Article

Fruity Ester-Rich Exotic Chemovars: Genome-Wide Identification and Transcriptional Architecture of the *Cannabis sativa* BAHD Superfamily

Jaap-Jan Roukens *

Posted Date: 12 February 2026

doi: 10.20944/preprints202602.1037.v1

Keywords: *cannabis sativa*; BAHD acyltransferases; volatile esters; glandular trichomes; functional genomics

 doi:10.20944/preprints202602.1037.v1

Article

Fruity Ester-Rich Exotic Chemovars: Genome-Wide Identification and Transcriptional Architecture of the *Cannabis sativa BAHD* Superfamily

Jaap-Jan Roukens

Affiliation: Independent Scientist & Cannabis R&D Consultant, 3014 Bern, Switzerland.
jaapjanroukens@gmail.com

Abstract

The contemporary breeding of *Cannabis sativa* L. has shifted toward exotic chemovars defined by fruity, floral, and sweet aromas, traits driven by minor volatile esters rather than quantitatively dominant terpenes. Despite their economic importance, the enzymatic machinery governing the synthesis of these high-value volatile esters remains uncharacterized. This study presents a genome-wide identification and tissue-specific expression landscape of the *CsBAHD* acyltransferase superfamily, the metabolic drivers of ester biosynthesis. Using a custom hidden Markov model (HMM), 108 high-confidence *CsBAHD* genes were identified in the *cs10* reference genome. Physical mapping reveals a non-uniform distribution characterized by dense telomeric gene clusters on chromosomes 1, 2, 4, 5, and 8. While these dynamic regions facilitate rapid chemical diversification via chromosomal recombination, their hyper-variability contributes to linkage drag and the phenotypic instability of exotic traits observed during hybridization. Integration with multi-tissue transcriptomic datasets identified distinct transcripts for the putative enzymes governing ester biosynthesis in the glandular trichome. *CsBAHD45* is a constitutively expressed high-abundance transcript (mean: 1953 TPM), whereas chemotypic diversity is generated by a subset of hyper-variable genes, including *CsBAHD19* and *CsBAHD16*, which exhibit extreme presence/absence variation across the plants. It is proposed that this strain-specific repertoire drives the accumulation of high-value odorants, including sulfur-containing esters responsible for tropical passionfruit notes, phenethyl esters driving honey and fruit nuances, and acetylated terpenes analogous to the floral profiles of lavender and rose. Structural validation via physicochemical fingerprinting and deep modeling with the ESM-2 protein language model confirmed a striking topological consensus with functionally reviewed reference enzymes. Beyond the floral sink, distinct vegetative gene clusters were identified that govern root-zone defense, stem fiber lignification, and seed coat maturation. Phylotranscriptomic analysis suggests that the aerial floral biosynthetic capacity evolved via the neofunctionalization of these ancestral core-fiber and root-defense genes. Finally, this study proposes a physicochemical sequestration model, hypothesizing that ancestral *CsBAHDs* catalyzed the formation of cannabinoid esters to stabilize volatile defensive terpenes as persistent contact insecticides. Consequently, the modern high-potency chemotype may be the product of anthropogenic selection, where the selection for free THC drove the loss of this ester-based defence system. Collectively, these findings provide a high-resolution genomic blueprint for the de-orphaning of the *CsBAHD* superfamily, establishing molecular targets for the marker-assisted breeding of premium aromatic chemovars, optimized industrial fiber properties, and robust root-zone defense systems.

Keywords: *Cannabis sativa*; BAHD acyltransferases; volatile esters; glandular trichomes; functional genomics

 doi:10.20944/preprints202602.1037.v1

1. Introduction

The breeding of *Cannabis sativa* L. has progressed through distinct phases, driven by changes in market regulations and demands. Historically, selection pressure prioritized the yield of the primary psychoactive cannabinoid, Δ^9-tetrahydrocannabinol (THC), often at the expense of broader secondary metabolite diversity [1]. Attention subsequently shifted to the terpene fraction, specifically mono- and sesquiterpenes, recognized for their distinct aroma and biological activities [2]. While the biosynthetic genes governing cannabinoid and terpene pathways have been largely elucidated and mapped, they fail to account for the full diversity of the plant's sensory profile and biological activity.

The contemporary frontier of cannabis breeding has moved toward exotic chemovars defined by unique, high-value aromas such as fruity, candy-like, or savory notes [1–5]. The traditional terpene profiles alone cannot explain the diverse and complex aromatic spectrum associated with the flowers of cannabis [2,6,7]. Instead, these exotic qualities are driven substantially by minor volatile organic compounds (VOCs), such as esters and volatile sulfur compounds (VSCs) [1]. These compounds possess exceptionally low odor thresholds, and thus exert a disproportionate influence on the plant's organoleptic properties even at trace concentrations [2,3]. Indeed, quantitative sensory mapping has demonstrated that the abundance of these minor heterologous volatiles—rather than the dominant terpene profile, correlates most strongly with the 'Exotic Score' and the desirable fruity or savory notes of modern germplasm [3].

These key agronomic traits likely depend on the functional diversity of the BAHD acyltransferase family. Named after the first four discovered members (BEAT, AHCT, HCBT, and DAT), this superfamily is renowned for its catalytic versatility in plants [8]. In *Lavandula angustifolia* (Lavender), specific BAHD enzymes catalyze the acetylation of linalool to produce linalyl acetate, a critical component of its characteristic floral aroma [9]. Beyond simple volatiles, the family modifies complex scaffolds; in *Papaver somniferum* (Opium poppy), the enzyme SalAT is essential for morphinan alkaloid synthesis [10], while in *Catharanthus roseus*, members of the family modify indole alkaloids utilized in chemotherapy [11]. Furthermore, BAHD enzymes mediate pigment modifications and defense mechanisms, such as the acylation of anthocyanins for color stability and the synthesis of specialized metabolites to protect subterranean tissues from pathogens as well as from abiotic stress [8].

The functional annotation of the *BAHD* family is frequently complicated by high sequence plasticity and the presence of non-functional paralogs. To achieve higher precision, an integrated workflow was engineered, starting with a custom hidden Markov model (HMM) trained exclusively on reviewed plant-specific BAHD sequences. This approach successfully identified 108 high-confidence genes in the *C. sativa* (cs10) genome. A critical component of this study was the integration of these candidates with a comprehensive multi-tissue transcriptomic atlas [12], which allowed the anchoring of genomic data to physiological expression patterns across 13 distinct tissues. To validate the structural integrity of these candidates without wet-lab testing, we employed diverse computational tools including profiling the conservation of essential catalytic (HXXXD) and structural (DFGWG) motifs, physicochemical fingerprinting via Kyte-Doolittle hydrophobicity scales, and structural validation using the ESM-2 Protein Language Model to confirm 3D topological homology with known reference transferases. Collectively, this work provides a high-resolution genomic map of the *CsBAHD* family and identifies specific putative targets governing ester biosynthesis in the glandular trichomes. Moreover, this analysis highlights key candidates potentially involved in modulating stem fiber properties for industrial applications and rhizosphere defense mechanisms conferring pathogen resistance in the roots.

 doi:10.20944/preprints202602.1037.v1

2. Methods

2.1. Data acquisition and database construction

Reference sequences for the Transferase family (Pfam: PF02458) were retrieved from UniProtKB, restricted to Reviewed (Swiss-Prot) entries. A custom Python script filtered the dataset based on genus-level taxonomy, retaining only sequences from 37 specified plant genera. This reference set was augmented with five *Rosales* sequences from *Malus*, *Prunus*, and *Fragaria* utilizing identifier matching. The final dataset was deduplicated by sequence identity. The *Cannabis sativa* L. cultivar cs10 (CBDRx) proteome was retrieved from the NCBI RefSeq database (Assembly Accession: GCF_900626175.2) and extracted for local analysis.

2.2. Hidden Markov model construction and proteome mining

The computational workflow was executed in a cloud-based environment (Google Colab) utilizing the pyhmmer library. The 122 reference sequences were aligned using MUSCLE to generate the input for the plan7 builder. The resulting profile HMM was used to query the *Cannabis sativa* cs10 proteome with an E-value threshold of 1e-5. For each significant hit, the single highest-scoring domain was selected. Sequences were strictly trimmed to the HMM envelope coordinates defined by the env_from and env_to parameters. A length filtration step was applied to these extracted domains, retaining only sequences ≥ 350 amino acids.

2.3. Multiple sequence alignment and phylogenetic reconstruction

To resolve the evolutionary relationships of the *Cannabis* BAHD family, a cloud-based computational pipeline was employed. The full-length protein sequences were aligned using MAFFT (v7) with the --auto flag, which automatically selects the optimal alignment strategy (L-INS-i or FFT-NS-2) based on dataset size and complexity. This alignment served as the input for phylogenetic inference using FastTree, employing the Le-Gascuel (LG) model with the Gamma parameter to account for rate heterogeneity. The tree was rooted using the *Selaginella moellendorffii* HCT sequence (UniProt: D8S308) as a functional outgroup to orient the basal divergence of the family.

2.4. Physicochemical profiling and structural fingerprinting

To verify that the identified candidates fall within the same physicochemical range as functional BAHD enzymes, global molecular properties were calculated using the ProtParam module of Biopython. Full-length sequences were utilized for this analysis to capture the properties of the complete nascent protein. Molecular weight (MW) and isoelectric point (pI) were computed for each candidate and projected onto a virtual 2D gel landscape.

To assess structural folding topology, hydrophobicity fingerprints were generated using the Kyte-Doolittle scale. This analysis was performed on the global multiple sequence alignment of the full-length proteins. By utilizing the aligned sequences, positional equivalence was ensured across the dataset while retaining the complete structural context. The average hydrophobicity for each aligned column was smoothed using a 15-residue moving average. Structural homology was validated by calculating the Pearson correlation coefficient (r) and Root Mean Square Deviation (RMSD) against the reference profile.

2.5. Motif conservation profiling

To characterize the structural diversity of the Cannabis BAHD candidates, the conservation of the canonical HXXXD and DFGWG motifs was analyzed across the entire candidate set. Unlike standard pipelines that filter for perfect motif retention, all identified sequences were retained to capture the full evolutionary spectrum of full domain CsBAHDs, including potential pseudogenes or variants with divergent active sites. The global multiple sequence alignment was interrogated using a custom Python script to calculate positional bit scores for the motif windows and four flanking

 doi:10.20944/preprints202602.1037.v1

residues. This approach allowed for a quantitative comparison of motif conservation between the *Cannabis* candidates and the validated reference core without excluding atypical sequences.

2.6. Deep structural validation via protein language modeling

To interpret the latent structural constraints governing the BAHD family, the ESM-2 Protein Language Model (esm2_t33_650M_UR50D), a transformer architecture trained on 65 million protein sequences, was employed. This analysis was executed in a Google Colab) where full-length sequences were tokenized and passed through the model to predict pairwise amino acid contact maps, which serve as high-fidelity proxies for 3D protein folding.

To compare the structural core of the Cannabis candidates against the reference architecture, a custom matrix alignment algorithm was implemented. A standardized core window of 350 residues was defined based on a random reference template. For every target sequence, the predicted contact map was spatially scanned using a sliding window approach (range -50 to +200 positions) to identify the region of maximal structural overlap with the template core. Validated cores were aggregated to compute a consensus contact map for both the reference and *Cannabis* BAHD populations. The structural identity of the two groups was quantified by calculating the Pearson correlation coefficient (r) between their flattened average contact matrices. Finally, a composite overlay visualization was generated to map the conserved enzymes internal binding interactions common to both groups.

2.7. Quantitative alignment visualization

To analyze the structural continuity and sequence conservation relative to the catalytic center, the global multiple sequence alignment was computationally partitioned into Reference and *Cannabis* sub-alignments. A binary transformation was applied to both datasets, converting amino acid sequences into residue occupancy matrices (1 for residues, 0 for gaps) to visualize the preservation of the structural scaffold independent of sequence identity.

The alignment was anchored by algorithmically detecting the column with the highest occupancy of the HXXXD motif. Two quantitative profiles were computed relative to this anchor. First, an intra-group conservation score was calculated as the frequency of the most abundant residue per column, normalized by the total number of sequences to penalize gaps. Second, a group similarity index was derived to quantify shared consensus; this metric assigned a score based on the inverse gap frequency (1-max_gap_freq) exclusively at positions where the consensus residue was identical between the Reference and *Cannabis* populations. The resulting profiles were smoothed using a 5-residue moving average.

2.8. Candidate refinement and systematic nomenclature

To establish a consistent naming convention, the initial set of HMM-identified candidates underwent a redundancy filtration step to remove alternative splice variants. This analysis utilized the official NCBI RefSeq genomic annotation file (GFF3) corresponding to the *Cannabis sativa* cs10 assembly (GCF_900626175.2). A custom Python script parsed these annotations to group protein hits mapping to the same gene locus. Only the longest protein isoform per locus was retained to represent the unique gene.

The resulting non-redundant candidates were then physically mapped to the genome to assign systematic identifiers. Using the coordinate data extracted from the RefSeq GFF3 file, the candidates were sorted first by chromosome (Chr01 through ChrX, followed by unplaced scaffolds) and subsequently by their physical start position in ascending order. Sequential identifiers (*CsBAHD01* to *CsBAHD108*) were assigned based on this physical distribution, ensuring that the nomenclature reflects the genomic organization of the family.

 doi:10.20944/preprints202602.1037.v1

2.9. Transcriptomic data integration and reciprocal homology mapping

To contextualize the identified BAHD candidates within a comprehensive physiological landscape, we integrated data from the cannabis expression atlas, a recent meta-analysis compiling 394 high-quality RNA-seq samples across 13 tissues and 55 cultivars [12]. As the Atlas standardized all public expression data onto the *Jamaican Lion* reference genome (GCA_012923435.1), a computational bridging strategy was required to link these profiles to our *cs10*-based candidates.

The Atlas dataset, originally stored in serialized R-data format (.rda), was extracted using a custom Python pipeline utilizing the rdata library. Gene identifiers were recovered from the object metadata and mapped to their corresponding NCBI Protein Accessions to enable nucleotide sequence analysis. To link the cs10 candidates to the Jamaican Lion gene models used in the Atlas [12], a reciprocal BLASTP pipeline was implemented. The *cs10* BAHD candidates were queried against the complete *Jamaican Lion* protein dataset (Forward BLAST; E-value $< 1e^{-10}$). Top hits were subsequently queried back against the *cs10* proteome (Reverse BLAST). A candidate was classified as a verified ortholog only if the reverse search returned the original *cs10* query as the primary hit. To ensure robust data linkage, NCBI version suffixes (e.g., the .1 in XP_0000.1 denoting database history) were normalized prior to comparison. This step prevented false negatives caused by minor annotation updates without compromising the specificity of the ortholog assignment. This approach successfully anchored the diverse transcriptomic data from the expression atlas to our specific genomic targets.

2.10. Integrated phylogenomic visualization and statistical profiling

To analyze the transcriptomic landscape within an evolutionary context, a unified computational workflow linking phylogeny, orthology, and gene expression was developed. The maximum likelihood tree served as the structural scaffold for data harmonization; rows within the expression matrix were reordered to strictly match the tip topology of the phylogeny, ensuring a direct correspondence between evolutionary position and expression profiles. To prevent artifacts from ambiguous orthology, a masking filter was applied: expression values (TPM) for candidates classified as non-reciprocal hits (see Section 2.9) were computationally nullified (set to NaN). This ensured that quantitative analyses relied exclusively on high-confidence, sequence-validated orthologs.

The harmonized data was visualized using a custom phylogenetic-heatmap plotting that synchronizes the tree with a binary orthology heatmap and a quantitative expression map. Bubble dimensions were scaled logarithmically (size $\propto \log_2[TPM + 1]$) to accommodate the dynamic range of transcript abundance, with values exceeding a saturation threshold of 500 TPM capped to preserve visual resolution for specialized metabolites.

Functional specialization was further quantified using a suite of statistical metrics. The Tissue Specificity Index (r) and Shannon Entropy (H) were calculated for each gene to distinguish between ubiquitous housekeeping transcripts and those exhibiting strict tissue compartmentalization. Transcriptional variability was assessed via the Coefficient of Variation (CV) across all samples, classifying candidates into stability quantiles (high, medium, low). Finally, to identify dominant candidates in agriculturally relevant sinks, a cumulative abundance analysis was performed for root, stem, and seed tissues.

3. Results

3.1. Genome-wide identification and chromosomal distribution

The Hidden Markov Model (HMM) based screening of the *Cannabis sativa* cs10 proteome initially yielded 137 putative acyltransferase sequences. To distinguish functional BAHD candidates from fragmented pseudogenes or partial domains, a structural filtering step was applied. Analysis of the domain length distribution revealed a distinct population of truncated sequences; consequently,

 doi:10.20944/preprints202602.1037.v1

a cutoff of 350 amino acids was implemented (Figure 1A), removing 29 fragmented entries. This filtering, combined with the removal of redundant isoforms, resulted in a final high-confidence set of 108 BAHD candidates.

These genes were physically mapped to the *cs10* reference assembly and assigned systematic identifiers (*CsBAHD01–CsBAHD108*) based on their genomic order. The chromosomal distribution is highly non-uniform and exhibits a strong telomeric bias **(Figure 1B)**. A spatial clustering analysis revealed that 55% of the family (59 genes) is organized into 10 high-density tandem duplication arrays (TDAs) (defined as 4 genes within 200 kb). The most significant expansion occurs on chromosome 4, which hosts a large, continuous array of 14 genes (*CsBAHD41–CsBAHD54*) spanning a 437-kb region at the distal end. Notably, this locus harbors *CsBAHD45*, a candidate identified in downstream analyses as the dominant transcript in glandular trichomes, suggesting this genomic expansion drives high metabolic ester output. Similarly, chromosome 8 contains a hyper-dense cluster of 9 genes within a 106-kb window, while chromosome 1 features a terminal array of 6 genes. In contrast, the interstitial and centromeric regions are largely devoid of such structures, indicating that the evolution of the *Cannabis BAHD* superfamily has been driven by, sequential duplication events at the chromosomal termini.

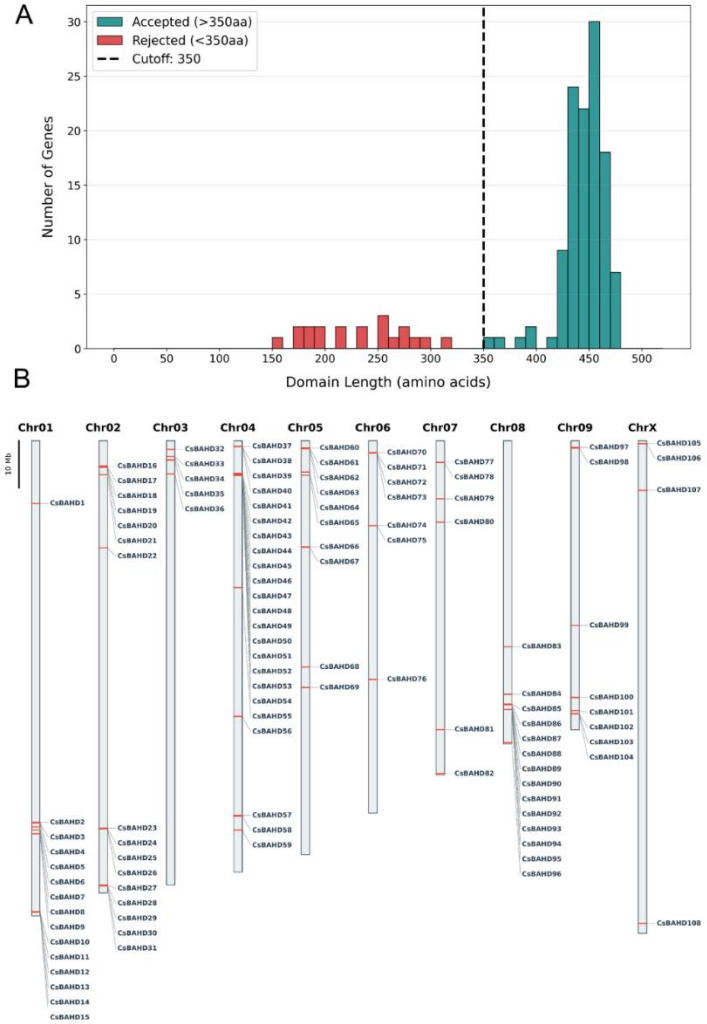

Figure 1. Genome-wide identification and chromosomal distribution of the *Cannabis sativa BAHD* superfamily. (A) Domain length distribution of putative acyltransferase sequences identified via HMM screening. The dashed vertical line indicates the structural quality threshold (350 aa) applied to filter truncated or fragmented domains. **(B)** Physical map of the *CsBAHD* gene family on the ten *C. sativa* chromosomes (cs10 assembly; RefSeq GCF_900626175.2). Systematic identifiers (*CsBAHD01–CsBAHD108*) were assigned sequentially based on physical gene order, traversing from the top of chromosome 1 through chromosome X. The distribution highlights distinct telomeric clustering, particularly on chromosomes 4 and 8.

 doi:10.20944/preprints202602.1037.v1

3.2. Structural conservation and physicochemical landscape

To validate the structural integrity of the identified candidates, their primary features were compared against a curated plant only reference set of functionally characterized plant BAHD enzymes. Analysis of the sequence length distribution **(Figure 2A)** confirmed that the *Cannabis* candidates fall strictly within the canonical size range of the superfamily. The *Cannabis* repertoire exhibits a mean length of 456.0 ± 22.9 aa, highly comparable to the reference population (444.6 ± 17.7 aa), indicating that the core domain architecture is strictly preserved in the *cs10* lineage without aberrant truncation or extension.

Structural conservation was further assessed via hydrophobicity fingerprint analysis **(Figure 2B)**, which maps the hydropathic profile across the multiple sequence alignment. Despite sequence diversification, the CsBAHD family maintains a global folding pattern nearly identical to the reference enzymes, evidenced by a striking positive Pearson correlation ($r = 0.96$) and a minimal root mean square deviation (RMSD) of 0.091 in the smoothed hydropathy scores. These hydropathic profiles were generated using a 15-residue sliding window average, a technique that filters high-frequency sequence noise to reveal the underlying secondary structure tendencies. This visualization effectively demarcates the alternating hydrophobic core regions and hydrophilic surface loops that define the conserved BAHD fold and confirms that the core hydrophobic packing and solvent-channel architecture required for transferase activity are strictly conserved in the *Cannabis* lineage.

Finally, the global physicochemical properties were mapped using a virtual 2D gel simulation **(Figure 2C)**. The analysis revealed that the CsBAHD family occupies a defined physicochemical space that overlaps extensively with the reference enzymes. Both populations are predominantly acidic, with nearly identical isoelectric point (pI) centroids (Reference pI: 6.37 ± 1.13; Cannabis pI: 6.44 ± 1.10) and similar molecular weight distributions, further supporting the classification of these candidates as functional acyltransferases.

 doi:10.20944/preprints202602.1037.v1

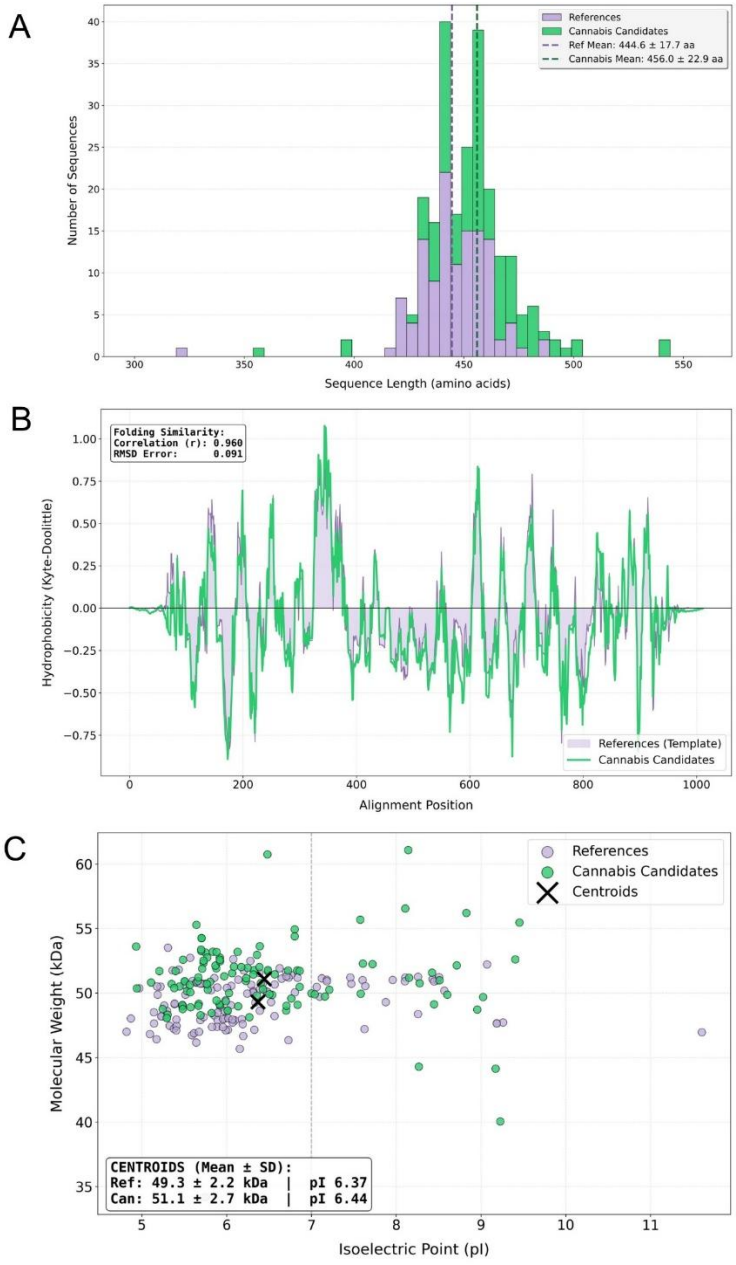

Figure 2. Structural and physicochemical profiling of the *Cannabis sativa* BAHD superfamily. (A) Sequence length distribution comparison between the CsBAHD candidates ($n = 108$) and characterized reference BAHD enzymes ($n = 121$). **(B)** Hydrophobicity fingerprint analysis comparing the consensus Kyte-Doolittle hydropathy profile of the CsBAHD family against the reference set. Hydropathy scores were calculated across the alignment and smoothed using a sliding window average to highlight secondary structure trends. **(C)** Virtual 2D gel electrophoresis simulation plotting molecular weight (MW) against isoelectric point (pI). The X markers indicate the population centroids.

3.3. Conservation of catalytic motifs and deep structural modeling

To assess the functional potential of the CsBAHD repertoire, the positional entropy of the canonical catalytic motifs required for acyl-transfer activity was analyzed. The HXXXD motif, which houses the general base catalyst, exhibited near-perfect conservation in the *Cannabis* lineage **(Figure 3A)**. The invariant histidine yielded a bit score of 4.10, while the stabilizing aspartate reached the theoretical maximum of 4.32 bits. Similarly, the structural DFGWG motif maintained high level of preservation, ensuring the integrity of the surface level CoA binding pocked near active site cleft.

 doi:10.20944/preprints202602.1037.v1

To move beyond linear sequence analysis, the ESM-2 (650M parameter) protein language model was utilized to predict the deep structural contact maps of the entire family. By generating a consensus contact geometry for both the reference and *Cannabis* populations, the 3D structural constraints governing the family's fold were visualized **(Figure 3B)**. Intriguingly, the *Cannabis* consensus map displayed a striking topological overlap with the reference family, with preserved long-range contact patterns which seem to be characteristic of the distinct BAHD solvent channel architecture. A Pearson correlation analysis comparing these population-aggregated contact densities revealed a high degree of structural homology ($r = 0.947$), confirming that the evolutionary constraints defining the BAHD scaffold are strictly maintained across the *Cannabis* family despite extensive diversification in the loops.

To spatially resolve this balance between conservation and diversification, a global alignment landscape was constructed **(Figure 4)**. By converting the multiple sequence alignment into a pixelated residue matrix, the distribution of insertions and deletions (indels) across the entire *Cannabis* repertoire was visualized. The analysis reveals a distinct block-modular architecture: the Cannabis candidates **(Figure 4B)** maintain the same rigid, high-consensus core blocks as the reference enzymes **(Figure 4A)** but exhibit lineage-specific variation in the inter-domain linkers. The column-wise consensus score **(Figure 4C)** fluctuates rhythmically, marking the boundaries between the solvent-protected core (α-barrels and β-sheets) and the surface-exposed variable loops where a high degree of variability is encoded.

3.4. Phylotranscriptomic landscape and chemotypic functionalization

To dissect the functional specialization of the *CsBAHD* superfamily, the maximum likelihood phylogeny was integrated with a multi-tissue transcriptomic atlas [12] covering root, stem, leaf, seed, and glandular trichomes and nine other tissues. To ensure robustness, this integration was restricted to high-confidence reciprocal orthologs, effectively filtering noise and preventing cross-mapping artifacts between closely related paralogs. The resulting phylotranscriptomic map **(Figure 5)** reveals a striking modular architecture where expression patterns are inextricably linked to genomic topology. While the deep phylogenetic backbone of the family largely exhibits basal, ubiquitous expression levels (<10 TPM), likely representing housekeeping functions, the terminal lineage-specific expansions have evolved into high-titer, tissue-specific metabolic modules. This dichotomy suggests that *Cannabis* has recruited specific *BAHD* subclades from a generalist ancestry to drive the massive, specialized biosynthetic output required by distinct vegetative and reproductive organs.

The most significant feature of this landscape is the convergence of three phylogenetically distinct clades toward glandular trichome specificity **(Figure 6)**, identifying them as the putative machinery for volatile ester biosynthesis**.** The segregation of these targets into disparate branches of the phylogenetic tree indicates a history of multiple independent recruitment events. It appears that distinct genomic lineages were parallelly co-opted from the ancestral repertoire to fuel the specialized metabolic demands of the glandular trichome sink. Clade A1, corresponding to the expanded chromosome 4 tandem array **(Figure 6A)**, represents based on expression data a key loci for ester biosynthesis in the trichome. This cluster is dominated by *CsBAHD45*, the single most abundant transcript in the entire dataset. With a median expression of 1999.6 TPM (Mean: 1953.0 TPM) and a trichome/tissue specificity ratio of 133.9, its constitutive high-level expression across all genotypes suggests it functions as the primary acetyltransferase responsible for the baseline flux of esters **(Figure 8)**. The sheer magnitude of this transcriptional output is notable, as the metabolic load of this single locus within the glandular trichome exceeds the cumulative expression of the entire remaining superfamily combined. This overwhelming dominance confirms its role as the invariant metabolic engine of the pathway; while its absolute abundance varies quantitatively across the germplasm, it remains functionally ubiquitous, effectively distinguishing it from the stochastic presence-absence variation that defines the specialized chemotype drivers. Clade C on chromosome 2 **(Figure 6C)** has diverged to form a highly variable, strain-specific trichome and root module. This group is anchored by *CsBAHD16* (Median: 433.8 TPM), but is most notable for *CsBAHD19* which exhibits negligible

 doi:10.20944/preprints202602.1037.v1

median trichome expression (1.2 TPM), with an explosive dynamic range, reaching 1451.4 TPM in specific cultivars. This extreme skew between typical (median) and potential (max) output identifies it as a chemotype-defining switch. Finally, the clade B lineage **(Figure 6B)** contributes a third tier of trichome-exclusive activity, driven by *CsBAHD97* (median: 196.6 TPM) and *CsBAHD21* (median: 60.8 TPM), which display intermediate expression profiles distinct from the constitutive A and variable C modules.

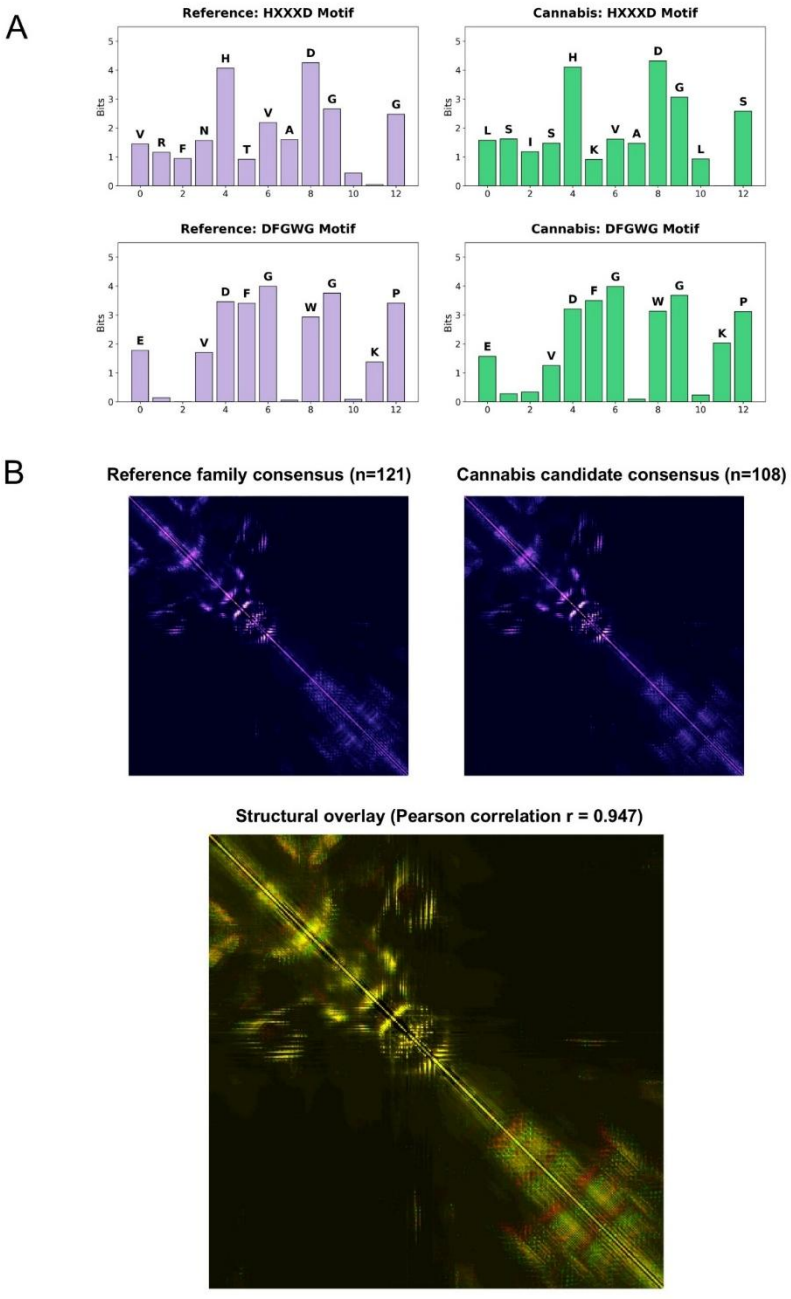

Figure 3. Active site conservation and deep structural consensus modeling of the *Cannabis* BAHD superfamily. (A) Positional entropy analysis of the canonical HXXXD and DFGWG motifs. The bar heights represent the information content for the reference and *Cannabis* populations. Bit scores were calculated using Shannon entropy ($R = \log_2(20) - H$) on the multiple sequence alignment, where a theoretical maximum of 4.32 bits indicates absolute identity (0% entropy) and 0 bits indicates random distribution. **(B)** Deep structural consensus modeling using the ESM-2 (evolutionary scale modeling) protein language model (650M parameters). The heatmaps represent the population-averaged residue-residue contact probabilities, generated by extracting the pairwise attention maps from the model's final transformer layers for every sequence and averaging them to

 doi:10.20944/preprints202602.1037.v1

produce a consensus geometry for each lineage. Structural overlay of the consensus contact maps, where the Pearson correlation coefficient quantifies the linear relationship between the pixel intensities of the two matrices, serving as a global metric for 3D topological similarity.

Mapping these functional modules to the genome reveals clear difference between the two trichome clades A and C. While the constitutive trichome ester driver (*CsBAHD45*; clade A1) is located on chromosome 4, the variable trichome module (clade C) on chromosome 2, also includes a root-specific gene. This specific *CsBAHD16-20* array is positioned on the same chromosome as the root-specific *CsBAHD23-27* cluster. This topological linkage suggests that the high-variability trichome drivers may have evolved via local duplication and neofunctionalization of ancestral root defense genes, and selection by humans for those desirable aromatic traits.

3.5. Stochastic regulation drives chemotypic diversity

To elucidate the regulatory logic governing these expression patterns, a multi-dimensional variability analysis of the top six trichome candidates was performed across the 59 analyzed transcriptomes **(Figure 7)**. The distribution of transcript abundance **(Figure 7A)** reveals a fundamental dichotomy in the biosynthetic apparatus. *CsBAHD45* exhibits a constitutive profile, maintaining a tight, high-normal distribution across the entire population (CV = 39.3%; **Figure 7B**). In stark contrast, *CsBAHD19* and *CsBAHD17* display bimodal distributions, appearing functionally silenced in the majority of strains while surging to extreme outliers in specific genotypes (CV > 175%). This regulatory inequality is quantified by the Gini Index **(Figure 7C)**, where *CsBAHD19* (Gini = 0.87) ranks as the single most polarized transcript in the entire superfamily (Rank #1 of 43 expressed genes), approaching theoretical exclusivity.

 doi:10.20944/preprints202602.1037.v1

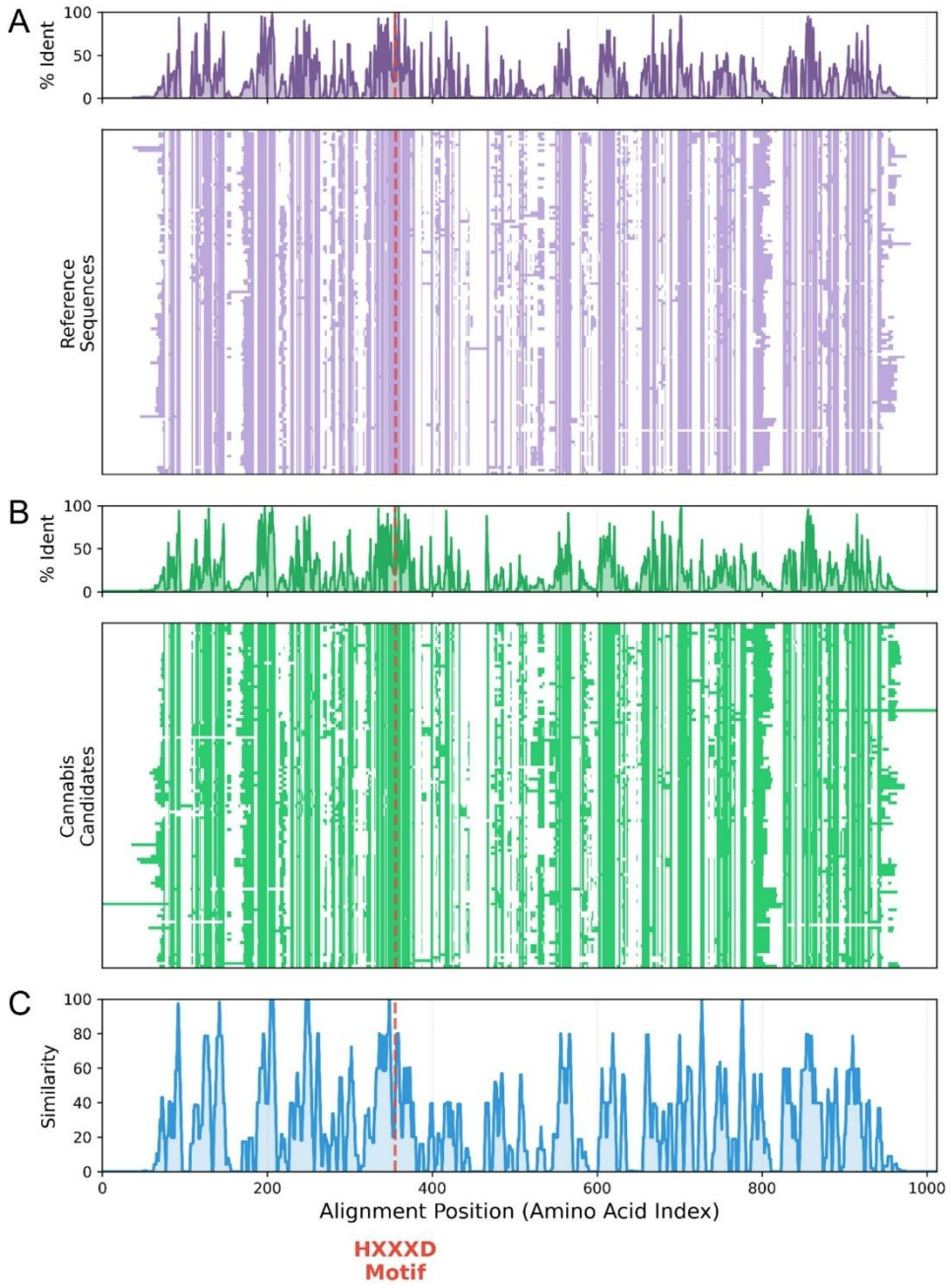

Figure 4. Global alignment landscape and indel distribution of the *Cannabis* BAHD repertoire. (A and B) Pixelated alignment matrices for the Reference (n = 121) and *Cannabis* (n = 108) populations. Sequences were digitized into a residue matrix where the X-axis represents the alignment position (0–1200) and the Y-axis represents individual gene entries. Regions of high structural homology appear as continuous vertical bands, while evolutionary plasticity is visualized as staggered gaps (white space) representing insertions/deletions (indels). The *Cannabis* landscape mirrors the reference block structure, confirming global domain integrity. **(C)** The consensus score plot function represents the positional homogeneity (S) calculated for each alignment column i as $S_i = (n_{max}/N)$, where n_{max} is the frequency of the most abundant residue and is the total number of non-gap sequences.

This binary expression profile of *CsBAHD19* is visualized by the empirical cumulative distribution function (ECDF) **(Figure 7D)**, which maps the proportion of the population exceeding specific expression thresholds. For *CsBAHD45*, the curve is sigmoidal and right-shifted: 0% of the population falls below 10 TPM, while 86.4% of all strains sustain very high expression levels >1000 TPM. Conversely, the *CsBAHD19* curve is heavily left-skewed: 69.5% of the population fails to reach

 doi:10.20944/preprints202602.1037.v1

the 10 TPM detection threshold, confirming its status as a latent gene. However, in the 5.1% of strains where it is fully activated (>1000 TPM), it rivals the output of the housekeeping machinery. Indeed, *CsBAHD19* exceeds the detection threshold (>10 TPM) in only 18 of the 59 strains (30.5%), effectively functioning as a binary presence/absence switch. In this active subset ($n = 18$), transcript levels averaged 424.8 ± 505.1 TPM, reflecting significant amplitude variation even among expressing genotypes.

Crucially, the cumulative metabolic load analysis **(Figure 7E)** demonstrates that the total acyltransfer capacity of the glandular trichome is not fixed but highly plastic. Total expression levels vary dramatically across the population, ranging from basal baselines to hyper-productive states exceeding 5,000 TPM. This surge in metabolic output is not merely a scaling of the housekeeping machinery; rather, it is driven by the recruitment of specific driver genes. While *CsBAHD45* provides the stable metabolic backbone, accounting for approximately 53% of the *CsBAHD* transcript abundance in the trichomes, the distinct high-titer chemotypes are defined by the selective activation of the variable *CSBAHD* candidates. Principal Component Analysis **(Figure 7F)** confirms that these drivers exert opposing forces on the chemotypic landscape: *CsBAHD16* (PC1: +0.79) and *CsBAHD19* (PC1: -0.53) vectors pull in divergent directions. Collectively, these data indicate that the qualitative diversity of *Cannabis* resin is generated not by reinventing the biosynthetic pathway, but by selectively toggling these strain-specific acyltransferases atop the constitutive biosynthetic baseline.

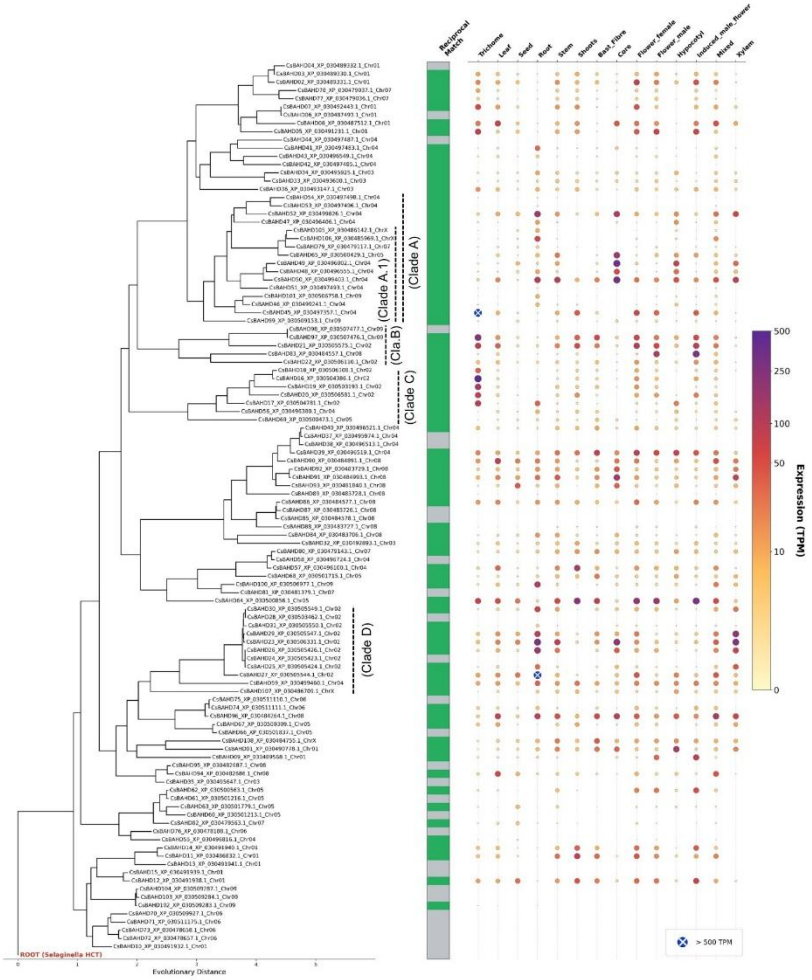

Figure 5. Reciprocal phylogenomic integration and tissue-specific resolution of the *Cannabis BAHD* superfamily. Maximum likelihood phylogenetic reconstruction of the *CsBAHD* repertoire (n = 108), annotated with a concentric heatmap displaying tissue-specific transcript abundance derived from the *Cannabis* Expression Atlas [12]. The tree was inferred using FastTree (JTT+CAT model) from a MAFFT multiple sequence alignment and rooted with the basal *Selaginella moellendorffii* HCT (UniProt D8S308). Orthology between the *cs10* reference

 doi:10.20944/preprints202602.1037.v1

genome and the *Jamaican Lion* anchored expression atlas was established via a Reciprocal Best Hit (RBH) strategy. Markers (green) denote high-confidence orthologs identified by a bidirectional BLASTp search ($E < 1e^{-50}$), ensuring robust functional assignment. Gray nodes indicate candidates where a reciprocal top hit was not found; these are excluded from the expression heatmap to prevent cross-mapping artifacts. For validated orthologs, heatmap display log2-transformed (TPM + 1) transcript abundance ($n = 59$ for trichomes) across vegetative and reproductive tissues. The phylogeny resolves four primary functional clades: A (trichome, root and core), clade B (trichome), clade C (trichome and root), and clade D (root). The comprehensive expression atlas encompasses 394 RNA-seq libraries distributed across 14 distinct tissues. Expression data includes the primary metabolic sinks, leaf ($n = 151$), trichome ($n = 59$), female flower ($n = 34$), root ($n = 32$), and stem ($n = 26$). The dataset is further resolved by specialized developmental and vascular tissues comprising Hypocotyl ($n = 24$), mixed inputs ($n = 19$), male flower ($n = 18$), bast fiber ($n = 12$), seed ($n = 10$), induced male flower ($n = 6$), and single-replicate reference samples for Core, Xylem, and Shoots (each).

3.6. Subterranean and reproductive specialization: The functional differentiation of vegetative modules

Beyond the glandular trichome, the *CsBAHD* superfamily exhibits a distinct modular architecture tailored to the physiological demands of vegetative and reproductive tissues **(Figure 9)**. Quantitative analysis of total transcript abundance identifies the root system as the second-largest metabolic sink for the family, driven primarily by two distinct genomic loci.

The primary root module (clade D; **Figure 5**) is located on chromosome 2, where *CsBAHD27* (644.8 TPM), *CsBAHD23* (289.0 TPM), and *CsBAHD26* (170.4 TPM) form a tight physical cluster. This module likely governs the acylation of suberin monomers, providing the chemical barrier essential for rhizosphere defense. Crucially, clade A, previously characterized by the highly expressed trichome gene *CsBAHD45* and several root dominant genes, reveals a broader function along the vertical axis of the plant. *CsBAHD27* (349.8 TPM), *CsBAHD49* (349.8 TPM), *CsBAHD50* (321.5 TPM), and *CsBAHD48* (59.8 TPM), are preferentially expressed in the lignified stem core while remaining significantly lower in the outer cortex. This localization places the genes in clade A physically and functionally at the intersection of root, core and trichome physiology.

Finally, the reproductive tissues exhibit a phylogenetically distinct metabolic program **(Figure 9)**. Within a larger gene array on chromosome 8, *CsBAHD93* emerges as seed-specific candidate (42.6 TPM). Unlike the vegetative modules which primarily function as O-acyltransferases (forming esters), seed-specific *BAHDs* frequently function as N-acyltransferases. *CsBAHD93* may catalyze the conjugation of hydroxycinnamic acids to polyamines (e.g., spermidine) to form hydroxycinnamic acid amides (HCAAs). Unlike the labile volatile esters of the trichome, these stable amide conjugates cross-link to the cell wall, reinforcing the seed coat against oxidation and physical damage to ensure embryo viability.

 doi:10.20944/preprints202602.1037.v1

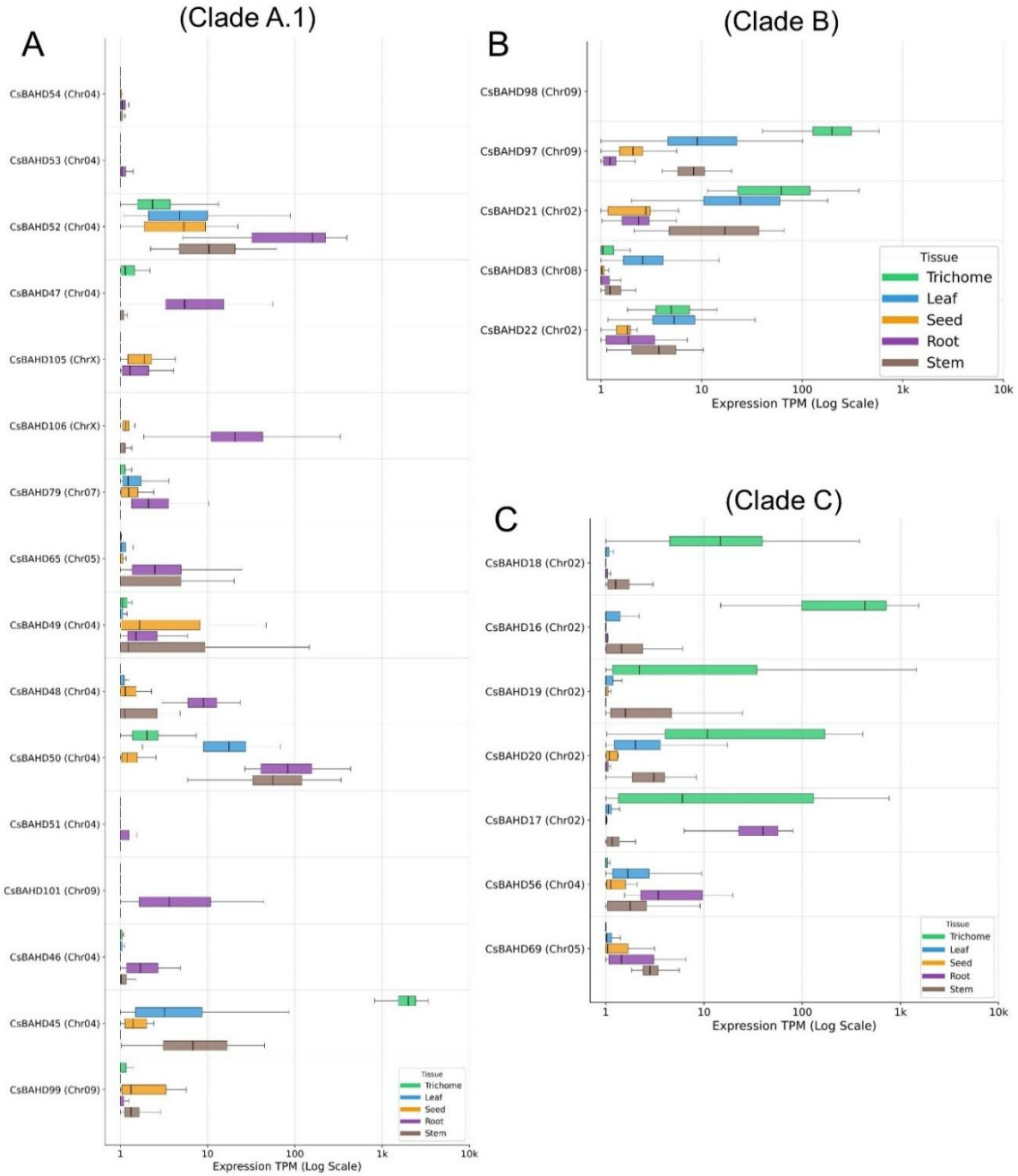

Figure 6. Transcriptional specialization of the primary *CsBAHD* trichome active clades. Comparative expression profiling of the three phylogenetic lineages (clades A, C, and D) across five key tissues. Data is presented as box plots representing transcript abundance $[\log_{10}(TPM + 1)]$. The center line denotes the median, box limits indicate the upper and lower quartiles, and whiskers extend to 1.5x the interquartile range. **(A)** Clade A exhibits a broad expression profile along the plant's vertical axis (root and stem), reflecting its structural role in the lignified core. The expression pattern is skewed by the hyper-expressed trichome outlier *CsBAHD45*. **(B)** In clade B lie two genes with relative high expression in the trichome. **(C)** Clade C displays a highly variable trichome dominant expression pattern.

 doi:10.20944/preprints202602.1037.v1

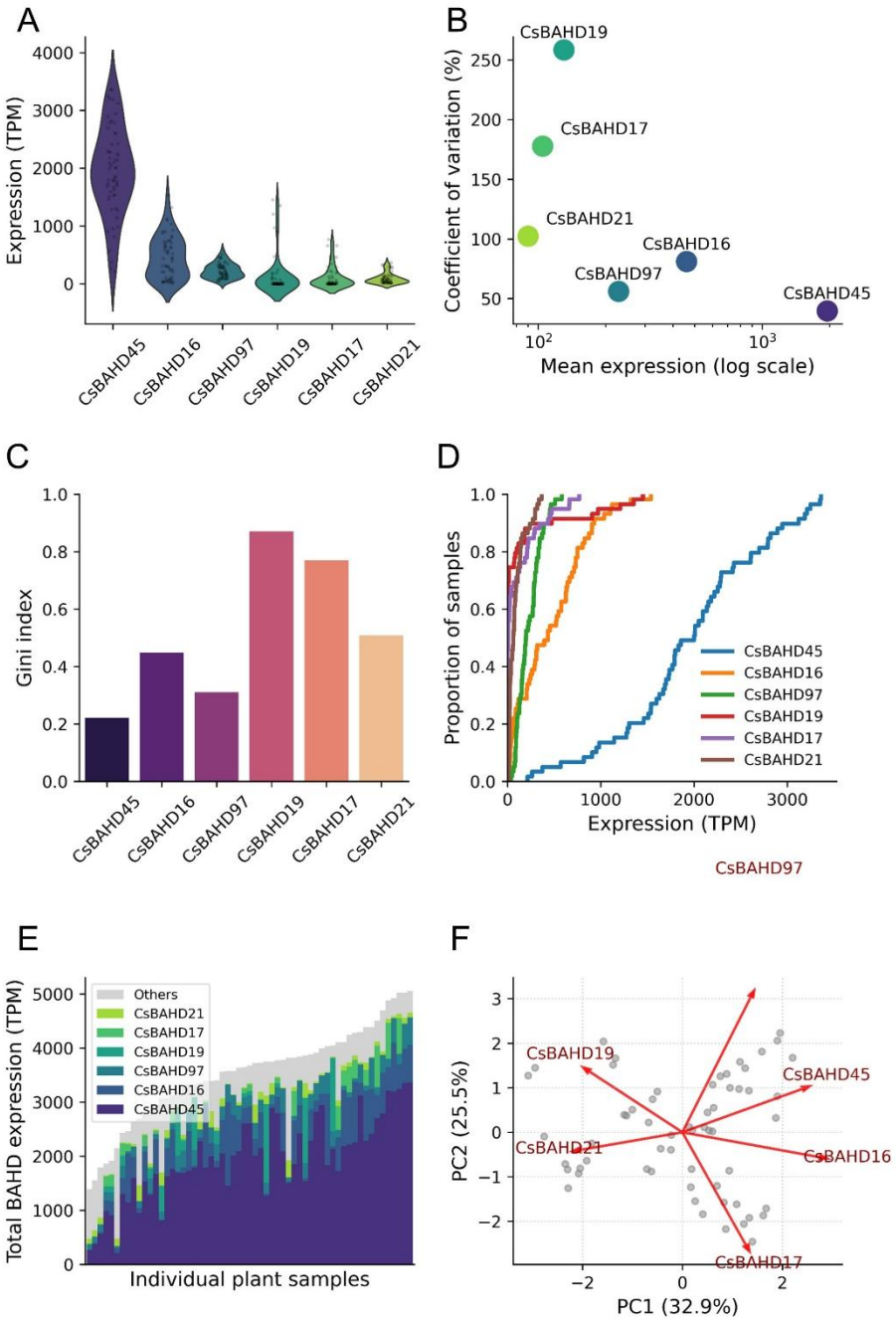

Figure 7. Statistical quantification of transcriptional plasticity and chemotypic diversity in the *CsBAHD* superfamily. Multi-panel statistical analysis of the six *CsBAHD* genes highest expressed in the trichomes. **(A)** Violin plots visualizing the probability density of transcript abundance [$\log_{10}(TPM + 1)$]. The contour represents a kernel density estimate (KDE), which smooths the discrete sampling distribution ($n = 59$; overlaid points) to model the probability of observing specific expression levels. This visualization effectively deconvolutes the population structure: unimodal distributions indicate constitutive expression (e.g., *CsBAHD45*), whereas bimodal distributions reveal distinct sub-populations (presence/absence variations), identifying genes that function as binary chemotypic switches. **(B)** Coefficient of Variation (CV) analysis. Scatter plot correlating mean expression ($\log_{10}$ scale) with the coefficient of variation, calculated as $CV = \frac{\sigma}{\mu} \times 100$. Unlike standard deviation, which scales naturally with transcript abundance, the CV provides a normalized measure of dispersion relative to the mean. This effectively decouples biological noise from absolute expression levels, distinguishing constitutive stability (low CV; e.g., *CsBAHD45*) from chemotypic plasticity (high CV), where expression is driven

 doi:10.20944/preprints202602.1037.v1

by chemotype-specific regulatory switches rather than basal metabolic demands. **(C)** Candidates ranked by statistical dispersion (Gini coefficient), calculated as $G = \frac{\sum_{i=1}^{n}(2i-n-1)x_i}{n\sum_{i=1}^{n}x_i}$, where represents transcript abundance sorted in ascending order and is the sample size. While standard variance metrics (e.g., SD) conflate random noise with biological variation, the Gini coefficient specifically quantifies transcriptional inequality. A coefficient approaching 0 indicates ubiquitous, housekeeping-like expression (perfect equality), whereas a value approaching 1 identifies hyper-specialized alleles (maximal inequality) that are transcriptionally silent in the majority of the population but highly active in specific chemotypes. **(D)** Empirical cumulative distribution function (ECDF) characterizing the transcriptional penetrance of the family. Defined mathematically as , this function plots the proportion of the population failing to reach a specific expression threshold $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \le x}$. The curve's starting height (vertical intercept at $x = 0$) is biologically critical: it quantifies the zero-inflation fraction, representing the exact percentage of the germplasm in which the gene is transcriptionally silent. This metric reveals a fundamental bifurcation in gene regulation. *CsBAHD19* exhibits a high intercept (> 0.75), indicating it is functionally absent in the majority of strains and acts as a conditional on/off switch. In stark contrast, the *CsBAHD45* curve begins at zero and follows a sigmoidal trajectory to the right, confirming its status as a constitutive metabolic enzyme active in nearly all *Cannabis* chemotypes. **(E)** Cumulative transcriptional load of CsBAHDs in the trichomes. Stacked bar chart displaying the total *CsBAHD* transcript abundance per individual plant sample. Samples are ordered along the X-axis by increasing total expression, demonstrating that the total acyl-transferase capacity of the trichome is highly plastic. The top 6 candidates (colored segments) account for the majority of the metabolic flux, validating their prioritization over the remaining gene pool (Others genes in gray). **(F)** The principal component analysis biplot. provides a dimensionality reduction of the z-score standardized expression matrix. Loading vectors (arrows) illustrate the direction and magnitude of each gene's contribution to the principal components. The distinct angles between these vectors demonstrate the orthogonality (statistical independence) of the six major genes, confirming that they drive mutually exclusive chemotypic spaces, as indicated by the divergent trajectories of the *CsBAHD19* and *CsBAHD16* vectors.

 doi:10.20944/preprints202602.1037.v1

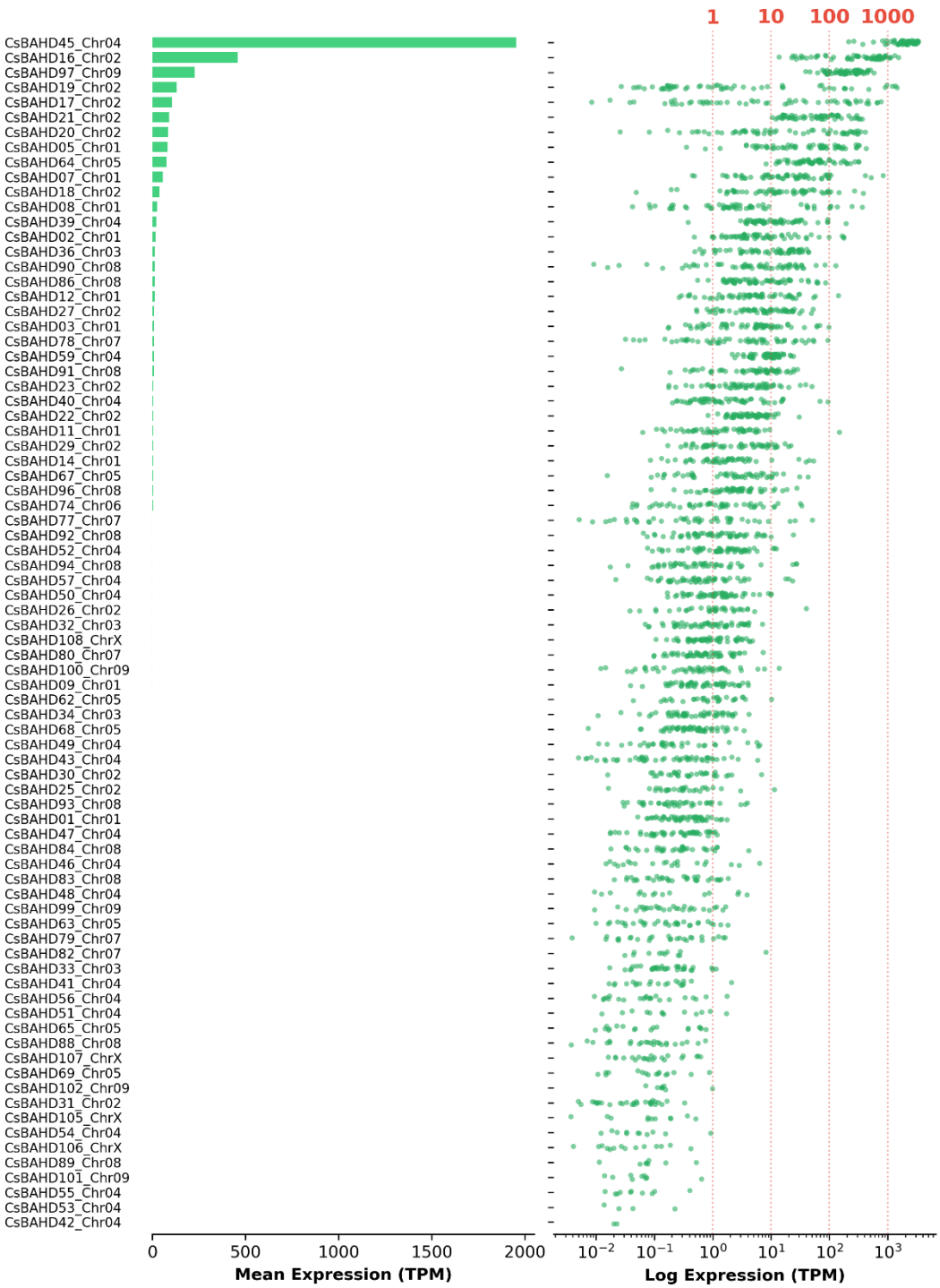

Figure 8. Quantitative expression ranking of the *CsBAHD* superfamily in glandular trichomes. Transcriptional profiling of the complete *CsBAHD* superfamily ($n = 108$) within the trichome transcriptome ($n = 59$). A total of 28 candidates were identified as transcriptionally silent (0 TPM) and excluded. **(A)** The 80 expressed candidates are ranked by mean abundance on a linear scale to visualize absolute expression. The distribution is characterized by the extreme dominance of *CsBAHD45* (1932 ± 773 TPM), which exhibits a transcriptional load exceeding the cumulative expression of all other family members combined. **(B)** The same dataset is plotted on a logarithmic scale ($\log_{10}$) to resolve the expression dynamics of lower-abundance candidates.

 doi:10.20944/preprints202602.1037.v1

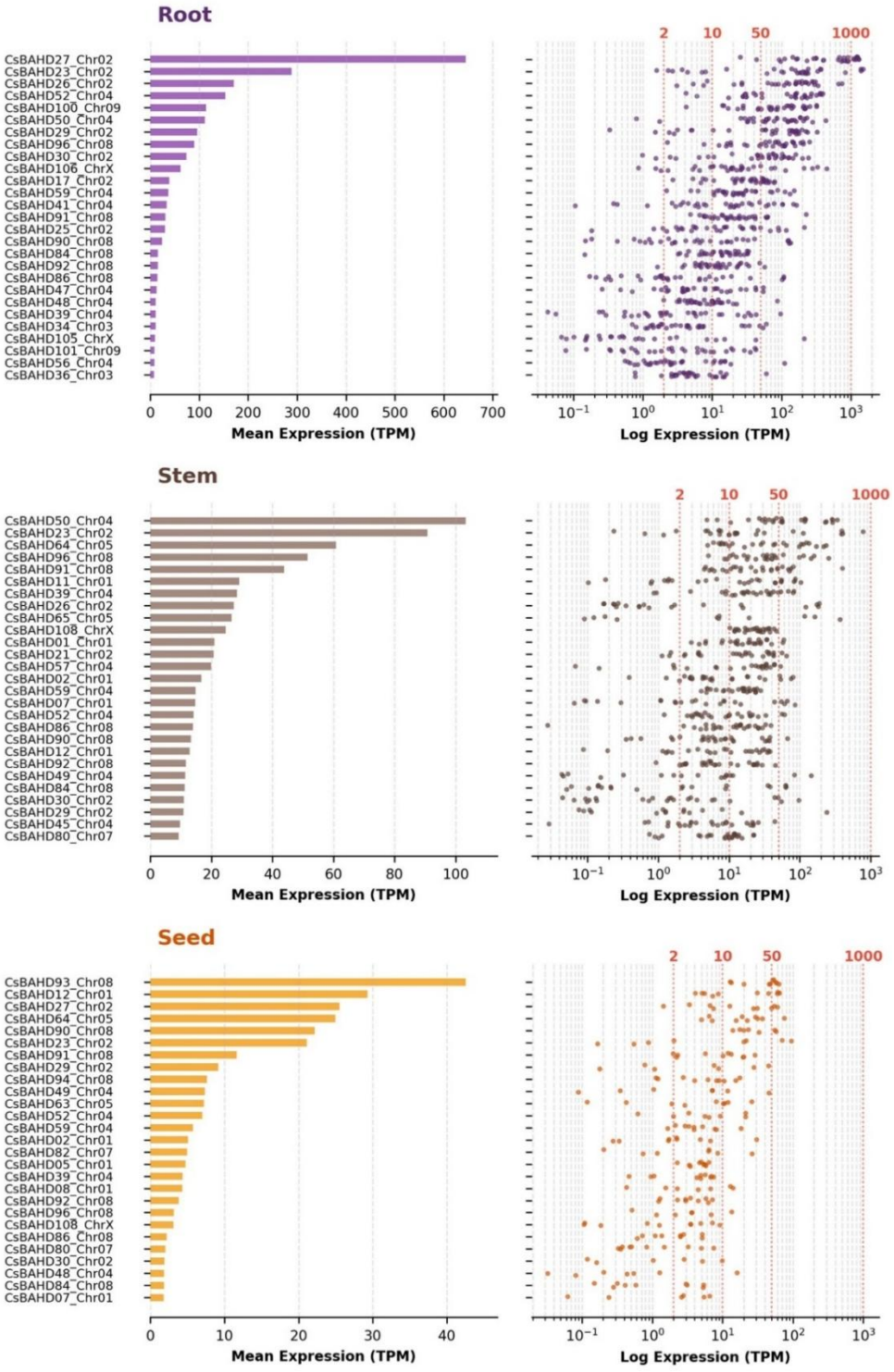

Figure 9. Tissue-specific transcriptional hierarchies in vegetative and reproductive sinks. Comparative expression ranking of the active *CsBAHD* repertoire across root ($n = 32$), stem ($n = 26$), and seed ($n = 10$) transcriptomes. Data is presented as paired linear (left) and logarithmic (right). In contrast to the trichome's singular singular enzyme dominated transcript profile, these vegetative and reproductive tissues display more distributed transcriptional profiles, reflecting multi-genic regulation of putative rhizosphere defense (*CsBAHD27*, root), lignification (*CsBAHD50*, stem), and seed development (*CsBAHD93*, seed).

 doi:10.20944/preprints202602.1037.v1

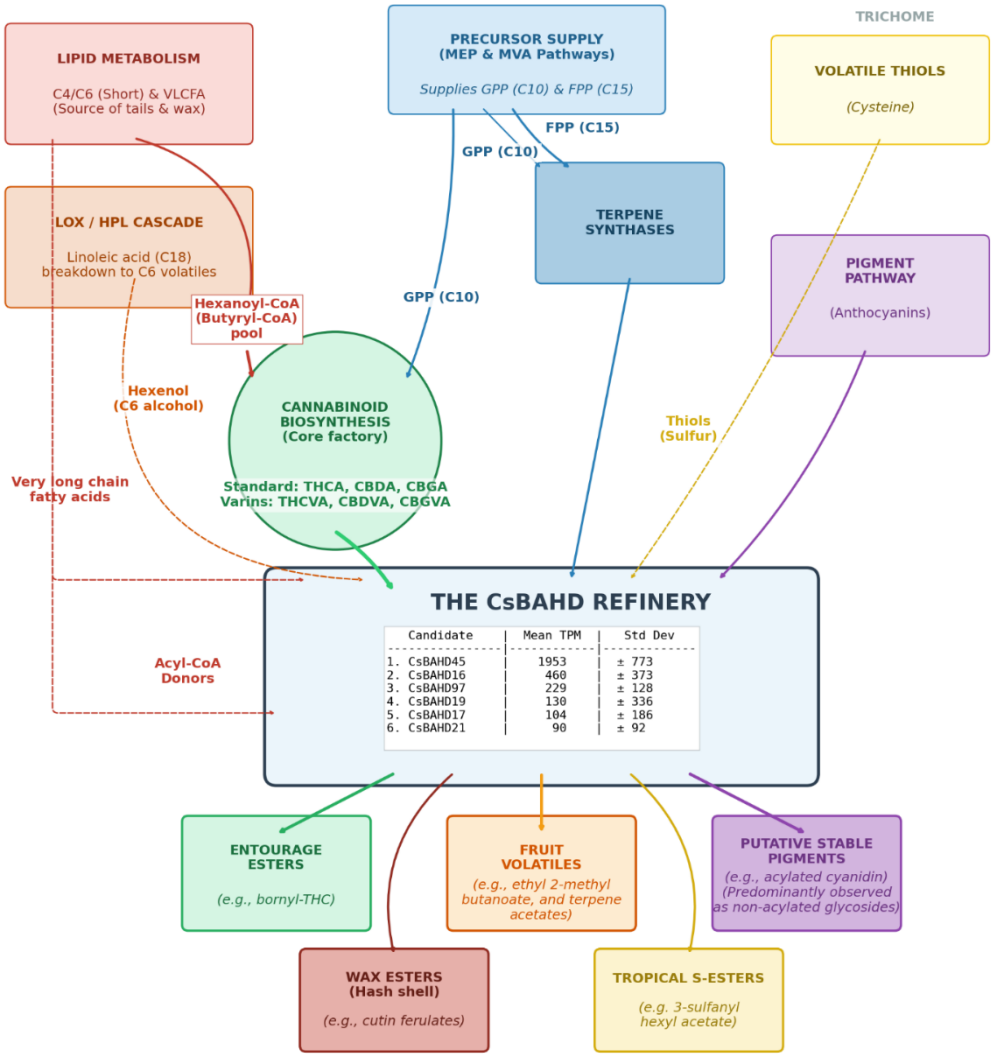

Figure 10. Hypothetical model of *CsBAHD* functional integration in the glandular trichome. Schematic representation of the proposed acyl-transferase network interfacing with lipid, terpene, and cannabinoid biosynthesis. The model posits that the *CsBAHD* superfamily contributes to chemical diversity through four primary acylation modules, inferred from phylogenetic homology and the tissue-specific context in *Cannabis*. **(i)** Structural wax esters: A maintenance role supporting the trichome cuticle, phylogenetically inferred from conserved cutin synthases. May play a role in trichome tensile strength, important for hashish and storage in general. **(ii)** Volatile (sulfanyl) esters: These add floral, fruity and exotic notes to the flower. For example the confirmed metabolite 3-sulfanylhexyl acetate imparting passionfruit notes. This study proposes that this compound arises from the acylation of thiol-bearing alcohols (e.g., 3-sulfanylhexan-1-ol), where the enzyme targets the hydroxyl moiety while preserving the sulfur group. This metabolomic evidence supports a broader acylation module that likely extends to the acetylation of terpene alcohols (e.g., geranyl acetate). This biochemical transformation is directly analogous to the volatile biosynthesis found in floral tissues such as *Rosa hybrid*, where the BAHD enzyme *RhAAT1* catalyzes the acetylation of geraniol to produce the characteristic rose aroma [13]. In the context of *Cannabis*, the recruitment of such a pathway would fundamentally alter the resin's aroma profile, softening the sharp notes of free terpene alcohols into their more complex, floral-fruity ester conjugates. Crucially, the formation of these volatile conjugates requires alcohol acyltransferase (AAT) activity, a canonical function of the *BAHD* superfamily. Phylogenetic analysis identifies specific *CsBAHD* candidates clustering with known AATs from fruit systems, supporting the hypothesis that these genes serve as the enzymatic drivers of the resin's fruity esterified aroma profile. **(iii)** Cannabinoid esters: A hypothetical biosynthetic route for identified esterified cannabinoids (e.g., fenchyl-THC), linking CsBAHD activity to the expansion of the chemotype specific entourage effects. **(iv)** Acylated pigments: A hypothetical mechanism for anthocyanin stabilization. A theoretical module for the acylation of flavonoid pigments. Although *Cannabis*

 doi:10.20944/preprints202602.1037.v1

anthocyanins are predominately glycosylated, this pathway suggests a latent capacity for aromatic acylation, a mechanism utilized by homologous BAHDs to enhance chromatic stability [8]. *Figure note:* While most of these metabolite classes have been chemically characterizied in *Cannabis*, their direct enzymatic synthesis by specific *CsBAHD* enzymes represents a theoretical framework for future biochemical elucidation.

4. Discussion

4.1. Genomic landscape and telomeric plasticity for accelerated precision breeding

The identification of 108 *CsBAHD* genes positions *Cannabis sativa* squarely within the typical expansion range of dicotyledonous specialized metabolism, comparable to *Fragaria vesca* (89 members) [14], *Prunus avium* (sweet cherry) (125 members) [14], *Piper nigrum L.* (110 members) [15], and *Lavandula* (166 members) [9].

Physical mapping reveals a non-uniform distribution characterized by dense telomeric gene clusters. This topological organization presents a double-edged sword for breeding programs. While these dynamic regions facilitate rapid chemical diversification, they are also hotspots for chromosomal recombination and hemizygosity. Recent pangenomic analyses demonstrate that while the canonical cannabinoid synthase loci are genetically conserved, the broader specialized metabolic architecture is characterized by extensive structural variation [16].

The telomeric *CsBAHD* gene clusters identified here likely reside within these hyper-variable accessory regions, subjecting them to significant copy number variation (CNV) and presence/absence variation (PAV) across the germplasm. This genomic instability provides a molecular mechanism for the elusive nature of exotic fruity traits and their tendency to segregate unpredictably during hybridization. Consequently, stabilizing these chemotypes necessitates moving beyond phenotypic selection to precision marker-assisted strategies that specifically target these dynamic telomeric arrays to mitigate linkage drag. Defining the global *CsBAHD* repertoire through future pangenomic interrogation is now critical to mining the rare, cultivar-specific alleles responsible for these high-value aroma profiles, a challenge that demands scalable, high-throughput genomic tools.

To this end, the characterisation of the *CsBAHD* superfamily is a prerequisite that now enables the high-resolution mining of pangenomic datasets for rare aroma alleles. Beyond this immediate application, the computational workflow established here, anchored by curated HMM profiling and transcriptomic filtering, demonstrates the efficacy of a purely *in silico* strategy utilizing public datasets. This scalable framework provides a robust blueprint for de-orphaning other agronomically critical gene families in *Cannabis sativa* and accelerates the transition from traditional phenotypic selection to precision molecular breeding.

4.2. Functional specialization in vegetative and reproductive sinks: lignification, rhizosphere defense, and seed maturation as evolutionary scaffolds

In the root system, the transcriptional landscape is defined by the high-magnitude expression of *CsBAHD27* (644.8 TPM) and *CsBAHD23* (289.0 TPM). The extreme specificity of *CsBAHD27* (Specificity Ratio $\approx$ 13) suggests a dedicated function in the subterranean interface. Given the well-established role of BAHD acyltransferases in the modification of root exudates (e.g., triterpenes, suberin components) in other species [17], it is proposed that one on the highly expressed root *CsBAHDs* mediates the acylation of defensive phytochemicals secreted into the rhizosphere. These acyl modifications presumably potentiate the bioactivity and chemical stability of antimicrobial compounds, thereby establishing a chemical barrier against soil-borne pathogens while simultaneously modulating the rhizosphere community. Consequently, the specific role of these modifications warrants critical investigation within the context of cannabis–pathogen interactions.

Expression in the stem is characterized by a moderate but consistent profile dominated by *CsBAHD50* (103.3 TPM). Phylogenetic analysis places this candidate within clade A together with CsBAHD45 which is highly expressed in the trichomes. The differential expression of specific *CsBAHDs* within the stem suggests a mechanism for tuning cell wall properties. In model systems

 doi:10.20944/preprints202602.1037.v1

such as *Populus* [18] and *Oryza* [19], BAHD acyltransferases catalyze the esterification of p-coumarate to monolignols prior to their incorporation into the lignin polymer. This modification introduces ester linkages into the lignocellulosic matrix which significantly impacts cell wall digestibility and fiber processing efficiency [20]. Consequently, the presence or absence of these specialized acyltransferases in *Cannabis* could serve as a molecular determinant for fiber quality, potentially differentiating and optimizing cultivars suited for fine textile applications from those optimized for structural biomass.

The members of Clade A constitute a phyletic lineage, with its principal expansion mapping to a dense tandem duplication arrays on chromosome 2 and 4. The majority of genes within this array, such as *CsBAHD50* and *CsBAHD52*, retain ancestral expression patterns localized to the vegetative leaf, stem, core and roots. The emergence of *CsBAHD45* as a dominant trichome transcript within this otherwise vegetative cluster points to a mechanism of neofunctionalization. Rather than originating from a distinct floral-specific clade, high-titer resin production appears to have evolved via the regulatory divergence of a vegetative paralog, effectively shifting its transcriptional domain to the glandular trichome, which would be an interesting area for future research. This trajectory mirrors Clade C, where the trichome-specific drivers (*CsBAHD16–CsBAHD20*) form a physical cluster on Chromosome 2 alongside their phylogenetic homologs, the root-specific *CsBAHD17 and CsBAHD56*. Collectively, these genomic and transcriptomic patterns indicate that the fruity chemotypic traits of modern *Cannabis* may have evolved through the recruitment of ancestral core-fiber and subterranean acyltransferases into the aerial floral sink.

In the reproductive sink, a smaller set of *CsBAHD* were expression at functional levels. Among these are *CsBAHD93* and *CsBAHD12* as the primary seed-resident transcripts. The stable expression of these genes suggests a role in the chemical maturation of the seed coat or embryo. Potential functions include the conjugation of polyamines (e.g., spermidine conjugates), which are known BAHD substrates critical for regulating seed dormancy, longevity, and defense against predation during the vulnerable germination phase [21]. The identification of these seed-specific markers provides new targets for investigating the metabolic regulation of seed vigor in *Cannabis*.

4.3. Transcriptional architecture of the glandular trichome: biosynthesis of fruit esters and the modulation of terpene profiles

In the context of the trichome one candidate stands out as major *BAHD* transcript, *CsBAHD45* (1932 ± 773 TPM). Furthermore, a relatively small subset of trichome specific *BAHDs* were found identified yet found to be high variable expressed. Beyond these primary drivers, the trichome transcriptome reveals a large set of functionaly expressed genes (20 genes with mean > 10 TPM), including *CsBAHD16* (460.0 TPM) and *CsBAHD97* (228.8 TPM). These candidates maintain a more robust, consistent expression across the trichome samples ($n = 59$), distinguishing them from the hyper-variable trichome specific *CsBAHD19 and CsBAHD17*. This provides a molecular explanation for the elusive nature of tropical fruity chemotypes found in specific chemovars.

4.3.1. Biosynthesis of short-chain fruit esters for floral, sweet and exotic aroma

The superfamily here identified is likely responsible for the production of many volatile organic compounds (VOCs), such as methyl anthranilate and 3-sulfanylhexyl acetate which impart fruity notes such as grape-like and floral or passionfruit respectively [1]. While metabolomic surveys have identified over 30 distinct esters, including methyl and dimethyl anthranilates (grape), ethyl hexanoate (green apple), and n-propyl hexanoate (blackberry) [3], targeted analytics across diverse cannabis pool may reveal and even more diverse ester volatilome. Notably, recent research has identified phenethyl alcohol and its esters (e.g., phenethyl acetate and phenethyl butyrate) as the primary drivers of the 'honey-like', floral, and sweet nuances.

The *CsBAHD* genes described in this study, particularly those highly expressed in the trichomes, likely facilitate the biosynthesis of these fruity, exotic and floral odorants **(Figure 10)**. Based on the expression profile being dominated by *CsBAHD45* (1932 TPM), it is likely this enzyme functions as the promiscuous alcohol acyltransferase (AAT), analogous to SAAT in *Fragaria* (strawberry), which

 doi:10.20944/preprints202602.1037.v1

accepts a broad range of substrates [22]. These enzymes catalyze the esterification of short-chain alcohols to produce volatile conjugates such as isoamyl acetate (banana), ethyl 2-methylbutyrate (apple), and ethyl hexanoate (pineapple). Furthermore, this promiscuous activity likely extends to sulfur-containing precursors, mediating the acetylation required to synthesize 3-sulfanylhexyl acetate (passionfruit). Consequently, *CsBAHD45* could be driving the fruity floral and tropical-sulfur notes in *Cannabis*, yet could still be dependent on the available substrate pool which is not addressed in this study, but requires future investigations.

4.3.2. Acetylation of the terpene pool to create soft delicate floral notes

The *CsBAHD* superfamily likely retains its ancestral role in the modification of the volatile terpene pool. As described in fundamental fragrance chemistry, the enzymatic acetylation of terpene alcohols alters the human sensory experience via polar capping of the hydroxyl group. This structural modification consistently shifts the odor profile from sharp, herbal, or earthy to sweeter, fruitier, or more polished notes, depending on the molecular weight. Comparative biochemical analysis with *Lavandula angustifolia* (lavender) provides a compelling model, where specific BAHD acyltransferases (e.g., *LaAAT*) catalyze the acetylation of linalool to linalyl acetate, a conversion that is critical for the plant's characteristic floral aroma profile [9]. *Cannabis* resin is abundantly rich in analogous terpenoid alcohols, specifically linalool, fenchol, geraniol, and α-terpineol, which serve as direct substrates for this class of enzymes.

High-resolution metabolomic profiling of *Cannabis* dried inflorescences has definitively confirmed the presence of the corresponding esters, including fenchyl acetate and bornyl acetate, in the volatile fraction [2]. Although these conjugates typically appear in lower stoichiometric abundance than their alcohol precursors, their sensory impact is potent. For example, the acetylation of geraniol to geranyl acetate introduces a distinct rose-like, fruity nuance [13]. Similarly, the capping of heavier sesquiterpene alcohols removes earthy or damp soil aspects, resulting in a more transparent and refined woody profile. This chemical distinction supports recent findings that the "Exotic Score" of modern cultivars is driven not by dominant terpenes, but by the abundance of minor, non-terpenoid volatiles [3].

Beyond sensory modulation, this enzymatic derivatization may impact the pharmacological potential of the resin via the entourage effect. While sesquiterpene hydrocarbons such as β-caryophyllene are established ligands for the cannabinoid receptor 2 (CB2) [23], the acetylation of such sesquiterpene alcohols significantly alters their lipophilicity and bioavailability. This chemical modification may modulate the molecule's ability to cross physiological barriers or interact with lipid-bound receptors. From an odorant perspective this acetylation affects the interaction between the modified terpenes and the sensory receptors. Therefore, the *CsBAHD* superfamily functions as the essential flavor refinement machinery, overlaying a complex, high-value ester profile onto the generic terpene background.

4.3.3. Structural modeling and physicochemical sequestration of volatile terpenes

A more speculative yet biochemically significant frontier involves the potential for CsBAHD enzymes to catalyze the formation of terpene-cannabinoid esters. While early chromatographic evidence suggests the existence of these metabolites [24], their biosynthetic origin remains unresolved. A two-step biosynthetic route is proposed that necessitates the coordination of the *CsBAHD* superfamily with an upstream, as-yet-unidentified activating enzyme. The carboxylic acid moiety of the cannabinoid precursor (e.g., THCA) is thermodynamically inert toward esterification in aqueous cellular conditions. Therefore, the pathway must initiate with the ATP-dependent activation of the cannabinoid acid by a putative cannabinoyl-CoA synthetase (Enz_X), analogous to the 4-coumarate:CoA ligases (4CL) of phenylpropanoid metabolism:

$$Cannabinoid_{acid} + CoASH + ATP \xrightarrow{Enz_X} Cannabinoyl\text{-}S\text{-}CoA + AMP + PP_i$$

 doi:10.20944/preprints202602.1037.v1

Following activation, it is hypothesized that a specialized *CsBAHD* isoform catalyzes the transfer of the bulky cannabinoid moiety onto a terpenoid alcohol acceptor. Consistent with the conserved catalytic logic of the *BAHD* superfamily, this reaction is predicted to proceed via an ordered Bi-Bi kinetic mechanism. The cannabinoyl-S-CoA donor binds first, anchoring its CoA moiety to the conserved DFGWG (or DXXXG) structural motif located at the periphery of the active site tunnel. This initial anchoring serves as a molecular pivot; it secures the coenzyme while allowing the bulky cannabinoid tail to remain transiently mobile at the solvent interface. It is proposed that the dynamic flexibility of the enzyme's surface loops is then critical to accommodate the large substrate, permitting the cannabinoid moiety to rotate and navigate the steric constriction of the tunnel entry. This conformational sampling eventually guides the reactive thioester bond into precise alignment with the catalytic HXXXD motif deep within the pocket. The reaction cycle concludes with the entry of the alcohol acceptor (e.g., fenchol, guaiol) through the opposing solvent channel, facilitating a histidine-mediated nucleophilic attack on the carbonyl carbon of the thioester to yield the final ester.

$$Cannabinoyl\text{-}S\text{-}CoA + sesquiterpene\text{-}OH \xrightarrow{CsBAHD} Cannabinoid\text{-}O\text{-}sesquiterpene + CoASH$$

The feasibility of this reaction rests entirely on the biophysical plasticity of the active site and surrounding domains. Preliminary rigid-body docking simulations indicated severe steric incompatibility when attempting to position the bulky, tricyclic cannabinoid core within the canonical acyl-donor tunnel. Specifically, the rigid docking protocol resulted in unresolvable steric clashes where the cannabinoid moiety of the acyl donor substrate is sandwiched between the central β-sheet core and the α-helix immediately adjacent to the catalytic HXXXD motif. This docking trajectory was intended to recapitulate the binding pose observed in the crystal structure of *Arabidopsis thaliana* HCT in complex with p-coumaroyl-CoA [25]. However, this exploratory modeling was necessarily limited to truncated cannabinoyl-CoA constructs within a static enzyme conformation. The computational resources required to simulate the full flexibility of the long CoA backbone, combined with the extensive conformational sampling needed to capture the transient opening of the active site for such a massive thioester complex, were beyond the capacity of the current analysis.

Crucially, this observation supports a dynamic binding model consistent with X-ray crystal structures of related BAHD enzymes, which have experimentally resolved the p-coumaroyl-CoA binary complex [25]. These crystallographic snapshots confirm that the coenzyme A moiety binds first, acting as a fixed anchor deep within the donor tunnel, while the bulky acyl tail remains mobile and disordered at the solvent and active site interface. Therefore, it is proposed that CsBAHDs must undergo a specific breathing motion, temporarily opening the active site cleft via surface loop displacement, to allow this mobile cannabinoyl moiety to rotate and lock into position against the HXXXD catalytic motif. Such active site flexibility has been explicitly identified as a determinant of substrate permissiveness in other clade members [25], indicating that future validation must utilize high-performance molecular dynamics (MD) simulations to capture this transient opening event, which static docking cannot resolve. Ultimately enzyme assays would be the most clear and direct way to address this question.

The biochemical capacity of the BAHD superfamily to accommodate massive, complex scaffolds is well-established, although typically observed in the alcohol acceptor binding pocket. For instance, in *Taxus* species, enzymes like taxadien-5α-ol O-acetyltransferase (TAT) have evolved expansive active site cavities to acylate the bulky, tricyclic taxane core. While these classic examples involve large acceptors, the superfamily also exhibits significant plasticity in the donor tunnel. The hydroxycinnamoyl-CoA transferase (HCT) subfamily, the reference used in this docking, naturally accommodates aromatic donors (e.g., p-coumaroyl-CoA) rather than simple acetyl groups [25].

 doi:10.20944/preprints202602.1037.v1

4.3.4. The evolutionary enigma of cannabinoid biosynthesis

The evolutionary forces driving the massive accumulation of secondary metabolites in the glandular trichomes of *Cannabis sativa* have long remained a subject of debate. To date, three primary hypotheses have dominated the literature, each attempting to explain the physiological utility of cannabinoids and terpenes through the lens of abiotic stress or general toxicity.

The first, the UV-B radiation shield hypothesis, posits that cannabinoid production is an adaptation to the high-altitude, high-irradiance environments of the Central Asian steppe where the species originated [26,27]. As phenolic compounds, cannabinoids (specifically THCA and CBDA) possess chromophores capable of absorbing harmful ultraviolet radiation, theoretically functioning as a chemical sunscreen to protect the developing embryo and genomic integrity of the seed. While plausible, this model fails to account for the substantial metabolic investment in the volatile terpene fraction, which accumulates in the same cellular compartment yet offers negligible photoprotection. Terpenes lack the conjugated -electron systems required for UV absorption; if the resin's primary function were photoprotection, the co-accumulation of these energetically costly non-absorbing volatiles would represent a significant evolutionary inefficiency. Furthermore, superior UV-screening compounds, such as flavonoids, are synthesized ubiquitously in the leaf epidermis at a fraction of the metabolic cost, rendering the specialized trichome machinery redundant for this specific purpose.

The second, the desiccation tolerance hypothesis, suggests that the resinous exudate functions primarily as a physical sealant [28]. In this model, the massive accumulation of lipophilic compounds creates a hydrophobic barrier that prevents cuticular transpiration, ensuring floral hydration during the arid autumn maturation phase. However, this hypothesis contradicts the fundamental morphology of the glandular trichome. In *Cannabis*, the resin is elevated on a multicellular stalk (capitate-stalked trichomes), effectively lifting the lipophilic payload away from the epidermal surface rather than forming a cohesive, sealing film. A true anti-desiccant strategy typically involves the thickening of the planar cuticle with long-chain waxes, not the secretion of discrete, fragile globules. Additionally, a significant fraction of the resin consists of mono- and sesquiterpenes, which act as solvents and increase the mixture's fluidity and vapor pressure, properties that are antithetical to the formation of a stable, occlusive seal.

Finally, the general toxicity hypothesis argues that the resin serves as a broad-spectrum chemical defense [29]. Free cannabinoids are known to be cytotoxic to insect larvae and exhibit antimicrobial activity against Gram-positive bacteria. Similarly, monoterpenes such as -pinene and limonene are potent insecticides and antifungals. While this theory correctly identifies the defensive potential of these compounds, it fails to address the physicochemical paradox of their co-accumulation. Monoterpenes are highly volatile; upon trichome rupture, they undergo rapid phase transition rapid evaporation, dissipating into the atmosphere rather than persisting on the attacker. Without a mechanism to fix these volatiles to the herbivore's cuticle, their contact toxicity is fleeting. Existing models provide no biochemical explanation for how the plant harnesses the synergy between the volatile terpenes and the non-volatile cannabinoids, leaving the functional relationship between these two massive metabolic sinks unresolved.

4.3.5. Physicochemical sequestration of volatile terpenes: esters as a mechanism for contact toxicity

The identification of the cannabinoid esters [24] and the CsBAHD superfamily, facilitates a novel evolutionary hypothesis: the physicochemical sequestration hypothesis (Figure 11). This study postulates that the ancestral function of cannabinoids was not merely to serve as free bioactive agents, but rather as non-volatile bioactive scaffolds for the stabilization of potent antimicrobial volatile defensive terpenes.

Monoterpenes act as potent contact toxins, yet their efficacy is severely compromised by their high vapor pressure. While the glandular trichome cuticle provides physical containment during storage, this integrity is lost immediately upon rupture by herbivory or mechanical abrasion. In the absence of a chemical fixative, the volatile payload undergoes rapid phase transition, dissipating into

 doi:10.20944/preprints202602.1037.v1

the atmosphere as a transient fumigant rather than persisting on the herbivore. Consequently, the massive metabolic investment in terpene synthesis yields diminishing returns in the absence of a stabilization mechanism.

Consequently, the *CsBAHD* superfamily may have evolved to mitigate this volatility, catalyzing the esterification of volatile terpenes to the lipophilic cannabinoid core (e.g., forming fenchyl-cannabinolate). This conjugation dramatically lowers the vapor pressure of the defensive payload. Upon trichome rupture, the esterified defense complex does not volatilize and instead functions as a viscous bioactive occlusive agent. This adherent resin coats the herbivore's mouthparts and spiracles, ensuring that the toxic terpene moiety persists long enough to penetrate the chitinous cuticle. Thus, the cannabinoid-ester linkage represents an evolutionary optimization for persistence, transforming a fleeting gas into a durable contact toxin.

Critically, this model addresses the energetic paradox of utilizing complex secondary metabolites as scaffolds. While esterification to simple fatty acids could theoretically achieve volatility reduction, the utilization of a cannabinoid anchor creates a synergistic bipartite defense system. Previous biochemical characterizations indicate that cannabinoid esters exhibit significant antimicrobial potency [24]. The ester linkage therefore serves a dual function: physicochemical stabilization and enhanced lipophilicity, facilitating the transport of the toxic terpene moiety across the lipophilic insect cuticle or fungal membrane. Support for this evolutionary strategy necessitates comparative entomological and antimicrobial bioassays. Future investigations could quantify the toxicity of these purified esters against specific herbivore and pathogen models, contrasting their efficacy with non-esterified cannabinoid-terpene mixtures. Such mechanistic studies would determine whether the conjugates function as stable, lipophilic contact toxins that penetrate the cuticle or fungal cell wall, or as metabolic pro-drugs that release their volatile payload via hydrolytic cleavage within the alkaline environment of the insect gut or microbial cytoplasm.

4.3.6. Chemotypic divergence trough domestication and the uncoupling of ancestral defense pathways

The genomic architecture of the *CsBAHD* superfamily, characterized by a prevalence of orphan genes and latent switches, provides the molecular footprint of a metabolic domestication event. Anthropogenic selection has historically imposed intense directional pressure on two specific traits: psychotropic potency (dependent on free, unesterified THC) and olfactory intensity (dependent on high-vapor-pressure terpenes and small volatile esters).

 doi:10.20944/preprints202602.1037.v1

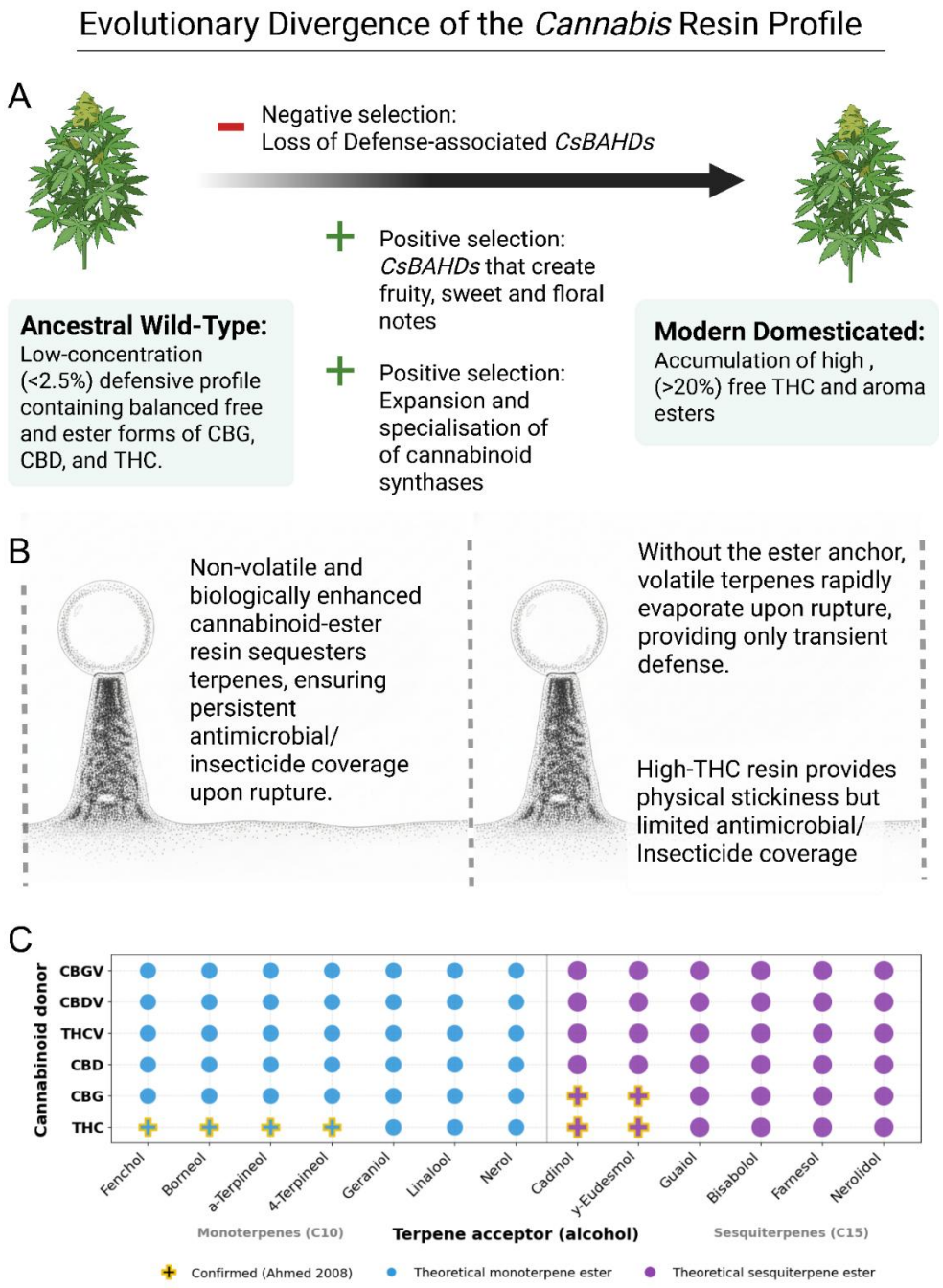

Figure 11. The physicochemical sequestration hypothesis and the anthropogenic selection of *Cannabis sativa*. **(A)** Evolutionary divergence of the resin profile. Schematic model of the proposed trajectory from a defense-oriented wild ancestor to a specialized, high-potency modern chemovar. Anthropogenic domestication likely involved the simultaneous upregulation of upstream cannabinoid synthases and the selective negative selection of specific cannabinoid ancestral ester deference associated *CsBAHD* acyltransferases responsible for creating non-volatile cannabinoid esters. In parallel, human selection preserved and enhanced those *CsBAHDs* responsible for volatile aroma esters. This selective pressure effectively uncoupled the ancestral defense pathway to maximize the accumulation of free psychoactive cannabinoids. **(B)** Physicochemical Mechanism: Contact Retention vs. Rapid Volatilization. Thermodynamic consequences of the uncoupled esterification pathway upon trichome rupture. In the ancestral phenotype (left), cannabinoid biosynthesis is driven by promiscuous, low-fidelity synthases [30], which we model here as operating within a low-titer (<2.5%) physiological context. In this energetically constrained system, cannabinoids function as non-volatile molecular anchors, sequestering highly volatile antimicrobial monoterpenes into stable, lipophilic esters. This linkage dramatically reduces vapor pressure, ensuring the cytotoxic payload persists on the target organism (insect cuticle or microbial cell wall) to

 doi:10.20944/preprints202602.1037.v1

exert prolonged contact toxicity. In the modern phenotype (right), the domestication-driven selection for high-efficiency synthases, coupled with the loss of the esterification valve, results in the rapid volatilization of defensive terpenes immediately upon rupture, yielding a resin rich in free cannabinoids but depleted of its synergistic permeabilizers. **(C)** Combinatorial matrix of potential cannabinoid esters. Gold crosses indicate conjugates chemically validated by Ahmed [24]. Blue (monoterpene) and purple (sesquiterpene) markers represent the latent chemical space, a theoretical library of metabolites derived from the combinatorial coupling of terpene alcohols with both canonical pentyl (e.g., THC, CBD) and propyl varin (e.g., THCV, CBDV) scaffolds. This matrix illustrates the full structural potential of the *CsBAHD* superfamily assuming substrate promiscuity. These compounds, which may remain undetected in modern germplasm due to absence, low abundance, or the lack of targeted screening standards, represent defined molecular targets for future synthetic reconstruction to investigate their broad biological potential, ranging from ecological defense against insects and microorganisms to novel therapeutic applications in human pharmacology.

The expansion of this aroma profile presents a fundamental paradox for an anemophilous (wind-pollinated) species. Unlike entomophilous crops that require complex floral scents to attract pollinators, *Cannabis* has no direct reproductive requirement for such olfactory signaling. While it is theoretically possible that these volatiles function as synomones to recruit beneficial predatory insects in a complex wild ecosystem, this ecological explanation is likely secondary. It is far more plausible that the floral, fruity and exotic profiles of modern germplasm are not adaptations to the environment, but artifacts of intense human selection for sensory pleasure.

This interpretation is reinforced by recent phylogenomic reconstructions, which demonstrate that the highly specialized THC and CBD synthases found in modern germplasm evolved from a single promiscuous ancestral enzyme [30]. Consequently, the evolution of the modern chemotype was driven by two parallel genomic events: the neofunctionalization of this inefficient generalist into high-fidelity specialists and a gene copy number expansion to increase metabolic flux.

We posit that this supply-side optimization necessitated a critical trade-off regarding the putative ancestral *CsBAHD* esterification defense pathway. The retention of the ester moiety (e.g., in cannabinoid-terpene conjugates) would have effectively lowered the yield of the free cannabinoid fraction. Consequently, domestication may have exerted negative selection against the functional *CsBAHD* enzymes responsible for this coupling. Crucially, however, domestication resulted in functional divergence rather than the total erasure of the superfamily. While the defense-associated sequestration modules were likely suppressed, a distinct subset of *CsBAHD* enzymes responsible for aroma esters appears to have been maintained or enhanced via positive selection.

Data Availability: The raw genomic data for *Cannabis sativa* cultivar cs10 (CBDRx) analyzed in this study are openly available in the NCBI RefSeq database (Assembly Accession: GCF_900626175.2). The transcriptomic datasets utilized to construct the expression atlas were derived from the *Cannabis* Expression Atlas [12] and which is publicly available. Supplementary File 1 provides the comprehensive master matrix containing the 108 identified *CsBAHD* genes, their systematic identifiers, original NCBI accession numbers, and the integrated tissue-specific expression values (TPM) used for phylotranscriptomic analysis. The physical chromosomal coordinates, strand orientation, and gene cluster assignments for all candidates are detailed in Supplementary File 2. The full-length amino acid sequences for the 108 *CsBAHD* candidates and the 122 functionally characterized reference plant BAHDs used for phylogenetic inference are provided in Supplementary File 3 and Supplementary File 4, respectively. The raw phylogenetic tree file is available as Supplementary File 5 (Newick format). The custom hidden Markov model (HMM) profile used for candidate identification is available from the author upon reasonable request.

Conflicts of Interest: The author declares no conflicts of interest.

Funding: This research received no external funding.

 doi:10.20944/preprints202602.1037.v1

Abbreviations

4CL, 4-coumarate:CoA ligase
AAT, alcohol acyltransferase
ATP, adenosine triphosphate
BLAST, basic local alignment search tool
CBD, cannabidiol
CBDA, cannabidiolic acid
CBDV, cannabidivarin
CBG, cannabigerol
CBGA, cannabigerolic acid
CBGV, cannabigerovarin
CoA, coenzyme A
CNV, copy number variation
CV, coefficient of variation
DAT, deacetylvindoline 4-O-acetyltransferase
ECDF, empirical cumulative distribution function
ESM, evolutionary scale modeling
FPP, farnesyl pyrophosphate
GPP, geranyl pyrophosphate
HCBT, anthranilate N-hydroxycinnamoyl/benzoyltransferase
HCT, hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyltransferase
HMM, hidden Markov model
HPL, hydroperoxide lyase
KDE, kernel density estimate
LG, Le-Gascuel
LOX, lipoxygenase
MD, molecular dynamics
MEP, methylerythritol phosphate
MVA, mevalonate
MW, molecular weight
PAV, presence/absence variation
PC, principal component
PCA, principal component analysis
pI, isoelectric point
RBH, reciprocal best hit
RMSD, root mean square deviation
RNA-seq, RNA sequencing
TDA, tandem duplication array
THC, tetrahydrocannabinol
THCA, tetrahydrocannabinolic acid
THCV, tetrahydrocannabivarin
TPM, transcripts per million
UV-B, ultraviolet B
VOC, volatile organic compound
VSC, volatile sulfur compound

References

1. T.K.L. Tran, T. Avellaneda, A. André, E. Gillich, M. Steinhaus, D.Á. Carrera, L. Katsir, I. Chetschik, The Plant of Many Scents: Unraveling the Odorant Composition of Selected CBD Hemp Cultivars, J. Agric. Food Chem. 73 (2025) 24314–24325. https://doi.org/10.1021/acs.jafc.5c07208.

 doi:10.20944/preprints202602.1037.v1

2. S. Rice, J.A. Koziel, Characterizing the Smell of Marijuana by Odor Impact of Volatile Compounds: An Application of Simultaneous Chemical and Sensory Analysis, PLOS ONE 10 (2015) e0144160. https://doi.org/10.1371/journal.pone.0144160.
3. I.W.H. Oswald, T.R. Paryani, M.E. Sosa, M.A. Ojeda, M.R. Altenbernd, J.J. Grandy, N.S. Shafer, K. Ngo, J.R. Peat, B.G. Melshenker, I. Skelly, K.A. Koby, M.F.Z. Page, T.J. Martin, Minor, Nonterpenoid Volatile Compounds Drive the Aroma Differences of Exotic Cannabis, ACS Omega 8 (2023) 39203–39216. https://doi.org/10.1021/acsomega.3c04496.
4. I.W.H. Oswald, M.A. Ojeda, R.J. Pobanz, K.A. Koby, A.J. Buchanan, J. Del Rosso, M.A. Guzman, T.J. Martin, Identification of a New Family of Prenylated Volatile Sulfur Compounds in Cannabis Revealed by Comprehensive Two-Dimensional Gas Chromatography, ACS Omega 6 (2021) 31667–31676. https://doi.org/10.1021/acsomega.1c04196.
5. K.P. Kaminski, J. Hoeng, K. Lach-Falcone, F. Goffman, W.K. Schlage, D. Latino, Exploring Aroma and Flavor Diversity in *Cannabis sativa* L.—A Review of Scientific Developments and Applications, Molecules 30 (2025) 2784. https://doi.org/10.3390/molecules30132784.
6. T.R. Paryani, M.E. Sosa, M.F.Z. Page, T.J. Martin, M.V. Hearvy, M.A. Ojeda, K.A. Koby, J.J. Grandy, B.G. Melshenker, I. Skelly, I.W.H. Oswald, Nonterpenoid Chemical Diversity of Cannabis Phenotypes Predicts Differentiated Aroma Characteristics, ACS Omega 9 (2024) 28806–28815. https://doi.org/10.1021/acsomega.4c03225.
7. P. Janta, S. Vimolmangkang, Chemical profiling and clustering of various dried cannabis flowers revealed by volatilomics and chemometric processing, J. Cannabis Res. 6 (2024) 41. https://doi.org/10.1186/s42238-024-00252-w.
8. D. Xu, Z. Wang, W. Zhuang, T. Wang, Y. Xie, Family characteristics, phylogenetic reconstruction, and potential applications of the plant BAHD acyltransferase family, Front. Plant Sci. 14 (2023). https://doi.org/10.3389/fpls.2023.1218914.
9. W. Zhang, J. Li, Y. Dong, Y. Huang, Y. Qi, H. Bai, H. Li, L. Shi, Genome-wide identification and expression of BAHD acyltransferase gene family shed novel insights into the regulation of linalyl acetate and lavandulyl acetate in lavender, J. Plant Physiol. 292 (2024) 154143. https://doi.org/10.1016/j.jplph.2023.154143.
10. A. Sharafi, H. Hashemi Sohi, A. Mousavi, P. Azadi, B. Dehsara, B. Hosseini Khalifani, Enhanced morphinan alkaloid production in hairy root cultures of Papaver bracteatum by over-expression of salutaridinol 7-o-acetyltransferase gene via Agrobacterium rhizogenes mediated transformation, World J. Microbiol. Biotechnol. 29 (2013) 2125–2131. https://doi.org/10.1007/s11274-013-1377-2.
11. B. St-Pierre, P. Laflamme, A.-M. Alarco, V. D, E. Luca, The terminal O-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer, Plant J. 14 (1998) 703–713. https://doi.org/10.1046/j.1365-313x.1998.00174.x.
12. K. Barbosa-Xavier, F. Pedrosa-Silva, F. Almeida-Silva, T.M. Venancio, Cannabis Expression Atlas: a comprehensive resource for integrative analysis of *Cannabis sativa* L. gene expression, Physiol. Plant. 176 (2024) e70010. https://doi.org/10.1111/ppl.70010.
13. M. Shalit, I. Guterman, H. Volpin, E. Bar, T. Tamari, N. Menda, Z. Adam, D. Zamir, A. Vainstein, D. Weiss, E. Pichersky, E. Lewinsohn, Volatile Ester Formation in Roses. Identification of an Acetyl-Coenzyme A. Geraniol/Citronellol Acetyltransferase in Developing Rose Petals, Plant Physiol. 131 (2003) 1868–1876. https://doi.org/10.1104/pp.102.018572.
14. C. Liu, X. Qiao, Q. Li, W. Zeng, S. Wei, X. Wang, Y. Chen, X. Wu, J. Wu, H. Yin, S. Zhang, Genome-wide comparative analysis of the BAHD superfamily in seven Rosaceae species and expression analysis in pear (Pyrus bretschneideri), BMC Plant Biol. 20 (2020) 14. https://doi.org/10.1186/s12870-019-2230-z.
15. S. Yadav, W. Maurya, R. Kumari, P. Rangan, A.B. Gaikwad, Genome-wide identification and characterization of BAHD acyltransferases in black pepper (Piper nigrum L.) reveals their important role in piperine biosynthesis, Sci. Rep. 15 (2025) 39415. https://doi.org/10.1038/s41598-025-22795-5.
16. R.C. Lynch, L.K. Padgitt-Cobb, A.R. Garfinkel, B.J. Knaus, N.T. Hartwick, N. Allsing, A. Aylward, P.C. Bentz, S.B. Carey, A. Mamerto, J.K. Kitony, K. Colt, E.R. Murray, T. Duong, H.I. Chen, A. Trippe, A. Harkess,

 doi:10.20944/preprints202602.1037.v1

S. Crawford, K. Vining, T.P. Michael, Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome, Nature 643 (2025) 1001–1010. https://doi.org/10.1038/s41586-025-09065-0.

17. I. Molina, Y. Li-Beisson, F. Beisson, J.B. Ohlrogge, M. Pollard, Identification of an Arabidopsis Feruloyl-Coenzyme A Transferase Required for Suberin Synthesis, Plant Physiol. 151 (2009) 1317–1328. https://doi.org/10.1104/pp.109.144907.
18. L. de Vries, H.A. MacKay, R.A. Smith, Y. Mottiar, S.D. Karlen, F. Unda, E. Muirragui, C. Bingman, K. Vander Meulen, E.T. Beebe, B.G. Fox, J. Ralph, S.D. Mansfield, pHBMT1, a BAHD-family monolignol acyltransferase, mediates lignin acylation in poplar, Plant Physiol. 188 (2021) 1014–1027. https://doi.org/10.1093/plphys/kiab546.
19. L.P.Y. Lam, Y. Tobimatsu, S. Suzuki, T. Tanaka, S. Yamamoto, Y. Takeda-Kimura, Y. Osakabe, K. Osakabe, J. Ralph, L.E. Bartley, T. Umezawa, Disruption of p-coumaroyl-CoA:monolignol transferases in rice drastically alters lignin composition, Plant Physiol. 194 (2024) 832–848. https://doi.org/10.1093/plphys/kiad549.
20. R.A. Smith, E.T. Beebe, C.A. Bingman, K. Vander Meulen, A. Eugene, A.J. Steiner, S.D. Karlen, J. Ralph, B.G. Fox, Identification and characterization of a set of monocot BAHD monolignol transferases, Plant Physiol. 189 (2022) 37–48. https://doi.org/10.1093/plphys/kiac035.
21. J. Luo, C. Fuell, A. Parr, L. Hill, P. Bailey, K. Elliott, S.A. Fairhurst, C. Martin, A.J. Michael, A Novel Polyamine Acyltransferase Responsible for the Accumulation of Spermidine Conjugates in Arabidopsis Seed, Plant Cell 21 (2009) 318–333. https://doi.org/10.1105/tpc.108.063511.
22. J. Beekwilder, M. Alvarez-Huerta, E. Neef, F.W.A. Verstappen, H.J. Bouwmeester, A. Aharoni, Functional Characterization of Enzymes Forming Volatile Esters from Strawberry and Banana, Plant Physiol. 135 (2004) 1865–1878. https://doi.org/10.1104/pp.104.042580.
23. J. Gertsch, M. Leonti, S. Raduner, I. Racz, J.-Z. Chen, X.-Q. Xie, K.-H. Altmann, M. Karsak, A. Zimmer, Beta-caryophyllene is a dietary cannabinoid, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 9099–9104. https://doi.org/10.1073/pnas.0803601105.
24. S.A. Ahmed, S.A. Ross, D. Slade, M.M. Radwan, F. Zulfiqar, R.R. Matsumoto, Y.-T. Xu, E. Viard, R.C. Speth, V.T. Karamyan, M.A. ElSohly, Cannabinoid ester constituents from high-potency Cannabis sativa, J. Nat. Prod. 71 (2008) 536–542. https://doi.org/10.1021/np070454a.
25. O. Levsh, Y.-C. Chiang, C.F. Tung, J.P. Noel, Y. Wang, J.-K. Weng, Dynamic Conformational States Dictate Selectivity toward the Native Substrate in a Substrate-Permissive Acyltransferase, Biochemistry 55 (2016) 6314–6326. https://doi.org/10.1021/acs.biochem.6b00887.
26. D.W. Pate, Possible role of ultraviolet radiation in evolution ofCannabis chemotypes, Econ. Bot. 37 (1983) 396–405. https://doi.org/10.1007/BF02904200.
27. J. Lydon, A.H. Teramura, C.B. Coffman, UV-B RADIATION EFFECTS ON PHOTOSYNTHESIS, GROWTH and CANNABINOID PRODUCTION OF TWO *Cannabis sativa* CHEMOTYPES, Photochem. Photobiol. 46 (1987) 201–206. https://doi.org/10.1111/j.1751-1097.1987.tb04757.x.
28. D.W. Pate, Chemical ecology of Cannabis, (1994). https://www.druglibrary.net/olsen/HEMP/IHA/iha01201.html (accessed February 9, 2026).
29. V. Mediavilla, S. Steinemann, Essential oil of *Cannabis sativa* L. strains, (n.d.). https://www.druglibrary.net/olsen/HEMP/IHA/jiha4208.html (accessed February 9, 2026).
30. C. Villard, I. Baser, A.C. van de Peppel, K. Cankar, M.E. Schranz, R. van Velzen, Resurrected Ancestral Cannabis Enzymes Unveil the Origin and Functional Evolution of Cannabinoid Synthases, Plant Biotechnol. J. n/a (n.d.). https://doi.org/10.1111/pbi.70475.