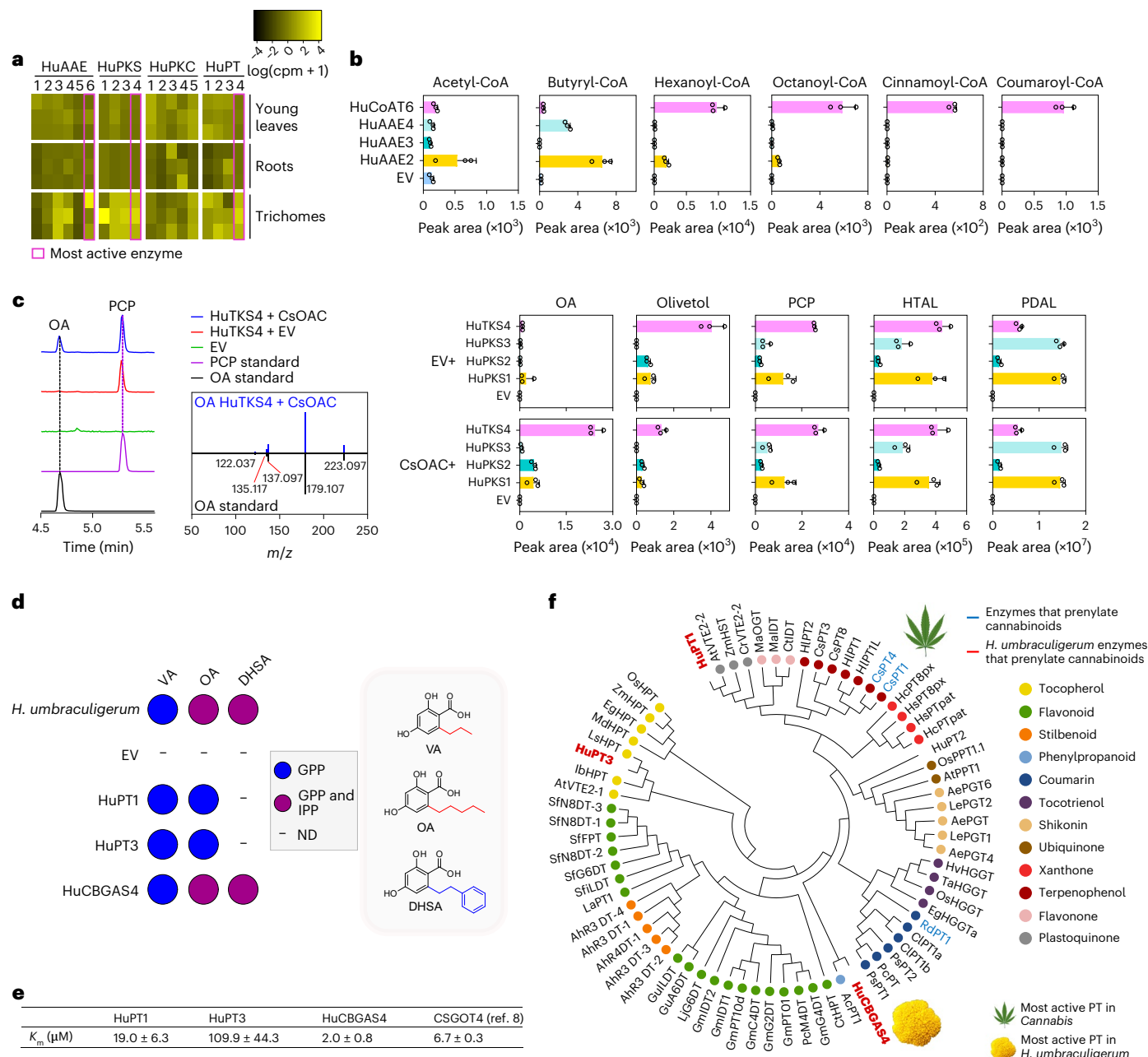
Fig. 3 | Discovery of the core cannabinoid biosynthetic pathway enzymes.
a, Gene expression in young leaves, roots and trichomes of the putative enzymes characterized in this study (log(cpm + 1), n = 3); cpm, counts per million. **b**, Products of recombinant enzyme assays of purified HuAAE proteins using various alkyl (short- and medium-chain FAs) and aromatic (cinnamic and coumaric acids) substrates. Peak areas of ion abundance were used for the comparisons (mean ± s.d.; n = 3 technically independent experiments). CoAT, acyl-CoA-transferase; EV, empty vector. **c**, Products of coupled recombinant enzyme assays of HuPKSs with either an EV or CsOAC, in the presence of hexanoyl-CoA and malonyl-CoA. Peak areas of ion abundance were used for the comparisons (mean ± s.d.; n = 3 technically independent experiments). **92** and **95** were identified using analytical standards ([M − H]⁻ = 223.097 Da). **d**, Activity assay of microsomal fractions expressing PTs using an array of

aromatic substrates and either GPP or IPP as the isoprenoid donors. The circles represent observed mono- or iso-prenylated products in *H. umbraculigerum* or in vitro assays. VA, divarinolic acid; CBGAS, CBGA synthase. **e**, Steady-state kinetic analysis of HuPT1, HuPT3 and HuCBGAS4 with **92** and GPP. The Michaelis–Menten K_m values were calculated using varying (0.5 μM–3 mM) and constant (1 mM) concentrations of each substrate (mean ± s.d.; n = 3 technically independent experiments). The literature K_m value of *Cannabis* CsGOT4 (ref. 8) was added for comparison. **f**, Phylogenetic analysis of PT proteins from *H. umbraculigerum* and other plants. The selection of the proteins was based on functionally characterized enzymes as described by de Bruijn et al.[22]. The clades according to the different substrates are marked in coloured circles. A full list of protein IDs is available in Supplementary Table 15.

In the next step, **92** or OA derivatives are prenylated by aromatic PTs to form **1** and its derivatives. We expressed four enzymes in yeast and purified the microsomal fractions used for enzymatic assays (HuPT1–4; Fig. 3d). We examined an array of aromatic substrates and either geranyl pyrophosphate (GPP) or isopentenyl pyrophosphate

(IPP) as the isoprenoid donors. All the HuPTs geranylated **92** and divarinolic acid to yield **1** and **9**, respectively. HuPT4 also geranylated the aromatic dihydrostilbenic acid (DHSA) (**93**), and was the only enzyme that isoprenylated **92** and **93** (Fig. 3d). HuPT4 was also active with farnesyl pyrophosphate, yielding sesquicannabigerolic acid (Extended

Data Fig. 4d). Kinetic assays of the HuPTs with GPP and **92** revealed that HuPT4 (HuCBGAS4) exhibited a smaller Michaelis−Menten K_m value than the reported one from *Cannabis* CsPT4 (ref. 8) (Fig. 3e and Extended Data Fig. 4e). Interestingly, none of the HuPTs prenylated the phloroglucinoid or chalcone intermediates, and none of their sequences clustered with previously known terpenophenolic PTs[22] (Fig. 3f).

To get more insight into the evolution of the pathway, we searched for orthologous enzymes in *Cannabis* and in all other Asteraceae species with annotated genomes. To the best of our knowledge, these species do not accumulate terpenophenols. Similar to the phylogenetic relationships observed for functionally tested enzymes (that is, AAEs, PKSs and PTs; Extended Data Fig. 5), the enzymes that enabled *H. umbraculigerum* to produce cannabinoids evolved independently in this lineage (Supplementary Orthology Data 1−4 and Supplementary Tables 16−18). Particularly for the PKS-type enzymes, multiple instances of gene duplication and subsequent specialization are likely to have occurred within this family (Supplementary Orthology Data 3 and Supplementary Table 17). Interestingly, the PTs from *Cannabis* and those from *H. umbraculigerum* did not cluster in the same orthogroup (Supplementary Orthology Data 4 and Supplementary Table 18), suggesting that they are derived from evolutionarily distant ancestors.

Uridine-diphosphate-glycosyltransferase-type and alcohol acyl-transferase-type enzymes decorate cannabinoids. Glycosylated cannabinoids have not been reported to occur naturally in planta. Here we identified glucosylated OA (Glc-OA) (**102**) (Supplementary NMR Data 11) and Glc-DHSA (**103**) (Supplementary NMR Data 12) as well as glucosylated C_3-C_6 alkyl-chain intermediates (**104**−**108**), Glc-CBGA (**109**) and Glc-heliCBGA (**110**), and their isoprenylated forms (Glc-CBPA (**111**) and Glc-heliCBPA (**112**)) (Supplementary Figs. 20 and 21 and Supplementary Table 6). All these metabolites exhibited neutral losses of 162.053 Da corresponding to hexose and similar fragments as the non-glucosylated compounds. No di-glucosylated metabolites were identified in the extracts. In *Arabidopsis thaliana*, uridine 5′-diphospho-glucuronosyltransferases (AtUGT89B1, AtUGT71B1, AtUGT75B1 and AtUGT71B2; the accession numbers are provided in Supplementary Table 15) catalyse the glycosylation of several hydroxybenzoic acids (HBA and DHBAs)[23], which are structurally like **92** (Fig. 4a). We selected 13 gene candidates in *H. umbraculigerum* on the basis of sequence similarity to these proteins and positive correlations between gene expression and the accumulation of glucosylated metabolites (HuUGT1-13; Fig. 4a,b and Supplementary Table 19).

Of the 13 UGTs from *H. umbraculigerum*, 11 were expressed in *E. coli* and examined for enzyme activity using **92**, **1** and **2** in a reaction including uridine diphosphate glucose (UDP-Glc) as the sugar donor. Eight of the 11 enzymes showed activity on the different substrates, including HuUGT1-2, HuUGT4-7, HuUGT11 and HuUGT13 (Extended Data Fig. 6a). The production of **102** in the enzyme assays with

UDP-Glc was supported by the NMR assignment of the glucose moiety (Supplementary NMR Data 11). We next purified the four most active enzymes (HuUGT1, HuUGT6, HuUGT11 and HuUGT13) and performed in vitro assays with an array of cannabinoid substrates, both natural and unnatural to *H. umbraculigerum* (Fig. 4d, Extended Data Fig. 6b, Supplementary Figs. 22 and 23, and Supplementary Table 20). We also included enzymes from stevia and rice (SrUGT and OsUGT, respectively) reported to possess cannabinoid glycosylation activity despite these plants not producing cannabinoids[24,25]. All enzymes were active with varying substrate specificity and products. For example, HuUGT1 and HuUGT6 were most active on the cannabinoids (HuCBUGT1 and HuCBUGT6), while HuUGT11 (HuOAUGT11) and HuUGT13 were highly active on the cannabinoid intermediates but almost inactive on the prenylated metabolites. Di-glucosylation of acid metabolites was observed only in HuCBUGT6, while olivetol, cannabidiol and cannabigerol were di-glucosylated by different HuUGTs depending on the metabolite (Supplementary Fig. 23). Interestingly, the UGTs from *H. umbraculigerum* also glucosylated the phloroglucinoid and flavonoid precursors naturally present in the plant (Extended Data Fig. 6b), even though the glucosylated forms were not observed in the plant extracts. Kinetic assays of HuOAUGT11, HuUGT13, OsUGT and SrUGT showed substantial catalytic activity of HuOAUGT11 with **92** and UDP-Glc compared with all other enzymes (Fig. 4c and Extended Data Fig. 6c). HuOAGT11 was also co-expressed with other cannabinoid-related enzymes (Supplementary Table 12) and is therefore the most likely enzyme responsible for the large quantities of **102** and **103** produced in *H. umbraculigerum* (Supplementary Fig. 20).

Previous reports identified isoprenylated *O*-acylated amorfrutins in *H. umbraculigerum*[12,13] but not geranylated or alkyl-type ones, which are also not found in *Cannabis*. Here we identified a diverse group of *O*-acylated cannabinoids and amorfrutins including the *O*-acylated alkyl (**113**−**130**) and aralkyl (**131**−**141**) metabolites (Supplementary Figs. 24−26 and Supplementary Tables 1 and 2). We hypothesized that the acyl group is derived from short- or medium-chain FAs (Extended Data Fig. 3) and verified this using a precursor isotope-labelling approach (Extended Data Fig. 7a). Most of the alkyl cannabinoids in this group had five-carbon-atom tails (according to labelling with hexanoic-D_{11} acid), and both alkyl and aralkyl metabolites comprised iso- or monoprenyls and linear or branched short-chain *O*-acyl groups, as shown by the specific labelling (Supplementary Figs. 24−26). To confirm the identification of this group of metabolites, we purified *O*-methylbutyryl-cannabigerolic acid (**120**) and *O*-methylbutyryl-helicannabigerolic acid (**138**) and confirmed their structure by NMR (Supplementary NMR Data 13 and 14).

O-Acylation of specialized metabolites in plants is frequently catalysed by BAHD-type (named according to the first letter of each of the first four biochemically characterized enzymes of this family (BEAT, AHCT, HCBT, and DAT)[26]) alcohol acyl-transferase (AAT) enzymes. We therefore selected 15 *H. umbraculigerum* BAHD
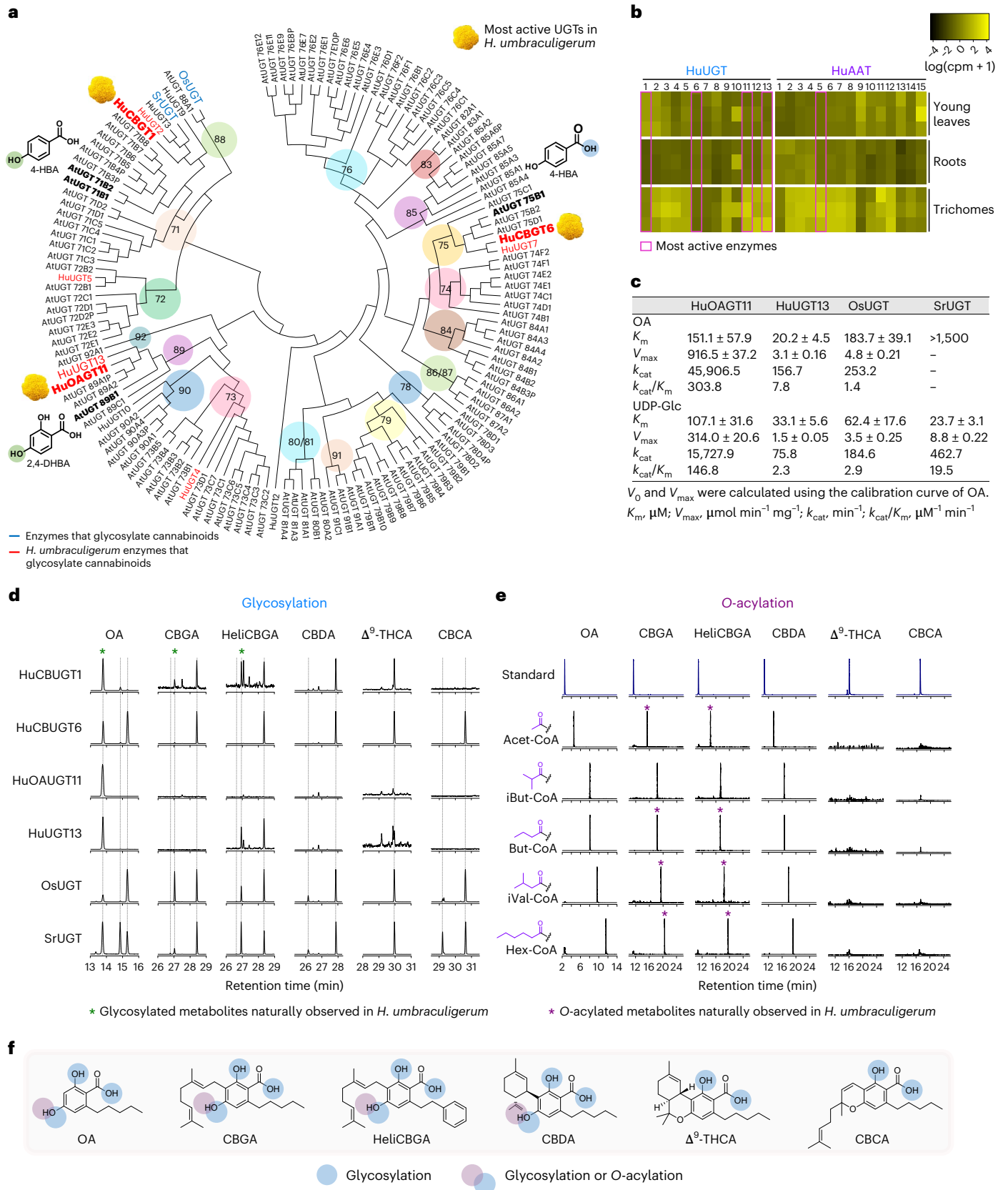

Fig. 4 | Functional characterization of cannabinoid tailoring enzymes.
a, Phylogenetic analysis of selected UGT proteins from *H. umbraculigerum*, *Arabidopsis thaliana*, *Oryza sativa* and *Stevia rebaudiana*. The clades were annotated according to the *A. thaliana* UGT family classification (the numbers in the coloured circles). The full list of protein IDs is available in Supplementary Table 15. 4-Hydroxybenzoic acid (4-HBA) and 2,4-dihydroxybenzoic acid (2,4-DHBA), which have structural similarities to **92** and **1**, are located next to the UGT enzymes that glycosylate them[23]. Glycosylated hydroxyls are highlighted. **b**, Gene expression in young leaves, roots and trichomes of the putative UGT and AAT enzymes characterized in this study (log(cpm+1), n = 3); cpm, counts per million. **c**, Comparison of steady-state kinetic analysis of HuOAGT11 and HuUGT13 versus OsUGT and SrUGT, with **92** and UDP-Glc. The assays were performed using varying (0.5 μM–3 mM) and constant (1 mM) concentrations of each substrate (mean ± s.d.; n = 3 technically independent experiments). **d**, Extracted ion chromatograms of monoglucosides according to the theoretical *m/z* values, following enzymatic assays with the purified enzymes in the presence of UDP-Glc

and an array of aromatic substrates (additional assays are shown in Extended Data Fig. 6b). One to three glucosylated compounds were observed for each substrate. The peaks were putatively assigned by MS/MS fragmentation patterns (Supplementary Fig. 22 and Supplementary Table 20). The chromatograms were normalized to the highest value. **e**, Extracted ion chromatograms of the *O*-acylated cannabinoids following enzymatic assays with purified HuCBAT5 in the presence of different acyl donors and aromatic substrates as acceptors. Major ion products were selected in each liquid chromatography−MS/MS (LC−MS/MS) chromatogram. A single peak was observed for each pair of substrates. The detected analogue peaks shifted in retention time depending on their change in hydrophobicity relative to the acyl group. Identification was performed according to MS/MS fragmentation (Extended Data Fig. 8 and Supplementary Table 21) and retention time. The chromatograms were normalized to the highest value. **f**, Potential glycosylation and observed *O*-acylation sites are highlighted in blue and/or purple on each chemical structure, respectively.

homologues, 4 of them co-expressed with other cannabinoid-related enzymes (Figs. 2e and 4b and Supplementary Tables 12 and 19). Of the 15 AATs, 12 were expressed in *E. coli* and examined for their activity with butyryl- and hexanoyl-CoA as acyl donors and **1** and **2** as acceptors. Only HuAAT5 and HuAAT14 showed activity towards these substrates

(Extended Data Fig. 7b). Phylogenetic analysis showed that these two enzymes clustered in clade IIIa, representing BAHDs of diverse catalytic functions (Extended Data Fig. 7c)[26]. HuAAT5 (HuCBAT5) produced larger amounts of products and was therefore selected for in-detail characterization with an array of acyl donors and acceptors. It accepted
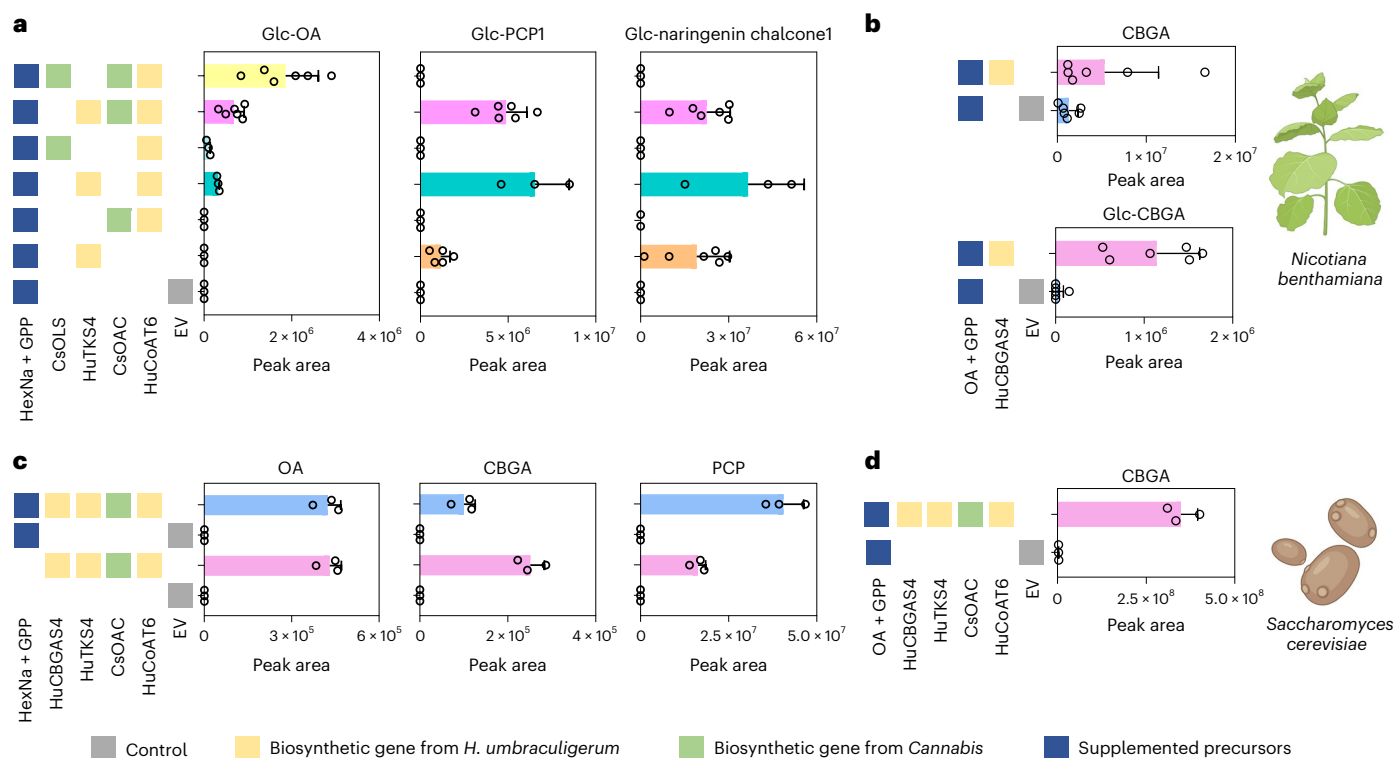

c

	HuOAGT11	HuUGT13	OsUGT	SrUGT
OA				
K_m	151.1 ± 57.9	20.2 ± 4.5	183.7 ± 39.1	>1,500
V_{max}	916.5 ± 37.2	3.1 ± 0.16	4.8 ± 0.21	–
k_{cat}	45,906.5	156.7	253.2	–
k_{cat}/K_m	303.8	7.8	1.4	–
UDP-Glc				
K_m	107.1 ± 31.6	33.1 ± 5.6	62.4 ± 17.6	23.7 ± 3.1
V_{max}	314.0 ± 20.6	1.5 ± 0.05	3.5 ± 0.25	8.8 ± 0.22
k_{cat}	15,727.9	75.8	184.6	462.7
k_{cat}/K_m	146.8	2.3	2.9	19.5

V_0 and V_{max} were calculated using the calibration curve of OA.
K_m, μM; V_{max}, μmol min^{-1} mg^{-1}; k_{cat}, min^{-1}; k_{cat}/K_m, μM^{-1} min^{-1}

Fig. 5 | In vivo reconstruction of the core cannabinoid pathway in heterologous systems. a–d, Co-expression of different combinations of HuCoAT6, HuTKS4 and HuCBGAS4 along with CsOAC and CsOLS from *Cannabis* in *N. benthamiana* leaves (**a,b**) and *S. cerevisiae* yeasts (**c,d**). Peak areas of ion abundance were used for the comparisons (mean ± s.d.; n = 3–6 biological replicates). *N. benthamiana* produced mainly glycosylated products identified according to the previously conducted in vitro UGT enzyme assays (Fig. 4d and Extended Data Fig. 6b). All the metabolites were identified by exact mass, retention time and MS/MS spectra (Extended Data Fig. 9).

all acyl donors tested and acylated **92**, **1**, **2** and CBDA, giving rise to a single *O*-acyl-cannabinoid from each pair of substrates (Fig. 4e,f, Extended Data Fig. 8 and Supplementary Table 21). Many of the cannabinoids produced were naturally observed in the plant (marked with asterisks in Fig. 4e). However, the enzyme was inactive on Δ^9-THCA and **15**. It is therefore likely that it only acylates the hydroxyl in C_5 (according to the atom numbering in Supplementary NMR Data 13 and 14). In addition, *O*-acyl esterification in *H. umbraculigerum* was observed only on prenylated cannabinoids and amorfrutins and not on their intermediates.

In vivo reconstruction of the core cannabinoid pathway. We verified the in planta activity of the enzymes towards **1** by transiently co-expressing different combinations of HuCoAT6, HuTKS4 and HuCB-GAS4 and the *Cannabis* CsOAC and CsOLS in *Nicotiana benthamiana* leaves. Following leaf infiltration with sodium hexanoate and GPP, we observed the production of glycosylated forms of **92** (HuTKS4 + CsOAC or CsOLS + CsOAC) and **95** (only with HuTKS4; Fig. 5a and Extended Data Fig. 9a,b). This was consistent with previous studies reporting **92** glycosylation by endogenous enzymes in this plant[25]. Interestingly, we also observed glycosylated products of naringenin chalcone (**97**) with HuTKS4, suggesting that this enzyme can accept aromatic substrates in addition to aliphatic types (Fig. 5a and Extended Data Fig. 9a,b). However, we did not observe **1** or its glycosylated forms with HuCBGAS4, probably due to the low availability of **92** and its rapid glycosylation in planta. When leaves expressing HuCBGAS4 were infiltrated with **92** and GPP, **1** and **109** were observed (Fig. 5b and Extended Data Fig. 9a,c).

We also reconstituted the cannabinoid pathway by expressing the genes encoding HuCoAT6, HuTKS4, CsOAC and HuCBGAS4 in *Saccharomyces cerevisiae*. We observed the production of **92**, **1** and **95** without precursor feeding (Fig. 5c and Extended Data Fig. 9d,e). Similar to the in vitro assays, we also observed peaks of HTAL and PDAL,

which were not present in planta (Extended Data Fig. 9f). When cells were supplemented with **92** and GPP, substantially larger amounts of **1** were produced (Fig. 5d).

Discussion

Mounting accessibility, expanding legalization and multiple applications of cannabinoids drive worldwide attention to these molecules. Because it is the sole source of cannabinoids to date, ongoing research and breeding are focused on *Cannabis*, where there has been a huge genetic bottleneck due to domestication and selection towards the psychotropic effect of (−)-Δ^9-*trans*-tetrahydrocannabinol (Δ^9-THC)[27]. Our research highlights a second, yet-untapped and genetically distant species as a source for cannabinoid genetics and chemistry (Fig. 6). Cannabigerol, the better-known neutral form of **1**, is experiencing a huge rise in interest, as it shows therapeutic potential for diverse conditions including neurologic disorders, colon carcinogenesis and many others[2,28–30]. Here we found that *H. umbraculigerum* accumulates **1** in glandular trichomes of its aerial parts to the maximum levels observed in inflorescences of over 320 *Cannabis* chemotypes[14]. This characteristic, in addition to its easy propagation and fast growth rate, makes *H. umbraculigerum* a potential all-year-round supply of Δ^9-THCA-free **1** without the need for flower induction or the use of extraordinary growing facilities.

In addition to high amounts of **1**, the versatility of *H. umbraculigerum* chemistry expands the repertoire of modulators of the endocannabinoid system and of other therapeutic targets. Glycosylation and acylation are major decorations of cannabinoids that have not been previously identified in *Cannabis* (Fig. 6). In the case of glucosylation, the increase in aqueous solubility has been proposed as a strategy to enhance the bioavailability and absorption of cannabinoids into the bloodstream[25,31]. The *O*-acyl cannabinoids and
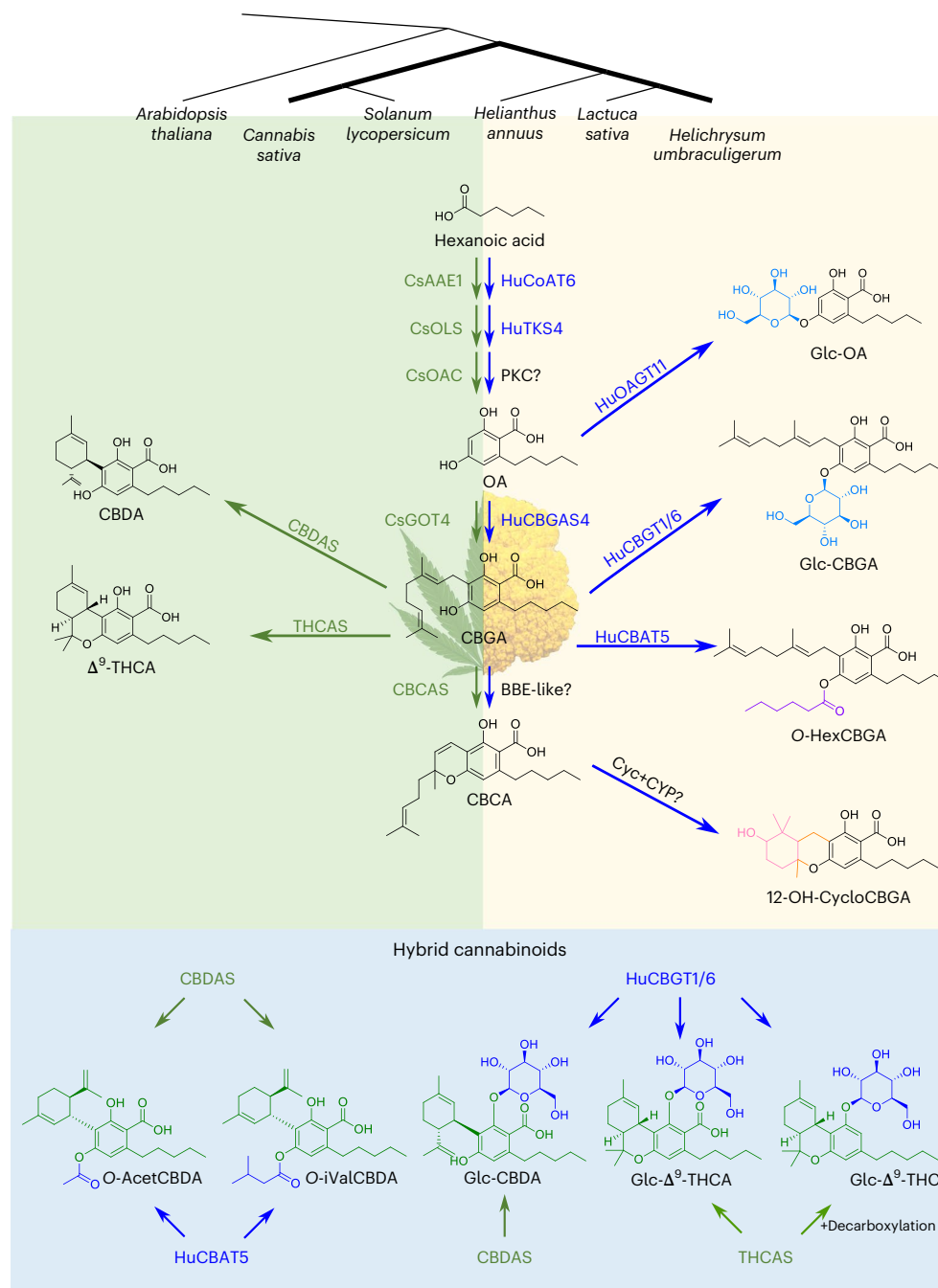
Fig. 6 | Parallel and divergent evolution of the cannabinoid biosynthetic pathway. Side-by-side comparison of the cannabinoid biosynthetic routes in *H. umbraculigerum* and *Cannabis*. On top, the phylogenetic relationship among *Arabidopsis thaliana*, *Solanum lycopersicum*, *Helianthus annuus*, *Lactuca sativa*, *Cannabis sativa* and *Helicrysum umbraculigerum* illustrates the evolutionary distances between *Cannabis* and *H. umbraculigerum*. The tree was constructed on the basis of the whole proteomes of each species using the word-based software Prot-SpaM. Hybrid, yet-unreported metabolites were produced in this study by reacting cannabinoids naturally biosynthesized in *Cannabis* (green)

with uridine diphosphate glucose (UDP-Glc) or acyl-CoAs in the presence of HuCBAT5, HuCBGT1 or HuCBGT6 enzymes from *H. umbraculigerum* (blue). CsAAE, *Cannabis* AAE; CsGOT4, *Cannabis* geranylpyrophosphate:olivetolate geranyltransferase; CBDAS, CBDA synthase; CBCAS, CBCA synthase; THCAS, Δ^9-THCA synthase. The active enzymes identified in this study are marked by their names. TKS, tetraketide synthase; OAGT, OA UGT; CBGT, cannabinoid UGT; CBAT, cannabinoid acyl-transferase; BBE-like, berberine bridge enzyme-like; Cyc, cyclase; CYP, cytochrome P450; 12-OH-cycloCBGA, 12-OH-cyclocannabigerolic acid.

their neutral analogues may target the ionotropic thermo-TRP ion channels as reported for *O*-acyl-prenyl amorfrutins[13].

In terms of physiological relevance, modifications of the backbone may have a role in the biosynthesis, storage, stability and transport of cannabinoids. Both glycosylation and acylation were found to be active in planta only on cannabinoids and amorfrutins and not on the other terpenophenols. Interestingly, in planta glucosylation, which increases

the hydrophilicity, was found to occur mainly on the cannabinoid precursors (prior to prenylation), whereas acylation, which increases the lipophilicity, was found to occur only on the prenylated cannabinoids. Hitherto, in vitro assays have shown that the active enzymes work both on the precursors and on the prenylated cannabinoids, suggesting an orchestrated compartmentalization of both pathway substrates and enzymes. Considering the complex subcellular

organization observed in DCs (Extended Data Fig. 2d–f), acylation may occur within plastids or mitochondria, inside vesicles, or following secretion from DCs as in the case of **1** cyclization to Δ^9-THCA in the apoplast of *Cannabis* trichomes[10].

The production of six identical cannabinoids in two distant species such as *Cannabis* and *H. umbraculigerum* provides a case of parallel evolution where the pathway architecture is the same and involves paralogous enzymes that evolved independently in the two plant lineages (Fig. 6). In contrast, downstream products of the pathway show notable divergence between the two species (Fig. 6). Other *Helichrysum* species have been reported to biosynthesize prenylated acylphloroglucinoid, chalcone, and flavanone derivatives[32–34] but no cannabinoids or amorfrutins, meaning that they probably lack the capability to perform aldol cyclization. It is possible that, within the *Helichrysum* genus or the wider Gnaphalieae tribe, some species acquired this ability through mutations of a CHS-type enzyme and a yet-to-be-identified helper protein (Supplementary Table 17). In addition, the lack of HuCBGAS4 activity with phloroglucinoid and chalcone derivatives (both in vitro and in vivo) further supports the hypothesis that *H. umbraculigerum* acquired the capacity to prenylate cannabinoids later in evolution.

Our inability to identify an equivalent OAC-type enzyme indicates that, as suggested for *Rhododendron dauricum*[20], *H. umbraculigerum* developed a mechanism different from *Cannabis* to produce **92**. However, the fact that in both in vitro and in vivo assays the addition of CsOAC boosts the production of **92** supports the hypothesis of a helper protein. Potential proteins of this type may be chalcone-isomerase-like or BetV1-like enzymes as previously investigated for the cyclization of **92** in *Cannabis*[7]. It is also feasible that a functional CsOAC homologue exists, but it has not been discovered yet. The identification of the missing PKC-type enzyme would probably increase the rate of the aldol reaction in the coupled assays, shedding light on the evolutionary aspects of the cannabinoid pathway in this species.

The independent evolution of cannabinoids in the Asteraceae and Cannabaceae families suggests that these molecules have important ecological functions selected across different plant lineages. Studies in *Cannabis* propose that they might play a role in plant–animal interactions and/or provide protection against UV light and desiccation[10], but further research is needed to fully understand the evolutionary advantage provided by cannabinoids in these and other plants. Yet, the elucidation of the entire cannabinoid repertoire in *H. umbraculigerum* and the biosynthetic enzymes substantially enhances our understanding of the pathway and our future capacity to overcome the synthetic biology challenge of engineering complex cannabinoids with diverse pharmacological activities.

Methods

Materials

The list of chemicals is provided in the Supplementary Methods. Seeds of *H. umbraculigerum* (Silverhill Seeds) were germinated and were grown in a greenhouse in a long-day photoperiod. The plants were propagated by cuttings.

Feeding experiments

Feeding solutions were prepared as aqueous solutions of 0.5 mg ml^{-1} of the precursor. The pH of the FA solutions was adjusted to 5.5–6.0. The phenylalanine feeding experiments were performed on leaves from young mother plants excised by cutting at the proximal side of the pedicel with scissors under water, leaving 1–2 cm of the pedicel attached. For the FA feeding experiments, 10 cm young cuttings were obtained from mother plants. The lower leaves were removed, leaving four or five leaves on each stem, and the stem was peeled to increase the intake of the labelled solutions. Three or four leaves of the young cuttings were immersed in aqueous solutions (DDW (control) or unlabelled or labelled precursors; each group consisted of a minimum

of three biological replicates). All feeding experiments were performed in a controlled environment for 48–96 h at 25 °C and under constant fluorescent illumination and humidity. The tubes were periodically refilled. Upon termination, the fresh leaves were rinsed with a small amount of water, dried gently, flash frozen and stored at −80 °C for extraction.

LC–MS chemical analysis

Unless otherwise stated, 100 mg of frozen powdered plant tissue was extracted with 300 µl of ethanol, sonicated for 15 min, agitated for 30 min and centrifuged at 14,000 *g* for 10 min. The supernatant was filtered through a 0.22 µm syringe filter and analysed in the obtained concentration. Detection was performed using both targeted and non-targeted approaches as described in the Supplementary Methods using an ultra-high-performance liquid chromatography–tandem quadrupole time-of-flight (UPLC–qTOF) system comprising a UPLC (Waters Acquity) with a diode array detector connected either to a XEVO G2-S QTof (Waters) or to a Synapt HDMS (Waters). Data acquisition and analysis were performed using MassLynx v.4.2 (Waters). The chromatographic separation was performed on a 100 mm × 2.1 mm i.d. (internal diameter), 1.7 µm UPLC BEH C18 column (Waters Acquity). The mobile phase consisted of 0.1% formic acid in acetonitrile:water (5:95, v/v; phase A) and 0.1% formic acid in acetonitrile (phase B). Terpenophenols were analysed using UPLC Method 1, and intermediates and glucosylated metabolites were analysed using UPLC Method 2 (Supplementary Methods). Electrospray ionization was used in either positive or negative ionization modes at an *m/z* range of 50–1,000 Da. Masses were detected with the following settings: capillary, 1 kV; source temperature, 140 °C; desolvation temperature, 450 °C; desolvation gas flow, 800 l h^{-1}. Argon was used as the collision gas. Data acquisition for untargeted analysis was in negative ionization using the MS^E mode. The collision energy was 4 eV for the low-energy function and 15–50 eV ramp for the high-energy function. For data analysis of the isotope labelling data, we used the R package Miso (version 0.2.1)[35]. Briefly, differential metabolites were selected if the fold change was greater than or equal to 10 and the *P* value was less than 0.05. MS/MS experiments were performed in positive or negative ionization modes according to the specific protonated or deprotonated masses with the following settings: capillary spray, 1 kV; cone voltage, 30 eV; collision energy ramps, 10–45 eV for positive mode and 15–50 eV for negative mode.

Absolute quantification of CBGA

Fresh samples of leaves (young and old), flowers, stems and roots were collected from a plant at the flowering stage (*n* = 3 biological replicates). Florets and flower receptacles were detached using a scalpel and analysed separately. Tissues were flash-frozen in liquid N$_2$ and ground into fine powder. To measure **1** content in dry tissue, fresh leaves were flash-frozen, ground and lyophilized (*n* = 5 biological replicates). For the extraction, 100 mg of the frozen powders were accurately weighed, extracted with 1 ml ethanol and prepared as previously described. Samples were injected in several dilutions to fit into the linear range of the calibration curves. The injections were performed on a UPLC (Waters) connected to a Triple Quad detector (TQ-S, Waters) in multiple reaction monitoring mode. The system was operated with a similar column and mobile phase as for UPLC–qTOF analysis according to UPLC Method 3 (Supplementary Methods). The instrument was operated in negative mode with a capillary voltage of 1.5 kV and a cone voltage of 40 V. Absolute quantification of **1** was performed by external calibration using two different transitions (359.3 to 191.2, 32 V for quantification; and 359.3 to 315.4, 21 V for qualification).

Metabolite purification for NMR analysis

A total of 86 g of fresh leaves were flash-frozen in liquid N$_2$, ground into fine powder using an electrical grinder, extracted with 600 ml of ethanol, sonicated for 20 min and agitated for 30 min. The supernatant

was filtered, evaporated using a rotary evaporator at 40 °C and lyophilized. The extract was reconstituted in 25 ml of acetonitrile and used for either direct purification (following ten times dilution) or pre-fractionation via medium-pressure LC (Büchi Sepacore). The metabolites were purified using either an Agilent 1290 Infinity II UPLC system or a Waters Acquity UPLC system with a similar mobile phase as for the UPLC–qTOF. Triggering was performed using specific UV wavelengths according to the metabolite. General instrument setups and operation are described in the Supplementary Methods. Eluents containing the same metabolite were pooled, dried under a stream of N_2 and stored at −20 °C until NMR analysis. NMR analyses were performed as described in the Supplementary Methods and Supplementary NMR Data.

MALDI imaging

For the peeling experiment, two whole fresh leaves from a young plant were attached to glass slides using double-sided tape on either the abaxial or adaxial surface, gently peeled above or below the midrib using duct tape, and desiccated overnight under moderate vacuum. Images were taken using a digital camera. For the localization of metabolites to individual trichomes, fresh leaves and flowers ($n = 2$ biological replicates each) were sectioned, and matrix was sprayed using a TM sprayer (HTX Technologies) with 2,5-dihydroxybenzoic acid (DHB; 40 mg ml^{-1} in 70% containing 0.2% trifluoroacetic acid). The nozzle temperature was set to 70 °C, and the DHB matrix solution was sprayed in 16 passes over the tissue sections at a linear velocity of 120 cm min^{-1} with a flow rate of 50 µl min^{-1}. The sections were imaged with a Nikon DS-Ri2 microscope. MALDI imaging was performed using a 7T Solarix Fourier transform ion cyclotron resonance mass spectrometer (Bruker Daltonics). Data acquisition and analysis were performed using flexImaging v.4.1 (Bruker). The datasets were collected in positive ionization using lock mass calibration (DHB matrix peak: [3DHB + H − 3H$_2$O]$^+$, m/z 409.055408 Da) at a frequency of 1 kHz and a laser power of 40%, with 200 laser shots per pixel and a pixel size of 50, 15 or 25 µm for the peeled trichomes and for the sectioned leaves and flowers, respectively. Each mass spectrum was recorded in the range of m/z 150–3,000 in broadband mode with a time domain for acquisition of 1 M, providing an estimated resolving power of 115,000 at m/z 400. The spectra were normalized to root-mean-square intensity, and MALDI images were plotted at theoretical $m/z \pm 0.005\%$ with pixel interpolation on.

Cryo-SEM, TEM and confocal microscopy

For the cryo-SEM analyses, frozen samples were attached to a holder either by mechanical clamping (leaves) or by a glue made of a concentrated PVP solution. The holder with the samples was then plunge-frozen in liquid N_2, transferred to a BAF 60 freeze fracture device (Leica Microsystems) using a VCT 100 Vacuum Cryo Transfer device (Leica) and sublimed for 30 min at −95 °C. The samples were transferred to an Ultra 55 cryo-SEM (Zeiss) using a VCT 100 shuttle and were observed at −95 °C without coating using mostly mixed mode of InLens + SE detectors at 1–1.3 kV. Multiple leaf and flower areas were sampled ($n > 3$). For TEM analysis, young and old leaves of *H. umbraculigerum* were fixed with 4% paraformaldehyde and 2% glutaraldehyde in 0.1 M cacodylate buffer containing 5 mM CaCl$_2$ (pH 7.4), then postfixed with 1% osmium tetroxide supplemented with 0.5% potassium hexacyanoferrate trihydrate and potassium dichromate in 0.1 M cacodylate (1 h), stained with 2% uranyl acetate in water (1 h), dehydrated in graded ethanol solutions and embedded in Agar 100 epoxy resin (Agar Scientific). Ultrathin sections (70–90 nm) were viewed and photographed with a FEI Tecnai SPIRIT (FEI) transmission electron microscope operated at 120 kV and equipped with an OneView Gatan Camera. Multiple leaf sections were sampled showing trichomes at different developmental stages ($n = 2$–5 for each stage). Confocal microscopy of trichomes was carried out on a Nikon Eclipse A1 microscope. Transmitted light was used to image the trichomes since

they lack fluorescence. Autofluorescence of chlorophyll (chloroplasts) was used as a contrast for better visualization of the trichomes. Far-red laser was used to detect autofluorescence of chlorophyll (excitation, 640 nm; emission, 663–738 nm). Multiple leaf areas were sampled ($n > 3$).

Genome assembly

High-molecular-weight DNA was extracted from young frozen leaves and sequenced at the UC Davis Genome Center. Sequencing was done in a Pacbio Sequel II platform with ~12-kilobase DNA SMRT bell library preparation according to the manufacturer's protocol. Three different SMRT 8 M cells were used, yielding a total of 57.8 Gb of HiFi data (~44× haploid coverage). In addition to Pacbio HiFi data, 200 M reads of PE 2 × 150 Illumina Hi-C data were obtained by the company Phase Genomics. Hifiasm[36] software was used to integrate both Pacbio HiFi and HiC data to produce chromosome-scale and haplotype-resolved assemblies. Further scaffolding was performed with the Hi-C data, mapping the reads following the Arima Genomics pipeline and the SALSA software[37,38]. Visualizations of Hi-C heat maps were performed with Juicer, and quality metrics were obtained with Assemblathon 2 script[39]. Finally, the assembly was softmasked for repetitive elements using EDTA[40] with the cds flag to incorporate coding sequences (CDSs) from the transcriptomic data.

RNA sequencing and genome annotation

RNA was extracted from seven tissues ($n = 3$ biological replicates): young leaves, old leaves, florets and receptacles of flowers, stems, roots and trichomes (Supplementary Table 8). The protocol for trichome enrichment is provided in the Supplementary Methods. RNA integrity was checked using a TapeStation instrument. Paired-end Illumina libraries were prepared for five of the tissues, sequenced on an Illumina HiSeq 3000 instrument (PE 2 × 150, ~40 M reads per sample) and processed following the guidelines of Freedman and Weeks[41]. The details about the genome annotation process and other types of RNA-seq processing and data visualization can be found in the Supplementary Methods.

Orthology and synteny analyses

Proteomes were obtained from all available annotated Asteraceae genomes provided by the National Center for Biotechnology Information (NCBI) (Supplementary Table 22). Orthogroups and their phylogenetic relationships were inferred with Orthofinder[42]. Genomic positions and putative functions of all the genes belonging to the orthogroups of HuCoAT6 (OG0014461), HuTKS4 (OG0000313) and HuCBGAS4 (OG0002538) were determined using the corresponding GFF files, and the plots were produced with the gggenomes package (version 0.9.5.9000)[43]. Phylogenetic gene trees generated by Orthofinder were plotted with MEGA11.

Recombinant expression in *E. coli* and protein purification

Hu*AAE1-6*, Hu*UGT1-13* and Hu*AAT1-15* CDSs from *H. umbraculigerum* and previously characterized sequences from rice (Os*UGT*) and stevia (Sr*UGT*) were individually cloned into the pET28b vector digested with EcoRI using the ClonExpress II One Step Cloning kit (Vazyme). Hu*PKS1-4*, Hu*PKC1-5*, Cs*OLS* and Cs*OAC* were ligated into the pOPINF vector (digested with HindIII and KpnI) using the ClonExpress II One Step Cloning kit (Vazyme). Due to the high sequence similarity of the CDSs, Hu*PKS2-4* were synthesized by the company Twist Biosciences. All constructs were expressed in *E. coli* BL21 (DE3) cells (a complete list of the primers is provided in Supplementary Table 23). Bacterial starters were grown overnight in LB medium at 37 °C, diluted in fresh LB 1:100 and reincubated at 37 °C. When the cultures reached A600 = 0.6, protein expression was induced with 400 µM isopropyl-1-thio-β-D-galactopyranoside overnight at 15 °C. Bacterial cells were lysed by sonication in 50 mM Tris–HCl (pH 8.0), 0.5 mM

phenylmethylsulfonyl fluoride (Sigma Aldrich) solution in isopropanol, 10% glycerol and protease inhibitor cocktail (Sigma Aldrich), and 1 mg ml⁻¹ lysozyme (Sigma Aldrich). The whole-cell extract was either kept for functional activity or used for protein purification. Purification of hexahistidine-tagged proteins was performed on Ni-NTA agarose beads (Adar Biotech). The proteins were eluted with 200 mM imidazole (Fluka) in buffer containing 50 mM NaH_2PO_4 (pH 8.0) and 0.5 M NaCl. The protein concentration of the eluted fractions was measured with Pierce 660 nm protein assay reagent (Thermo Scientific). In vitro and kinetic enzyme assays are described in the Supplementary Methods ($n = 3$ technically independent experiments).

Transient expression of selected genes in *N. benthamiana*

Overexpression constructs of GFP (as a negative control), Cs*OLS* and Cs*OAC* were generated using GoldenBraid cloning to a final vector of pAlpha2–Ubq10–CCD–Ter10. Hu*CoAT6*, Hu*TKS4* and Hu*CBGAS* were amplified and cloned in pAlpha2–NPT II–Ubq10–CCD–Ter10 vector digested with BsaI using the ClonExpress II One Step Cloning kit (Vazyme). A complete list of the primers is provided in Supplementary Table 23. All plasmids were sequenced and transformed into *Agrobacterium tumefaciens* strain GV3101 by electroporation. *A. tumefaciens* harbouring the overexpression constructs were grown overnight at 28 °C in LB medium in the presence of kanamycin and gentamycin. Bacterial cells were collected by centrifugation, washed and resuspended in infiltration buffer (10 mM MES, 2 mM $MgCl_2$, 2 mM Na3PO4, 0.5% glucose and 100 mM acetosyringone) to $OD_{600} = 0.3$. Equal volumes of *A. tumefaciens* suspension with different expression vectors were combined to obtain the desired gene combinations and incubated for 2 h at room temperature. The solutions were infiltrated into four- or five-week-old *N. benthamiana* leaves from the abaxial side using a 1 ml needleless syringe ($n = 3–6$ biological replicates). Substrates (0.5 mM each) were infiltrated into the same leaf areas two days after the initial infiltration, and the leaves were collected for metabolite analysis after 24 h. The leaf samples were flash-frozen and extracted as previously described with 300 µl of methanol, and analysed on a similar UPLC system connected to an Orbitrap IQ-X Tribrid MS (Thermo Scientific) using UPLC Method 2 in negative mode. Data acquisition and analysis were performed using Xcalibur v.4.5 (Thermo). The operational parameters are provided in the Supplementary Methods. The metabolites were identified using analytical standards and/or products from in vitro UGT enzyme assays.

Heterologous expression in *S. cerevisiae*

For the expression of Hu*CoAT6*, Hu*TKS4*, Cs*OAC* and Hu*CBGAS* in *S. cerevisiae*, the CDSs were amplified, and the purified amplicons were inserted into series of pESC (Amp^R) plasmids allowing simultaneous expression of two genes from one plasmid. Hu*CoAT6* and Hu*TKS4* were inserted using the ClonExpress II One Step Cloning kit (Vazyme) into pESC-HIS plasmid linearized with SalI and SacI restriction enzymes, respectively. Hu*CBGAS* and Cs*OAC* were cloned in the same way into pESC-TRP plasmid linearized with SalI and SacI restriction enzymes, respectively. The full list of primers is provided in Supplementary Table 23. pESC constructs were transformed into *S. cerevisiae* WAT11 using the Yeastmaker yeast transformation system (Clontech). We transformed yeast cells with combinations of pESC vectors allowing the expression of all four genes at once. Transformed yeast were grown on SD minimal media supplemented with the appropriate amino acids and 2% glucose. The colonies were screened, and the presence of the transgene was confirmed by colony PCR. For the induction of gene expression, transformed cells were grown in 2 ml of minimal medium with 2% glucose and after 24 h were transferred to a minimal medium with 2% galactose without additional supplementation or supplemented with GPP (0.21 mM) and either sodium hexanoate (1 mM) or **92** (0.2 mM); they were then grown for an additional 24 h at 30 °C. The

cultures were transferred to 2 ml Eppendorf tubes and centrifuged at 8,000 *g* for 1 min. The cell pellet was weighed, double the amount of glass beads (diameter 500 µm) and 500 µl of methanol were added, and the cells were lysed using a beadbeater at 22 Hz for 6 min. Lysed cells were centrifuged at 21,000 *g* for 5 min and clear supernatant was collected and dried using a SpeedVac. Dry residues were dissolved in 100 µl of methanol, filtered through a 0.22 µm filter and analysed using LC–MS as detailed for the *N. benthamiana* samples.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All NGS raw sequencing data as well as the primary genome assembly version Humb_v1 can be found in ENA under the accession number PRJEB52026. The sequences of the active genes reported in this Article have been deposited in NCBI GenBank (accessions OQ673107–OQ673113). The genomic position, functional annotation, CDSs and protein sequences of these genes are available in Supplementary Tables 14 and 19. The protein databases used in this study are Uniprot Swiss-Prot release-2022_02, sunflower UP000215914_4232, *Arabidopsis* UP000006548_3702, tomato UP000004994_4081, rice UP000059680_39947 and *Cannabis* NCBI GCF_900626175.1_cs10. The Pfam database was Pfam-A.hmm release 34.0. Source data are provided with this paper.

Code availability

All the code used in this study along with a description of the scripts can be found at https://github.com/Luisitox/Helichrysum_paper.

References

1. di Marzo, V. & Piscitelli, F. The endocannabinoid system and its modulation by phytocannabinoids. *Neurotherapeutics* **12**, 692–698 (2015).
2. Ligresti, A., de Petrocellis, L. & di Marzo, V. From phytocannabinoids to cannabinoid receptors and endocannabinoids: pleiotropic physiological and pathological roles through complex pharmacology. *Physiol. Rev.* **96**, 1593–1659 (2016).
3. Turner, S. E., Williams, C. M., Iversen, L. & Whalley, B. J. Molecular pharmacology of phytocannabinoids. *Prog. Chem. Org. Nat. Prod.* **103**, 61–101 (2017).
4. Morales, P., Hurst, D. P. & Reggio, P. H. Molecular targets of the phytocannabinoids: a complex picture. *Prog. Chem. Org. Nat. Prod.* **103**, 103–131 (2017).
5. Stout, J. M., Boubakir, Z., Ambrose, S. J., Purves, R. W. & Page, J. E. The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in *Cannabis sativa* trichomes. *Plant J.* **71**, 353–365 (2012).
6. Taura, F. et al. Characterization of olivetol synthase, a polyketide synthase putatively involved in cannabinoid biosynthetic pathway. *FEBS Lett.* **583**, 2061–2066 (2009).
7. Gagne, S. J. et al. Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique catalytic route to plant polyketides. *Proc. Natl Acad. Sci. USA* **109**, 12811–12816 (2012).
8. Luo, X. et al. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* **567**, 123–126 (2019).
9. Hanuš, L. O., Meyer, S. M., Muñoz, E., Taglialatela-Scafati, O. & Appendino, G. Phytocannabinoids: a unified critical inventory. *Nat. Prod. Rep.* **33**, 1357–1392 (2016).
10. Gülck, T. & Møller, B. L. Phytocannabinoids: origins and biosynthesis. *Trends Plant Sci.* **25**, 985–1004 (2020).
11. Appendino, G., Taglialatela-Scafati, O. & Muñoz, E. Cannabidiol (CBD) from non-cannabis plants: myth or reality? *Nat. Prod. Commun.* https://doi.org/10.1177/1934578X221098843 (2022).

12. Bohlmann, F. & Hoffmann, E. Cannabigerol-ähnliche verbindungen aus Helichrysum umbraculigerum. *Phytochemistry* **18**, 1371–1374 (1979).

13. Pollastro, F. et al. Amorfrutin-type phytocannabinoids from *Helichrysum umbraculigerum*. *Fitoterapia* **123**, 13–17 (2017).

14. Procaccia, S. et al. Cannabis for medical use: versatile plant rather than a single drug. *Front. Pharmacol.* https://doi.org/10.3389/fphar.2022.894960 (2022).

15. Berman, P. et al. A new ESI-LC/MS approach for comprehensive metabolic profiling of phytocannabinoids in *Cannabis*. *Sci. Rep.* **8**, 14280 (2018).

16. de Meijer, E. P. M. & Hammond, K. M. The inheritance of chemical phenotype in *Cannabis sativa* L. (V): regulation of the propyl-/pentyl cannabinoid ratio, completion of a genetic model. *Euphytica* **210**, 291–307 (2016).

17. Kearsey, L. J. et al. Structure of the *Cannabis sativa* olivetol-producing enzyme reveals cyclization plasticity in type III polyketide synthases. *FEBS J.* **287**, 1511–1524 (2020).

18. Yamaguchi, T. et al. Cross-reaction of chalcone synthase and stilbene synthase overexpressed in *Escherichia coli*. *FEBS Lett.* **460**, 457–461 (1999).

19. Parvez, A., Giri, S., Bisht, R. & Saxena, P. New insights on cyclization specificity of fungal type III polyketide synthase, PKSIIINc in *Neurospora crassa*. *Indian J. Microbiol.* **58**, 268–277 (2018).

20. Taura, F. et al. A novel class of plant type III polyketide synthase involved in orsellinic acid biosynthesis from *Rhododendron dauricum*. *Front. Plant Sci.* **7**, 1452 (2016).

21. Okada, Y. & Ito, K. Cloning and analysis of valerophenone synthase gene expressed specifically in lupulin gland of hop (*Humulus lupulus* L.). *Biosci. Biotechnol. Biochem.* **65**, 150–155 (2001).

22. de Bruijn, W. J. C., Levisson, M., Beekwilder, J., van Berkel, W. J. H. & Vincken, J. P. Plant aromatic prenyltransferases: tools for microbial cell factories. *Trends Biotechnol.* **38**, 917–934 (2020).

23. Lim, E. K. et al. The activity of *Arabidopsis* glycosyltransferases toward salicylic acid, 4-hydroxybenzoic acid, and other benzoates. *J. Biol. Chem.* **277**, 586–592 (2002).

24. Geissler, M., Volk, J., Stehle, F., Kayser, O. & Warzecha, H. Subcellular localization defines modification and production of Δ9-tetrahydrocannabinolic acid synthase in transiently transformed *Nicotiana benthamiana*. *Biotechnol. Lett.* **40**, 981–987 (2018).

25. Gülck, T. et al. Synthetic biology of cannabinoids and cannabinoid glucosides in *Nicotiana benthamiana* and *Saccharomyces cerevisiae*. *J. Nat. Prod.* **83**, 2877–2893 (2020).

26. Tuominen, L. K., Johnson, V. E. & Tsai, C. J. Differential phylogenetic expansions in BAHD acyltransferases across five angiosperm taxa and evidence of divergent expression among *Populus* paralogues. *BMC Genomics* **12**, 236 (2011).

27. Ren, G. et al. Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Sci. Adv.* **7**, eabg2286 (2021).

28. Nachnani, R., Raup-Konsavage, W. M. & Vrana, K. E. The pharmacological case for cannabigerol. *J. Pharmacol. Exp. Ther.* **376**, 204–212 (2021).

29. Anokwuru, C. P. et al. Cannabigerol: a bibliometric overview and review of research on an important phytocannabinoid. *Phytochem. Rev.* https://doi.org/10.1007/s11101-021-09794-w (2022).

30. Russo, E. B. et al. Survey of patients employing cannabigerol-predominant cannabis preparations: perceived medical effects, adverse events, and withdrawal symptoms. *Cannabis Cannabinoid Res.* https://doi.org/10.1089/CAN.2021.0058 (2021).

31. Hardman, J. M., Brooke, R. T. & Zipp, B. J. Cannabinoid glycosides: in vitro production of a new class of cannabinoids with improved physicochemical properties. Preprint at *bioRxiv* https://doi.org/10.1101/104349 (2017).

32. Lourens, A. C. U., Viljoen, A. M. & van Heerden, F. R. South African *Helichrysum* species: a review of the traditional uses, biological activity and phytochemistry. *J. Ethnopharmacol.* **119**, 630–652 (2008).

33. Jakupovic, J. et al. Twenty-one acylphloroglucinol derivatives and further constituents from South African *Helichrysum* species. *Phytochemistry* **28**, 1119–1131 (1989).

34. Jakupovic, J., Kuhnke, J., Schuster, A., Metwally, M. A. & Bohlmann, F. Phloroglucinol derivatives and other constituents from South African *Helichrysum* species. *Phytochemistry* **25**, 1133–1142 (1986).

35. Dong, Y., Feldberg, L. & Aharoni, A. Miso: an R package for multiple isotope labeling assisted metabolomics data analysis. *Bioinformatics* **35**, 3524–3526 (2019).

36. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

37. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).

38. *Mapping Pipeline for Data Generated Using HiC* (Arima Genomics, 2019); https://github.com/ArimaGenomics/mapping_pipeline

39. Bradnam, K. R. et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).

40. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

41. Freedman, A. & Weeks, N. Best practices for de novo transcriptome assembly with Trinity. *Harvard FAS Informatics* https://informatics.fas.harvard.edu/best-practices-for-de-novo-transcriptome-assembly-with-trinity.html (2020).

42. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

43. Hackl, T. & Ankenbrand, M. gggenomes: a grammar of graphics for comparative genomics. *GitHub* https://github.com/thackl/gggenomes (2022).

44. Shockey, J. M., Fulda, M. S. & Browse, J. *Arabidopsis* contains a large superfamily of acyl-activating enzymes: phylogenetic and biochemical analysis reveals a new class of acyl-coenzyme A synthetases. *Plant Physiol.* **132**, 1065–1076 (2003).

45. Naake, T., Maeda, H. A., Proost, S., Tohge, T. & Fernie, A. R. Kingdom-wide analysis of the evolution of the plant type III polyketide synthase superfamily. *Plant Physiol.* **185**, 857–875 (2021).

Acknowledgements

Author contributions

P.B., L.A.d.H., A.J., P.D.S. and A.A. designed the study. P.B., L.A.d.H. and A.A. wrote the manuscript with the assistance and input of all coauthors. P.B. led the chemical profiling, feeding and MALDI imaging experiments with the aid of A.J., J.C., Y.D. and I.R. L.A.d.H. performed all the bioinformatic analyses. P.B., L.A.d.H., A.J., P.D.S., R.B. and R.L. cloned and performed the in vitro enzymatic assays. A.J. performed the confocal measurements and the heterologous expression of genes in *S. cerevisiae*. S.P. prepared the libraries for Transeq sequencing. S.L.-Z., E.S. and N.D. performed the electron microscopy studies. Z.P. performed the transient co-expression of genes in *N. benthamiana*. P.B. and S.L.-Z. analysed the electron microscopy data. T.S. acquired and analysed the NMR data. E.P.-K. performed the flow-cytometry-based genome size estimation. S.M. cultivated and propagated the plants. A.A. supervised the study. All the authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | LC-MS/MS fingerprinting of CBGA 1, heliCBGA 2 and APHA 3 in *H. umbraculigerum*. a. Extracted ion chromatograms and MS/MS spectral matching of cannabigerolic acid (CBGA **1** [M-H]⁻ = 359.222 Da), heli-cannabigerolic acid (heliCBGA **2** [M-H]⁻ = 393.206 Da), and pre-amorphastilbol (APHA **3** [M-H]⁻ = 391.191 Da) standards or authentic metabolites versus a *H. umbraculigerum* leaf extract. To confirm the assignment, CBGA **1** and heliCBGA **2** were purified and analyzed by NMR (Supplementary NMR Data 1, 2). **b**. Stable isotope labeling of CBGA **1**, heliCBGA **2** and APHA **3** via feeding of *H. umbraculigerum* leaves with hexanoic-D_{11} acid, phenylalanine-D_5 or phenylalanine-$^{13}C_9$. The MS/MS spectra of the non-labeled versus the labeled forms show similar fragmentation patterns with mass shifts corresponding with the labeled parts of the molecule.
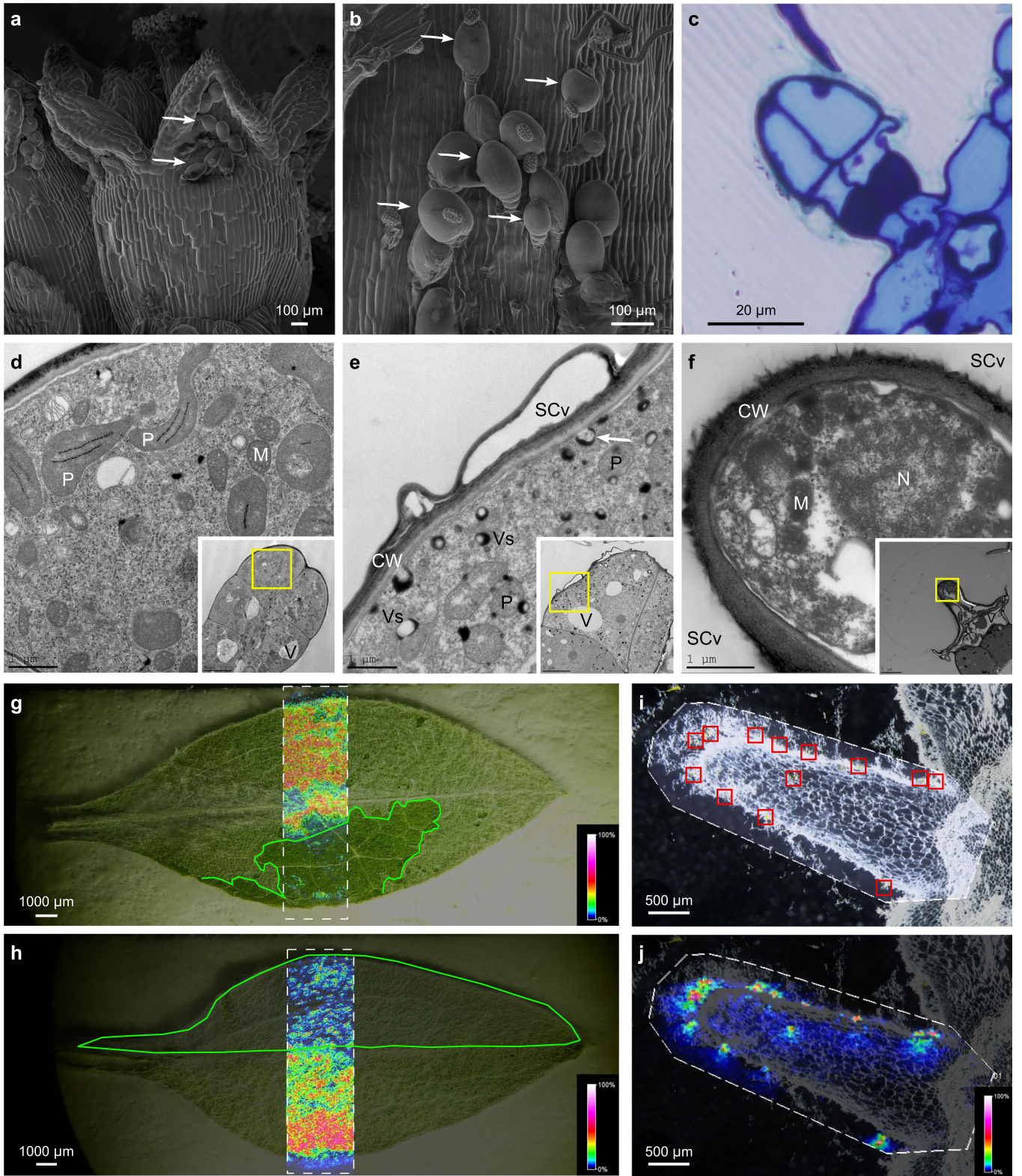
Extended Data Fig. 2 | See next page for caption.

**Extended Data Fig. 2 | Stalked glandular trichomes in leaves and flowers of *H. umbraculigerum*. a**., **b**. Representative cryo-SEM micrographs of the lateral view of flowers showing stalked glandular trichomes (marked by arrows). Micrographs are representative of multiple (n > 3) flower areas sampled on the same day. **c**. Light micrograph showing the biseriate structure of stalked glandular trichomes of leaves (n > 3). **d.-f**. Selected TEM micrographs of leaves' trichomes at different stages of secretion. Micrographs are representative of distinct trichomes (n = 2–5 for each developmental stage) from sections of young and old leaves. High magnification images show the ultrastructure of disk cells (DCs). CW, cell wall; M, mitochondria; N, nucleus; P, plastid; PSP, periplasmic space; SCv, secretory cavity; V, vacuole; Vs, vesicle. **d**. In the presecretory stage, DCs contained a very dense cytoplasm covered by ER and multiple ribosomes. There was no SCv or PSP and plastids were large and resembled pro-plastids. **e**. In the secretory stage, delamination of the apical DC wall led to the formation of the SCv. Electron transparent secretions were exuded out of plastids in vesicles delimited by an electron-dense layer. The vesicles released their contents to the PSP by exocytosis where the secretory product accumulated prior to secretion into the SCv (marked by arrow). **f**. DCs of mature trichomes post-secretion were largely vacuolated with a cytoplasm restricted to the small remaining area. Plastids at this stage had degenerated and no vesicles were observed. The cell wall had a largely cutinized layer with a large SCv. MALDI-MSI of m/z 361.23 ± 0.01 Da signals of the **g**. abaxial and **h**. adaxial leaf domains (n = 1), following partial removal of trichomes by duct tape (the peeled area is outlined by green line). The areas with partially/fully removed trichomes show less or no signals compared to the untouched parts. **i**. Optical image and **j**. MALDI-MSI of m/z 361.23 ± 0.01 Da of a cross-sectioned flower receptacle (n = 2). Glandular trichomes in **i** are marked to improve interpretation. The signals in **g**., **h**. and **j**. correspond with the protonated m/z of CBGA **1** and geranylphlorocaprophenone **4**. The white broken lines in **g.-j**. mark the regions analyzed.
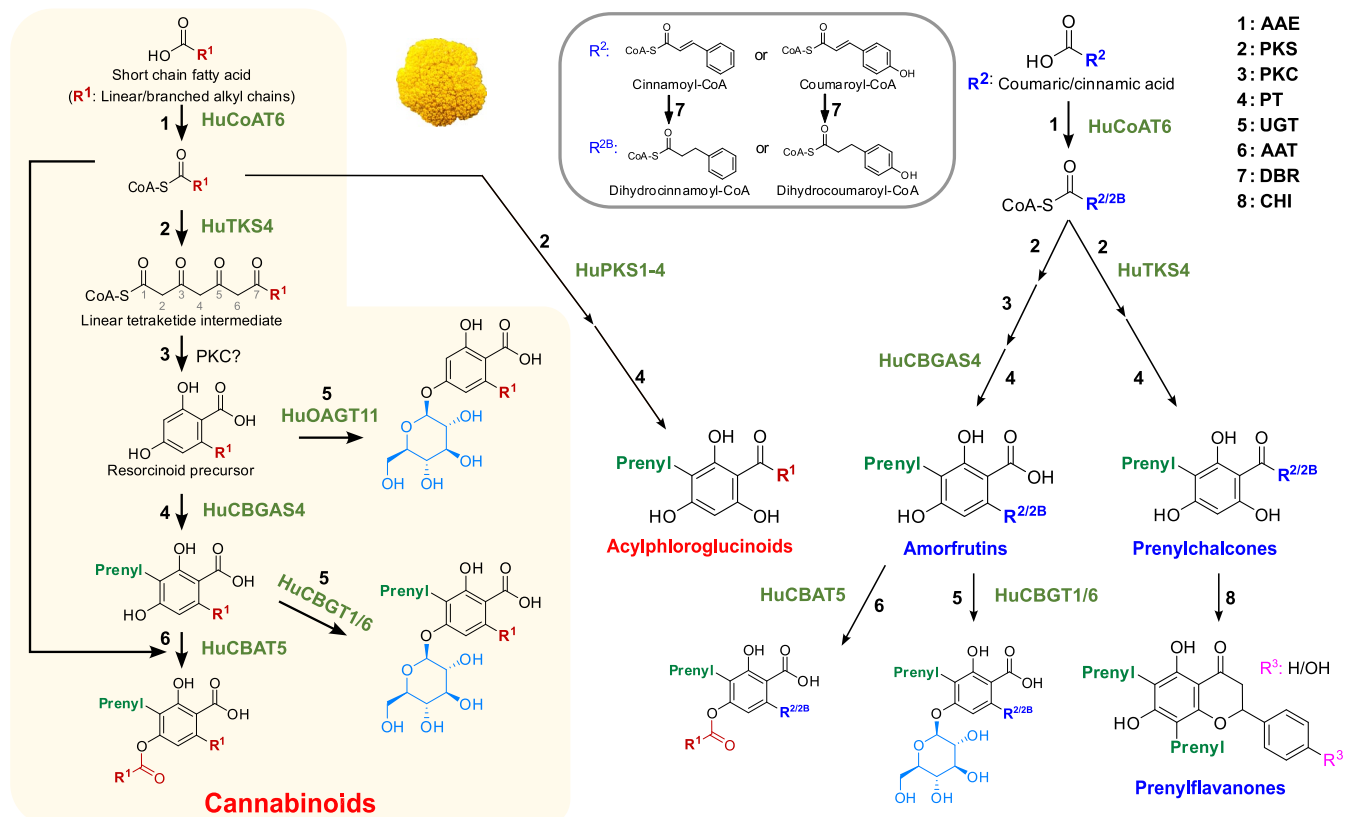
Extended Data Fig. 3 | Predicted parallel metabolic pathways for the biosynthesis of cannabinoids and other terpenophenols present in *H. umbraculigerum*. The predicted types of enzymes catalyzing each reaction are marked by 1–8. Additional functional groups and rearrangements include hydroxylation, double bond isomerization or reduction, cyclization and others. Alkyl chains can be linear/branched with one to seven carbons length; AAE, acyl activating enzyme; PKS, type III polyketide synthase; PKC, polyketide cyclase; PT, prenyl-transferase; UGT, uridine diphosphate-glycosyltransferase; AAT, alcohol acyl transferase; DBR, double bond reductase; CHI, chalcone isomerase. The active enzymes identified in this study are marked by their names. CoAT, acyl-CoA-transferase; TKS, tetraketide synthase; CBGAS, cannabigerolic acid synthase; OAGT, olivetolic acid UGT; CBGT, cannabinoid UGT; CBAT, cannabinoid acyl-transferase.
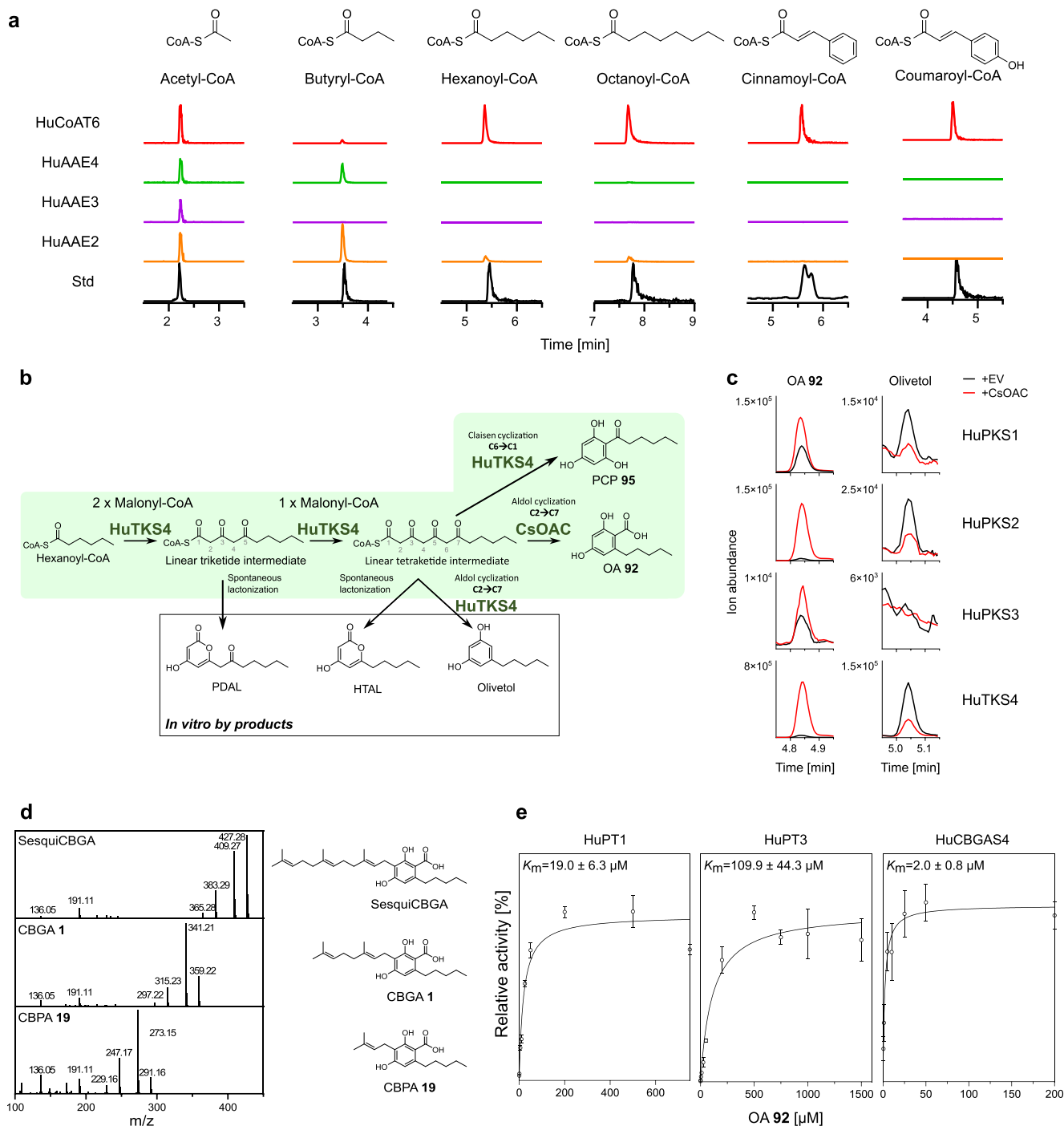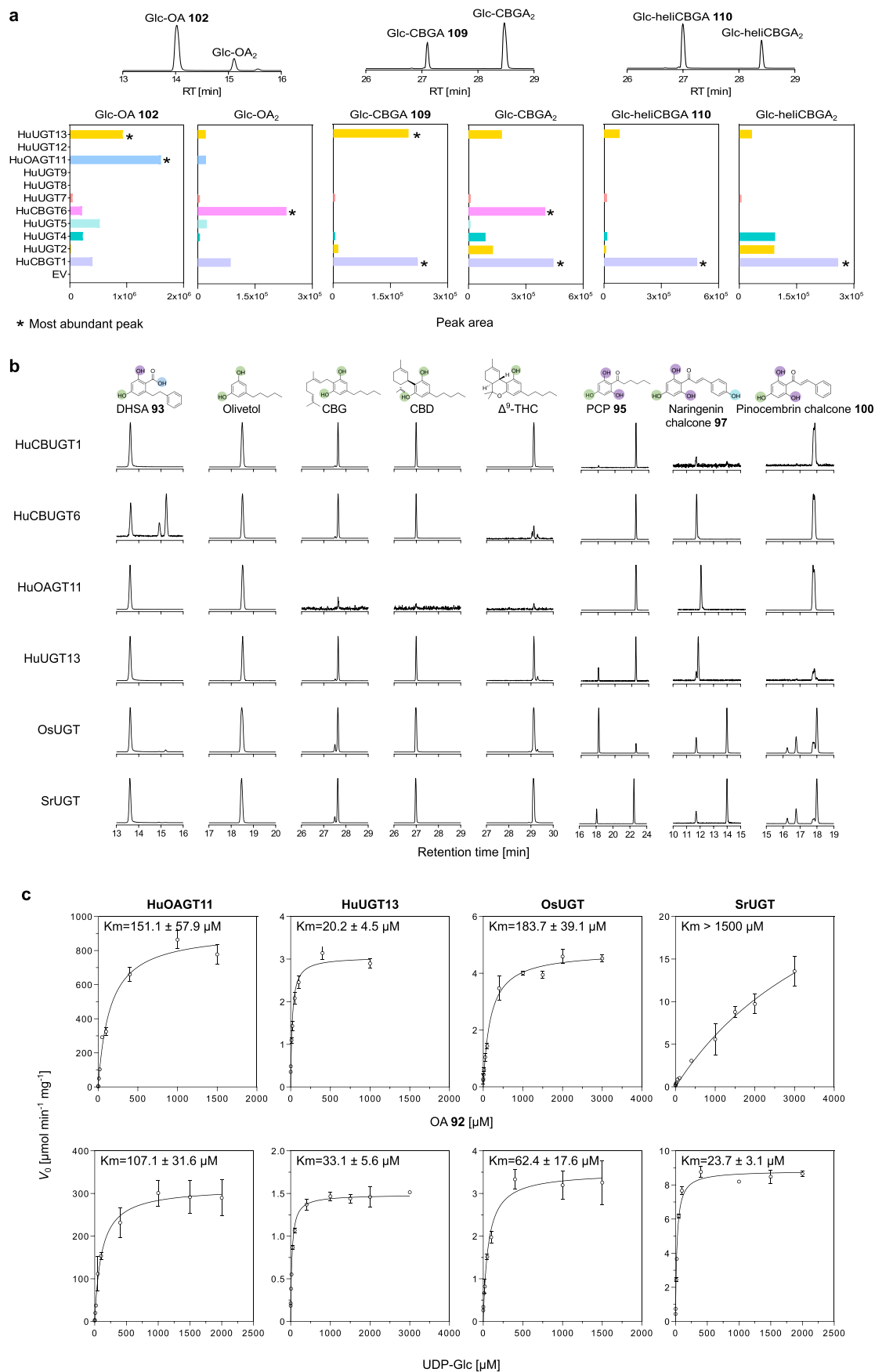
Extended Data Fig. 4 | Functional characterization of HuAAE, HuPKS and HuPTs. a, Ion abundances from triple-Quad analyses of acyl-CoAs produced *in vitro* by the HuAAEs versus analytical standard (Std). **b**, A scheme showing the steps and types of products and by-products synthesized *in vitro* by the recombinant HuPKSs with or without the *Cannabis* olivetolic acid cyclase (CsOAC). **c**, Ion abundances from triple-Quad analyses of OA **92** and olivetol products from coupled recombinant enzyme assays of HuPKSs with either an empty vector (EV) or *Cannabis* olivetolic acid cyclase (CsOAC), in the presence of hexanoyl-CoA and malonyl-CoA. **d**, MS/MS spectra of prenylated OA **92** products with cannabigerolic acid synthase (HuCBGAS4) and either isopentenyl pyrophosphate (IPP), geranyl pyrophosphate (GPP) or farnesyl pyrophosphate (FPP) as the prenyl donors. CBPA **19**, cannabiprenylic acid; CBGA **1**, cannabigerolic acid; sesquiCBGA, sesqui cannabigerolic acid (MS/MS spectrum corresponds to published data from *Cannabis*[15]). **e**, Steady state kinetic analysis of *H. umbraculigerum* prenyl-transferases HuPT1, HuPT3 and HuCBGAS4 with OA **92** and GPP. The Michaelis–Menten *K*m value of each enzyme was calculated using varying (0.5 µM–3 mM) and constant (1 mM) concentrations of each substrate (mean ± s.d.; n = 3 technically independent samples).
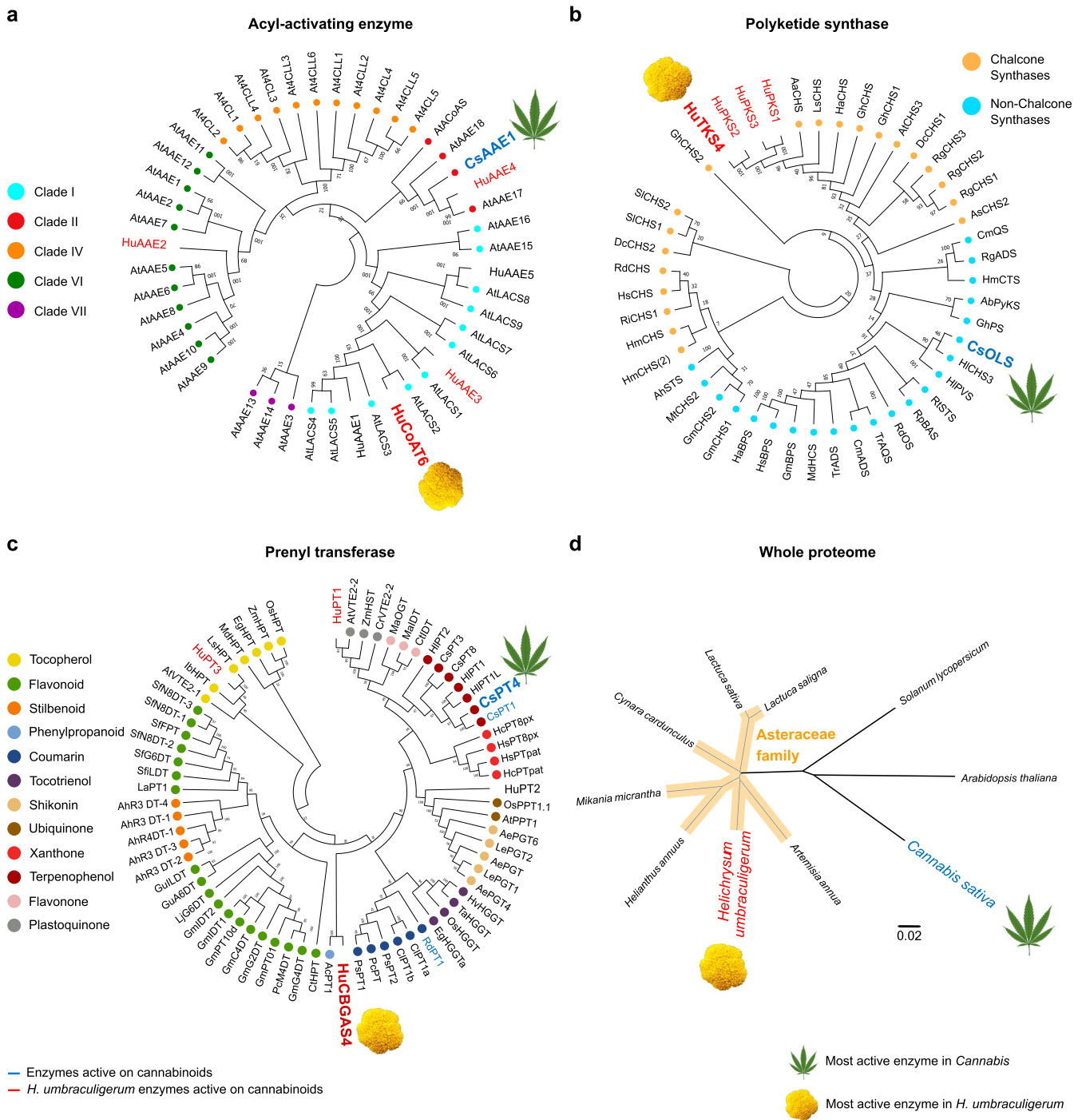
Extended Data Fig. 5 | Phylogenetic analyses of enzymes and whole proteome from *H. umbraculigerum* and different plant species. Phylogenetic analysis of **a**. acyl-activating enzyme (AAE), **b**. polyketide synthase (PKS) and **c**. prenyl transferase (PT) proteins from *H. umbraculigerum* and other plants. Full list of protein IDs is available in Supplementary Table 15. Bootstrap values are indicated at the nodes of each branch. The selection of the proteins was based on **a**. *Arabidopsis thaliana* enzymes[44] or **b**.,**c**. functionally tested enzymes[22,45]. Clades according to substrates or functionalities are marked by different colors. None of the active *H. umbraculigerum* enzymes clustered with any of the known *Cannabis* proteins. **d**. Phylogenetic relationship between *Arabidopsis thaliana*, *Solanum lycopersicum*, *Helianthus annuus*, *Lactuca sativa*, *Cannabis sativa* and *Helicrysum umbraculigerum* illustrate the evolutionary distance between the last two species. The tree was constructed based on the whole proteomes of each species using the word-based software Prot-SpaM.
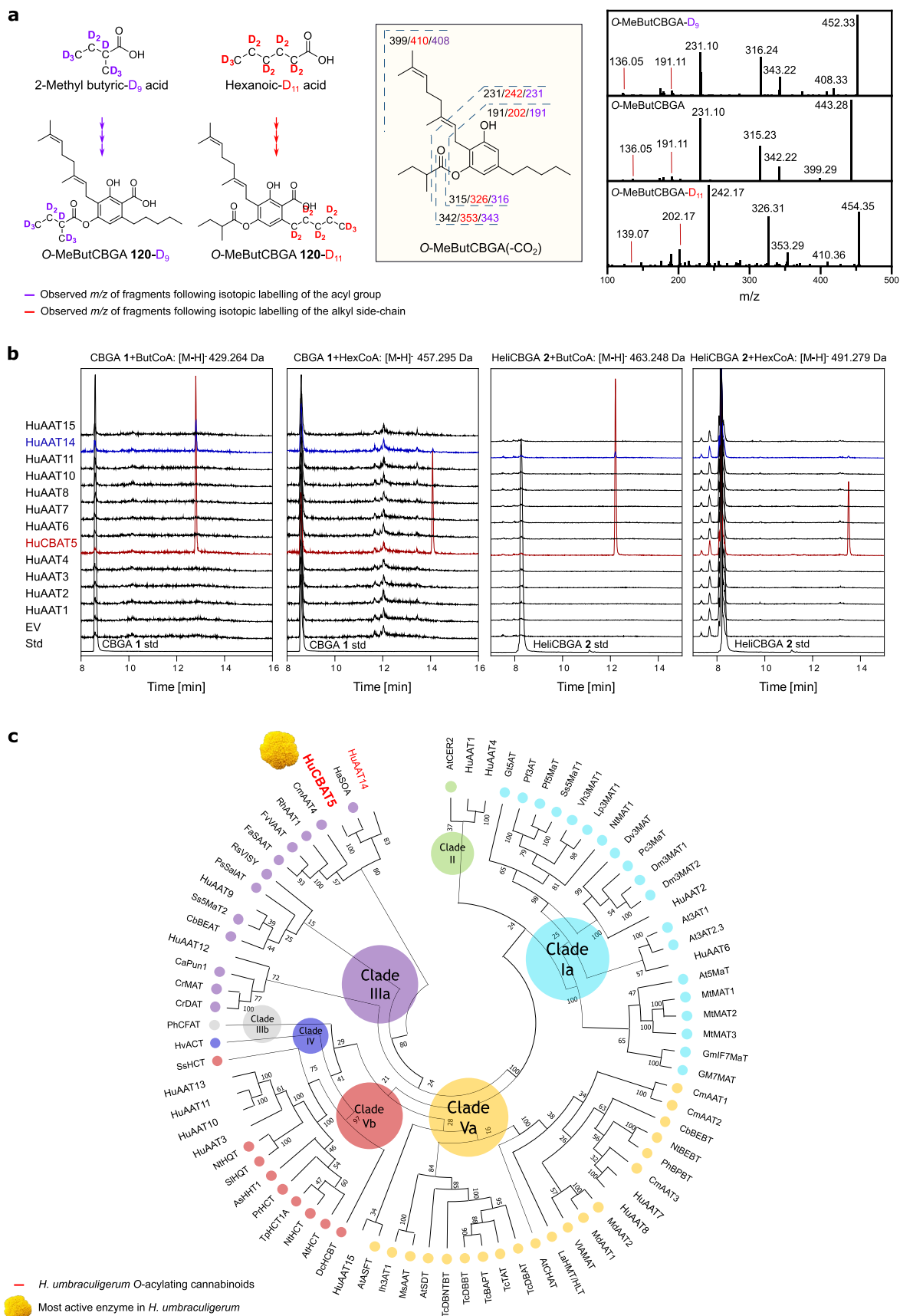
a

* Most abundant peak

b

c

Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Functional characterization of HuUGTs. a. Activities of lysates containing HuUGTs with OA **92**, CBGA **1** and heliCBGA **2** as substrates and uridine diphosphate glucose (UDP-Glc) as the sugar donor (n = 1). Reactions show differing substrate specificities and type of products. Representative peaks correspond to chromatograms obtained for HuCBUGT1. EV, empty vector. **b.** *In vitro* production of monoglucosides with the purified UGTs and additional substrates. Extracted ion chromatograms of the observed monoglucosides using UDP-Glc and either DHSA **93**, olivetol, CBG, CBD, Δ⁹-THC, PCP **95**, naringenin chalcone **97** or pinocembrin chalcone **100**. The substrates naringenin chalcone **97** and pinocembrin chalcone **100** contained mixtures of the chalcones and respective flavanones. All LC-MS chromatograms were selected for the theoretical *m/z* values of the respective metabolites of interest. **c.** Comparison of steady state kinetics of UGTs with OA **92** and UDP-Glc. HuOAUGT11 and HuUGT13 were compared with UGTs from rice (OsUGT) and stevia (SrUGT). Kinetic values were calculated using varying (0.5 μM–3 mM) and constant (1 mM) concentrations of each substrate (mean ± s.d.; n = 3 technically independent samples; measurements were plotted individually). V_0 and Vmax were calculated using the calibration curve of OA **92** since there was no analytical standard available for Glc-OA **102**.

a

— Observed *m/z* of fragments following isotopic labelling of the acyl group
— Observed *m/z* of fragments following isotopic labelling of the alkyl side-chain

b

c

— *H. umbraculigerum O*-acylating cannabinoids

Most active enzyme in *H. umbraculigerum*

Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Functional characterization of HuAATs. a. Stable dual isotope labeling of *O*-MeButCBGA **120** via feeding of *H. umbraculigerum* leaves with either 2-methyl butyric-D_9 acid or hexanoic-D_{11} acid. The MS/MS spectra of the non-labeled versus the two-labeled forms show fragmentation patterns with mass shifts corresponding with the labeled parts of the molecule. **b**. Activities of lysates containing HuAATs with different acyl donors and cannabinoid acceptors. Extracted ion chromatograms were selected for the theoretical *m/z* values of the respective metabolites. Only HuCBAT5 and HuAAT14 (red and blue, respectively) acylated CBGA **1** and heliCBGA **2** with both acyl-CoAs. EV, empty vector; Std, standard; ButCoA, butyryl-CoA; HexCoA, hexanoyl-CoA. **c**. Phylogenetic analysis of HuAAT proteins and identified BAHD AATs from other plants. The Maximum Likelihood tree was constructed with 100 bootstrap tests based on a MUSCLE multiple alignment using the MEGA11 software. The evolutionary distances were computed using the JTTmatrix-based method. Bootstrap values are indicated at the nodes of each branch. The clades of the different AAT types are marked in circles based on Tuominen et al.[26]. The active HuCBAT5 and HuAAT14 were clustered in clade IIIa which represents BAHDs of diverse catalytic functions. Full list of protein IDs is available in Supplementary Table 15.
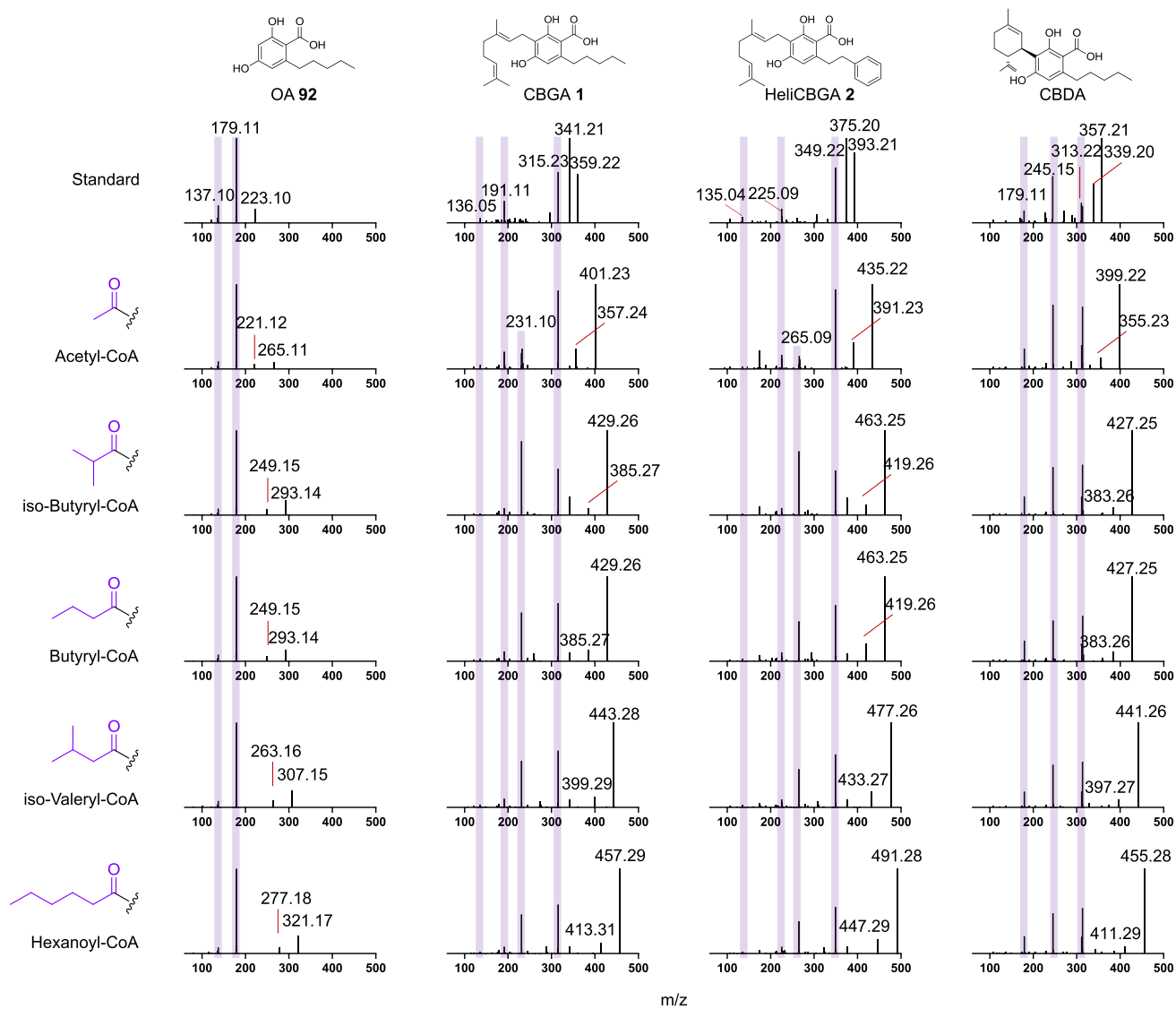
Extended Data Fig. 8 | MS/MS spectra of observed O-acylated cannabinoids following enzymatic assays with the purified HuCBAT5. OA **92**, olivetolic acid; CBGA **1**, cannabigerolic acid; HeliCBGA **2**, helicannabigerolic acid; CBDA, cannabidiolic acid. Full data of MS/MS products appears in Supplementary Table 21. MS/MS fragmentation and retention times correspond to the O-acylated cannabinoids found in the plant.

Extended Data Fig. 9 | Reconstruction of the core cannabinoid pathway in heterologous systems. Schematic representation of products observed in **a**. *N. benthamiana* leaves and **d**. *S. cerevisiae* yeasts following co-expression of different combinations of HuCoAT6, HuTKS4, and HuCBGAS4, along with CsOAC from *Cannabis*. NbUGT, putative *N. benthamiana* uridine diphosphate-glycosyltransferase; HexNa, sodium hexanoate; GPP, geranyl pyrophosphate; OA **92**, olivetolic acid. Extracted ion chromatograms and MS/MS spectra showing **b**. glycosylated OA (Glc-OA **102**), glycosylated polycaprophenone (Glc-PCP1/2) and glycosylated naringenin chalcone (Glc-Naringenin chalcone 1/2) following feeding with HexNa and GPP (I); and **c**. glycosylated cannabigerolic acid (Glc-CBGA **109**) following feeding with OA **92** and GPP (II). Glycosylated metabolites synthesized by the recombinant stevia (SrUGT) or rice (OsUGT) enzymes were used as reference for identification of *N. benthamiana* products according to exact mass, retention time and MS/MS spectra. EV, empty vector; UDP-Glc, uridine diphosphate glucose. **e**. Extracted ion chromatograms of OA **92**, PCP **95** and CBGA **1** products observed in yeasts without any feeding. Identification was according to analytical standards. **f**. Summary of the observed products in each assay. PDAL, pentyl acyl diacetic acid lactone; HTAL, hexanoyl acyl triacetic acid lactone.

Corresponding author(s): Asaph Aharoni
Prashant Sonawane
Paula Berman

Last updated by author(s): Mar 22, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	qTOF MS data was acquired and analyzed using MassLynx V4.2 (Waters). IQ-X data was acquired and analyzed using Xcalibur V4.5 (Thermo). MALDI-MSI data was acquired and analyzed using flexImaging V4.1 (Bruker). NMR data was analyzed using TopSpin 4.1.1 software (Bruker)
Data analysis	For NGS data analyses, details about the parameters used for each software can be found in https://github.com/Luisitox/Helichrysum_paper.

Software versions:
hifiasm v0.15.5-r350
SALSA2 commit v15/06/21
EDTA v1.9.4
Rcorrector v1.0.4
trim_galore v0.6.6
Bowtie2 v2.3.5.1
STAR v2.7.9a
minimap2 v2.24-r1122
Trinity v2.13.2
IsoSeq v3.4.0
Braker v2.1.6
PASA v2.5.1
UMI-tools v1.1.2
TransDecoder v5.5.0
TMHMM v2.0c |

```
SignalP v4.1
TargetP v2.0
HMMER v3.3.2
ncbi-blast v2.10.1+
OrthoFinder v2.5.4

Rpackages
CEMiTool_1.10.2
WGCNA_1.70-3
DESeq2_1.26.0
edgeR_3.28.1
gggenes_0.4.1
circlize_0.4.14
gggenomes_0.9.5.9
```

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All NGS raw sequencing data and genome assembly version Humb_v1 can be found in ENA under the accession number PRJEB52026. The sequences of the active genes reported in this article have been deposited in NCBI GenBank (accessions OQ673107– OQ673113). The genomic position, functional annotation, CDS and protein sequences of these genes are available in Supplementary Table 14 and 19.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical based sample size calculation was not applicable to our study. Details of technical/biological replicates related to various experiments are provided in the manuscript wherever necessary.
Data exclusions	No data was excluded
Replication	Details of technical/biological replicates used in various experiments are provided in the Methods section as well as in the Main, Extended and Supplementary Figure legends, wherever necessary.
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒ ☐	Antibodies
☒ ☐	Eukaryotic cell lines
☒ ☐	Palaeontology and archaeology
☒ ☐	Animals and other organisms
☒ ☐	Clinical data
☒ ☐	Dual use research of concern

Methods

n/a	Involved in the study
☒ ☐	ChIP-seq
☒ ☐	Flow cytometry
☒ ☐	MRI-based neuroimaging