Ancestral class-promiscuity as a driver of functional diversity in the BAHD acyltransferase family in plants

Lars H. Kruse[1], Austin T. Weigle[3], Jesús Martínez-Gómez[1,2], Jason D. Chobirko[1,5], Jason E. Schaffer[6], Alexandra A. Bennett[1,7], Chelsea D. Specht[1,2], Joseph M. Jez[6], Diwakar Shukla[4], Gaurav D. Moghe[1*]

Footnotes:

[1] Plant Biology Section, School of Integrative Plant Sciences, Cornell University, Ithaca, NY, 14853, USA

[2] L.H. Bailey Hortorium, Cornell University, Ithaca, NY, 14853, USA

[3] Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

[4] Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

[5] Present address: Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, 14853, USA

[6] Department of Biology, Washington University in St. Louis, St. Louis, MO, 63130, USA

[7] Present address: Institute of Analytical Chemistry, Universität für Bodenkultur Wien, Vienna, 1190, Austria

* Corresponding author: gdm67@cornell.edu

ABSTRACT

Gene duplication-divergence and enzyme promiscuity drive metabolic diversification in plants, but how they contribute to functional innovation in enzyme families is not clearly understood. In this study, we addressed this question using the large BAHD acyltransferase family as a model. This fast-evolving family, which uses diverse substrates, expanded drastically during land plant evolution. *In vitro* characterization of 11 BAHDs against a substrate panel and phylogenetic analyses revealed that the ancestral enzymes prior to origin of land plants were likely capable of promiscuously utilizing most of the substrate classes used by current, largely specialized enzymes. Motif enrichment analysis in anthocyanin/flavonoid-acylating BAHDs helped identify two motifs that potentially contributed to specialization of the ancestral anthocyanin-acylation capability. Molecular dynamic simulations and enzyme kinetics further resolved the potential roles of these motifs in the path towards specialization. Our results illuminate how promiscuity in robust and evolvable enzymes contributes to functional diversity in enzyme families.

KEY WORDS

Evolutionary biochemistry, enzyme family, comparative genomics, gene duplication, promiscuity, protein structure analysis, BAHD acyltransferase

41 **INTRODUCTION**

42 Enzymes involved in plant specialized metabolism often belong to enzyme
43 families, some of whom (e.g. cytochrome P450s, lipases, acyltransferases,
44 dioxygenases) have several hundred members in angiosperm genomes. Such enzyme
45 families are characterized by frequent gene duplication, functional divergence, and
46 promiscuity, all of which contribute to metabolic diversification. For duplication, various
47 models explaining duplicate gene evolution have been proposed, such as neo-
48 functionalization, sub-functionalization, escape from adaptive conflict, dosage balance,
49 and pseudogenization (reviewed in Panchy et al., 2016). Due to advances in sequencing
50 technologies, much is known today about how duplicates evolve at the genomic,
51 epigenetic and transcriptomic levels (Ganko et al., 2007; Zou et al., 2009; Schnable et
52 al., 2011; Moghe et al., 2014; J. Wang et al., 2014); however, our understanding of how
53 substrate preference evolves in duplicate enzymes, especially in large enzyme families
54 is lacking. Specifically, while it is common knowledge that different members of large
55 enzyme families use substrates containing very different chemical/structural scaffolds,
56 the extent to which this "family multi-functionality" or functional diversity is due to ancestral
57 promiscuity vs. neo-functionalization after duplication is not clear.

58 Promiscuity refers to the ability of an enzyme to catalyze multiple reactions, either
59 by using different substrates (substrate promiscuity), producing multiple products from
60 the same substrate (product promiscuity), or performing secondary reactions that cause
61 different chemical transformations (catalytic promiscuity) (Copley, 2015). Here, we do not
62 consider the physiological relevance of the secondary products but only study an
63 enzyme's ability to use multiple substrates – a definition of promiscuity typically used by
64 molecular/structural biologists (Copley, 2015; Kreis and Munkert, 2019). We also define
65 a special type of substrate promiscuity called "class-promiscuity", referring to the ability
66 of an enzyme to use substrates containing very different structural scaffolds e.g. aliphatic
67 alcohol vs. anthocyanin. In contrast, the term "multi-functionality" is used here in the
68 context of the collective enzyme family using multiple substrates e.g. multi-functionality
69 of the BAHD family. Even if the product at first is irrelevant in a physiological context, the
70 promiscuous reaction may still continue to occur and may get selected upon if the product
71 directly or indirectly increases organismal fitness. Existence of such promiscuity-driven

72 “underground metabolism” can occur via drift and contributes to the standing natural
73 variation of metabolites (Notebaart et al., 2014).
74 In this study, we address the question of how gene duplication and promiscuity
75 contribute to plant specialized metabolic diversity using the large BAHD acyltransferase
76 family (referred to as BAHDs hereafter) as a model. The ease of heterologous protein
77 expression in *Escherichia coli*, intronless nature of many BAHD genes, their ability to use
78 structurally diverse substrates, and availability of functional data from multiple species
79 make this an attractive family to address the above question. Named after the four first
80 discovered enzymes of this family – benzyl alcohol O-acetyltransferase (BEAT)
81 (Dudareva et al., 1998), anthocyanin O-hydroxycinnamoyltransferase (AHCT) (Fujiwara
82 et al., 1997; H. Fujiwara et al., 1998; Hiroyuki Fujiwara et al., 1998), N-
83 hydroxycinnamoyl/benzoyltransferase (HCBT) (Yang et al., 1997), and deacetylvindoline
84 4-O-acetyltransferase (DAT) – members of this large family (referred to as BAHDs
85 hereafter) catalyze the transfer of an acyl group from a coenzyme A (CoA) conjugated
86 donor to a –OH or $–NH_2$ group on an acceptor (D’Auria, 2006). BAHDs play important
87 roles in the biosynthesis of several phenylpropanoids, amides, volatile esters, terpenoids,
88 alkaloids, anthocyanins, flavonoids, and acylsugars (D’Auria, 2006; Tuominen et al.,
89 2011). Although >150 members of this family have been experimentally characterized
90 across the plant kingdom, transfer of known functions using sequence similarity to these
91 characterized enzymes is difficult owing to their rapid sequence divergence, substrate
92 promiscuity and functional divergence. For example, the 4-5 acylsugar acyltransferases
93 involved in acylsugar biosynthesis in Solanaceae trichomes are BAHDs (Moghe et al.,
94 2017) but are only 40-50% identical, and yet all of them use sucrose/acylated sucrose as
95 substrates. In contrast, a single amino acid change is sufficient to convert a BAHD from
96 preferentially using phenylpropanoid substrates to using phenolic amine substrates
97 (Levsh et al., 2016). Compared to the scale of BAHD acceptor diversity, there is an
98 incomplete understanding of BAHD sequence-function and structure-function
99 relationships, which is representative of a similar lack of knowledge in other large enzyme
100 families generated via gene duplication.
101 Previous studies on BAHDs have revealed existence of substrate promiscuity
102 (Aharoni et al., 2000; Aymerick Eudes et al., 2016; Levsh et al., 2016; Moghe et al., 2017;

103 Chiang et al., 2018a), but compared to the known number of BAHD substrates, the extent
104 of our knowledge about BAHD class-promiscuity is still limited. Furthermore, it is unclear
105 whether BAHD multi-functionality is a result of multiple rounds of neo-functionalization –
106 where new activities emerged completely afresh in the BAHD family (also referred to
107 below as "innovation") – vs. specialization of activities that were already possible in the
108 common ancestor.

109 The BAHD family is speculated to have arisen from carnitine acyltransferases
110 involved in fatty acid metabolism (St Pierre and De Luca, 2000; D'Auria, 2006), however,
111 their evolution across plants has not been studied. We were interested in characterizing
112 evolution of the *capability*/*potential* of BAHDs to use different substrate classes rather
113 than their *actual in vivo* substrates, since the inherent capability of an enzyme can be a
114 starting point for selection to act and fix diversified enzyme activities in different *in vivo*
115 contexts. Although it may be argued that BAHDs might use any substrates with hydroxyl
116 or amine groups, previous results provide clear evidence of specialization (D'Auria, 2006),
117 and it is unclear how these specializations emerged. We first characterized the known
118 substrate space of extant characterized BAHDs and used it as a template in the context
119 of BAHD phylogeny to delineate the putative ancestral substrate space. Prediction of the
120 ancestral state helped us differentiate between neo-functionalization vs. ancestral
121 promiscuity, and the sequence and structural features that enabled specialization of
122 ancestrally accessible functions. Overall, this study provides a template to assess
123 functional evolution after duplication in large enzyme families, and generates resources
124 foundational for rational prediction of BAHD function in plant genomes.
125

126 ## RESULTS

127 ### *The BAHD enzyme family occupies a wide substrate space*

128 BAHDs have been experimentally characterized across land plants (Sander and
129 Petersen, 2011; Aymerick Eudes et al., 2016; Levsh et al., 2016; Moghe et al., 2017;
130 Chiang et al., 2018a). To get a complete picture of the range of known BAHD substrates,
131 we first compiled a database of 136 biochemically characterized BAHDs that used a total
132 of 187 acceptor substrates and ~30 acyl donor substrates from 64 species across the

133 green plants **(File S2)**. BAHDs characterized solely using other types of experimental
134 evidence such as gene knock-out and knock-down, or gene expression analysis were
135 excluded from this study, because of the remaining uncertainty about the actual substrate
136 the enzyme is acting on. These substrates were empirically classified into thirteen

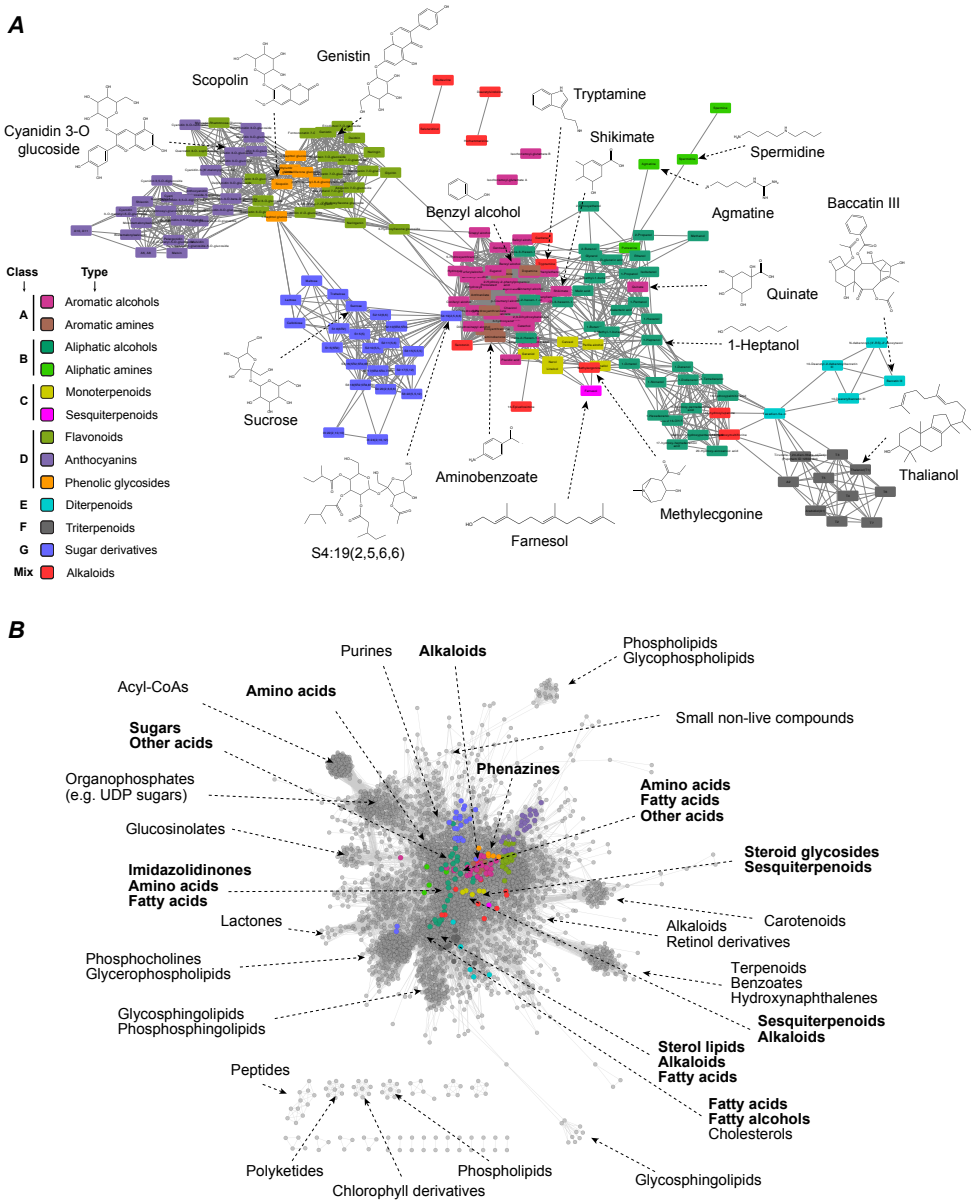


Figure 1. BAHD substrate space. (A) Substrate similarity network of BAHD acceptor substrates (**File S2B**), shown as the "Prefuse Force Directed" layout in Cytoscape. An MCS-Tanimoto similarity cutoff of 0.5 used to draw edges between two substrate nodes. Substrates are colored based on the substrate type they belong to. **(B)** BAHD substrates in the context of a larger network of plant compounds. Plant compounds with KNApSAcK ID were gathered from the ChEBI database. The network was visualized as in part A and BAHD substrates are highlighted by colors corresponding to their substrate class. For each labeled cluster/region in the network, five compounds were chosen randomly and the compound class for each compound was determined based on the ChEBI Ontology (see **File S2**). Representative classes are shown for each analyzed region. Names in bold are the classes closest to the known BAHD substrates.

137 different types based on similarity of their chemical scaffolds and functional groups **(Table**
138 **S1; Fig. S1A)** (Wang et al., 2013), and organized into a network for visualization of the
139 already characterized BAHD substrate space **(Fig. 1A; Fig. S1B)**.

140 Distinct clusters obtained in this visualization suggested that BAHDs have evolved
141 to accept at least eight different structural scaffolds **(classes A-G-Mix; Fig. 1A**). A
142 majority (109, 80%) of the characterized enzymes use substrates in four classes – class
143 A comprising of aromatic alcohol and amines (69 enzymes, 51%), class B containing
144 aliphatic alcohols and amines (38 enzymes, 28%), class C that includes monoterpenoids
145 and a sesquiterpenoid (7 enzymes, 5%), and class D comprising flavonoids,
146 anthocyanins, and phenolic glycosides (38 enzymes, 28%) **(Fig. 1A; File S2B)** – roughly
147 indicative of the degree of research attention on lignins, cuticular lipids, floral volatiles and
148 pigments, respectively. BAHDs can also transform a wider array of terpenoids including
149 monoterpenoid (e.g. geraniol), diterpenoid (e.g. taxol intermediates), and triterpenoid
150 alcohols (e.g. thalianol) in classes E and F. Some polyamines (spermine, spermidine)
151 were more distant from aliphatic alcohols primarily due to different functional groups (-OH
152 vs. $-NH_2$), but still were classified into the same class B due to overall scaffold relatedness
153 **(Fig. S1A; Table S1)**. Other substrates such as alkaloids (class Mix) and sugar
154 derivatives (class G), represented smaller and independent classes in the network.
155 Alkaloids themselves are a loosely defined compound type, and hence class Mix
156 comprises of alkaloids that could actually be assigned to other classes.

157 Although large, the characterized BAHD substrate space still only represents a
158 small proportion of the larger phytochemical space. As described in a later section, only
159 ~50% of all multi-species BAHD orthologous groups (OGs) have characterized activities,
160 which suggests that many substrate classes remain undiscovered. We thus used the
161 substrate network to obtain a bird's eye view of known BAHD substrates in the context of
162 the wider phytochemical space. Using 7128 additional plant-specific compounds, we
163 mapped a broader phytochemical similarity network **(Fig. 1B).** Since BAHDs can only use
164 substrates containing amine or hydroxyl groups, only ~70% of the compounds from this
165 pool are actually available to them.

166 Previous studies recognize the role of substrate ambiguity and structural motifs in
167 generating secondary reactions (Bar-Even et al., 2011), which, under appropriate

168 conditions, can be selected upon. An investigation of the global "underground
169 metabolism" found that promiscuously used substrates of *E. coli* enzymes tend to be
170 structurally similar (Notebaart et al., 2014). Based on these observations, we divided
171 compounds in the above phytochemical network **(Fig. 1B)** into three groups: (1) known
172 BAHD substrate classes; these classes are spread out over a large region of the
173 phytochemical space **(colored nodes, Fig. 1B)**, which potentially allows for transitions to
174 new substrate classes through mutational changes, (2) compound classes that are
175 theoretically accessible – due to similarity to known substrates – but not yet known as
176 BAHD substrates, such as small organic acids in central metabolism, some amino acids,
177 nitrogen bases/nucleosides, many alkaloids, sesquiterpenoids, peptides, polyphenols,
178 and oligosaccharides **(Fig. 1B; File S3)**, and (3) compound classes whose utilization
179 cannot be inferred due to absence of prior data or are unlikely to be BAHD substrates. As
180 far as we could determine, no biochemical or genetic studies have identified BAHDs using
181 sulfur/phosphorus containing compounds (e.g. glucosinolates, nucleotides) and long-
182 chain lipid types (e.g. sulfolipids, sphingolipids, carotenoids/tetraterpenoids) as acyl
183 acceptor substrates – despite some members having hydroxyl groups – and hence, these
184 substrate classes were categorized into group 3. From groups 2 and 3, some substrate
185 types may be inaccessible to BAHDs due to their lipophilic nature or their absence in
186 cytoplasmic environments where most BAHDs are known to exist. Nonetheless, such
187 visualization of the larger network shows that several metabolite classes are structurally
188 very similar to existing BAHD substrates, and represent the latent catalytic potential of
189 BAHDs that could be selected upon after duplication-divergence or could be used for *in*
190 *vitro* enzyme engineering. These activities may still lie undetected in uncharacterized
191 enzymes or as secondary activities of characterized enzymes.

192 To address our questions about class-promiscuity, we further assessed enzymes
193 accepting class A, B and D substrates, which are structurally very distinct **(Fig. 1A)**. Three
194 enzymes utilizing class A substrates (AtHCT, SmHCT, PsHCT2) were previously shown
195 in one study to use naringenin, a flavonoid (Chiang et al., 2018a). No class D substrate
196 utilizing enzyme was shown to use class A/B substrates **(File S2A)**. While this difference
197 could represent true functional differentiation, it may likely be a result of experimenter-
198 bias (i.e. researchers studying pigmentation may not assess lignin or volatile esters (class

199 A/B substrates), and vice versa). To minimize the effect of such a bias on our later
200 evolutionary inferences, we performed extensive characterization of 11 BAHDs that use
201 class A, B and D substrates and obtained better insights into promiscuity of BAHD
202 enzymes.
203
204 ***Determining the class-promiscuity of characterized BAHDs***

205 We tested a total of 11 novel and previously characterized BAHD enzymes against
206 an acceptor substrate array **(Fig. S2)**, using optimal conditions described for those
207 enzymes. For donors, we selected the most preferred donor of each tested enzyme,
208 which, for most enzymes was coumaroyl-CoA or malonyl-CoA. Cocaine synthase from
209 *Erythroxylum coca* (EcCS) was described to use benzoyl-CoA and cinnamoyl-CoA to
210 produce cocaine from methylecgonine (Schmidt et al., 2015; A. Eudes et al., 2016), but
211 in our enzyme assays, we discovered that it is also able to use coumaroyl-CoA (**Fig. S2**).
212 For the hydroxyacid/alcohol hydroxycinnamoyl transferase of the liverwort *Marchantia*
213 *emarginata* (MeHFT), we used its preferred donor feruloyl-CoA. Based on the substrate
214 networks **(Fig. 1A; Fig. S1),** we selected 11 substrates from six substrate types in classes
215 A, B and D as acceptors for initial analysis, and based on these results, further assayed
216 some enzymes with additional anthocyanin and terpenoid substrates **(Fig. S2B,C)** to
217 confirm additional hypotheses.
218 Of the eleven enzymes, two used only one tested substrate, six showed substrate
219 promiscuity within the same class (which we also refer to as “specialized” below), while
220 three showed class-promiscuity (**Fig. 2A; Fig. S2**) According to our assays and gathered
221 literature information **(Fig. 4)**, 75% (103 enzymes) of BAHDs analyzed in this study can
222 use more than one acceptor substrate, 27% (37 enzymes) use $\geq$5 substrates, and 9%
223 (12 enzymes) can use $\geq$10 substrates, highlighting the considerable substrate promiscuity
224 in the family. One of the two enzymes using only one substrate was a previously
225 uncharacterized enzyme from the outgroup species selected for this study *Chara braunii*
226 – representing charophytic algae, the most closely related sister lineage to land plants
227 (Cheng et al., 2019; Donoghue and Paps, 2020; Vries and Rensing, 2020). Under the
228 testing conditions, this enzyme exclusively accepted quinate (an aromatic alcohol) **(Fig.**
229 **2A,C; Fig. S2A)**, leading us to rename this enzyme as CbHQT-like**.** This observation,

230 coupled with the evolutionary analysis performed below, suggests that activities essential
231 for monolignol biosynthesis in land plants (Weng and Chapple, 2010; Renault et al., 2019)

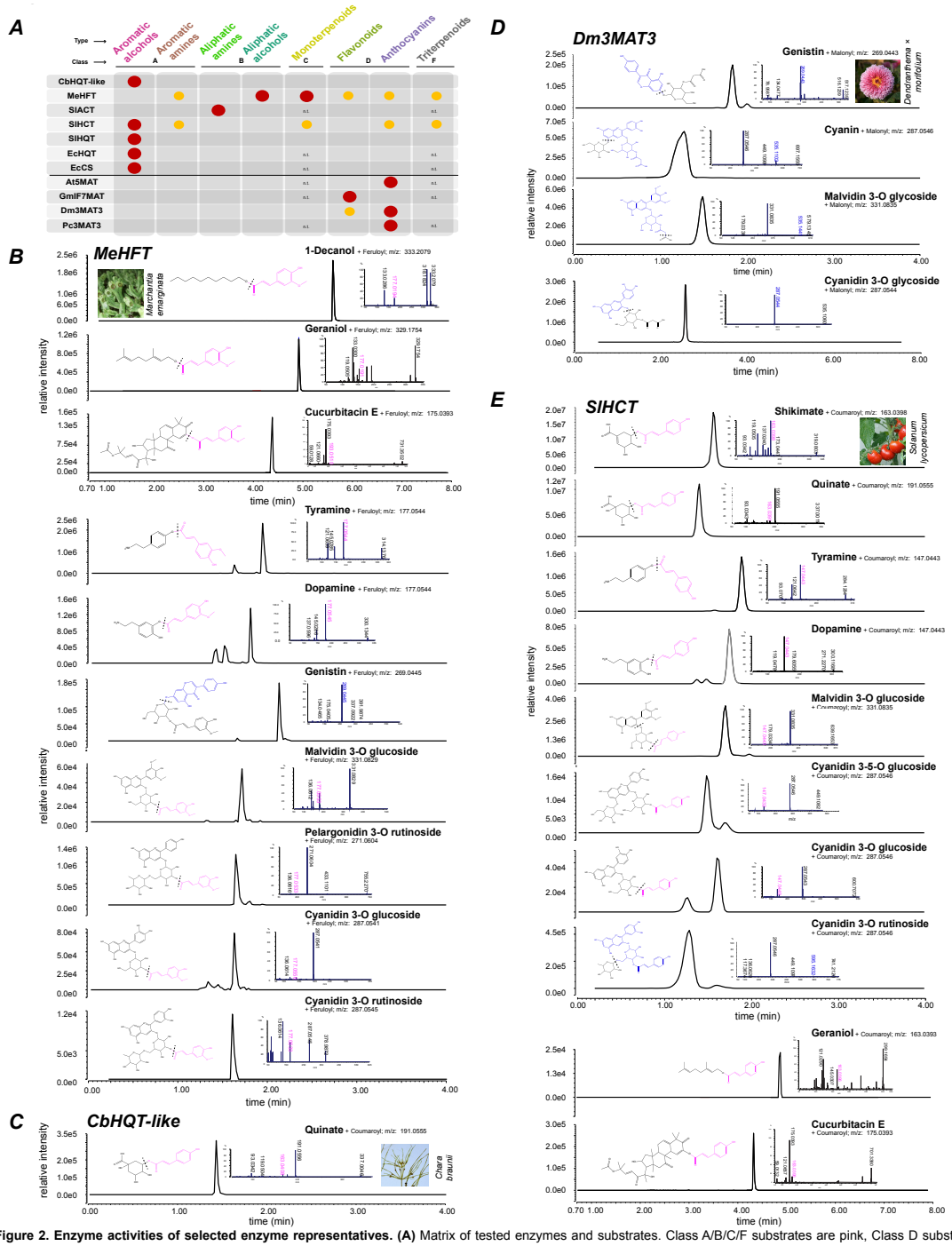


Figure 2. Enzyme activities of selected enzyme representatives. (A) Matrix of tested enzymes and substrates. Class A/B/C/F substrates are pink, Class D substrates are blue. Larger red circle corresponds to major enzyme activity and smaller orange circles indicate moderate activity. See **Fig. S2** for more detailed calculations. n.t. = not tested. **(B-E)** extracted ion chromatograms of the quantifier ions of different enzymatic products of **(B)** MeHFT with its preferred donor feruloyl-CoA (**C)** CbHQT-like using quinate and coumaroyl-CoA **(D)** Dm3MAT3 using malonyl-CoA, and **(E)** SlHCT using coumaroyl-CoA. Each product was measured using product-specific PRM methods (see **Table S4**) and the most abundant fragment ion was used for quantification. This ion is noted in the upper right corner of each chromatogram. Structures represent the best-inference based on previously reported structures and the observed fragmentation patterns. Species photographs are from following sources: *Chara braunii* (Picture by Rob Palmer released under the CC BY-NC-SA license), *Marchantia emarginata* (Picture by Boon-Chuan Ho), *Dendranthema × morifolium* (Wikimedia Commons released into the public domain), and *Solanum lycopersicum* (no license attached).

232 existed before the last common ancestor (LCA) of land plants. Given the timescale of
233 hundreds of millions of years representing the split between charophytes and land plants,
234 it is possible CbHQT-like specialized for utilizing quinate after the split from land plants.
235 Alternatively, CbHQT-like might natively use substrates not tested in this study. Further
236 *in vivo* experiments coupled with untargeted metabolomics could provide further insights
237 into the functional capacity of algal BAHDs.

238 Three out of eleven tested BAHDs were class-promiscuous. The previously
239 uncharacterized *Solanum lycopersicum* hydroxycinnamoyl CoA transferase (SlHCT) was
240 able to acylate shikimate and quinate (aromatic alcohols, class A), dopamine and
241 tyramine (aromatic amines, class A), geraniol (monoterpenoids, class C), curcubitacin E
242 (triterpenoids, class F) as well as malvidin 3-O glucoside, cyanidin-3,5-O diglucoside,
243 cyanidin 3-O glucoside, and cyanidin 3-O rutinoside (anthocyanins, class D) **(Fig. 2A,E;**
244 **Fig. S2)**. Aromatic alcohol, amine and flavonoid use has been described for HCT-type
245 enzymes before (Sander and Petersen, 2011; Aymerick Eudes et al., 2016; Levsh et al.,
246 2016; Peng et al., 2016; Chiang et al., 2018a), but not anthocyanin acylation. No activity
247 was found with free sucrose or glucose for any of the 11 enzymes, despite evidence of
248 acylation on the glycoside of the anthocyanin. The liverwort enzyme MeHFT showed a
249 similarly diverse substrate utilization pattern **(Fig. 2A,E; Fig. S2)**. We note that
250 Cucurbitacin E is not known as a substrate for any known BAHD, but still, both SlHCT
251 and MeHFT showed significant activity with it. Another enzyme, EcCS which has only
252 been tested with its native substrate methylecgonine (alkaloid, mix class) (Schmidt et al.,
253 2015; A. Eudes et al., 2016), showed high specific activities with shikimate and quinate
254 (class A) – two substrates that are structurally not very similar to methylecgonine (MCS-
255 Tanimoto = 0.47, **File S2C**)**.** Methylecgonine is most similar to aliphatic alcohols (class B)
256 – and thus, EcCS can also be considered a class-promiscuous enzyme. These findings
257 show that some BAHDs have specialized *in vivo* substrate profiles, but have the capability
258 to explore a large region of the phytochemical space through secondary activities.

259 As opposed to the relatively promiscuous HCT/HQT-type enzymes, the AnAT-type
260 enzymes show a more specialized substrate usage. All four AnATs, exclusively used
261 other flavonoid and anthocyanin substrates under the testing conditions **(Fig. 2A,D; Fig.**
262 **S2A,C)**. A previously characterized BAHD (*Dendranthema morifolium* 3-O

263 malonyltransferase, Dm3MAT3) known to malonylate cyanidin 3-O glucoside acylated
264 other anthocyanins and a flavonoid containing a 3-O glycosylation, and to a lesser extent,
265 3,5-O glycosylation. At5MAT (5-O malonyltransferase from *Arabidopsis thaliana*) and
266 Pc3MAT (3-O malonyltransferase from *Pericallis cruenta*) showed the same specificity
267 with 5-O glycosylated and 3-O glycosylated anthocyanins, respectively. These results
268 suggest that enzymes that have specialized for class D substrate acylation may have
269 undergone adaptations constraining them from using class A/B substrates. Since we
270 tested AnATs using malonyl-CoA donor and class A/B-utilizing enzymes with coumaroyl-
271 and feruloyl-CoA, we also tested whether these enzymes can acylate aromatic alcohols
272 and anthocyanins respectively, using non-native donors. We incubated Dm3MAT3 with
273 coumaroyl-CoA and shikimate or quinate. For comparison, SlHQT and SlACT were
274 incubated with malonyl-CoA using five different anthocyanin substrates. In both cases,
275 no product formation was detected **(Fig. S2A,B)**. These results suggest that enzymes
276 transforming substrates in classes A, B and D form two separate groups. While some
277 class A/B enzymes are able to acylate class D substrates, no evidence was obtained for
278 the reverse being true. This finding was robust to changes in donor CoAs, which may be
279 expected considering AnAT-BAHDs are postulated to carry out an ordered bi-bi type of
280 reaction, with the donor binding first (Shaw and Leslie, 1991; Tanner et al., 1999; Suzuki
281 et al., 2003).
282 These observations of promiscuity raise questions about how substrate specificity
283 – especially of broad substrate range enzymes – is maintained *in vivo*. One mechanism
284 is restricting gene expression – EcCS is specifically expressed in palisade parenchyma
285 of *E. coca,* which coincides with the highest concentrations of its product cocaine in the
286 plant (Schmidt et al., 2015). Highly specific expression patterns have been shown before
287 for BAHDs and other enzymes (e.g. St-Pierre et al., 1999; Kruse et al., 2017). Another
288 HCT from *Plectranthus scutellarioides* (PsHCT2, previously *Coleus blumei*) maintains
289 sufficient substrate specificity *in vivo* by using a conserved Arg residue near the active
290 site as a handle that regulates substrates entry into the active site (Levsh et al., 2016)
291 through inducing conformational change (Chiang et al., 2018a). Authors also found
292 evidence for alternative substrate binding sites that increased BAHDs substrate
293 permissiveness through diffusion of the non-native substrate towards the active site, while

294 maintaining its specialized native function (Chiang et al., 2018a). Other BAHDs may show
295 similar mechanistic adaptations to maintain *in vivo* substrate specificity.
296 Through these enzyme assays, we found existence of alternate, previously
297 undocumented scale of BAHD class-promiscuity, which ranged from strong (EcCS using
298 shikimate) to moderate (SlHCT using malvidin-3-O-glucoside) to weak (MeHFT using
299 cyanidin-3-O-glucoside) **(Fig. S2)**. Such latent capacity of BAHDs, like all enzyme
300 families, is critical for emergence of new reactions and new metabolites in cellular
301 environments. We only assayed BAHD promiscuity among the known substrate classes;
302 however, additional promiscuity existing for structurally related classes **(Fig. 1B)** or under
303 other conditions may influence BAHD incorporation into new pathways. Enzyme assays
304 of *C. braunii* HQT-like and MeHFT, coupled with the knowledge of lignin and cutin/suberin
305 as ancestral polymers in land plants (Renault et al., 2019; Philippe et al., 2020), also
306 suggested that enzymes utilizing substrates in class A/B were members of an ancestral
307 clade. To address this hypothesis and determine origins of other activities, we next
308 studied BAHD family evolution.
309
310 ***Functional innovation and expansion in the BAHD acyltransferase family in plants***
311 Through an evolutionary analysis, we asked two questions. First, what was the
312 ancestral state of the BAHD family in plants? Second, how did the BAHD family expand
313 and diversify in the plant kingdom? We first identified BAHDs from 49 sequenced plant
314 genomes using the acyltransferase domain model (PF02458). BAHDs were detected in
315 1-5 copies in multiple green algal genomes, however, angiosperm and gymnosperm
316 genomes contain dozens to hundreds of copies, with non-seed plants showing
317 intermediate BAHD counts **(Fig. 3A,C)**. Ancestral state reconstruction of normalized
318 BAHD counts using a Bounded Brownian Motion model (Boucher and Démery, 2016)
319 revealed that the relative BAHD gene content began to increase upon origin of the land
320 plants **(Fig. 3A,B)**. The modal BAHD count per thousand genes rose from <1 in algae to
321 ~1.3 in the ancestor of land plants, increasing to ~3 in the ancestor of seed plants. While
322 the values for vascular plants and euphyllophytes were intermediate, there was also less
323 confidence in the specific modal value at these internal nodes, shown by the broad spread
324 of their distributions. These patterns suggest a gradual increase of relative BAHD count

325 from the origin of land plants to the origin of seed plants, after which the relative counts
326 generally stabilized in the genome with some lineage-specific exceptions **(Fig. 3).** We
327 thus sought to determine the ancestral state of BAHD activities in the last common
328 ancestor of land plants, prior to their expansion.

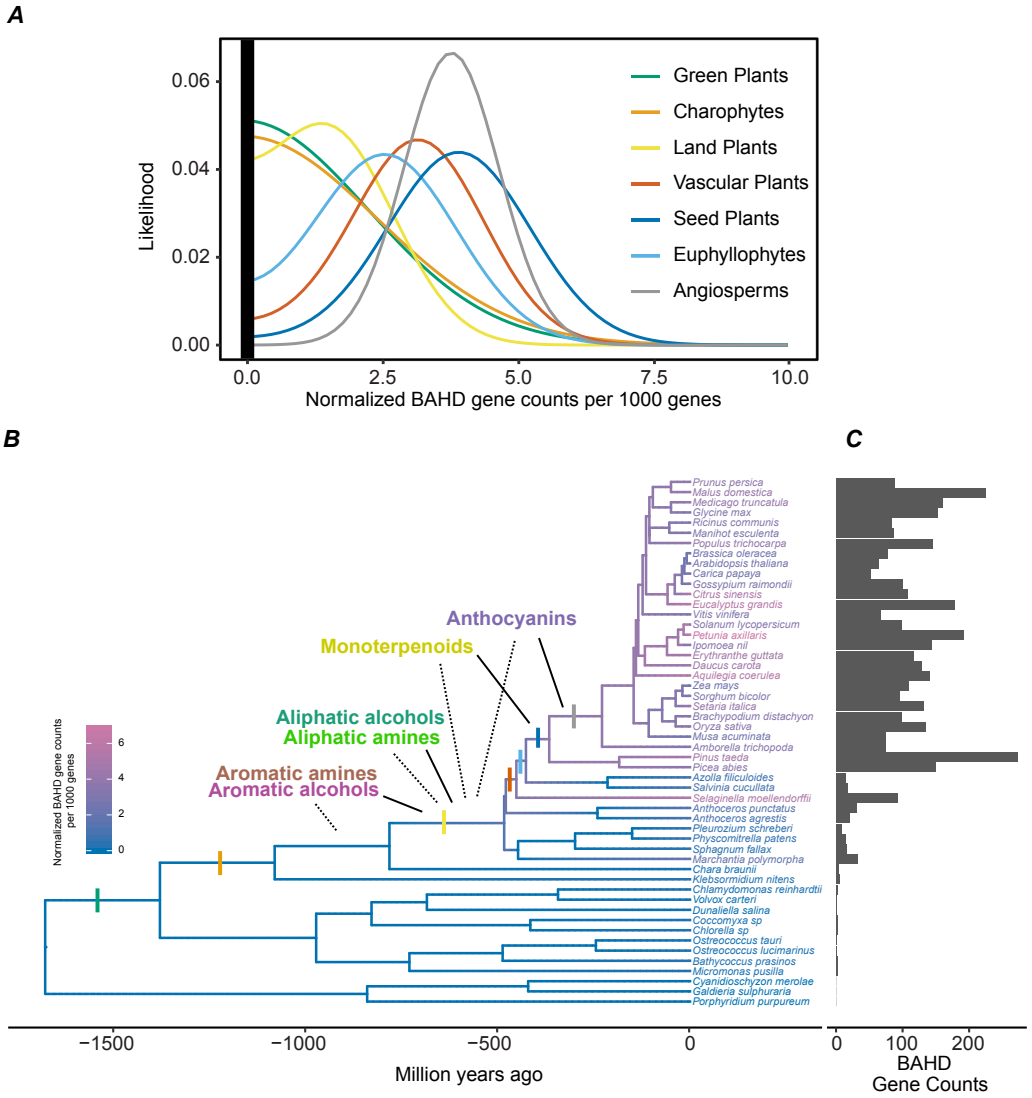


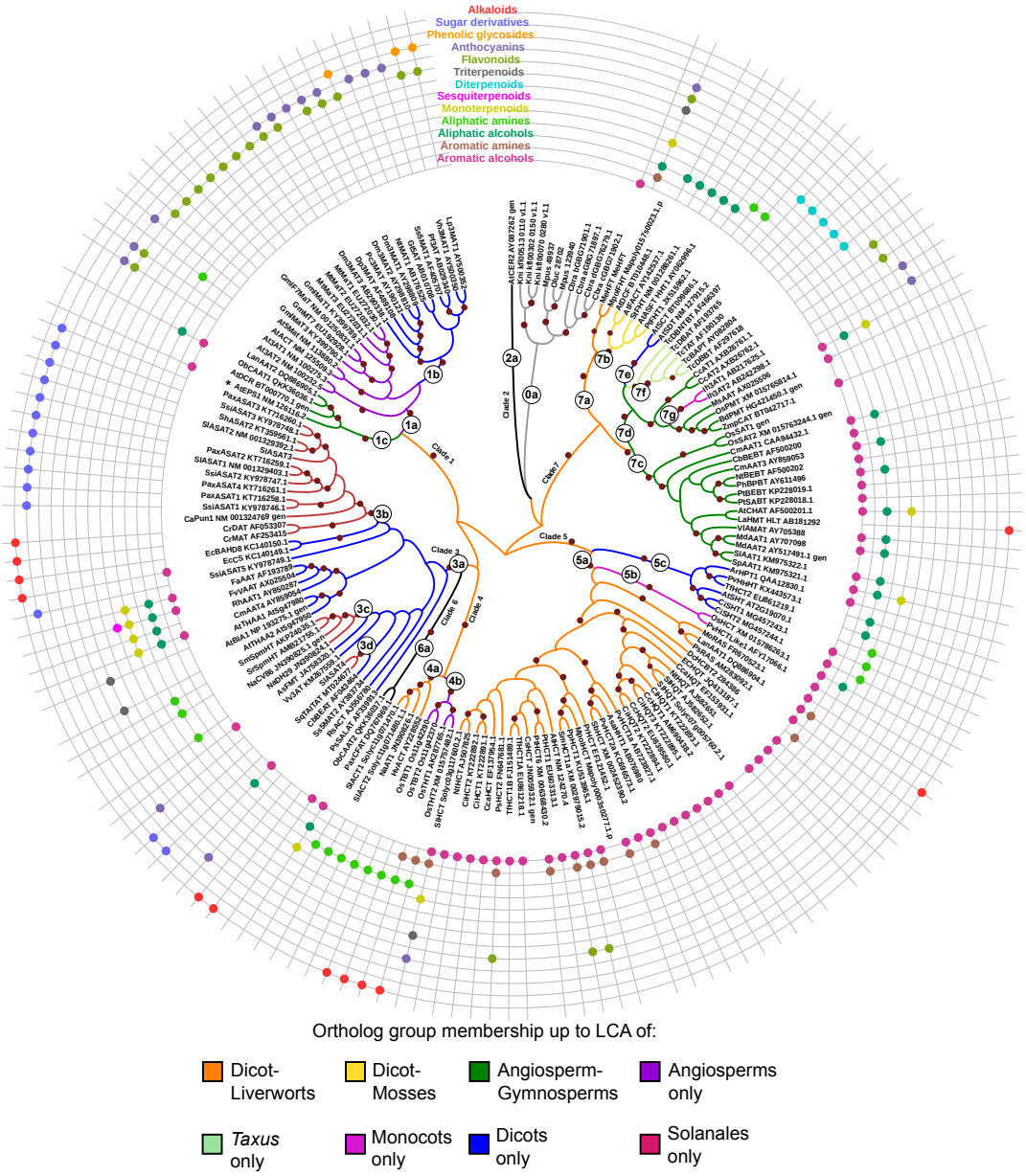
Figure 3. Number of BAHD acyltransferases in different across the plant phylogeny. (A) Likelihood of the number of BAHDs per 1000 genes in the genome of different phylogenetic groups, as per the bound Brownian motion model. The corresponding nodes are indicated in **Fig. 3B**. **(B)** Species tree of the analyzed species with sequenced genomes. Color scale of the branches indicates the normalized BAHD count. Maximum extent of OG conservation for each substrate class is highlighted by a solid black line, while tentative assignment of emergence of activities is indicated by the dashed line. See **Fig. 4** and **Fig. 8** for more details. **(C)** Bar graph showing BAHD count per analyzed genome.

329 Ancestral state is typically obtained in two different ways. Performing activity
330 assays after ancestral sequence resurrection is one approach (Huang et al., 2016).
331 However, BAHDs are fast evolving enzymes that maintain activity and structural folds
332 despite undergoing large-scale sequence evolution. Due to such rapid sequence change,
333 there is very little confidence in amino acid state prediction in deep nodes **(**example: **File**
334 **S6)**, and thus, ancestral sequence reconstruction or resurrection for deep nodes is not
335 possible. The second approach – predicting ancestral state based on analyses of
336 activities of extant enzymes – is also restricted due to poor confidence in many internal
337 nodes of the BAHD phylogeny and absence of accurate knowledge about extant
338 character states i.e. absence of knowledge about a substrate's utilization by an enzyme
339 could be truly an inability to use the substrate or simply an untested interaction. Thus, we
340 used a different approach, where we first constructed 765 BAHD OGs across the 49
341 sequenced plant genomes, of which 132 comprised of >2 members and 89 comprised of
342 members from >2 species (multi-species OGs). We then used sequence similarity to
343 assign biochemically characterized enzymes to OGs, thus roughly predicting the OGs
344 biochemical function at the substrate class utilization level. Depending on the breadth of
345 conservation of a specific OG, we inferred the deepest internal node in the species tree
346 likely housing the OG's associated function. The 136 BAHDs were mapped to only 47
347 OGs, suggesting that ~50% of the multi-species OGs still have unidentified functions.
348 While this method enables assignment of discrete states to specific ancestral nodes,
349 there are no probability estimates associated with the predictions, which is a caveat of
350 this approach. However, most internal node functional inferences were supported by
351 multiple characterized BAHDs with the same function **(File S4)**, thus providing confidence
352 in assigning the phylogenetic extent of each clade in the BAHD gene tree **(Fig. 4)**.
353 The characterized enzymes can be divided into seven clades with high bootstrap
354 support, of which four (clades 1-4) are the same as defined previously (D'Auria, 2006).
355 Clade V in D'Auria, 2006 was divided into three separate clades 5-7 **(Fig. S3)**. Three
356 clades containing HCT/HQT enzymes (clade 5a), alcohol acyltransferases (clade 7a) and
357 polyamine acyltransferases (clade 4a) were the most widely conserved, with orthologs
358 extending from angiosperms to liverworts **(Fig. 4)**. Clade 5a, most of whose members are
359 involved in phenylpropanoid pathway and lignin biosynthesis, is – based on branch

360 lengths – under the most purifying selection of all BAHDs **(Fig. S3)**, suggesting that
361 sequence similarity-based, substrate class-level functional prediction of unknown BAHDs
362 mapping to this clade will likely be accurate. Accurate class-level predictions may also be
363 possible for clades 1a/b (all but one anthocyanin/flavonoid acylating despite long branch
364 lengths) and 7a/7b (cuticular wax biosynthesis and slow-evolving), however, other clades
365 appear to have diverged rapidly at the sequence level as well as for substrate utilization
366 **(Fig. S3)**. A deeper analysis of the sub-clades will be needed to determine their predictive
367 potential.

368 Combined with substrate preference patterns from our enzyme assays and
369 previous studies **(Figs. 2,4)**, these results suggest that aliphatic and aromatic alcohol and
370 amine acylating activities (clades 4a, 5a, 7a) were already established in the ancestor of
371 land plants **(Fig. 3)**. At ~1.3 BAHDs per 1000 genes in this ancestor, and given the
372 predicted gene content of *Chara*, mosses and liverworts, it is possible that ~15-30 BAHDs
373 may have existed in this ancestor, housing the three clades described above (with
374 additional uncharacterized/lineage-specific activities). Of these, only the aromatic alcohol
375 acylating activity was detected in the outgroup species *C. braunii*, providing more support
376 to the inference that at least this activity was present in the ancestor of all land plants and
377 perhaps in the LCA of charophytes-land plants. Given the possibility of specialization of
378 the enzyme in the lineage leading to *C. braunii*, occurrence of other activities (aromatic
379 amine, aliphatic alcohol and amine acylation) in this ancestor cannot be ruled out.

380 The evolution of the AnAT and terpenoid acylation activities is slightly more
381 complex. First, the AnAT- and terpenoid-specialized clades have OGs only containing
382 angiosperm species, but these activities exist in multiple distantly related BAHD clades
383 (clades 1,3,5,7 and 3,5,6,7, respectively). Second, these activities also exist in MeHFT
384 and SlHCT clades, whose OGs extend up to liverworts **(Fig. 2B,E, Fig. S2A,B)**. Both
385 observations combined together suggest that AnAT and terpenoid acylation activities may
386 have been accessible to BAHDs for a long time before their specialization in angiosperms.



Ortholog group membership up to LCA of:

Dicot-Liverworts | Dicot-Mosses | Angiosperm-Gymnosperms | Angiosperms only

Taxus only | Monocots only | Dicots only | Solanales only

Figure 4. Defining BAHD clades. The tree was rooted using the algal enzyme clade (Clade 0a). Maroon circles on branches refer to clades with bootstrap values > 70. Clades were first defined based on deepest, high-confidence monophyletic clades. Clades 1-4 are same as D'Auria et al, 2006 definitions **(Fig. S3)**, while Clade V from that study is divided into Clades 5-7 here, based on the above criterion. Sub-clades were further defined within each clade based on the extent of conservation of OGs across the plant phylogeny **(File S3)**. These clade/sub-clade definitions can be further expanded as new, uncharacterized BAHDs are characterized in the future. Solid circles in the concentric circles around the tree represent activities characterized in this and previous studies, with "gen" referring to genetically characterized BAHDs whose substrates were not considered. * AtEPS1 showes unusual reaction mechanisms for BAHDs and has an isochorismoyl-glutamate A pyruvoyl-glutamate lyase activity that produces salicylic acid.

387

388 ***Evolution of specialized BAHD activities***

389 Orthology-based ancestral state reconstruction suggested that BAHDs accessed
390 new substrate classes or specialized in ancestrally accessible classes over their
391 expansion through gene duplication. Comparisons of OG sequences could thus help
392 identify specific residues that contributed to functional specialization for a given substrate
393 class. We first compared the HCT/HQT-like enzymes (clade 5a/b) to amine-acylating
394 enzymes (clade 4a), both clades being conserved across land plants. We identified 35
395 residues that were present in >70% of the tested sequences in clade 5a/b but had
396 completely switched to a different residue in >70% of the tested sequences in clade 4a
397 **(File S5)**. Of these, only 2 – AtHCT F303L and R356D **(Fig. 5A)** – were close to the active
398 site, with R356 present in ~90% of clade 5a/b but 0% of clade 4a, being replaced by Asp
399 or Glu in a majority of clade 4a sequences. The effect of R356D switch was previously
400 described (Levsh et al., 2016; Chiang et al., 2018b) in several HCTs, which use shikimate
401 as its primary substrate but do not show activity with aromatic amine substrates. It was
402 found that this switch was able to convert this enzyme into using amine-containing
403 substrates. While this promiscuous activity exists in at least some *O*-acylating enzymes
404 **(Fig. 2, Fig. S2),** the ubiquity of the R356D mutation in *N*-acylating enzymes suggests
405 that this residue was critical for specialization towards positively charged substrates
406 despite alternative binding sites in the protein contributing to increased promiscuity
407 (Chiang et al., 2018a). For F303L, we found that while 95% of the HCT/HQT-like enzymes
408 have the Phe residue, this is completely reversed, with 95% of the amine-acylating
409 enzymes having a Leu residue. Although this residue was not experimentally tested, we
410 hypothesize that this position also plays an important role in specialization towards the
411 respective substrates.

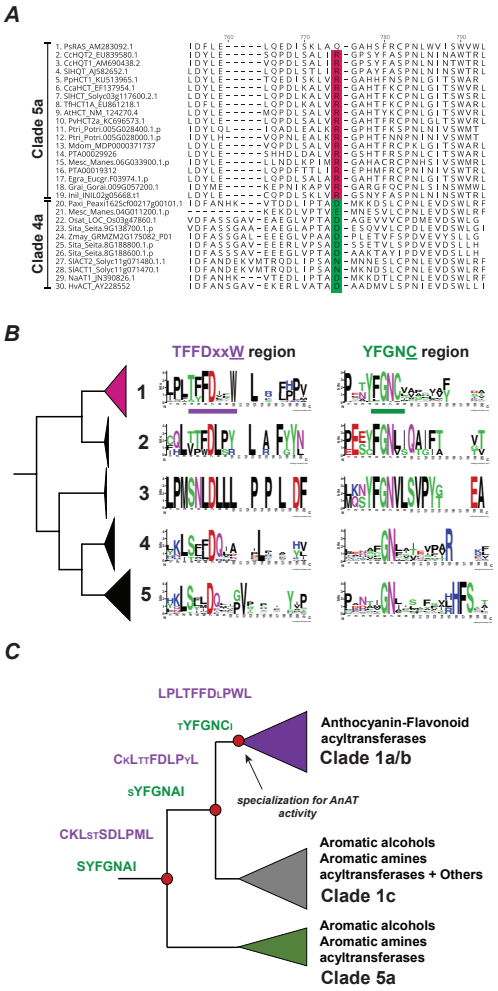


Figure 5. Conserved residues in different clades and ancestral state reconstruction. (A) Alignment of clade 5a and clade 4a OG sequences. Highlighted in red is the Arg residue conserved in ~90% of clade 5a BAHDs that predominantly use aromatic alcohols. In green, the corresponding Asp, Glu or Asn residues in amine acylating BAHDs are shown, of which Asp is present in ~90% of clade 4a sequences. **(B)** Conserved TFFDxxW and YFGNC motif region in different clades of the BAHD phylogeny. Groups 1-5 are defined in **Figure S4B**. A 20 amino acid broad window is shown. **(C)** Ancestral sequence reconstruction of the conserved TFFDxxW (purple) and YFGNC (green) region. Large letters are residues with a posterior probability >80%. Smaller letters represent residues with a posterior probability <80%. Additional information can be found in **File S6**.

Using OG comparisons, we also looked at sequence differences between HCT/HQT-like enzymes (clade 5a/b) and AnAT-like enzymes (clades 1a/b) by supplementing the single-residue analysis with enrichment analysis of larger motifs given the larger structural difference between the substrates. We focused on AnAT-like enzymes, since they are the most widespread and well-characterized clades among plants **(Fig. 4)**. The OGs of clades 1a/1b extend farthest back to angiosperms, which suggests that the fixation of this activity occurred only in this lineage. This spread corroborates with previous knowledge about evolution of the core anthocyanidin pathway in seed plants (Davies et al., 2020; Piatkowski et al., 2020), which is further extended via

421 glycosylation, methylation and acylation. Anthocyanin acylation can improve the stability
422 of the molecule to heat, higher pH, UV light, also providing an evolutionary advantage to
423 its fixation in flowering plants.

424 Motif enrichment analysis revealed among others, two large, over-represented
425 motifs in these clades (TFFDxxW: E-value=1.9e-248; YFGNC: E-value=4.5e-221**, Fig.**
426 **5B, Fig. S4A)**. Single-residue analysis also confirmed that two residues Trp36 and
427 Cys320 contained in the conserved motifs, were highly conserved AnAT enzymes in
428 comparison to other biochemically characterized BAHDs (100% vs 0% and 95% vs 4%,
429 respectively) **(Fig. S4; Fig. S5A,B)**. Both these residues were closer to the catalytic His
430 than other identified residues in the crystallized structure of Dm3MAT3 in complex with
431 malonyl-CoA (PDB: 2E1T) (Unno et al., 2007). Ancestral sequence reconstruction was
432 performed to determine when these residues appeared in the BAHD phylogeny. Over
433 80% of residues in the ancestor of all AnAT-type enzymes could not be predicted
434 confidently (posterior probability<0.8), owing to the rapid sequence evolution of BAHDs
435 **(File S6)**; however, both Trp36 and Cys320 were confidently placed in the ancestral node
436 of clade 1a/b **(Fig. 5C**; posterior probability>0.95**)**. Emergence of these two residues was
437 most likely preceded by a Tyr (<80% posterior probability) and an Ala, respectively, in the
438 prior ancestral node **(Fig. 5C)**. These results suggest that the acquisition of Trp36 and
439 Cys320 was important for the angiosperm-restricted specialization of the ancestrally
440 accessible AnAT activity. We next performed molecular dynamic (MD) simulations to
441 determine the role of these residues in the AnAT activity. MD simulations were performed
442 using the wild-type Dm3MAT3 enzyme and by replacing the two residues with Ala, a
443 catalytically neutral residue.

444

445 ***The role of Trp36 and Cys320 in anthocyanin malonyltransferase catalysis***

446 The first step of the acyltransferase reaction involves proton abstraction from the
447 cyanidin 3-*O*-glucoside (C3G) 6”-hydroxyl by the deprotonated, basic nitrogen of the
448 His170 imidazole (**Fig. 6B**) (Unno et al., 2007). For successful intermolecular proton
449 transfer to occur, the distance between these two atoms should be less than 4 Å to
450 account for the longest possible hydrogen forming (Harris and Mildvan, 1999).
451 Simulations revealed that the distance for C3G 6”-hydroxyl proton abstraction fell within

452 a 4 Å threshold for WT Dm3MaT3 for the entire 1 μs of production runs. After maintaining
453 a catalytically competent distance for the first 100 ns of simulation, the distance for proton
454 abstraction in the C320A mutant exceeded the 4 Å threshold. The W36A mutant never
455 achieved a distance satisfactory for catalysis to proceed **(Fig. 6A)**. This result suggested
456 that the W36A mutant would have a lower catalytic efficiency than wild-type and the
457 C320A mutant for both C3G and malonyl-CoA.
458

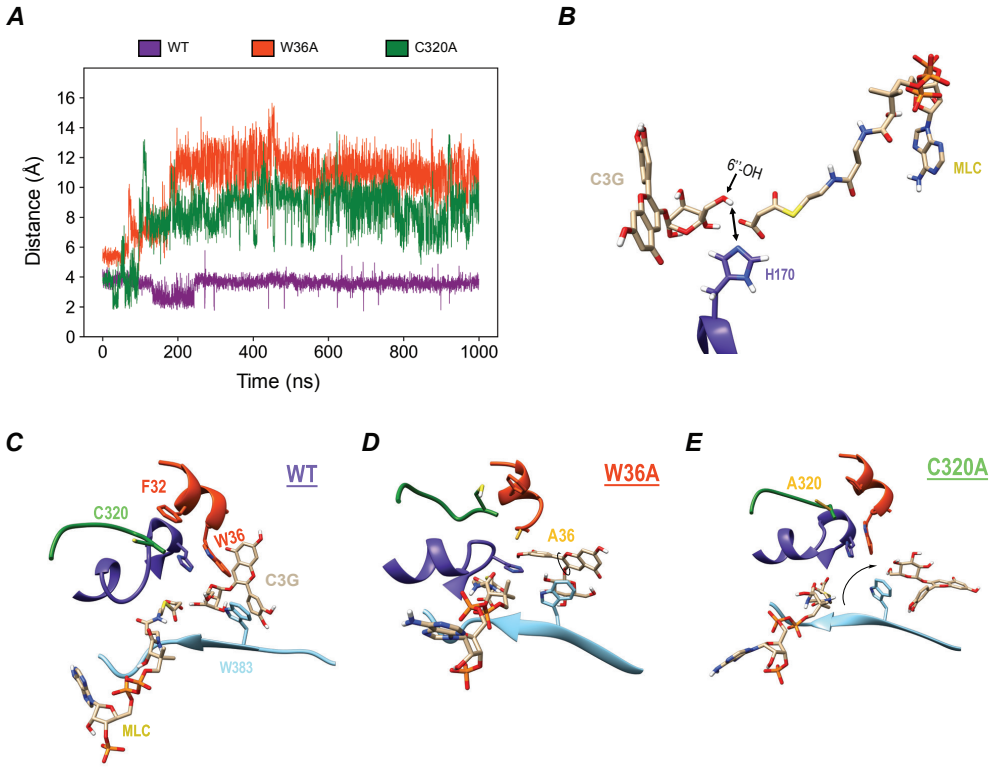


Figure 6. Distance calculations between the cyanidin 3-O glucosides 6''-OH hydrogen and the deprotonated, basic nitrogen of the His170 imidazole. (A) Distance calculation for WT (purple), W36A (red) and C320A (green) over 1000 ns of simulation time. **(B)** Illustration of the distance measured during the simulation. Active site organization is altered upon replacement of Trp36 and Cys320. Active site organization in the **(C)** WT protein, **(D)** the W36A variant, and **(E)** C320A variant during catalysis. Ala substitutions are colored and labeled as goldenrod in the appropriate systems.

459 We next assessed how C3G is repositioned by each of the Dm3MaT3 variants
460 **(Fig. 6C-E)**. In wild-type Dm3MaT3, the glucose moiety of C3G maintained a tight
461 proximity to His170 within the active site, as it was flanked by bulky residues Trp36 and
462 Trp383 **(Fig. 6C)**. The benzene diol moiety of C3G stacked with Trp383 and the
463 chromenylium moiety stacked with Trp36. Incorporation of these bulky residues in the
464 proximal active site imposed a restriction on the movement of C3G, encouraging catalysis

465 by keeping it close to the reactive His170. Upon W36A substitution, the bulky Trp residue
466 was no longer present, which allowed C3G to sample other conformations within a more
467 open active site, while still remaining bound to the enzyme. Trp383 initially stacked with
468 aromatic portions of C3G in the WT, as seen in the Dm3MaT3 WT simulation. However,
469 in W36A, the glucose moiety of C3G was flipped from its starting pose, where backbone
470 hydrogens on the hexose ring demonstrate C-H···pi interactions with Trp383, positioned
471 the 6”-hydoxyl far from His170 **(Fig. 6D)**. Thus, the W36A mutation directly altered active
472 site organization, providing a straightforward justification for this Dm3MaT3 variant’s
473 reduced k_{cat}/K_m.

474 While the C320A mutant initially maintained a native active site organization due
475 to the interactions between C3G and Trp36, C3G completely left and was excluded from
476 the active site by the end of the simulation **(Fig. 6A,E)**. Ala substitution altered the native
477 C-H···pi interactions between the backbone of residue 320 and the aromatic portion of
478 Phe32. In *apo* simulations, the distance between the backbone hydrogen of Ala320 and
479 the pi system of Phe32 was consistently less than 3 Å, but it experienced fluctuations
480 ranging between 2.5 and 6 Å during *holo* simulations **(Fig. S7A,B)**. For WT and W36A
481 variants, these distances were constant and unperturbed by the introduction of ligand,
482 suggesting that the basis of reduced activity in the C320A mutant is distinct from that
483 seen in the W36A mutant **(Fig. 6D)**.

484 Snapshots from simulation reveal that upon experiencing fluctuations in the
485 distance between the backbone alpha hydrogen of Ala320 and the aromatic portion of
486 Phe32, a series of stacking rearrangements occurred along the helix containing the
487 TFFDxxW motif **(Fig. S6C,G)**. First, Phe32 became destabilized due to inconsistent
488 backbone C-H···pi interactions with Ala320, resulting in increased edge-to-face stacking
489 between Phe31 and Phe32 **(Fig. S6A)**. Alternation of face-to-face and edge-to-face
490 stacking between Phe31 and Phe32 was seen throughout each of the *apo* and *holo*
491 systems simulated **(Fig. S7)**, but the extent of fluctuation between the two modes of
492 stacking, and the sampling of continuous no stacking interactions, was greatest for *holo*
493 C320A.

494 To recover stability within the C320A variant’s TFFDxxW motif, Phe35 broke its
495 edge-to-face stacking with Trp36 and began to stack with Phe31 and Phe32 **(Fig. S6D,**

496 **Fig. S7)**. While the W36A substitution prevented Phe35 from edge-to-face stacking with
497 an active site tryptophan, Phe35 never stacked with neither Phe31 nor Phe32 **(Fig. S8)**.
498 WT Dm3MaT3 also demonstrated virtually no Phe31-Phe35 nor Phe32-Phe35 stacking
499 **(Fig. S9)**, suggesting these interactions were developed in response to instabilities
500 caused by C320A substitution. Thus, as Phe35 sampled new stacking interactions with
501 Phe31 and Phe32, Phe35-Trp36 stacking interactions were momentarily lost **(Fig. S6D)**.
502 With the network of stacking interactions between Phe35, Trp36, and His170 disrupted,
503 the active site lost the tight organization which is requisite for maintaining a catalytically-
504 competent distance between His170 and the C3G 6”-hydroxyl (**Fig. S6E,F**). C3G then left
505 the immediate vicinity of His170 and the TFFDxxW motif to bind elsewhere within the
506 enzyme.
507 In contrast to the Ala substituted variants, wild-type Dm3MaT3 maintained an
508 ordered active site which better maintained proximity to C3G and overall acyltransferase
509 activity **(Fig. 6C)**. The stability and organization of the wild-type enzyme is underscored
510 by consistent Phe31-Phe32 face-to-face stacking and Phe35-Trp36 edge-to-face
511 stacking, neither of which were disrupted by Phe31-Phe35 or Phe32-Phe35 stacking at
512 any point during the *holo* wild-type simulation **(Fig. S9)**.
513 The MD simulation results thus suggested that that the W36A variant would have
514 inferior reaction kinetics in comparison to the wildtype due to absence of the bulky Trp36
515 that is important for keeping the C3G proximal to His170. Furthermore, we were
516 interested in examining how severely the C320A variant activity would be affected,
517 because the MD simulations demonstrated that C320A may position the acyl acceptor in
518 a catalytically competent position, suggesting the possibility for some functionality despite
519 mutation. In addition, for ~10% of the total simulation time (about 100 ns), the C320A
520 mutant maintained C3G 6”-hydroxyl and the His170 imidazole within a distance of 4 Å
521 (**Fig. 6A**). Therefore, we tested the effects of Ala replacement of Trp36 and Cys320 using
522 site specific mutagenesis and subsequent *in vivo* enzyme assays.
523
524 ***The catalytic importance of the two residues in anthocyanin acyltransferase***
525 ***activity***

526 We heterologously expressed and purified mutagenized Dm3MAT3 variants from
527 *E. coli*. These proteins showed similar folding behavior based on similar gel migration and
528 retention time in size-exclusion chromatography compared to the wild-type protein **(Fig.**
529 **S11A,B)**, suggesting that the mutation did not affect protein folding; however, both
530 mutants substantially affected enzyme reaction kinetics when comparing specific
531 activities (**Fig. 7A,B,C**). Comparing the pseudo-first-order reaction kinetics of the wild-
532 type Dm3MAT3 enzyme with the W36A mutant revealed that the mutation did not
533 influence the acceptor K_m value (**Fig. 7D**) but drastically reduced turnover number (k_{cat})
534 and catalytic rate (k_{cat}/K_m) by ~97% and 97.5%, respectively (**Fig. 7E,F**). All three kinetic
535 estimates were significantly reduced for the acyl donor malonyl-CoA (**Fig. 7D-F**).

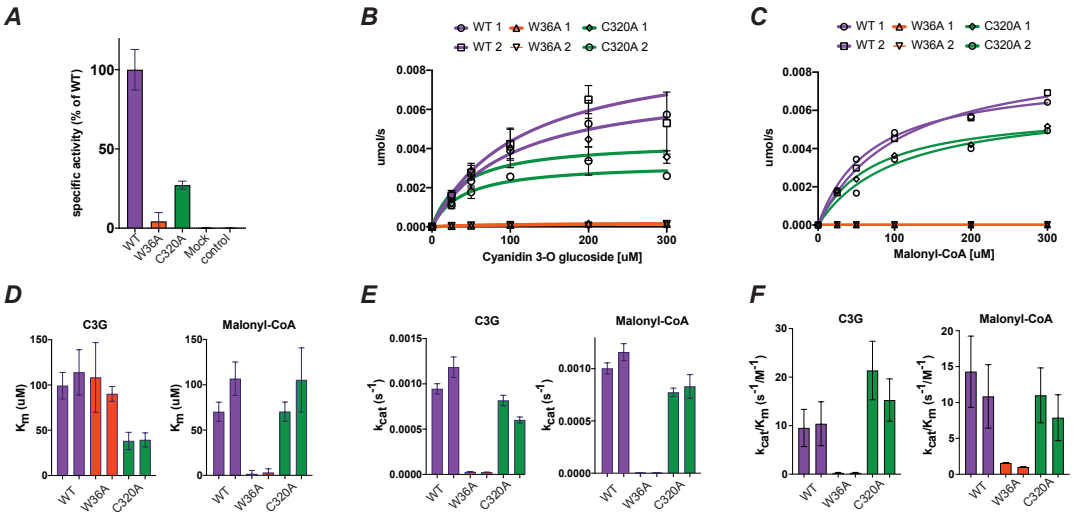


Figure 7. Enzyme activity and enzyme kinetics of Dm3MAT3 wildtype and mutants. (A) Specific activity of Dm3MAT3 wild-type, W36A and C320A mutant. The activity of the wild-type (WT) is set to 100 %. The error bar represents the standard deviation of three expression replicates which were each assayed in technical triplicates. **(B)** Michaelis-Menten curve for Dm3MAT3 WT (WT), Dm3MAT3 W36A (W36A), and Dm3MAT3 C320A (C320A) in dependence of cyanidin 3-O glucoside (C3G). **(C)** Michaelis-Menten curve for the malonyl-CoA donor. **(D)** Km values of WT, W36A and C320. **(E)** k_{cat} values. **(F)** k_{cat}/K_m values.

536 These changes in enzyme activity suggest that Trp36 is catalytically important for
537 the AnAT activity in clades 1a/b. The relatively similar acceptor K_m between wild-type and
538 W36A mutant suggested that the mutation did not affect acceptor substrate binding. Thus,
539 while the MD simulations suggested a quick C3G departure from the active site, the
540 molecule may still remain in the binding cavity due to interactions with other residues, and
541 exit normally. The kinetic assays – specifically the reduced k_{cat} and k_{cat}/K_m values for both

542 acceptor and donor – support MD simulation predictions that catalysis would be
543 drastically affected due to C3G departure from a catalytically competent distance with
544 His170. The K_m of the W36A mutant towards malonyl-CoA was also significantly reduced.
545 Considering AnAT-like acyltransferases are postulated to operate through a two-
546 substrate ordered bi-bi reaction mechanism with malonyl-CoA binding first (Suzuki et al.,
547 2003; Unno et al., 2007), we hypothesize that the reduced catalysis in W36A results in
548 the malonyl-CoA continually bound to the donor binding site, further resulting in a
549 saturation of the available donor sites at lower substrate concentration.

550 In the case of the C320A mutant, the observed effects were less severe than for
551 the W36A mutant (**Fig. 7A**), and the C320A still exhibited a k_{cat} ~47-74% of the WT (**Fig.**
552 **7E**). Compared to the wild-type, the C320A mutant showed an improved K_m towards C3G
553 (**Fig. 7D**). This lower K_m and the only slightly reduced k_{cat} resulted in a 50% improved
554 catalytic rate than WT (**Fig. 7F**). The C320A mutant showed an unchanged K_m and a
555 lower k_{cat} for malonyl-CoA. This resulted in a slightly less efficient enzyme with respect to
556 the donor (**Fig. 7F**).

557 These results support the simulation prediction that Cys320 plays a role in
558 optimizing the enzyme's activity rather than catalysis. While MD simulations indicated that
559 stability of the C3G bound form is reduced, this results in only a slightly decreased k_{cat}.
560 However, we postulate that the acceptor substrate remains in the substrate binding site
561 without catalysis, reducing the K_m to a greater extent than reduction in k_{cat}, thereby
562 mathematically increasing the catalytic efficiency. Combined together, these results
563 suggested that the presence of both Trp36 and Cys320 is necessary for optimal
564 anthocyanin malonyltransferase activity in clades 1a/b, explaining their conservation in
565 AnATs spread out over 150 million years of angiosperm evolution.
566

567 **DISCUSSION**

568 Evolution of functional diversity in large enzyme families that contribute to the
569 metabolic diversity in plants is still incompletely understood. In this study, we extensively
570 characterized the BAHD family to determine how gene duplication and promiscuity
571 contributed to the diversity of BAHD functions. Nine out of eleven BAHDs assayed in this

572 study were substrate-promiscuous **(Fig. 2; Fig. S2; File S2)**, and three (MeHFT, SlHCT,
573 EcCS) were class-promiscuous under the testing conditions **(Fig. 2A, Fig. 1)**. We only
574 tested nine substrate types (including glucose/sucrose), and the specialized enzymes
575 may still exhibit class-promiscuity with other, untested classes under different conditions.
576 More enzymes that cross class boundaries are known from previous studies **(Fig. 4)**.
577 Nonetheless, these observations suggest that while most BAHDs are not able to
578 discriminate between structurally related compounds (e.g. shikimate, quinate for SlHQT
579 or putrescine, agmatine for SlACT), they are able to differentiate between very different
580 chemical scaffolds. For individual enzymes, these adaptations for specialization may
581 constrain them and their duplicates from traversing large distances in the phytochemical
582 space. On the other hand, the class-promiscuity of multiple enzymes such as SlHCT,
583 AtHCT, SmHCT1, MeHFT, EcCS, alcohol acyltransferases provides a foundation on
584 which duplicates may traverse larger distances in the phytochemical space and "plug into"
585 emerging metabolic pathways. Since such class-promiscuity is not typically assayed, it
586 may be much more prevalent than currently known and may form an important basis for
587 metabolic diversification. Significant presence of class-promiscuity can also confound
588 evolutionary inferences and thus, functional annotation for enzyme family members.
589 Thus, more studies to determine if there are any rules for class-promiscuity are needed.
590 Such studies may also reveal new insights about the nature of selection on protein
591 structural features that enable promiscuity.
592 The integrative analysis of the AnAT activity can be interpreted following the
593 previously proposed potentiation-actualization-refinement model of emergence of new
594 functions (Blount et al., 2008). Our results suggest that the ability to acylate anthocyanins
595 was already actualized (first appeared/already existed) in ancestral BAHDs prior to land
596 plant evolution, or that the ancestral enzymes were potentiated to evolve the AnAT activity
597 through different routes. The refinement of the AnAT activity in clades 1a/b required
598 fixation of two residues, one of which (Trp36) is critical for positioning anthocyanins in the
599 active site and the second (Cys320) for positioning the first residue. Both these residues
600 together optimize the AnAT activity by affecting acceptor binding, however, they are likely
601 not sufficient to confer the AnAT activity. The median clade 1a/b AnAT identity is itself
602 ~50%, hence identifying all residues necessary to transform a class A/B utilizing enzyme

603 to class D substrate utilizing enzyme is challenging. Comparative sequence and structural
604 analysis can help further identify regions in the protein that could be tested by
605 mutagenesis.

606 An important question we asked was whether the structural diversity of substrates
607 used across BAHDs is related to an ancestral ability to use these substrates or
608 emergence of new activities through multiple rounds of duplication-neo-functionalization.
609 Our results can be discussed in the context of two models of BAHD evolution **(Fig. 8)**. In
610 both models, the most likely ancestral activity in the root node prior to BAHD expansion
611 is aromatic alcohol acylation. Based on the fact that many HCT/HQT enzymes also
612 acylate aromatic amines **(Fig. 4)**, class A substrates as a whole are the likely ancestral
613 BAHD substrate space **(Fig. 8A,B)**. For other activities, strong evidence for their
614 existence exists only in the ancestor of all land plants. For example, in our enzyme assays
615 **(Fig. 2A)** or the accumulated database **(Fig. 4)**, we do not see any enzyme from any
616 *orange/yellow* (extending to LCA of dicot-liverwort and dicot-mosses, respectively) clades
617 acylating both aliphatic and aromatic alcohols, but these activities are seen together in
618 seed plant clades 1c, 3a, 6a, 5b, 7c/d. Also, many aliphatic alcohol acyltransferases can
619 also acylate monoterpenoids. Thus, while it is possible that the 1-5 root node BAHDs
620 acylated aliphatic alcohols and thus, monoterpenoids, we cannot be certain about this
621 inference. Finally, the presence of anthocyanin acylation is seen in MeHFT, SlHCT and
622 multiple distantly related clades in the BAHD phylogeny **(Fig. 4)**. While this could
623 represent multiple instances of convergent evolution, a simpler explanation is existence
624 of this activity at least among the dozen-odd BAHDs in the land plant ancestor as a
625 promiscuous activity. No evidence exists for the assignment of this activity to the root
626 node, but given presence of class-promiscuous enzymes, this possibility cannot be
627 completely ruled out **(Fig. 8B)**. These inferences reveal that most of the currently known
628 BAHD acyltransferase substrate space was already accessible by the time BAHDs
629 started drastically expanding, either as fixed or unfixed, promiscuous activities. Compared
630 to simpler prior models of individual enzyme evolution such as the “patchwork model”
631 (Jensen, 1976; Matsumura and Ellington, 2001) and the “specialization-generalization-
632 specialization” model (Aharoni et al., 2005), our study of enzyme family reveals a complex
633 picture where ancestral promiscuity played a central role in incorporation of member

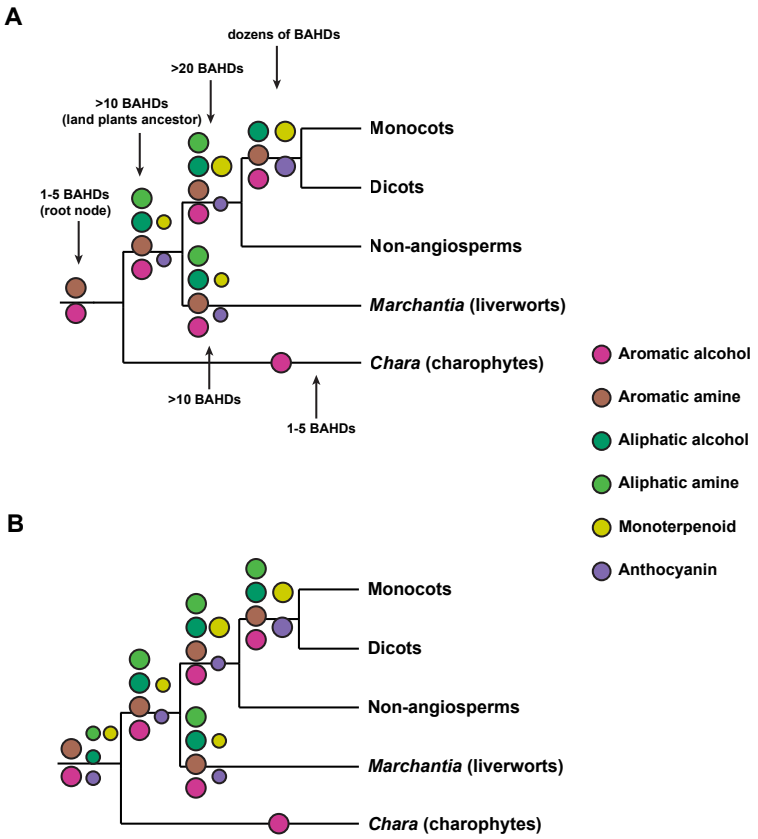


Figure 8. Two models of evolution of BAHD multi-functionality. (A) Conservative model: Only aromatic alcohols and amines are acylated in the root node. First steps of expansion in the branch towards the land plants ancestor leads to emergence of new activities that are fixed in the ancestor (big circles) as well as other promiscuous activities that remain unfixed (smaller circles) **(B)** Relaxed model: The 1-5 BAHDs in the LCA of charophytes and land plants are capable of utilizing most of the substrates in the current known substrate space. Specializations evolve in lineage-specific duplications. The aliphatic alcohol acyltransferase enzyme is likely a different one than the aromatic acyltransferase enzyme in the root node ancestor. The first detected monoterpenoid activities appear in the green angiosperm-gymnosperm ancestral clade **(Fig. 4)**, however, strong secondary activities were also detected in MeHFT and SIHCT.

634 enzymes into new pathways. Both of these models likely apply to evolution of individual
635 BAHD enzymes.

636 Our inferences are based on ~140 BAHDs, and activities of many BAHDs are
637 unknown. For example, to the best of our knowledge in tomato, only 10 out of 100 BAHDs
638 have an experimentally determined function (**Table S2**) (Niggeweg et al., 2004; Goulet et
639 al., 2015; Fan et al., 2016) (including this study), and >80% of BAHDs have a generic
640 domain annotation. It is possible that uncharacterized BAHDs may reveal new primary
641 activities *in vitro* and *in vivo*. However, our meta-analysis suggests that – despite the

642 dangers of class-promiscuity – some clade relationships **(Fig. 4)** can provide a strong
643 functional signal. Combined with other enzyme attributes, these clade relationships can
644 be useful in predicting putative BAHD substrate class utilizations with greater accuracy
645 than simply a similarity-based approach. e.g. using machine learning (Mahood et al.,
646 2020). At the same time, among the 42 multi-species OGs not represented among
647 characterized BAHDs, many show broad conservation across angiosperms and vascular
648 plants. These uncharacterized OGs may indeed reveal novel BAHD activities. As new
649 substrate classes are identified, it would be informative to test – using methods similar to
650 those employed in this study – whether the ability to use those classes also exists across
651 the BAHD family.

652 Findings from this study also provide insights into the role of enzyme families in
653 metabolic diversification in plants. BAHDs are fast-evolving enzymes; even in the well-
654 conserved clade 5a, where all known enzymes are associated with substrates related to
655 lignin biosynthesis (**Fig. 4**), the median percent identity is only ~60%. The ability to use
656 aromatic alcohols such as shikimate and quinate exists in BAHDs that are only 30-40%
657 identical to the clade 5a BAHDs. In other words, these enzymes are robust in retaining
658 their aromatic alcohol acyltransferase activity despite changes to ~70% of their sequence.
659 Similar behavior is observed for acylsugar acyltransferases (Moghe et al., 2017) and
660 aliphatic alcohol acyltransferases **(Fig. 4)**. While being robust, the class-promiscuity of
661 BAHDs and thus, the ability to specialize in one of the classes via duplication, would also
662 make them evolvable. The paradox of a biological system being both robust and
663 evolvable at the same time has previously been addressed in detail (Wagner, 2005, 2008;
664 Bloom et al., 2006; Pigliucci, 2008; Tokuriki and Tawfik, 2009; Payne and Wagner, 2019);
665 these properties may enable enzyme families to become involved in newly emerging
666 metabolic pathways or detoxify harmful metabolites without compromising their core
667 activities. The examples of aromatic alcohol, anthocyanin and terpenoid acylating
668 enzymes – whose OGs have a much narrower phylogenetic spread than the predicted
669 spread of the biochemical activity – also highlight a common theme in metabolism, where
670 enzyme activities may exist promiscuously for millions of years prior to the actual
671 genome-level signatures of their fixation. While presence of substrate promiscuity is
672 helpful in this regard, having class-promiscuous enzymes may also offer an added

673 benefit. Presence of robust evolvable enzymes is likely an important feature of metabolic
674 networks and needs to be studied in greater detail in the context of biochemical evolution.
675 The present study describes an integrative analysis of enzyme evolution,
676 biochemistry and structure-function relationships that captures potentially emergent
677 behaviors of enzyme families in the form of robustness and evolvability. The described
678 analysis can serve as a template for characterizing class-promiscuity in enzyme families,
679 although more high-throughput means of analysis are needed. Our results also identify
680 patterns in duplication-divergence of BAHDs that can be explored in other large enzyme
681 families to determine their involvement in metabolic diversification in plants.
682

683 ## MATERIALS AND METHODS

684 ***Creation of a database of biochemically characterized enzyme activities***

685 Biochemically characterized BAHD enzymes were gathered through an extensive
686 literature search. Only sequences belonging to the BAHD acyltransferase protein fold
687 (PFAM domain: PF02458) were considered. To ensure a high level of confidence for *in*
688 *vitro* activities and the resulting substrate-enzyme pairs, only enzymes that were subject
689 to *in vitro* biochemical assays were considered. For each enzyme, all tested acceptor and
690 donor substrates with their associated PubChem CID, the associated chemical structure,
691 cDNA and protein sequence as well as the species name from which the gene was
692 isolated were compiled **(File S2)**.
693

694 ***Generation of substrate similarity networks***

695 Structures of known BAHD substrates were downloaded from the PubChem
696 database using the PubChem ID as structure-data file (sdf) format. Substrates not found
697 in the PubChem database were created using ChemDraw, exported in the MOL data
698 format and manually brought into sdf format. To calculate substrate similarity based on
699 the maximum common substructure, overlap coefficient (MCS-Overlap), the R packages
700 ChemmineR v3.34.1 and fmcsR v1.24.0 were used (Cao et al., 2008; Wang et al., 2013)
701 with default values, except for the time threshold for the comparison of two molecules set

702 to 12s. The similarity network was visualized using Cytoscape v3.8.0 (Shannon et al.,
703 2003). Additional plant specific compounds were downloaded from the Chemical Entities
704 of Biological Interest (ChEBI) database (Hastings et al., 2013) and only compounds were
705 chosen that were also represented in the plant-centric KNApSAcK database (Afendi et
706 al., 2012),
707
708 ***Identification of BAHD acyltransferases and OG assignment***

709 For identification of BAHD acyltransferases from the analyzed genomes**,** we used
710 hmmsearch v3.1b2 (Potter et al., 2018) using the BAHD PFAM domain (PF02458) with
711 all default parameters except *cut_ga* as the hit significance threshold. OGs were
712 constructed using OrthoFinder v2.3.3 (Emms and Kelly, 2019) with default parameters
713 except an inflation parameter of 1.5 to make larger OGs. After defining OGs between
714 these species, we used blastp (Camacho et al., 2009) to assign biochemically
715 characterized enzymes to OGs. For each enzyme, its individual phylogenetic
716 conservation was determined based on the phylogenetic spread of its assigned OG **(File**
717 **S4)**. An internal node in the species tree was assigned an activity if multiple characterized
718 enzymes shared their conservation up to that node.
719
720 ***Creating a time-calibrated species phylogeny for ancestral state reconstruction***

721 While recent attempts to infer a well-sampled land plant phylogeny (Gitzendanner
722 et al. 2018) and estimate divergence times (Nie et al. 2019) of green plants exist, a time
723 calibrated phylogeny encompassing the breath of taxa included in our dataset does not.
724 In order to obtain one, we first obtained a time calibrated phylogeny from TimeTree, a
725 database of synthesized time calibrated studies (Kumar et al. 2017). Because not all taxa
726 were present in TimeTree we used 'proxy taxa' (closest relative represented in TimeTree)
727 to assign a position within the phylogenetic tree. **(File S7A)**. Using this approach, we
728 accounted for all taxa in our database with the exception of *Klebsormidium nitens*
729 (charophytic algae), which was manually added to the most likely position according to
730 (Leliaert et al., 2012). To obtain divergence time for this group, we recalibrated the
731 phylogeny using a penalized likelihood under a relaxed clock model (Sanderson, 2002),

732 a lambda=0 and 95% CI intervals for major clades, inferred from previous studies
733 (Magallón et al., 2015; Nie et al., 2020) **(File S7B)**, with the function chronos() in R
734 package ape v.5.4 (Sanderson, 2002; Paradis and Schliep, 2019).
735
736 ***BAHD family size evolution***

737 In order to determine the expansion dynamics of the BAHD gene family, we
738 modeled the evolution of normalized BAHD gene counts. Counts were normalized against
739 predicted total number of coding sequences in each genome with a methionine start
740 residue that were longer than 100bp. If >30% of the gene models did not fit these criteria,
741 these genomes (9 gymnosperm, 2 red algae) were eliminated from further analyses. We
742 fitted normalized counts and our phylogeny using Evolutionary Brownian Motion (BM) and
743 Bounded Brownian Motion (BBM) type models and performed model selection using the
744 R package BBMV v2.1 with default parameters except those specified below (Boucher et
745 al. 2016). BBM is a special case of Brownian motion (BM) where values are constrained
746 between a minimum and a maximum value (Boucher et al. 2016). We used a minimum
747 bound of 0, as negative gene copy is biologically nonsensical, and a maximum bound of
748 0.02914. It was suggested that the maximum bound should be set to the largest value in
749 the dataset (in our data set *Petunia axillaris*: 0.00679) (Boucher et al. 2016). However,
750 this value is value may not realistically represent the upper bound of the gene count for
751 an existing gene family. Therefore, we used *Arabidopsis thaliana*, the best annotated
752 plant genome, to identify the largest gene family present, which was the EAR repressome
753 family with 403 members (https://www.arabidopsis.org/browse/genefamily/index.jsp;
754 accessed August 20, 2020). We normalized this value against the total number of gene
755 in the Arabidopsis genome (https://www.arabidopsis.org/browse/genefamily/index.jsp;
756 accessed August 20, 2020) and set the upper bound of the model at 2x the normalized
757 gene number in this family (upper normalized limit = 0.02914). We acknowledge the
758 potential existence of gene families in nature that are larger but assume that this limit
759 should properly account for a majority of potential gene families.
760
761 ***Plant material, RNA extraction and cDNA synthesis***

762 Plant material from the cultivated tomato variety M82 was used for cloning tomato
763 BAHDs. Plants were grown in a growth chamber under constant light/dark (16 h/8 h)
764 regime at 24 ºC for 8 weeks until first flowers appeared. For RNA extraction, young leaves
765 were cut from a single plant using clean tweezers and scalpel, immediately flash frozen
766 using liquid nitrogen (liq. N2) and stored at -80 ºC until further use. Total RNA was
767 extracted using the E.Z.N.A Plant RNA Kit (Omega Bio-tek, Norcross, GA) as per
768 manufacturers protocol with on-column DNAse digestion. cDNA was synthesized using
769 the Protoscript II Reverse Transcriptase (New England Biolabs, Ipswich, MA [NEB]) with
770 Oligo-dT_{17} primer **(Table S3)** at 45 ºC for one hour. After heat inactivation for 20 minutes
771 at 80 ºC, the reaction mix was diluted with four volumes of nuclease-free water and stored
772 at -20 ºC until further use.
773

774 ***Amplification of candidate enzymes, cloning of expression constructs, and site-***
775 ***specific mutagenesis***

776 For amplifying cDNA sequences of candidate BAHD enzymes for cloning, the Q5
777 Hot Start High Fidelity DNA Polymerase (NEB) with gene specific primers was used
778 **(Table S3)**. To allow fast and easy cloning using the Gibson assembly, the primers
779 contained matching overlaps to the pET28b vector, with successful insertion resulting in
780 *N*-terminal fusion of the candidate enzyme with a 6x His tag. After successful
781 amplification, the PCR product was purified using the E.Z.N.A Cycle Pure Kit (Omega
782 Bio-tek, Norcross, GA) and eluted in 30 µl nuclease-free water. First, the pET28b vector
783 was linearized using BamHI and XhoI restriction enzymes (NEB) and purified using the
784 E.Z.N.A Cycle Pure Kit. Second, 50 ng of linearized vector was mixed with 100 ng of
785 purified PCR product and incubated with HiFi DNA Assembly Master Mix (NEB) according
786 the manufacturers protocol. After assembly, 3 µl of the reaction mix were used for
787 transformation of *E. coli* 10-beta (NEB) cells. The transformation mix was plated on Luria-
788 Bertani (LB) medium containing kanamycin (50 µg/ml) and streptomycin (50 µg/ml).
789 Grown colonies were screened using colony PCR with construct-specific primers. Positive
790 clones were confirmed using Sanger sequencing at the Cornell Institute of Biotechnology.
791 In case of Dm3MAT3, the full-length gene was codon-optimized, synthesized and cloned

792 into pET28b by Gene Universal (Newark, DE). For cloning the C. braunii BAHD, the *C.*
793 *braunii* genome sequence (Nishiyama et al., 2018) was used. Previous studies postulated
794 presence of a BAHD orthologous to HCT potentially involved in a progenitor of lignin
795 biosynthesis (de Vries et al., 2017; Renault et al., 2019) however, no experimentally
796 characterized functions are available. We chose two out of four *C. braunii* BAHD enzymes
797 (PFAM: PF02458) that contained an intact catalytic HxxxD motif. The gene was
798 synthesized in two overlapping (42 bp overlap) parts (gBLOCKS) (Integrated DNA
799 Technologies, Coralville, IA [IDT]) and assembled into the pET28b vector as described
800 above. All constructs were sequenced before transformation into *E. coli* Rosetta2 cells
801 (EMD-Millipore-Sigma, Burlington, MA) for heterologous expression. Site-specific
802 mutagenesis of Dm3MAT3 was conducted using the Q5 site-directed mutagenesis kit
803 (NEB) **(Table S3)**.
804
805 ***Heterologous expression* and *protein purification***

806 Protein expression was induced in mid-log phase growing *E. coli* cells containing
807 the respective expression construct after addition of 0.5 mM isopropyl-β-D-
808 thiogalactopyranoside (IPTG). The cultures (300 ml) were incubated for 16 hours at 25°C
809 in LB medium supplemented with 50 µg/ml kanamycin and 50 µg/ml chloramphenicol
810 while shaking at 250 rpm. For protein extraction, the overnight grown cells were pelleted
811 by centrifugation for 15 minutes at 5000g and re-suspended in lysis buffer (50 mM
812 NaH2PO4, 300 mM NaCl, 10 mM imidazole, pH 8.0), to which lysozyme was added
813 (~1 mg/ml) and incubated for 1 hour at 4°C. Subsequent sonification (intensity 90%,
814 Lawson Scientific, Harrogate, United Kingdom) for 10 minutes was conducted to shear
815 remaining intact cells. After centrifugation for 30 min at 21,000 g to remove cell debris
816 and insoluble proteins the cell-free lysate was incubated with Ni-NTA-Agarose matrix
817 (Qiagen, Germantown, MD [Qiagen]) for 30 min. After removing the lysate, the matrix was
818 washed with 10 column volumes (cv) wash buffer (50 mM NaH2PO4, 300 mM NaCl,
819 20 mM imidazole, pH 8.0) and then added to a 2 ml spin column (BioRad, Hercules, CA)
820 equipped with a frit. The matrix was washed again with 5 cv wash buffer after which
821 proteins were eluted three to four times using one cv elution buffer (50 mM NaH2PO4,
822 50 mM NaCl, 250 mM imidazole, pH 8.0). Fractions with similar protein content were

823 combined and the protein concentration was determined in triplicate using the Bradford
824 method (Bradford, 1976).
825
826 ***Size exclusion chromatography***
827 Protein was expressed from *E. coli* BL21 grown at 37°C with 50 µg/mL kanamycin
828 to select for desired cells. When A_{600} values reached approximately 0.4, cells were
829 induced with 0.1 M IPTG, cooled to 15°C, and incubated with shaking for 18 hours. Cells
830 were then pelleted with centrifugation, supernatant discarded, and cell pellets frozen at -
831 80°C until ready for next steps. Cell pellets were thawed on ice, then resuspended in lysis
832 buffer (50 mM tris pH 8, 20 mM imidazole, 500 mM NaCl, 10% glycerol, 1% Tween-20),
833 and lysed by sonication. Lysate was centrifuged to pellet debris, and the supernatant was
834 passed through 2 mL of Ni-NTA resin. Resin was washed with buffer (50 mM Tris pH 8,
835 20 mM imidazole, 500 mM NaCl, 10% glycerol), then protein eluted in a single fraction
836 (50 mM Tris pH 8, 250 mM imidazole, 500 mM NaCl, 10% glycerol). To the fraction
837 containing desired protein, 30 U of thrombin was added, the entire solution was
838 transferred to dialysis tubing, and dialyzed overnight in 1 L of wash buffer. The resulting
839 solution was removed from dialysis tubing and passed through a column containing 300
840 µL Ni-NTA resin and 200 µL Benzamidine-sepharose resin. The flowthrough was
841 collected and further purified by FPLC size-exclusion chromatography (Hi-Load 26/600
842 S-200 column, buffer contained 100 mM NaCl, 25 mM HEPES, pH 7.5).
843
844 ***Biochemical synthesis of p-coumaroyl-CoA and feruloyl-CoA***

845 For enzyme assays using p-coumaroyl-CoA and feruloyl-CoA as acyl donor, the
846 donor was synthesized biochemically using 4-coumaroyl-CoA ligase (4CL) from tomato
847 as described previously (Beuerle and Pichersky, 2002). The cDNA of tomato 4CL was
848 amplified using gene specific primers **(Table S3)**, cloned into pET28b, heterologously
849 expressed, and purified as described above. To generate the CoAs, 5 mM MgCl2, 2 mM
850 ATP, 2 mM p-coumaric acid (TCI America, Portland, OR) or ferulic acid (MP Biomedicals,
851 Solon, OH), and 3 mM coenzyme A (MP Biomedicals) in 50 mM sodium phosphate buffer
852 (pH 8.0) were incubated for 24 hours at room temperature (Beuerle and Pichersky, 2002).
853 The reaction was stopped by heating the reaction mix to 70 ºC for 20 min. After

854 centrifugation at 21,000 g for 30 min to remove precipitated protein the supernatant was
855 transferred into a new reaction tube and stored at -20 ºC until further use. The final
856 concentration of activated donor was estimated to be 1.6 mM based on the initial
857 substrate concentrations and an assumed conversion rate of 80% (Beuerle and
858 Pichersky, 2002).
859
860 ***Enzyme assays***

861 In general, enzyme assays used for screening the substrate promiscuity of
862 selected BAHD enzymes were performed in 50 µl reactions containing 50 mM sodium
863 phosphate buffer (pH 7.2), supplemented with 100 µM acceptor substrate, and 300 µM
864 donor substrate. The assay was started by the addition of 20 µg purified enzyme.
865 Preliminary enzyme assays with lower substrate concentrations did not show any
866 indications for substrate inhibition at such substrate concentrations. For each enzyme,
867 three replicates per substrate were performed and incubated for 1 hour. Reactions were
868 stopped by adding 100 µl of a mixture of isopropanol, acetonitrile, and water (ratio of 2:2:1
869 + 0.1% (v/v) formic acid) containing 15 µM of the internal standard telmisartan. The
870 reaction mix was centrifuged for 10 min at 21,000 g to remove precipitated proteins,
871 transferred into LC vials, and stored at -20 ºC until LC-MS analysis. Mock controls using
872 purified protein extracts from empty vector transformed *E. coli* were run to exclude *E. coli*
873 background activity, for each of the used acceptor and donor combinations.
874 To test whether SlACT, CbHCT, and MeHFT can use additional substrates from
875 the aromatic alcohol, aliphatic and aromatic amine class not captured using 100 uM of
876 substrate, a modified version with higher substrate concentrations were used (assay
877 version 2). A 50 µl reaction containing 50 mM sodium phosphate buffer (pH 7.2), 1 mM
878 acceptor substrate, 300 µM coumaroyl-CoA was started using 20 µg of the respective
879 enzyme, incubated for 1 hour, stopped, and subsequently run on LC-MS as described
880 above.
881 Enzyme assay reactions for screening for enzyme activities were measured
882 individually on the LC-MS and each product was detected with specific PRM parameters
883 (see below). In order to determine more detailed enzyme kinetics for anthocyanin
884 acylating enzymes, a modified version of the above described enzyme assays was used

885 (assay version 3). Varied concentrations of cyanidin-3-O-glucoside (25, 50, 100, 200,
886 300 μM) and malonyl-CoA (25, 50, 100, 200, 300 μM) were used to determine enzyme
887 kinetics. The counter substrate was kept at a saturating concentration of 300 μM. Two
888 biological replicates (independent enzyme expression and purification) were used and
889 Km, Kcat and Vmax were determined using non-linear curve fitting in Prism 8 (GraphPad,
890 San Diego, CA). All enzyme reactions were measured under initial rate conditions and
891 stopped as described above. The wildtype enzyme reactions were incubated for 10
892 minutes and mutant reactions for 20 and 30 minutes for C320A and W36A variants,
893 respectively.
894
895 ***LC-MS measurements***

896 LC-MS analysis was performed on a ThermoScientific Dionex Ultimate 3000 HPLC
897 equipped with an autosampler coupled to a ThermoScientific Q-Exactive HF Orbitrap
898 mass spectrometer using solvent A (water + formic acid (0.1% v/v)) and solvent B
899 (acetonitrile + formic acid (0.1%)) at a flow rate of 0.6 ml/min. Products of enzyme
900 reactions were detected with specific PRM methods using their predicted parent ion mass
901 in positive or negative ionization mode with an isolation window of 2 m/z **(Table S4)**.
902 Additional details of chromatographic and mass spectrometric methods are described in
903 **Table S4 and S5**. LC-MS data was analyzed with the ThermoScientific Dionex
904 Chromeleon 7 Chromatography Data System v7.2 software. Peaks were selected using
905 their specific masses (**Table S4**) and default peak detection parameters.
906
907 ***Phylogenetic analysis***

908 Protein sequence alignment was generated using MAFFT v.7.453-with-extensions
909 (Katoh et al., 2002) using following parameters: *--maxiterate 1000 --genafpair --thread*
910 *70*. The alignment was inspected manually to ensure proper alignment e.g. by inspecting
911 that known motifs like the HxxxD and DFGWG motif are aligned properly. Afterwards, IQ-
912 Tree v.1.6.10 (Nguyen et al., 2015) was used to infer a phylogenetic tree using following
913 parameters: *-st AA -nt AUTO -ntmax 70 -b 1000 -m TEST* after model selection

914 (LG+F+G4) using ModelFinder (Kalyaanamoorthy et al., 2017). Tree visualization was
915 created using iTol v.5.6.2 (Letunic and Bork, 2019).
916
917 ***Identification of enriched motifs in anthocyanin acylating enzyme***

918 For identifying enriched motifs in a specific clade of anthocyanin acylating
919 enzymes, we used MEME v.5.0.5 (Bailey et al., 2009) in discriminative mode using default
920 parameters but the maximum number of motifs to find set to 5. Positive and negative
921 examples were set as described in **Fig. S5**. The TFFDxxW and YFGNC sub-motifs were
922 selected for further analysis due to their high degree of conservation and their proximity
923 to the active site within the Dm3MAT3 protein. Clade-wise single residues were identified
924 using custom Python scripts.
925
926 ***Prediction of ancestral sequence of AnATs***

927 Ancestral sequence reconstruction was performed using IQ-TREE v1.6.10
928 (Nguyen et al., 2015). Twenty randomly selected sequences from each of the three
929 orthologous groups representing clades 1a-c and 5a (outgroup) were combined together
930 with previously characterized BAHDs from these clades. All protein sequences were
931 aligned using MAFFT v7.453 and provided as input to IQ-TREE, which was run with
932 model selection and ancestral state reconstruction with 1000 standard bootstraps. The
933 optimal tree was obtained using the JTT+I+G4 model. Per-site posterior probabilities of
934 the ancestral state prediction were filtered using a threshold of 0.95, and the resultant
935 FASTA sequence at each ancestral node was extracted using a custom Python script.
936
937 ***Preparation of PDB structures for docking and molecular simulations***

938 The *holo* structure of Dm3MaT3 bound to malonyl-CoA (MLC), the acyl-donor, was
939 retrieved from the Protein Data Bank (PDB: 2E1T) (Berman et al., 2000; Unno et al.,
940 2007).After treatment with the PROPKA-plugin in VMD to verify residue charge states at
941 pH 7.0, (Humphrey et al., 1996; Rostkowski et al., 2011) *holo*-Dm3MaT3 was then
942 submitted to the Solution Builder Input Generator in CHARMM-GUI (Jo et al., 2008).

943 Because 2E1T was crystallized as a dimer, only segment “A” of the protein was input to
944 Solution Builder. The N- and C-terminal residues were modeled and patched using the
945 ACE and CT3 terminal groupings, respectively. For the W36A and C320A Dm3MaT3
946 mutants, the respective point mutations were also made during pdb structure
947 manipulation. Resulting Dm3MaT3 models were solvated with 150 mM NaCl, neutralized,
948 and output in Amber format. PyMol 2.3.3 was then used to perform structural alignment
949 between the Dm3MaT3 models and the original 2E1T structure containing MLC so that
950 the donor molecule could be introduced in the bound pose to each Dm3MaT3 model
951 (Schrödinger, LLC, 2015). *Apo* simulations were prepared in a similar fashion but from
952 the 2E1U PDB file (Unno et al., 2007).
953 The MLC structure file was taken directly from the 2E1T PDB file and edited in the
954 Maestro 2017-3 Release for improved structural resolution by adding hydrogens not seen
955 in the crystal structure and revising charge states at pH 7.0 (Schrödinger, LLC, 2020).
956 Cyanidin 3-*O*-glucoside (C3G), the acyl-acceptor, was also edited in Maestro to reflect an
957 accurate charge state at pH 7.0 (Schrödinger, LLC, 2020). Both ligands were introduced
958 to antechamber and parameterized with the AM1-BCC charge model (Jakalian et al.,
959 2002). At pH 7.0, MLC bears a net charge of -5 while C3G has a net charge of 0, although
960 it is zwitterionic due to its oxonium moiety.
961
962 ***Docking of acyl acceptor into donor-containing Dm3MaT3 structures***

963 *Holo*-Dm3MaT3 was converted into a receptor pdbqt file for docking via AutoDock4
964 command-line (Morris et al., 1998). The box for docking C3G into each *holo* Dm3MaT3
965 variant (wild-type, WT; and mutants W36A and C320A) was determined in VMD
966 (Humphrey et al., 1996). Autodock Vina v1.1.2 was then used for docking. A docking
967 procedure was then performed using the previously determined box size and an
968 exhaustiveness score of 8 (Trott and Olson, 2010). A low exhaustiveness score was used
969 to obtain greater diversity of bound poses. This procedure was repeated 30 times in order
970 to generate a greater number of different starting seeds for C3G, where bound poses for
971 each iteration were evaluated based on the proximity of the reactive cyanidin 3-*O*-
972 glucoside 6”-hydroxyl to be ≤ 4.0 Å from the thioate moiety of MLC (Unno et al., 2007),

973 as proximity between the two ligands would serve as a proxy for an intermediate step
974 occurring within the reaction.
975 From this docking procedure, 268 C3G poses were generated for WT Dm3MaT3,
976 270 poses for W36A Dm3MaT3, and 261 poses for C320A Dm3MaT3. Regardless of the
977 orientation of the docked C3G, the pose selected for MD system assembly was that which
978 had the most favorable energy among those with the shortest distance (less than or equal
979 to 4 Å) from the MLC thioate. The C3G poses selected for simulation had docking scores
980 of -5.8 kcal/mol (WT), -7.6 kcal/mol (W36A), and -6.4 kcal/mol (C320A) **(Fig. S10)**.
981
982 ***Molecular simulation setup***

983 Each Dm3MaT3 system (WT, W36A, C320A) was then reassembled using the
984 *holo*-aligned protein and MLC structures, the AutoDock Vina output, 150 mM NaCl, and
985 ~18,500 water molecules in Packmol 18.169 (Martínez et al., 2009). Parameterization
986 and periodic box conditions were then applied using tleap in Amber18 (Case et al., 2018).
987 All ligands were parameterized with GAFF2 (Wang et al., 2004), Dm3MaT3 variants with
988 the AMBER-FB15 (Wang et al., 2017), and water with the TIP3PFB forcefields (L.-P.
989 Wang et al., 2014), respectively. Ions were parameterized using the 2008 parameter set
990 developed Joung and Cheatham (Joung and Cheatham, 2008). Hydrogen mass
991 repartitioning (Hopkins et al., 2015) was then applied to the resulting files in ParmEd 3.2.0
992 (Case et al., 2018). *Apo* systems were prepared in an identical fashion.
993 The following conditions were then applied to all Dm3MaT3 systems during
994 initialization stages. Minimization was performed for 50000 cycles, where steepest
995 descent was used for the first 5000 and then conjugate gradient for the remaining 45000
996 cycles. Each system was heated from 0 to 300 K in the NVT ensemble for 5 ns. The
997 systems were then held at 300 K for 5 ns at NPT. A Berendsen thermostat and barostat
998 were used throughout heating and equilibration stages for efficiency (Braun et al., 2018),
999 where the pressure was maintained at 1 bar using isotropic scaling and temperature at
1000 300 K (Berendsen et al., 1984). Each Dm3MaT3 variant and the bound ligands were
1001 restrained during the heating simulation. Restraints were removed during the 50 ns
1002 equilibration. The SHAKE algorithm was applied to all stages of initialization except for
1003 minimization (Krautler et al., 2001), while the Particle Mesh Ewald method used for

1004 treating long-range electrostatics at a 10 Å cutoff (Darden et al., 1993). A Langevin
1005 thermostat was implemented for temperature maintenance and a Monte Carlo barostat
1006 was implemented for pressure maintenance in the production runs (Loncharich et al.,
1007 1992; Åqvist et al., 2004). Simulations were performed for a total of 1 μs for each of the
1008 six systems (*holo* and *apo* versions of Dm3MaT3 WT, W36A, and C320A).
1009
1010 ***Data analysis***

1011 Trajectories were visualized using VMD 1.9.3 whereas images were rendered
1012 using Chimera 1.14 (Humphrey et al., 1996; Pettersen et al., 2004). Distance
1013 measurements between atoms were conducted using MDTraj 1.9.3 (McGibbon et al.,
1014 2015). Residue-residue face-to-face and edge-to-face stacking interactions were
1015 determined using a script which was adapted from a previously reported analysis
1016 (Ferreira de Freitas and Schapira, 2017). Briefly, the plane of an aromatic residue's pi
1017 system was determined using the same atoms as in GetContacts (*GetContacts*, 2020).
1018 The normal vector was then determined for each aromatic plane under inspection, where
1019 the intersection between the two vectors was solved for. Solving for the angle between
1020 the center of mass of one aromatic ring, the center of mass of another aromatic, and the
1021 intersection of the rings' normal vectors, allowed for the supplementary angle, θ, to be
1022 determined. Angle θ and distance cutoffs were then applied to determine the extent of
1023 aromatic stacking (Ferreira de Freitas and Schapira, 2017; "Schrödinger – Knowledge
1024 Base.," 2020), where exact stacking classifications are detailed in Supplementary
1025 Information. All graphs were generated using Matplotlib 3.2.0 (Hunter, 2007).
1026

ACKNOWLEDGEMENTS

We thank Dr. John D'Auria, Dr. Toru Nakayama and Dr. Sangeeta Dhaubhadel for providing BAHD constructs, Dr. Bennett Fox and Dr. Frank Schroeder for use of and assistance with the LC-MS experiments. We thank Dr. Thomas Stegemann for support in LC-MS method establishment, Elizabeth Mahood and Dr. Nicholas Santantonio for support in establishing the substrate similarity pipeline, Dr. Hening Lin for helpful advice regarding our biochemical experiments, Dr. Elisabeth Kaltenegger for helpful comments on the manuscript, and Anna-Lena Sprick and Dr. Kai Fan for support in the laboratory. We thank Dr. Florian Boucher for assistance with interpreting BBMV R package output.

COMPETING INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

FUNDING INFORMATION

This work was supported by the following funding sources:

- LK/GM: Deutsche Forschungsgemeinschaft award #411255989 (LK); Cornell startup funds (GM)
- JMG/CS: NSF award #DGE-1650441
- JC: NSF REU award #DBI-1850796 to the Boyce Thompson Institute/Georg Jander
- AW/DS: Blue Waters sustained-petascale computing project (NSF awards OCI-0725070 and ACI-1238993, the State of Illinois, and the National Geospatial-Intelligence Agency).

1051 **REFERENCES**

1052 Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S,
1053 Takahashi H, Altaf-Ul-Amin M, Darusman LK, Saito K, Kanaya S. 2012.
1054 KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases
1055 for Multifaceted Plant Research. *Plant Cell Physiol* **53**:e1–e1.
1056 doi:10.1093/pcp/pcr165

1057 Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. 2005. The
1058 "evolvability" of promiscuous protein functions. *Nat Genet* **37**:73–76.
1059 doi:10.1038/ng1482

1060 Aharoni A, Keizer LC, Bouwmeester HJ, Sun Z, Alvarez-Huerta M, Verhoeven HA, Blaas
1061 J, van Houwelingen AM, De Vos RC, van der Voet H. 2000. Identification of the
1062 SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays.
1063 *Plant Cell* **12**:647–661.

1064 Åqvist J, Wennerström P, Nervall M, Bjelic S, Brandsdal BO. 2004. Molecular dynamics
1065 simulations of water and biomolecules with a Monte Carlo constant pressure
1066 algorithm. *Chem Phys Lett* **384**:288–294. doi:10.1016/j.cplett.2003.12.039

1067 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS.
1068 2009. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res*
1069 **37**:W202–W208. doi:10.1093/nar/gkp335

1070 Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R. 2011. The
1071 Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping
1072 Enzyme Parameters. *Biochemistry* **50**:4402–4410. doi:10.1021/bi2002289

1073 Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. 1984. Molecular
1074 dynamics with coupling to an external bath. *J Chem Phys* **81**:3684–3690.
1075 doi:10.1063/1.448118

1076 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN,
1077 Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**:235–242.
1078 doi:10.1093/nar/28.1.235

1079 Beuerle T, Pichersky E. 2002. Enzymatic Synthesis and Purification of Aromatic
1080 Coenzyme A Esters. *Anal Biochem* **302**:305–312. doi:10.1006/abio.2001.5574

1081 Bloom JD, Labthavikul ST, Otey CR, Arnold FH. 2006. Protein stability promotes
1082 evolvability. *Proc Natl Acad Sci* **103**:5869–5874. doi:10.1073/pnas.0510098103

1083 Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a
1084 key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad*
1085 *Sci* **105**:7899–7906. doi:10.1073/pnas.0803151105

1086 Boucher FC, Démery V. 2016. Inferring Bounded Evolution in Phenotypic Characters from
1087 Phylogenetic Comparative Data. *Syst Biol* **65**:651–661.
1088 doi:10.1093/sysbio/syw015

1089 Bradford MM. 1976. A rapid and sensitive method for the quantitation of microgram
1090 quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*
1091 **72**:248–254. doi:10.1016/0003-2697(76)90527-3

1092 Braun E, Gilmer J, Mayes HB, Mobley DL, Monroe JI, Prasad S, Zuckerman DM. 2018.
1093 Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J*
1094 *Comput Mol Sci* **1**:5957. doi:10.33011/livecoms.1.1.5957

1095 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
1096 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.
1097 doi:10.1186/1471-2105-10-421

1098 Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T. 2008. ChemmineR: a compound mining
1099 framework for R. *Bioinformatics* **24**:1733–1734. doi:10.1093/bioinformatics/btn307

1100 Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TEI, Cruziero VWD, Darden
1101 TA, Duke RE, Ghoreishi D, Gilson MK, Gohlke H, Goetz AW, Greene D, Harris R,
1102 Homeyer N, Izadi S, Kovalenko A, Kurtzman T, Lee T-S, LeGrand S, Li P, Lin C,
1103 Liu J, Luchko T, Luo R, Mermelstein DJ, Merz KM, Miao Y, Monard G, Nguyen C,
1104 Nguyen H, Omelyan I, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C,
1105 Schott-Verdugo S, Shen J, Simmerling CL, Smith J, Salomon-Ferrer R, Swails J,
1106 Walker RC, Wang J, Wei H, Wolf RM, Wu X, Xiao L, York DM, Kollman PA. 2018.
1107 AMBER. San Francisco, CA: University of California, San Francisco.

1108 Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, Sun W, Li X, Xu Y, Zhang Y, Wittek S,
1109 Reder T, Günther G, Gontcharov A, Wang S, Li L, Liu X, Wang J, Yang H, Xu X,
1110 Delaux P-M, Melkonian B, Wong GK-S, Melkonian M. 2019. Genomes of Subaerial
1111 Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* **179**:1057-
1112 1067.e14. doi:10.1016/j.cell.2019.10.019

1113 Chiang Y-C, Levsh O, Kei Lam C, Weng J-K, Wang Y. 2018a. Active Site Dynamics and
1114 Substrate Permissiveness of Hydroxylcinnamoyltransferase (HCT). *Biophys J*
1115 **114**:583a–584a. doi:10.1016/j.bpj.2017.11.3193

1116 Chiang Y-C, Levsh O, Lam CK, Weng J-K, Wang Y. 2018b. Structural and dynamic basis
1117 of substrate permissiveness in hydroxycinnamoyltransferase (HCT). *PLOS*
1118 *Comput Biol* **14**:e1006511. doi:10.1371/journal.pcbi.1006511

1119 Copley SD. 2015. An evolutionary biochemist's perspective on promiscuity. *Trends*
1120 *Biochem Sci* **40**:72–78. doi:10.1016/j.tibs.2014.12.004

1121 Darden T, York D, Pedersen L. 1993. Particle mesh Ewald: An N·log(N) method for Ewald
1122 sums in large systems. *J Chem Phys* **98**:10089–10092. doi:10.1063/1.464397

1123 D'Auria JC. 2006. Acyltransferases in plants: a good time to be BAHD. *Curr Opin Plant*
1124 *Biol* **9**:331–340. doi:10.1016/j.pbi.2006.03.016

1125 Davies KM, Jibran R, Zhou Y, Albert NW, Brummell DA, Jordan BR, Bowman JL, Schwinn
1126 KE. 2020. The Evolution of Flavonoid Biosynthesis: A Bryophyte Perspective.
1127 *Front Plant Sci* **11**:7. doi:10.3389/fpls.2020.00007

1128 de Vries J, de Vries S, Slamovits CH, Rose LE, Archibald JM. 2017. How Embryophytic
1129 is the Biosynthesis of Phenylpropanoids and their Derivatives in Streptophyte
1130 Algae? *Plant Cell Physiol* **58**:934–945. doi:10.1093/pcp/pcx037

1131 Donoghue P, Paps J. 2020. Plant Evolution: Assembling Land Plants. *Curr Biol* **30**:R81–
1132 R83. doi:10.1016/j.cub.2019.11.084

1133 Dudareva N, D'Auria JC, Nam KH, Raguso RA, Pichersky E. 1998. Acetyl-
1134 CoA:benzylalcohol acetyltransferase – an enzyme involved in floral scent
1135 production in *Clarkia breweri*. *Plant J* **14**:297–304. doi:10.1046/j.1365-
1136 313X.1998.00121.x

1137 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
1138 genomics. *bioRxiv* 466201. doi:10.1101/466201

1139 Eudes A., Mouille M, Robinson DS, Benites VT, Wang G, Roux L, Tsai YL, Baidoo EE,
1140 Chiu TY, Heazlewood JL, Scheller HV, Mukhopadhyay A, Keasling JD, Deutsch
1141 S, Loque D. 2016. Exploiting members of the BAHD acyltransferase family to
1142 synthesize multiple hydroxycinnamate and benzoate conjugates in yeast. *Microb*
1143 *Cell Fact* **15**:198. doi:10.1186/s12934-016-0593-5

1144 Eudes Aymerick, Pereira JH, Yogiswara S, Wang G, Teixeira Benites V, Baidoo EEK,
1145 Lee TS, Adams PD, Keasling JD, Loqué D. 2016. Exploiting the Substrate
1146 Promiscuity of Hydroxycinnamoyl-CoA:Shikimate Hydroxycinnamoyl Transferase
1147 to Reduce Lignin. *Plant Cell Physiol* **57**:568–579. doi:10.1093/pcp/pcw016

1148 Fan P, Miller AM, Schilmiller AL, Liu X, Ofner I, Jones AD, Zamir D, Last RL. 2016. In
1149 vitro reconstruction and analysis of evolutionary variation of the tomato
1150 acylsucrose metabolic network. *Proc Natl Acad Sci* **113**:E239–E248.
1151 doi:10.1073/pnas.1517930113

1152 Ferreira de Freitas R, Schapira M. 2017. A systematic analysis of atomic protein–ligand
1153 interactions in the PDB. *MedChemComm* **8**:1970–1981.
1154 doi:10.1039/C7MD00381A

1155 Fujiwara Hiroyuki, Tanaka Y, Fukui Y, Ashikari T, Yamaguchi M, Kusumi T. 1998.
1156 Purification and characterization of anthocyanin 3-aromatic acyltransferase from
1157 *Perilla frutescens*. *Plant Sci* **137**:87–94. doi:10.1016/S0168-9452(98)00119-8

1158 Fujiwara H, Tanaka Y, Fukui Y, Nakao M, Ashikari T, Kusumi T. 1997. Anthocyanin 5-
1159 aromatic acyltransferase from *Gentiana triflora*: purification, characterization and
1160 its role in anthocyanin biosynthesis. *Eur J Biochem* **249**:45–51.

1161 Fujiwara H., Tanaka Y, Yonekura-Sakakibara K, Fukuchi-Mizutani M, Nakao M, Fukui Y,
1162 Yamaguchi M, Ashikari T, Kusumi T. 1998. cDNA cloning, gene expression and
1163 subcellular localization of anthocyanin 5-aromatic acyltransferase from *Gentiana*
1164 *triflora*. *Plant J* **16**:421–31. doi:10.1046/j.1365-313x.1998.00312.x

1165 Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in Expression between Duplicated
1166 Genes in Arabidopsis. *Mol Biol Evol* **24**:2298–2309. doi:10.1093/molbev/msm158

1167 GetContacts. 2020.

1168 Goulet C, Kamiyoshihara Y, Lam NB, Richard T, Taylor MG, Tieman DM, Klee HJ. 2015.
1169 Divergence in the Enzymatic Activities of a Tomato and *Solanum pennellii* Alcohol
1170 Acyltransferase Impacts Fruit Volatile Ester Composition. *Mol Plant*, Plant
1171 Metabolism and Synthetic Biology **8**:153–162. doi:10.1016/j.molp.2014.11.007

1172 Harris TK, Mildvan AS. 1999. High-Precision Measurement of Hydrogen Bond Lengths in
1173 Proteins by Nuclear Magnetic Resonance Methods. *Proteins Struct Funct*
1174 *Bioinforma* **35**:275–282. doi:https://doi.org/10.1002/(SICI)1097-
1175 0134(19990515)35:3<275::AID-PROT1>3.0.CO;2-V

1176 Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen
1177 G, Turner S, Williams M, Steinbeck C. 2013. The ChEBI reference database and
1178 ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids*
1179 *Res* **41**:D456–D463. doi:10.1093/nar/gks1146

1180 Hopkins CW, Le Grand S, Walker RC, Roitberg AE. 2015. Long-Time-Step Molecular
1181 Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput*
1182 **11**:1864–1874. doi:10.1021/ct5010406

1183 Huang R, O'Donnell AJ, Barboline JJ, Barkman TJ. 2016. Convergent evolution of
1184 caffeine in plants by co-option of exapted ancestral enzymes. *Proc Natl Acad Sci*
1185 **113**:10613–10618. doi:10.1073/pnas.1602575113

1186 Humphrey W, Dalke A, Schulten K. 1996. VMD: Visual molecular dynamics. *J Mol Graph*
1187 **14**:33–38. doi:10.1016/0263-7855(96)00018-5

1188 Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**:90–95.
1189 doi:10.1109/MCSE.2007.55

1190 Jakalian A, Jack DB, Bayly CI. 2002. Fast, efficient generation of high-quality atomic
1191 charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem*
1192 **23**:1623–1641. doi:10.1002/jcc.10128

1193 Jensen RA. 1976. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*
1194 **30**:409–425. doi:10.1146/annurev.mi.30.100176.002205

1195 Jo S, Kim T, Iyer VG, Im W. 2008. CHARMM-GUI: A web-based graphical user interface
1196 for CHARMM. *J Comput Chem* **29**:1859–1865. doi:10.1002/jcc.20945

1197 Joung IS, Cheatham TE. 2008. Determination of Alkali and Halide Monovalent Ion
1198 Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J Phys Chem*
1199 *B* **112**:9020–9041. doi:10.1021/jp8001614

1200 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017.
1201 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat*
1202 *Methods* **14**:587–589. doi:10.1038/nmeth.4285

1203 Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple
1204 sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**:3059–
1205 3066.

1206 Krautler V, van Gunsteren WF, Hunenberger PH. 2001. A fast SHAKE algorithm to solve
1207 distance constraint equations for small molecules in molecular dynamics
1208 simulations. *J Comput Chem* **22**:501–508.

1209 Kreis W, Munkert J. 2019. Exploiting enzyme promiscuity to shape plant specialized
1210 metabolism. *J Exp Bot* **70**:1435–1445. doi:10.1093/jxb/erz025

1211 Kruse LH, Stegemann T, Sievert C, Ober D. 2017. Identification of a Second Site of
1212 Pyrrolizidine Alkaloid Biosynthesis in Comfrey to Boost Plant Defense in Floral
1213 Stage. *Plant Physiol* **174**:47–55. doi:10.1104/pp.17.00265

1214 Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, Clerck OD.
1215 2012. Phylogeny and Molecular Evolution of the Green Algae. *Crit Rev Plant Sci*
1216 **31**:1–46. doi:10.1080/07352689.2011.615705

1217 Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new
1218 developments. *Nucleic Acids Res* **47**:W256–W259. doi:10.1093/nar/gkz239

1219 Levsh O, Chiang Y-C, Tung CF, Noel JP, Wang Y, Weng J-K. 2016. Dynamic
1220 conformational states dictate selectivity toward the native substrate in a substrate-
1221 permissive acyltransferase. *Biochemistry* **55**:6314–6326.

1222 Loncharich RJ, Brooks BR, Pastor RW. 1992. Langevin dynamics of peptides: the
1223 frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide.
1224 *Biopolymers* **32**:523–535. doi:10.1002/bip.360320508

1225 Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. 2015. A
1226 metacalibrated time-tree documents the early rise of flowering plant phylogenetic
1227 diversity. *New Phytol* **207**:437–453. doi:10.1111/nph.13264

1228 Mahood E, Kruse L, Moghe G. 2020. Machine learning: A powerful tool for gene function
1229 prediction in plants. *Appl Plant Sci* **8**:e11376–e11376. doi:10.1002/aps3.11376

1230 Martínez L, Andrade R, Birgin EG, Martínez JM. 2009. PACKMOL: A package for building
1231 initial configurations for molecular dynamics simulations. *J Comput Chem*
1232 **30**:2157–2164. doi:10.1002/jcc.21224

1233 Matsumura I, Ellington AD. 2001. *In vitro* evolution of beta-glucuronidase into a beta-
1234 galactosidase proceeds through non-specific intermediates. *J Mol Biol* **305**:331–
1235 339. doi:10.1006/jmbi.2000.4259

1236 McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX,
1237 Schwantes CR, Wang L-P, Lane TJ, Pande VS. 2015. MDTraj: A Modern Open
1238 Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **109**:1528–
1239 1532. doi:10.1016/j.bpj.2015.08.015

1240 Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S-H.
1241 2014. Consequences of Whole-Genome Triplication as Revealed by Comparative
1242 Genomic Analyses of the Wild Radish *Raphanus raphanistrum* and Three Other
1243 Brassicaceae Species. *Plant Cell Online* **26**:1925–1937.
1244 doi:10.1105/tpc.114.124297

1245 Moghe GD, Leong BJ, Hurney SM, Daniel Jones A, Last RL. 2017. Evolutionary routes
1246 to biochemical innovation revealed by integrative analysis of a plant-defense
1247 related specialized metabolic pathway. *eLife* **6**:e28468. doi:10.7554/eLife.28468

1248 Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. 1998.
1249 Automated docking using a Lamarckian genetic algorithm and an empirical binding

1250 free energy function. *J Comput Chem* **19**:1639–1662. doi:http://doi:
1251 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B

1252 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective
1253 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol*
1254 *Evol* **32**:268–274. doi:10.1093/molbev/msu300

1255 Nie Y, Foster CSP, Zhu T, Yao R, Duchêne DA, Ho SYW, Zhong B. 2020. Accounting for
1256 Uncertainty in the Evolutionary Timescale of Green Plants Through Clock-
1257 Partitioning and Fossil Calibration Strategies. *Syst Biol* **69**:1–16.
1258 doi:10.1093/sysbio/syz032

1259 Niggeweg R, Michael AJ, Martin C. 2004. Engineering plants with increased levels of the
1260 antioxidant chlorogenic acid. *Nat Biotechnol* **22**:746–754. doi:10.1038/nbt966

1261 Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas
1262 FB, Vanderstraeten L, Becker D, Lang D, Vosolsobě S, Rombauts S, Wilhelmsson
1263 PKI, Janitza P, Kern R, Heyl A, Rümpler F, Villalobos LIAC, Clay JM, Skokan R,
1264 Toyoda A, Suzuki Y, Kagoshima H, Schijlen E, Tajeshwar N, Catarino B,
1265 Hetherington AJ, Saltykova A, Bonnot C, Breuninger H, Symeonidi A,
1266 Radhakrishnan GV, Van Nieuwerburgh F, Deforce D, Chang C, Karol KG, Hedrich
1267 R, Ulvskov P, Glöckner G, Delwiche CF, Petrášek J, Van de Peer Y, Friml J, Beilby
1268 M, Dolan L, Kohara Y, Sugano S, Fujiyama A, Delaux P-M, Quint M, Theißen G,
1269 Hagemann M, Harholt J, Dunand C, Zachgo S, Langdale J, Maumus F, Van Der
1270 Straeten D, Gould SB, Rensing SA. 2018. The *Chara* Genome: Secondary
1271 Complexity and Implications for Plant Terrestrialization. *Cell* **174**:448-464.e24.
1272 doi:10.1016/j.cell.2018.06.033

1273 Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, Lázár V, Spohn R,
1274 Csörgő B, Wagner A, Ruppin E, Pál C, Papp B. 2014. Network-level architecture
1275 and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci*
1276 **111**:11762–11767. doi:10.1073/pnas.1406102111

1277 Panchy N, Lehti-Shiu M, Shiu S-H. 2016. Evolution of gene duplication in plants. *Plant*
1278 *Physiol* **171**:2294–2316. doi:10.1104/pp.16.00523

1279 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and
1280 evolutionary analyses in R. *Bioinformatics* **35**:526–528.
1281 doi:10.1093/bioinformatics/bty633

1282 Payne JL, Wagner A. 2019. The causes of evolvability and their evolution. *Nat Rev Genet*
1283 **20**:24–38. doi:10.1038/s41576-018-0069-z

1284 Peng M, Gao Y, Chen W, Wang W, Shen S, Shi J, Wang C, Zhang Y, Zou L, Wang S,
1285 Wan J, Liu X, Gong L, Luo J. 2016. Evolutionarily Distinct BAHD N-
1286 Acyltransferases Are Responsible for Natural Variation of Aromatic Amine
1287 Conjugates in Rice. *Plant Cell* **28**:1533–50. doi:10.1105/tpc.16.00265

1288 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE.
1289 2004. UCSF Chimera--a visualization system for exploratory research and
1290 analysis. *J Comput Chem* **25**:1605–1612. doi:10.1002/jcc.20084

1291 Philippe G, Sørensen I, Jiao C, Sun X, Fei Z, Domozych DS, Rose JK. 2020. Cutin and
1292 suberin: assembly and origins of specialized lipidic cell wall scaffolds. *Curr Opin*
1293 *Plant Biol*, Physiology and Metabolism **55**:11–20. doi:10.1016/j.pbi.2020.01.008

1294 Piatkowski BT, Imwattana K, Tripp EA, Weston DJ, Healey A, Schmutz J, Shaw AJ. 2020.
1295 Phylogenomics reveals convergent evolution of red-violet coloration in land plants
1296 and the origins of the anthocyanin biosynthetic pathway. *Mol Phylogenet Evol*
1297 **151**:106904. doi:10.1016/j.ympev.2020.106904

1298 Pigliucci M. 2008. Is evolvability evolvable? *Nat Rev Genet* **9**:75–82. doi:10.1038/nrg2278

1299 Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server:
1300 2018 update. *Nucleic Acids Res* **46**:W200–W204. doi:10.1093/nar/gky448

1301 Renault H, Werck-Reichhart D, Weng J-K. 2019. Harnessing lignin evolution for
1302 biotechnological applications. *Curr Opin Biotechnol*, Food Biotechnology • Plant
1303 Biotechnology **56**:105–111. doi:10.1016/j.copbio.2018.10.011

1304 Rostkowski M, Olsson MH, Søndergaard CR, Jensen JH. 2011. Graphical analysis of pH-
1305 dependent properties of proteins predicted using PROPKA. *BMC Struct Biol* **11**:6.
1306 doi:10.1186/1472-6807-11-6

1307 Sander M, Petersen M. 2011. Distinct substrate specificities and unusual substrate
1308 flexibilities of two hydroxycinnamoyltransferases, rosmarinic acid synthase and
1309 hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl-transferase, from *Coleus*
1310 *blumei* Benth. *Planta* **233**:1157–71. doi:10.1007/s00425-011-1367-2

1311 Sanderson MJ. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence
1312 Times: A Penalized Likelihood Approach. *Mol Biol Evol* **19**:101–109.
1313 doi:10.1093/oxfordjournals.molbev.a003974

1314 Schmidt GW, Jirschitzka J, Porta T, Reichelt M, Luck K, Torre JCP, Dolke F, Varesio E,
1315 Hopfgartner G, Gershenzon J, D'Auria JC. 2015. The Last Step in Cocaine
1316 Biosynthesis Is Catalyzed by a BAHD Acyltransferase. *Plant Physiol* **167**:89–101.
1317 doi:10.1104/pp.114.248187

1318 Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes
1319 by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad*
1320 *Sci* **108**:4069–4074. doi:10.1073/pnas.1101368108

1321 Schrödinger – Knowledge Base. 2020.

1322 Schrödinger, LLC. 2020. Schrödinger Release 2020-1: Maestro.

1323 Schrödinger, LLC. 2015. The AxPyMOL Molecular Graphics Plugin for Microsoft
1324 PowerPoint, Version 1.8.

1325 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski
1326 B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of
1327 Biomolecular Interaction Networks. *Genome Res* **13**:2498–2504.
1328 doi:10.1101/gr.1239303

1329 Shaw WV, Leslie AGW. 1991. Chloramphenicol Acetyltransferase. *Annu Rev Biophys*
1330 *Biophys Chem* **20**:363–386. doi:10.1146/annurev.bb.20.060191.002051

1331 St Pierre B, De Luca V. 2000. Evolution of acyltransferase genes: Origin and
1332 diversification of the BAHD superfamily of acyltransferases involved in secondary
1333 metabolism In: Romeo JT, Ibrahim R, Varin L, De Luca V, editors. Recent
1334 Advances in Phytochemistry, Evolution of Metabolic Pathways. Elsevier. pp. 285–
1335 315. doi:10.1016/S0079-9920(00)80010-6

1336 St-Pierre B, Vazquez-Flota FA, Luca VD. 1999. Multicellular compartmentation of
1337 *Catharanthus roseus* alkaloid biosynthesis predicts intercellular translocation of a
1338 pathway intermediate. *Plant Cell* **11**:887–900. doi:10.1105/tpc.11.5.887

1339 Suzuki H, Nakayama T, Nishino T. 2003. Proposed mechanism and functional amino acid
1340 residues of malonyl-CoA : anthocyanin 5-O-glucoside-6 ' ' '-O-malonyltransferase
1341 from flowers of *Salvia splendens*, a member of the versatile plant acyltransferase
1342 family. *Biochemistry* **42**:1764–1771. doi:10.1021/bi020618g

1343 Tanner KG, Trievel RC, Kuo M-H, Howard RM, Berger SL, Allis CD, Marmorstein R, Denu
1344 JM. 1999. Catalytic Mechanism and Function of Invariant Glutamic Acid 173 from
1345 the Histone Acetyltransferase GCN5 Transcriptional Coactivator. *J Biol Chem*
1346 **274**:18157–18160. doi:10.1074/jbc.274.26.18157

1347 Tokuriki N, Tawfik DS. 2009. Protein Dynamism and Evolvability. *Science* **324**:203–207.
1348 doi:10.1126/science.1169375

1349 Trott O, Olson AJ. 2010. AutoDock Vina: improving the speed and accuracy of docking
1350 with a new scoring function, efficient optimization and multithreading. *J Comput*
1351 *Chem* **31**:455–461. doi:10.1002/jcc.21334

1352 Tuominen LK, Johnson VE, Tsai C-J. 2011. Differential phylogenetic expansions in BAHD
1353 acyltransferases across five angiosperm taxa and evidence of divergent
1354 expression among *Populus* paralogues. *BMC Genomics* **12**:236.
1355 doi:10.1186/1471-2164-12-236

1356 Unno H, Ichimaida F, Suzuki H, Takahashi S, Tanaka Y, Saito A, Nishino T, Kusunoki M,
1357 Nakayama T. 2007. Structural and Mutational Studies of Anthocyanin
1358 Malonyltransferases Establish the Features of BAHD Enzyme Catalysis. *J Biol*
1359 *Chem* **282**:15812–15822. doi:10.1074/jbc.M700638200

1360 Vries J de, Rensing SA. 2020. Gene gains paved the path to land. *Nat Plants* **6**:7–8.
1361 doi:10.1038/s41477-019-0579-5

1362 Wagner A. 2008. Robustness and evolvability: a paradox resolved. *Proc R Soc B Biol Sci*
1363 **275**:91–100. doi:10.1098/rspb.2007.1137

1364 Wagner A. 2005. Robustness, evolvability, and neutrality. *FEBS Lett*, Systems Biology
1365 **579**:1772–1778. doi:10.1016/j.febslet.2005.01.063

1366 Wang J, Marowsky NC, Fan C. 2014. Divergence of Gene Body DNA Methylation and
1367 Evolution of Plant Duplicate Genes. *PLOS ONE* **9**:e110357.
1368 doi:10.1371/journal.pone.0110357

1369 Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. 2004. Development and testing
1370 of a general amber force field. *J Comput Chem* **25**:1157–1174.
1371 doi:10.1002/jcc.20035

1372 Wang L-P, Martinez TJ, Pande VS. 2014. Building Force Fields: An Automatic,
1373 Systematic, and Reproducible Approach. *J Phys Chem Lett* **5**:1885–1891.
1374 doi:10.1021/jz500737m

1375 Wang L-P, McKiernan KA, Gomes J, Beauchamp KA, Head-Gordon T, Rice JE, Swope
1376 WC, Martínez TJ, Pande VS. 2017. Building a More Predictive Protein Force Field:
1377 A Systematic and Reproducible Route to AMBER-FB15. *J Phys Chem B*
1378 **121**:4023–4039. doi:10.1021/acs.jpcb.7b02320

1379 Wang Y, Backman TWH, Horan K, Girke T. 2013. fmcsR: mismatch tolerant maximum
1380 common substructure searching in R. *Bioinformatics* **29**:2792–2794.
1381 doi:10.1093/bioinformatics/btt475

1382 Weng J-K, Chapple C. 2010. The origin and evolution of lignin biosynthesis. *New Phytol*
1383 **187**:273–285. doi:https://doi.org/10.1111/j.1469-8137.2010.03327.x

1384 Yang Q, Reinhard K, Schiltz E, Matern U. 1997. Characterization and heterologous
1385 expression of hydroxycinnamoyl/benzoyl-CoA:anthranilate N-
1386 hydroxycinnamoyl/benzoyltransferase from elicited cell cultures of carnation,
1387 *Dianthus caryophyllus* L. *Plant Mol Biol* **35**:777–789.
1388 doi:10.1023/A:1005878622437

1389 Zou C, Lehti-Shiu MD, Thomashow M, Shiu S-H. 2009. Evolution of Stress-Regulated
1390 Gene Expression in Duplicate Genes of Arabidopsis thaliana. *PLOS Genet*
1391 **5**:e1000581. doi:10.1371/journal.pgen.1000581

1392

1393

1394 # SUPPLEMENTARY TABLES

1395 **Table S1: Substrate class definitions.** Substrates were grouped into different classes based on their
1396 chemical properties and their membership in known compound classes. Also see **Fig. S1A**.

Scaffold	Class	Definition	Exemplary structures
Aromatic alcohols	A	Substrates were classified as aromatic alcohols when they contained a planar ring of carbon atoms and a hydroxy group attached to a sidechain. Compounds containing additional hydroxy groups directly attached to the ring were also classified into this category. All substrates in this class except 2-naphthaleneethanol had one ring.	Benzoyl alcohol, Coniferyl alcohol, Shikimate, Quinate
Anthocyanins	D	Substrates were grouped into the anthocyanins category if they fulfill the classical definition of an anthocyanin: contain an anthocyanidin (flavylium cation) with attached sugar group(s).	Cyanidin 3-O glucoside, Cyanidin 3-5-O glucoside, Pelargonidin 3-O rutinoside
Flavonoids	D	Substrates being a flavonoid or isoflavonoid were grouped into this category. Flavonoids a generally defined as having a 15-carbon skeleton which contains two benzene rings that are connected with a 3-carbon linking chain. Also, flavonoid glucosides were grouped into this class.	Naringin, Genistin, Quercetin 3-O- sophoroside
Aliphatic amines	B	Aliphatic amines were defined as amines not containing any aromatic rings but were allowed to contain more than one amine group and or have additional hydroxy groups.	Agmatine, Putrescine, Spermidine
Aliphatic alcohols	B	Aliphatic alcohols do not contain any aromatic rings but can range from short chain to very long chains. We also grouped fatty acids into this category because of their high chemical similarity.	1-Butanol, 1-Decanol, 16-Hydroxypalmitic acid
Monoterpenoids	C	Compounds formally containing of isoprene that do not contain aromatic rings or had branched chains were classified as terpenoids.	Geraniol, Nerol, Linalool
Sesquiterpenoids	C	Compounds build from three isoprene units.	Farnesol
Aromatic amines	A	Substrates containing an aromatic ring and an amine group were classified as aromatic amines. However, substrates containing an additional hydroxy group were also categorized as aromatic amine because their chemistry was expected to be more similar to amines than to aromatic alcohols, due to presence of the positively charged -NH2 group.	Tyramine, Dopamine, 3-Aminobenzoate
Alkaloids	Mix	Substrates were classified as alkaloids if they contained a heterocyclic bound nitrogen atom.	Methylecgonine, Tryptamine, Serotonin
Phenolic glycosides	D	We classified substrates as phenolic glycosides if they contained a sugar group bound via a glycosidic bond to a aromatic molecule that could not be classified as either flavonoids or anthocyanins.	Scopolin, 1-Naphtol glucoside, Umbelliferone glucoside
Sugar derivatives	G	Substrates classified as sugar derivatives are either mono- and disaccharides and their acylated variants.	Sucrose, Glucose, S1:5(5)
Diterpenoids	E	Substrates involved in the biosynthesis of taxenes were grouped into the diterpenoid class	Baccatin III, 10-Deacetylbaccatin III, Taxadien-5a-ol
Triterpenoids	F	Compounds that consisted of six isoprene units were grouped into the triterpenoid class.	Thalianol, Arabidiol, Tirucalla-7,24-dien-3beta-ol

1397

1398 **Table S2. BAHD acyltransferases in the *Solanum lycopersicum* (tomato) genome**. For characterized
1399 BAHDs their name and the respective reference is given.

Protein ID	PFAM domain name	Pfam ID	Name/ Annotation	Reference
Solyc08g005770.2.1	Transferase	PF02458.15	SlAAT1	(Goulet et al., 2015)
Solyc08g005760.1.1	Transferase	PF02458.15	SlAAT2	(Goulet et al., 2015)
Solyc11g071470.1.1	Transferase	PF02458.15	SlACT1	**this study**
Solyc11g071480.1.1	Transferase	PF02458.15	SlACT2	**this study**
Solyc12g006330.1.1	Transferase	PF02458.15	SlASAT1	(Fan et al., 2016)
Solyc04g012020.1.1	Transferase	PF02458.15	SlASAT2	(Fan et al., 2016)
Solyc11g067270.1.1	Transferase	PF02458.15	SlASAT3	(Fan et al., 2016)
Solyc01g105580.1.1	Transferase	PF02458.15	SlASAT4	(Fan et al., 2016)
Solyc03g117600.2.1	Transferase	PF02458.15	SlHCT	**this study**
Solyc07g005760.2.1	Transferase	PF02458.15	SlHQT	(Niggeweg et al., 2004)
Solyc05g052670.1.1	Transferase	PF02458.15		
Solyc05g052680.1.1	Transferase	PF02458.15		
Solyc11g008630.1.1	Transferase	PF02458.15		
Solyc07g015960.1.1	Transferase	PF02458.15		
Solyc09g014280.1.1	Transferase	PF02458.15		
Solyc03g097500.2.1	Transferase	PF02458.15		
Solyc02g079490.2.1	Transferase	PF02458.15		
Solyc07g014580.2.1	Transferase	PF02458.15		
Solyc05g014330.1.1	Transferase	PF02458.15		
Solyc07g049660.2.1	Transferase	PF02458.15		
Solyc08g005890.2.1	Transferase	PF02458.15		
Solyc07g049670.2.1	Transferase	PF02458.15		
Solyc03g025320.2.1	Transferase	PF02458.15		
Solyc04g078660.1.1	Transferase	PF02458.15		
Solyc05g015800.2.1	Transferase	PF02458.15		
Solyc02g093180.2.1	Transferase	PF02458.15		
Solyc04g079720.2.1	Transferase	PF02458.15		
Solyc06g074710.1.1	Transferase	PF02458.15		
Solyc07g006680.1.1	Transferase	PF02458.15		
Solyc12g096250.1.1	Transferase	PF02458.15		
Solyc06g051320.2.1	Transferase	PF02458.15		
Solyc04g082350.1.1	Transferase	PF02458.15		
Solyc01g105550.1.1	Transferase	PF02458.15		
Solyc01g107080.2.1	Transferase	PF02458.15		
Solyc02g062710.1.1	Transferase	PF02458.15		

Solyc11g066640.1.1	Transferase	PF02458.15
Solyc08g013830.1.1	Transferase	PF02458.15
Solyc07g006670.1.1	Transferase	PF02458.15
Solyc04g080720.2.1	Transferase	PF02458.15
Solyc02g081740.1.1	Transferase	PF02458.15
Solyc01g105590.2.1	Transferase	PF02458.15
Solyc06g051130.1.1	Transferase	PF02458.15
Solyc08g014490.1.1	Transferase	PF02458.15
Solyc07g043700.1.1	Transferase	PF02458.15
Solyc12g088170.1.1	Transferase	PF02458.15
Solyc07g043710.2.1	Transferase	PF02458.15
Solyc02g081760.1.1	Transferase	PF02458.15
Solyc12g096770.1.1	Transferase	PF02458.15
Solyc00g040390.1.1	Transferase	PF02458.15
Solyc07g043670.1.1	Transferase	PF02458.15
Solyc02g081800.1.1	Transferase	PF02458.15
Solyc12g010980.1.1	Transferase	PF02458.15
Solyc02g081750.1.1	Transferase	PF02458.15
Solyc08g075210.1.1	Transferase	PF02458.15
Solyc11g069680.1.1	Transferase	PF02458.15
Solyc12g096790.1.1	Transferase	PF02458.15
Solyc11g067290.1.1	Transferase	PF02458.15
Solyc12g096800.1.1	Transferase	PF02458.15
Solyc01g068140.2.1	Transferase	PF02458.15
Solyc11g067340.1.1	Transferase	PF02458.15
Solyc07g008380.1.1	Transferase	PF02458.15
Solyc07g008390.1.1	Transferase	PF02458.15
Solyc12g005430.1.1	Transferase	PF02458.15
Solyc11g067330.1.1	Transferase	PF02458.15
Solyc02g081770.1.1	Transferase	PF02458.15
Solyc12g005440.1.1	Transferase	PF02458.15
Solyc05g039950.1.1	Transferase	PF02458.15
Solyc00g040290.1.1	Transferase	PF02458.15
Solyc07g017320.1.1	Transferase	PF02458.15
Solyc01g008300.1.1	Transferase	PF02458.15
Solyc11g020640.1.1	Transferase	PF02458.15
Solyc10g079570.1.1	Transferase	PF02458.15
Solyc10g008680.1.1	Transferase	PF02458.15
Solyc04g009680.1.1	Transferase	PF02458.15

Solyc07g052060.2.1	Transferase	PF02458.15
Solyc12g087980.1.1	Transferase	PF02458.15
Solyc09g092270.2.1	Transferase	PF02458.15
Solyc08g075180.1.1	Transferase	PF02458.15
Solyc05g052650.2.1	Transferase	PF02458.15
Solyc10g055730.1.1	Transferase	PF02458.15
Solyc00g135260.1.1	Transferase	PF02458.15
Solyc01g107070.2.1	Transferase	PF02458.15
Solyc05g050760.1.1	Transferase	PF02458.15
Solyc01g005900.2.1	Transferase	PF02458.15
Solyc00g134620.1.1	Transferase	PF02458.15
Solyc01g107050.2.1	Transferase	PF02458.15
Solyc08g007210.2.1	Transferase	PF02458.15
Solyc08g078030.2.1	Transferase	PF02458.15
Solyc08g075200.1.1	Transferase	PF02458.15
Solyc07g026890.1.1	Transferase	PF02458.15
Solyc11g067350.1.1	Transferase	PF02458.15
Solyc05g015810.1.1	Transferase	PF02458.15
Solyc06g071940.1.1	Transferase	PF02458.15
Solyc03g078130.1.1	Transferase	PF02458.15
Solyc10g008670.2.1	Transferase	PF02458.15
Solyc04g009670.1.1	Transferase	PF02458.15
Solyc08g036440.1.1	Transferase	PF02458.15
Solyc04g078350.1.1	Transferase	PF02458.15
Solyc12g044660.1.1	Transferase	PF02458.15

1400
1401

1402 **Table S3. Primer sequences.** Sequences of Primers used to clone enzymes analyzed in this study.

Name	Sequence (5' to 3' end)	Used for
Oligo DT_{17}	GTCGACTCGAGAATTCTTTTTTTTTTTTTTTTT	cDNA synthesis
Gibson_N_6xHis4CL_F2	AGCATGACTGGTGGACAGCAAATGGGTCGGATGCCGAT GGATACCGAAACA	Cloning of Sl4CL
Gibson_4CL_R2	GCCGGATCTCAGTGGTGGTGGTGGTGGTGCATGCCGAT GGATACCGAAACA	Cloning of Sl4CL
Gibson_N_6xHis_HCT_F1	AGCATGACTGGTGGACAGCAAATGGGTCGGATGAAGAT CGAGGTGAAAAACTCA	Cloning of SlHCT
Gibson_HCT_R1	GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCAAATGT CATACAAGAACTTCTCAA	Cloning of SlHCT
SlHQT_Gibson_N_6xHis_F1	AGCATGACTGGTGGACAGCAAATGGGTCGGATGGGAAG TGAAAAAATGATGAAAATTAATATC	Cloning of SlHQT
SlHQT_Gibson_R1	GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCATAATT CATATAAATATTTTTCAAATAGTG	Cloning of SlHQT
Gibson_N_6xHis_ACT2_F2	AGCATGACTGGTGGACAGCAAATGGGTCGGATGAATGT GAAAATTGAGAGTTCAAAAATCATCAAGCCATTG	Cloning of SlACT
Gibson_ACT2_R2	GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCACTTTG CTTTCAAATCTAGAGAGTAACAAATTTTC	Cloning of SlACT
Dm3MAT3_W70A_F2	CCTTTTTCGATTTCTTTGCGCTGCGCAGTCCGCC	Mutation of Trp36 to Ala
Dm3MAT3_W70A_R2	GGCGGACTGCGCAGCGCAAAGAAATCGAAAAAGG	Mutation of Trp36 to Ala
Dm3MAT3_C322A_F	GCATATTTTGGCAATGCCGTGGGTGGTTGCGCC	Mutation of Cys320 to Ala
Dm3MAT3_C322A_R	GGCGCAACCACCCACGGCATTGCCAAAATATGC	Mutation of Cys320 to Ala

1403

1404 **Table S4: Masses of detected products.** Parent ion mass (bold) for each donor was used for isolation
1405 with product specific PRM method (see methods section) and the quantifier ion to quantify this product (if
1406 product was detected). For each product (if detected) the retention time is given in brackets. n.t = not tested;
1407 n.d. = not detected. LC methods are described in **Tab. S5**.

Substrate	+ Coumaroyl Product (m/z)	+ Feruloyl Product (m/z)	+ Malonyl Product (m/z)	Ionization mode	LC method
Shikimate	**319.0817** 163.0398 (1.6)	**349.0924** n.d.	**259.0454** n.d.	negative	4-min method
Quinate	**337.0927** 191.0555 (1.4)	**367.1034** n.d.	**277.0565** 191.0558 (1.4)	negative	4-min method
Tyramine	**284.1279** 147.0443 (1.85)	**314.1386** 177.0544 (1.66)	**224.0917** n.d.	positive	4-min method
Dopamine	**300.1229** 147.0443 (1.65)	**330.1336** 177.0545 (1.6)	**284.0764** n.d.	positive	4-min method
Agmatine	**277.1667** 147.0440 (1.1)	**307.1774** 177.0546 (1.3)	**217.1295** 131.1291 (1.1)	positive	4-min method
Putrescine	**235.1445** 147.0440 (0.6)	**265.1552** 177.0546 (1.0)	**175.1077** n.d.	positive	4-min method
1-Decanol	**303.1964** n.d.	**333.2071** 177.0194 (5.5)	**243.1601** n.d.	negative	8-min method
Naringin	**725.2091** 163.0401 (1.7)	**755.2198** 271.0603 (2.1)	**665.1723** 271.0608 (1.8)	negative	4-min method
Genistin	**577.1351** 269.0434 (1.65)	**607.1458** 269.0445 (2.1)	**517.0987** 269.0443 (1.8)	negative	4-min method
Cyanidin 3,5-O glucoside	**757.1974** 287.0546	**787.208** **n.d**	**697.1610** 287.0546	positive	4-min method

	(1.45)		(1.25)		
Malvidin 3-O glucoside	**639.1713** 331.0835 (1.65)	**669.1820** 331.0829 (1.6)	**579.1350** 331.0835 (1.45)	positive	4-min method
Pelargonidin 3-O rutinoside	**725.2080** n.d.	**755.2270** 271.0604 (1.5)	**665.1717** n.t.	positive	4-min method
Cyanidin 3-O glucoside	**595.1451** 287.0546 (1.6)	**625.1558** 287.0541 (1.5)	**535.1088** 287.0544 (2.5 (8-min method))	positive	4-min method 8-min method
Cyanidin 3-O rutinoside	**741.2029** 287.0546 (1.25)	**771.2136** 287.0545 (1.5)	**681.1666** n.t.	positive	4-min method
Linalool	**299.1725** n.d.	**329.183165** n.d.	**239.1362** n.d.	negative	8-min method
Geraniol	**299.1654** 163.0393 (4.9)	**329.1754** 329.1754 (4.9)	**239.1362** n.d.	negative	8-min method
Cucurbitacin E	**701.3380** 175.0393 (4.3)	**731.3502** 175.0393 (4.4)	**641.2967** n.d.	negative	8-min method

Table S5. LC methods used for analyzing enzyme assay reactions. All solvents and reagents used were LC-MS grade quality. For the 150 mm column, a 2 mm C18 guard column (AJ0-8782, SecurityGuard Ultra, 2.1 mm, Phenomenex, Torrance, CA) was used. The autosampler was kept at 20 ºC and the column compartment was heated to 40 ºC. For negative ion mode, the mass spectrometer parameters were as follows: spray voltage 3800V, capillary temperature 380C, sheath gas pressure 60 psi, auxiliary gas 20 psi, spare gas 2 psi, max. spray current 100eV, probe 400C, RF lens 50V, and a collision energy ramped from 20 to 40 eV. For positive ion mode MS parameters were the same except for the spray voltage that was set to 3500V.

4min-C18

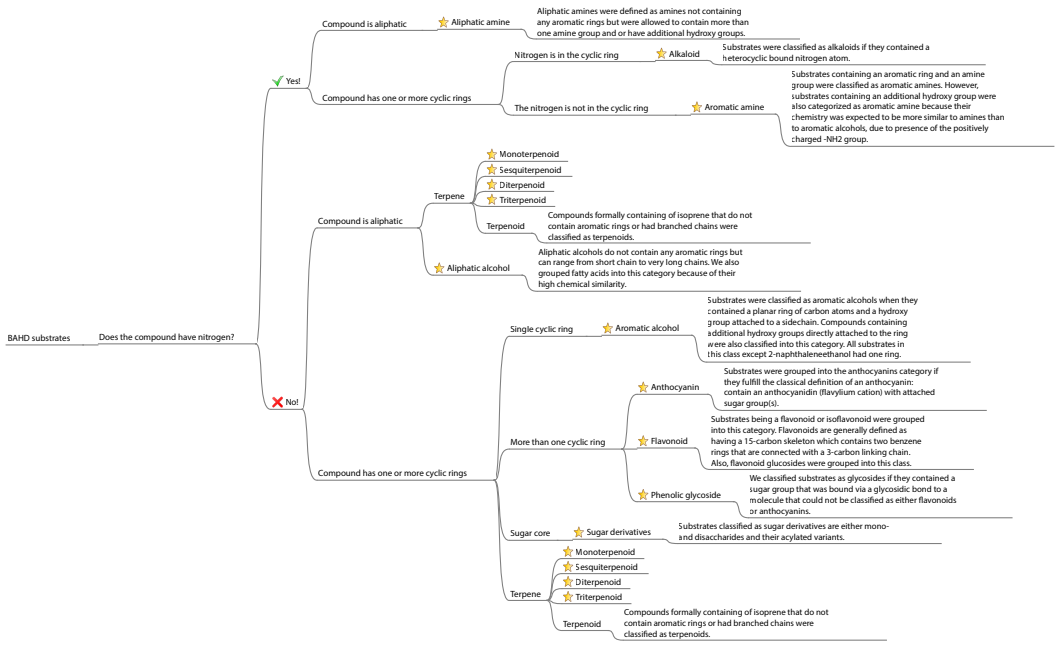
	Flow (mL/min)	%A	%B	Curve	Solvents	Column
					A: Water + 0.01% (v/v) Formic acid B: Acetonitrile + 0.01% (v/v) Formic acid	Kinetex PS C18 (00B-4780-AN, 2.6µm particle size, 100Å pore size, 50 x 2.10 mm, Phenomenex (Torrance, CA))
Initial	0.6	95	5	5		
0.5	0.6	95	5	5		
3.0	0.6	30	70	5		
3.3	0.6	95	5	5		
3.6	0.6	5	95	5		
4.0	0.6	5	95	5		

8min-C18

	Flow (mL/min)	%A	%B	Curve	Solvents	Column
					A: Water+0.01% (v/v) Formic acid B: Acetonitrile + 0.01% (v/v) Formic acid	Kinetex C18 (00F-4462-AN, 2.6µm particle size, 100Å pore size, 150 x 2.10 mm, Phenomenex (Torrance, CA))
Initial	0.6	95	5	5		
1	0.6	95	5	5		
5.6	0.6	5	95	5		
6.6	0.6	5	95	5		
6.61	0.6	5	95	5		
7.5	0.6	95	5	5		
8.0	0.6	95	5	5		

Figure S1

A



B

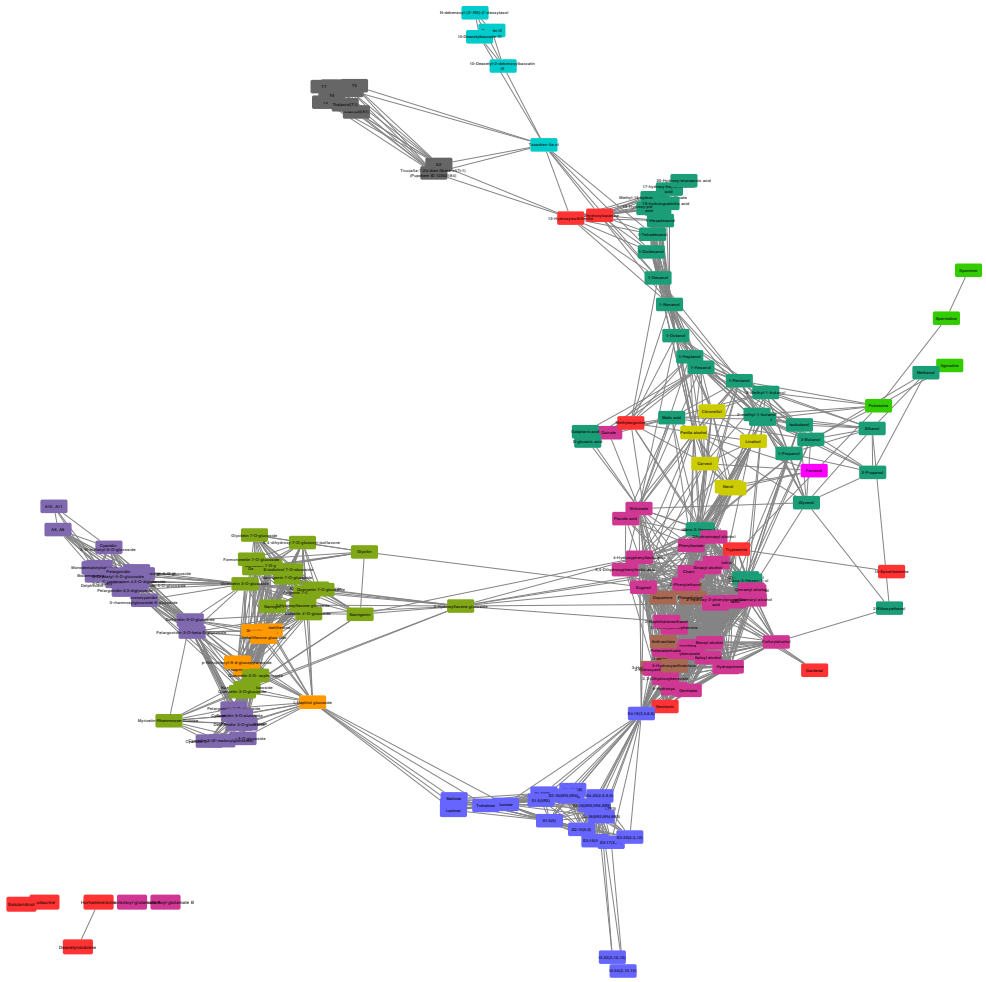
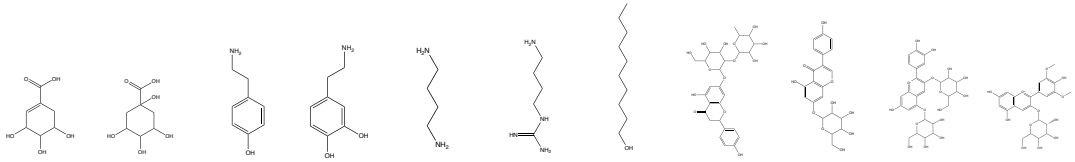


Figure S1: Substrate classification and BAHD substrates represented using an alternative layout. (A) Decision tree to group BAHD substrates into different substrate types. Further information about substrate class definitions can be found in **Table S1**. **(B)** The network was created in the same way and is colored the same way as **Fig. 1A**. The "Edge-weighted Spring Embedded" Layout using MCS-Tanimoto coefficient was used to create an edge-weighted network.

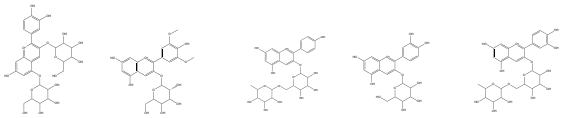
Figure S2

A



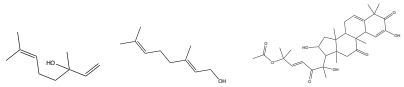
			aromatic alcohols		aromatic amines		aliphatic amines		aliphatic alcohols	flavonoids		anthocyanins	
Enzyme	Donor	acceptor concentration	Shikimate	Quinate	Tyramine	Dopamine	Putrescine	Agmatine	Decanol	Naringin	Genistin	Cyanidin 3,5-O diglucoside	Malvidin 3-O glucoside
CbHQT-like	Coumaroyl-CoA	100 µM	0.037 ±0.008	4.763 ±2.812	0.023 ±0.003	0.021 ±0.001	0.016 ±0.006	0.024 ±0.003	0.015 ±0.001	0.028 ±0.006	0.434 ±0.237	0.001 ±0.000	0.001 ±0.001
CbHQT-like*	Coumaroyl-CoA	1 mM	6.956 ±0.476	555.346 ±59.314	6.322 ±0.835	16.998 ±3.392	9.289 ±2.779	4.469 ±1.269	0.603 ±0.056	n.t.	n.t.	n.t.	n.t.
MeHFT	Coumaroyl-CoA	100 µM	0.052 ±0.008	0.003 ±0.001	0.083 ±0.026	0.134 ±0.015	0.037 ±0.006	0.025 ±0.007	0.000 ±0.000	0.067 ±0.013	0.193 ±0.028	0.009 ±0.001	0.180 ±0.014
MeHFT	Feruloyl-CoA	100 µM	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	8.332 ±0.000	0.041 ±0.004	0.336 ±0.040	0.191 ±0.173	2.588 ±0.083
MeHFT*	Feruloyl-CoA	1 mM	n.d.	n.d.	2.014 ±0.066	0.967 ±0.101	0.296 ±0.032	0.286 ±0.007	n.t.	n.t.	n.t.	n.t.	n.t.
SlACT	Coumaroyl-CoA	100 µM	0.006 ±0.001	0.009 ±0.001	0.050 ±0.007	0.054 ±0.002	22.673 ±3.164	6.476 ±0.455	0.018 ±0.003	0.022 ±0.006	0.065 ±0.113	0.002 ±0.001	0.008 ±0.002
SlACT*	Coumaroyl-CoA	1 mM	1.396 ±0.433	1.854 ±0.679	10.576 ±1.632	17.718 ±7.195	3,614.042 ±1,216.789	1,076.305 ±141.838	0.715 ±0.139	n.t.	n.t.	n.t.	n.t.
SlHCT	Coumaroyl-CoA	100 µM	65.378 ±14.824	70.896 ±17.602	2.238 ±0.292	1.093 ±0.136	0.057 ±0.014	0.027 ±0.007	n.t.	0.078 ±0.022	0.228 ±0.079	0.210 ±0.052	6.497 ±1.347
SlHQT	Coumaroyl-CoA	100 µM	21.608 ±1.838	160.828 ±27.189	0.041 ±0.009	0.093 ±0.033	0.056 ±0.006	0.030 ±0.004	n.t.	0.102 ±0.008	0.169 ±0.009	0.012 ±0.002	0.015 ±0.011
EcHQT	Coumaroyl-CoA	100 µM	9.695 ±1.284	132.568 ±7.869	0.034 ±0.004	0.055 ±0.008	0.111 ±0.091	0.035 ±0.015	n.t.	0.040 ±0.004	0.557 ±0.629	0.001 ±0.000	0.002 ±0.001
EcCS	Coumaroyl-CoA	100 µM	25.696 ±3.045	94.034 ±8.686	0.025 ±0.003	0.033 ±0.005	0.021 ±0.001	0.025 ±0.002	n.t.	0.031 ±0.021	0.188 ±0.049	0.001 ±0.000	0.002 ±0.000
At5MAT	Malonyl-CoA	100 µM	n.d.	n.d.	n.d.	0.005 ±0.005	n.d.	n.d.	n.t.	0.042 ±0.015	n.d.	1.876 ±0.317	0.943 ±0.147
GmIF7MAT	Malonyl-CoA	100 µM	n.d.	0.039 ±0.050	n.d.	n.d.	n.d.	0.017 ±0.015	n.t.	0.043 ±0.011	0.638 ±0.049	0.004 ±0.001	0.029 ±0.020
Dm3MAT3	Malonyl-CoA	100 µM	n.d.	n.d.	n.d.	0.002 ±0.003	n.d.	0.195 ±0.041	n.t.	0.000 ±0.000	1.229 ±0.116	2.683 ±0.163	13.098 ±4.598
Dm3MAT3	Coumaroyl-CoA	100 µM	n.d.	n.d.	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.
Pc3MAT3	Malonyl-CoA	100 µM	n.d.	0.016 ±0.007	n.d.	n.d.	n.d.	n.d.	n.t.	0.057 ±0.014	0.030 ±0.006	3.751 ±0.888	108.880 ±3.219
Mock	Coumaroyl-CoA	100 µM	0.597 ±0.049	0.940 ±0.367	0.047 ±0.001	0.093 ±0.026	0.034 ±0.002	0.023 ±0.003	0.001 ±0.002	0.071 ±0.011	0.296 ±0.200	0.017 ±0.004	0.018 ±0.007
Mock*	Coumaroyl-CoA	1 mM	22.757 ±3.350	0.212 ±0.084	6.018 ±0.530	14.886 ±4.448	3.156 ±2.660	4.539 ±0.643	0.111 ±0.000	n.t.	n.t.	n.t.	n.t.
Mock	Malonyl-CoA	100 µM	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.t.	0.039 ±0.005	0.000 ±0.000	0.003 ±0.000	0.006 ±0.001
MeHFT	Feruloyl-CoA	100 µM	n.d.	n.d.	0.045 ±0.004	n.d.	0.189 ±0.005	0.124 ±0.070	n.d.	n.t.	n.t.	n.t.	n.t.
Mock*	Feruloyl-CoA	1 mM	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	0.006 ±0.001	0.014 ±0.004	0.064 ±0.040	0.057 ±0.007

B



			anthocyanins				
Enzyme	Donor	acceptor concentration	Cyanidin 3,5-O diglucoside	Malvidin 3-O glucoside	Pelargonidin 3-O rutinoside	Cyanidin 3-O glucoside	Cyanidin 3-O rutinoside
CbHCT	Coumaroyl-CoA	100 µM	0.063 ±0.015	0.111 ±0.024	1.337 ±0.112	1.341 ±0.354	0.133 ±0.028
MeHFT	Feruloyl-CoA	100 µM	0.191 ±0.173	2.588 ±0.083	54.127 ±8.146	3.243 ±0.468	0.431 ±0.101
SlACT	Coumaroyl-CoA	100 µM	0.059 ±0.014	0.105 ±0.085	1.143 ±0.256	0.549 ±0.228	0.053 ±0.052
SlACT	Malonyl-CoA	100 µM	0.002 ±0.001	0.000 ±0.000	0.001 ±0.001	0.007 ±0.003	0.011 ±0.000
SlHCT	Coumaroyl-CoA	100 µM	1.127 ±0.332	38.458 ±19.705	0.908 ±0.313	2.445 ±0.955	1.550 ±2.472
SlHQT	Malonyl-CoA	100 µM	0.001 ±0.000	0.000 ±0.000	0.001 ±0.000	0.033 ±0.047	0.007 ±0.001
EcHQT	Coumaroyl-CoA	100 µM	0.037 ±0.064	0.060 ±0.103	0.443 ±0.768	0.726 ±1.258	0.080 ±0.070
Mock	Coumaroyl-CoA	100 µM	0.063 ±0.057	0.166 ±0.007	0.987 ±0.069	0.770 ±0.169	0.121 ±0.048
Mock	Feruloyl-CoA	100 µM	0.064 ±0.040	0.057 ±0.007	0.052 ±0.025	0.915 ±0.049	0.000 ±0.000
Mock	Malonyl-CoA	100 µM	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000

C



			Monoterpenoids		Triterpenoids
Enzyme	Donor	acceptor concentration	Linalool	Geraniol	Curcubitacine E
CbHQT-like	Coumaroyl-CoA	100 µM	3.383 ±0.358	0.000 ±0.000	0.436 ±0.042
MeHFT	Feruloyl-CoA	100 µM	0.034 ±0.011	45.255 ±20.153	10.589 ±0.421
SlHCT	Coumaroyl-CoA	100 µM	2.350 ±0.034	1.386 ±0.307	10.975 ±0.466
SlHQT	Coumaroyl-CoA	100 µM	2.893 ±0.163	0.000 ±0.000	0.529 ±0.029
Dm3MAT3	Malonyl-CoA	100 µM	3.952 ±0.643	3.900 ±0.334	0.000 ±0.000
Mock	Coumaroyl-CoA	100 µM	2.742 ±0.399	0.000 ±0.000	0.651 ±0.193
Mock	Feruloyl-CoA	100 µM	0.095 ±0.031	0.101 ±0.020	1.236 ±0.081
Mock	Malonyl-CoA	100 µM	4.217 ±0.580	4.497 ±0.689	0.000 ±0.000

Figure S2. Specific activities of selected enzymes against representative substrates. Enzymes are ordered by the phylogenetic relationship to each other, starting with the most basal at the top. The average enzyme activity and standard deviation of three technical replicates is given as nmol/mg/min. Each enzyme was tested with its preferred donor substrate (300 uM) and 100 uM of the acceptor substrate. Main activities for each enzyme are colored dark red, medium activities in orange, and low/trace activities in light red. Each activity considered had to be three times higher than the respective Mock control (empty vector). **(A)** Enzyme activities of selected enzymes against a large panel of substrates. In case of EcCS, coumaroyl-CoA was used instead of benzoyl-CoA which had been shown to be used previously. MeHFT was assayed with coumaroyl-CoA and feruloyl-CoA. n.d. = not detected n.t. = not tested. * = higher acceptor substrate concentration than other assays. In addition to the shown activities, all enzymes were tested with glucose and sucrose as acceptor substrates but no activities were observed. **(B)** Enzyme assays testing the ability of selected enzymes to acylate anthocyanins with their preferred donor. To exclude that such enzymes can acylate anthocyanins with a different donor, SlHQT and SlACT were also tested with malonyl-CoA, a donor not previously shown to be used by those enzymes. As expected, changing the donor did not allow those enzymes to acylate anthocyanins. **(C)** Enzyme assays testing the ability to use terpenoids. Two monoterpenoids and one triterpenoid were tested using different enzymes with their preferred donor. Mock controls for each donor were performed to exclude the unspecific formation of the analyzed reaction products.

Figure S3

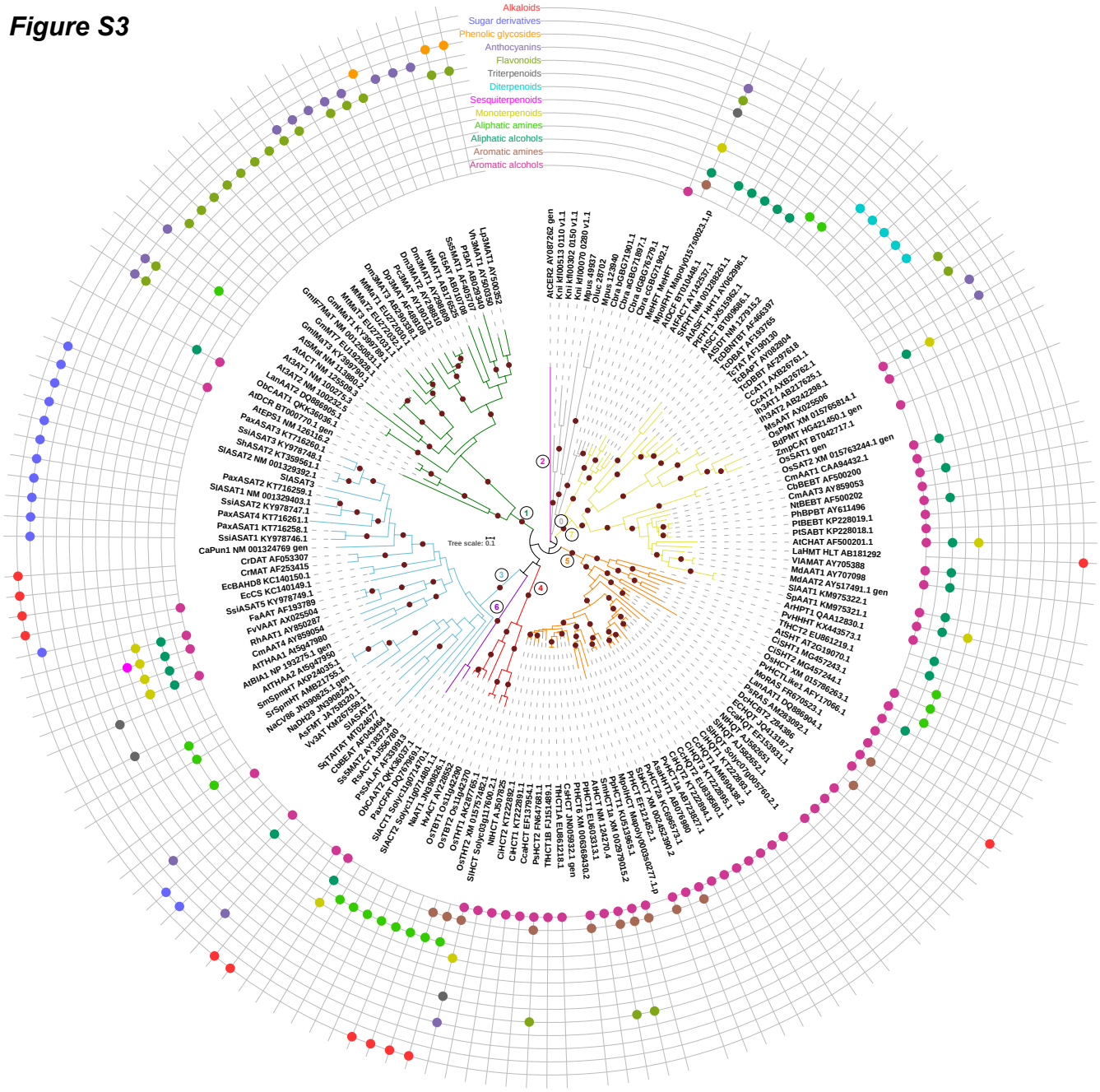


Figure S3. Defining BAHD clades. The tree was rooted using the algal enzyme clade (Clade 0a). Maroon circles on branches refer to clades with bootstrap values > 70. Clades were defined based on deepest, high-confidence monophyletic clades. Clades 1-4 are same as D'Auria et al, 2006 definitions, while Clade V from that study is divided into Clades 5-7 here, based on the above criterion. Branch colors indicate the different clades. Branch lengths indicate number of amino acid substitutions per site.

Figure S4

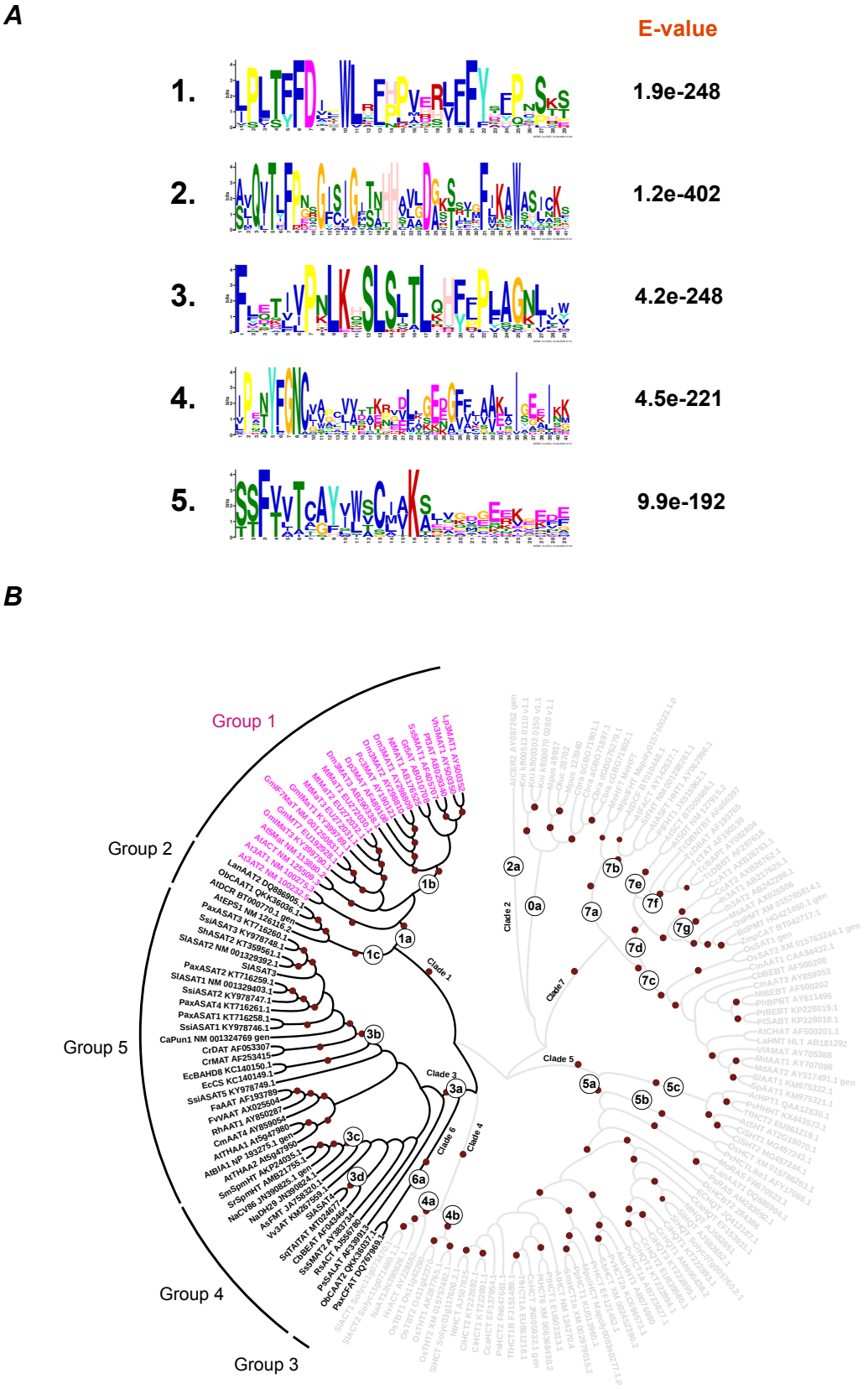


Figure S4. Sequences used for motif enrichment analysis. (A) Five top most enriched motifs in anthocyanin/flavonoid-acylating BAHD acyltransferases. **(B)** Groups of sequences that were further analyzed with respect to the TFFDxxW (1. motif) and YFGNC (4. motif). The same tree as in **Fig. 4** is shown. Black circles on branches refer to clades with bootstrap values > 70. Clades not analyzed in more detail are displayed in light gray and bootstrap values are not indicated. Groups of sequences correspond to **Fig. 4**. Group 5, highlighted in purple, contains the sequences of anthocyanin/flavonoid-acylating BAHDs that were analyzed in more detail.

Figure S5

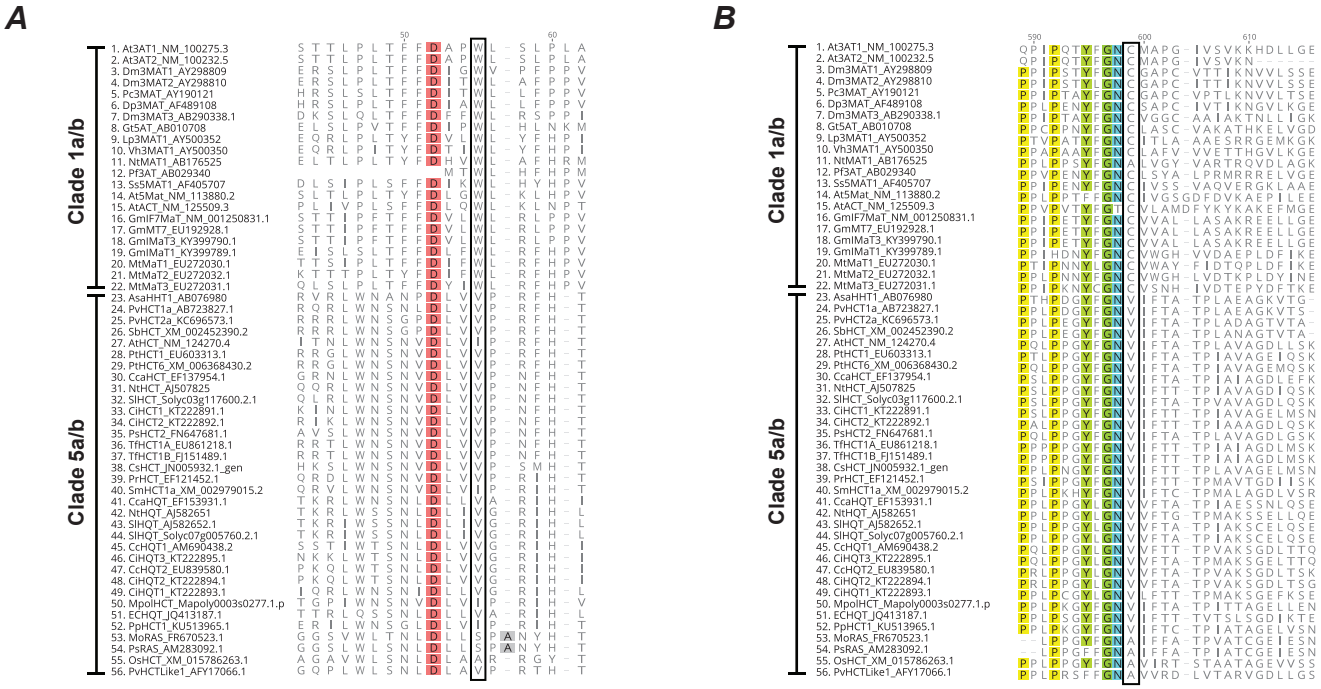


Figure S5. Alignment of HCT/HQT-like and anthocyanin/flavonoid-acylating BAHDs. (A) Alignment of the TFFDxxW region of sequences belonging to clade 1a/b (anthocyanin/flavonoid) and 5a/b (HCT/HQT). **(B)** Alignment of the same sequences for the YFGNC region. Conserved (95%) residues are highlighted in colors. The positions of the highly conserved Trp and Cys residues are highlighted with a box.

Figure S6

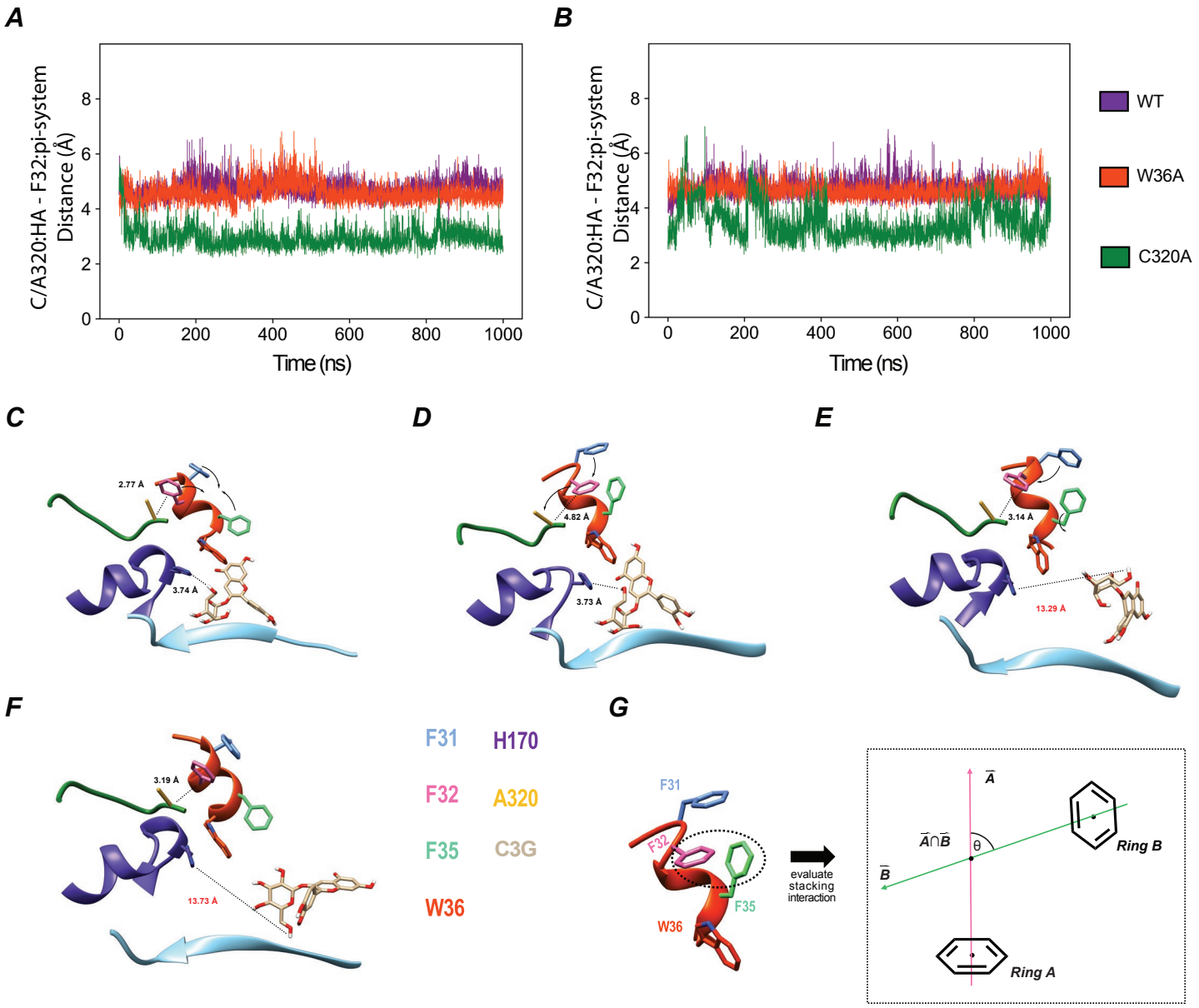
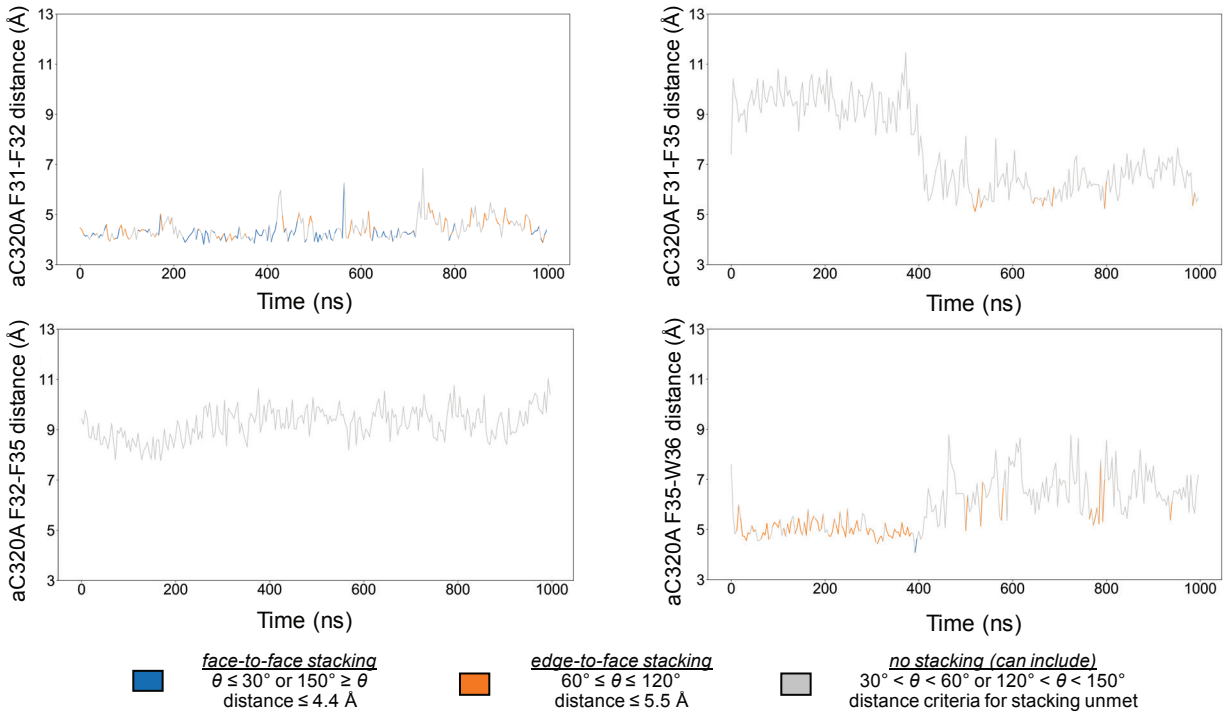


Figure S6. Analysis of distance fluctuations in Dm3MAT3 variants. (A) Distance calculations between backbone hydrogen of C/A320 and the center of mass of the aromatic ring of F32 for the apo simulation. WT is represented as purple, W36A is represented as orange-red, and C320A is represented as green. **(B)** Distance calculations for the holo simulation **(C)** Stacking rearrangements of phenylalanine aromatic sidechains within the TFFDXXW motif of C320A Dm3MaT3. Amino acid coloring: F31, cornflower blue; F32, hot pink; F35, spring green; W36, H170, A320, and C3G are colored as in previous images. Malonyl-CoA and W383 are hidden for simplicity. Distance fluctuations between A320 alpha hydrogen and F32 aromatic system cause aromatic stacking reorganization in the TFFDXXW motif. A catalytically competent distance between C3G-6"-OH and H170 is maintained. **(D)** C3G can still undergo reaction as F35 breaks its t-stacking with W36 to stabilize F31 and F32. **(E)** F31 and F32 have been stabilized. F32 moves closer to A320 while F35 returns to t-stack with W36, although C3G has distanced itself from H170. **(F)** Stacking interactions between residues in the active site and across the TFFDXXW motif have been reset, but now C3G is excluded from re-entering the active site due to steric and non-conventional bonding interactions with peripheral residues. **(G)** Schematic demonstrating θ calculations for discriminating between face-to-face and edge-to-face stacking interactions along the TFFDXXW motif. Figure and methodology as described in the Data Analysis subsection from the Main Text. Rings A and B represent the aromatic systems of two different residues. The normal vectors for each ring were then calculated so that the point of intersection between Rings A and B could be determined. Solving for the angle between the center of mass of Ring A, the center of mass of Ring B, and the intersection of the rings' normal vectors, allowed for the supplementary angle, θ, to be determined. When $\theta \leq 30°$ or $150° \geq \theta$ and the distance between the aromatic systems' centers of mass is ≤ 4.4 Å, face-to-face stacking is occurring. When $60° \leq \theta \leq 120°$ and the distance between the aromatic systems' centers of mass is ≤ 5.5 Å, edge-to-face stacking is occurring. When $30° < \theta < 60°$ or $120° < \theta < 150°$ or the distance requirements for either form of stacking are unmet, no stacking interaction is occurring.

Figure S7

A aC320 stackings



B hC320A stackings

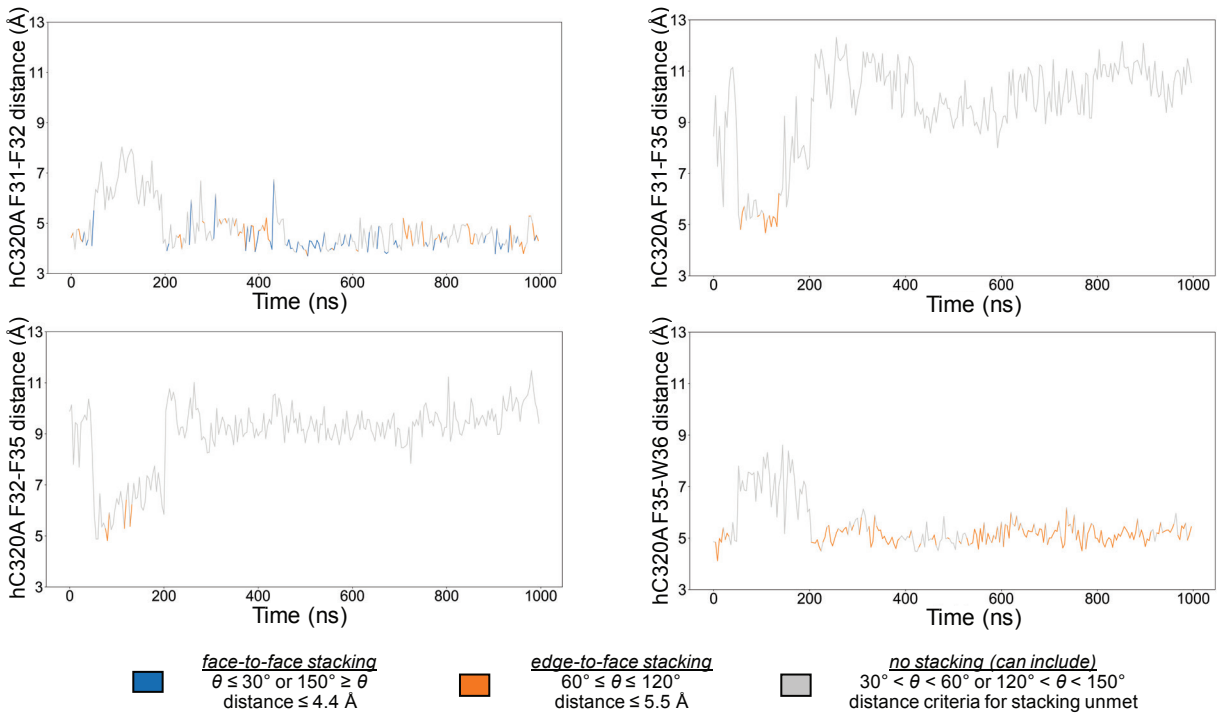
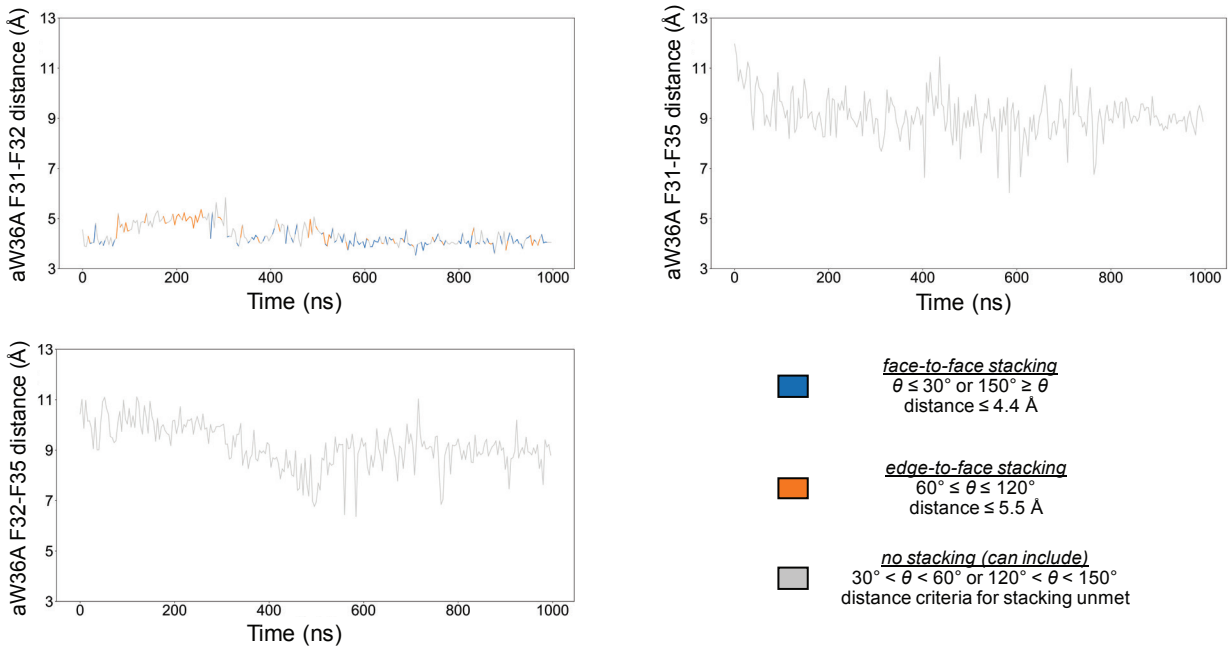


Figure S7. Stacking interactions for the C320A mutant. (A) Apo C320A (aC320A) stacking interactions categorized as face-to-face stacking, edge-to-face stacking, or no stacking interactions. **(B)** Holo C320A (hC320A) stacking interactions. Face-to-face stacking is represented as blue, edge-to-face stacking as orange, and no stacking as gray. Distance measurements were taken between the centers of mass of each residue listed on the y-axis of each panel.

Figure S8

A aW36A stacking



B hW36A stacking

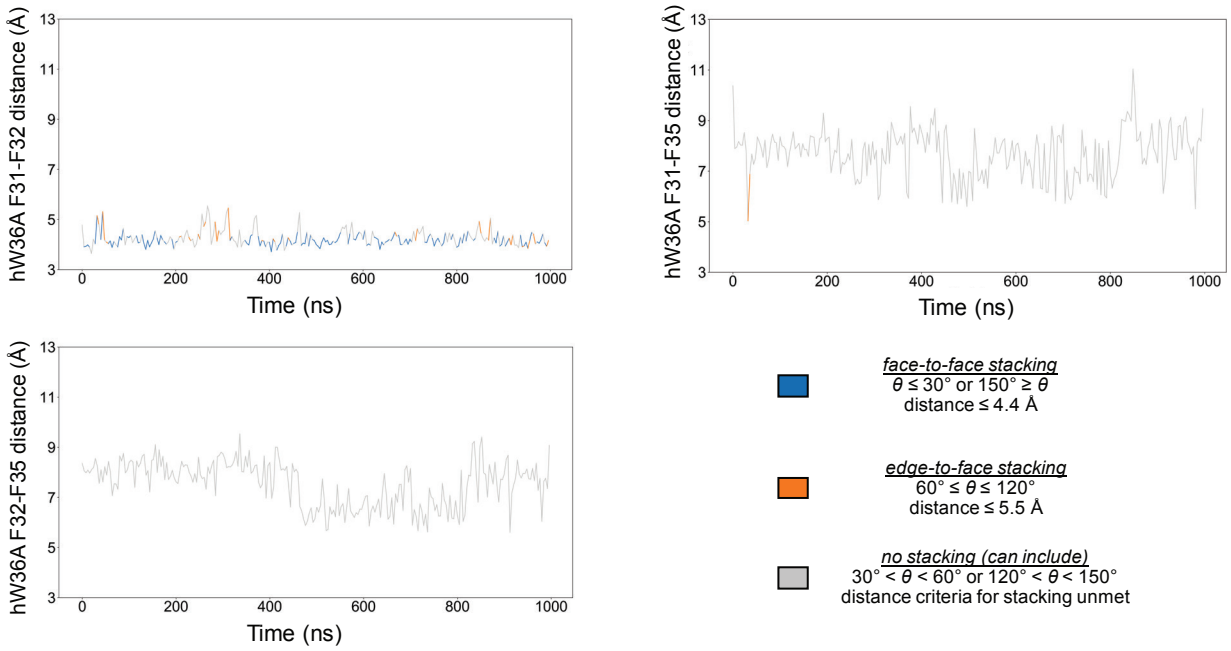


Figure S8. Stacking interactions for the W36A mutant. (A) Apo W36A (aW36A) stacking interactions categorized as face-to-face stacking, edge-to-face stacking, or no stacking interactions. **(B)** Holo W36A (hW36A) stacking interactions. Face-to-face stacking is represented as blue, edge-to-face stacking as orange, and no stacking as gray. Distance measurements were taken between the centers of mass of each residue listed on the y-axis of each panel. Only three panels are provided because W36A substitution prevents F35-W36 stacking from occurring.

Figure S9

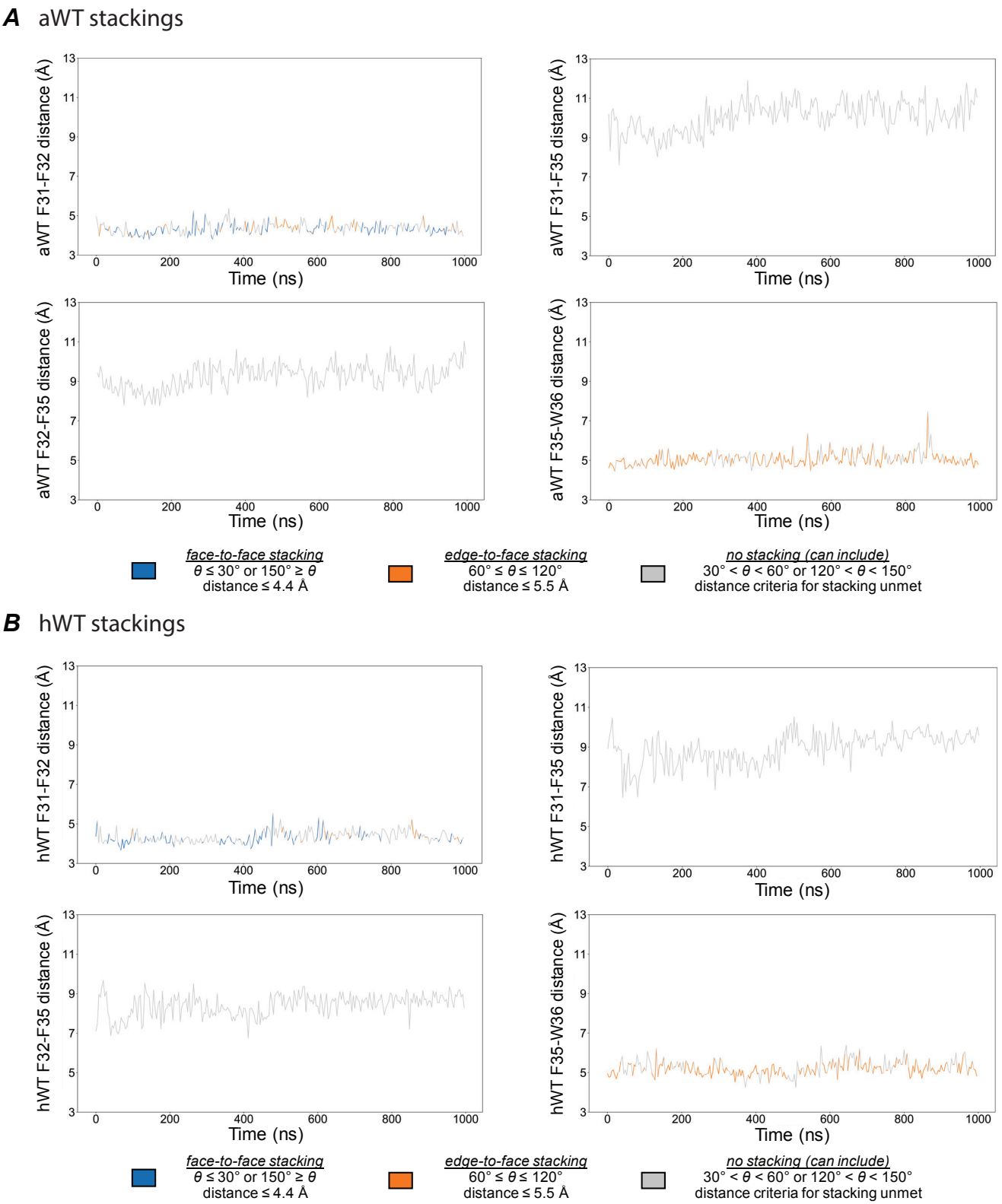


Figure S9. Stacking interactions for the WT Dm3MAT3. (A) Apo WT (aWT) stacking interactions categorized as face-to-face stacking, edge-to-face stacking, or no stacking interactions. **(B)** Holo WT (hWT) stacking interactions. Face-to-face stacking is represented as blue, edge-to-face stacking as orange, and no stacking as gray. Distance measurements were taken between the centers of mass of each residue listed on the y-axis of each panel.

Figure S10

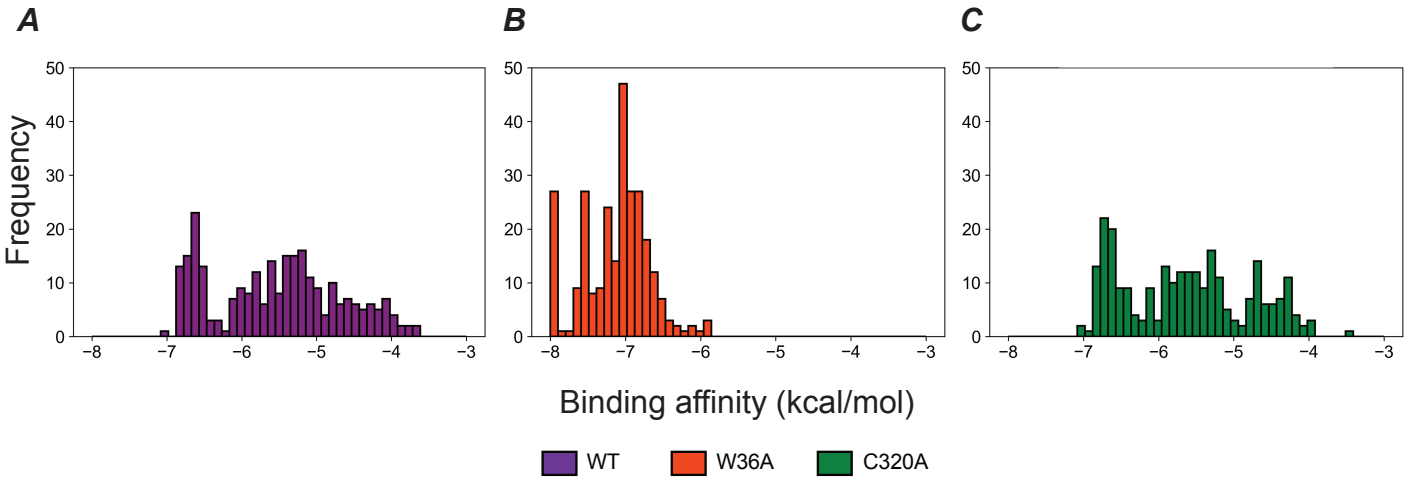
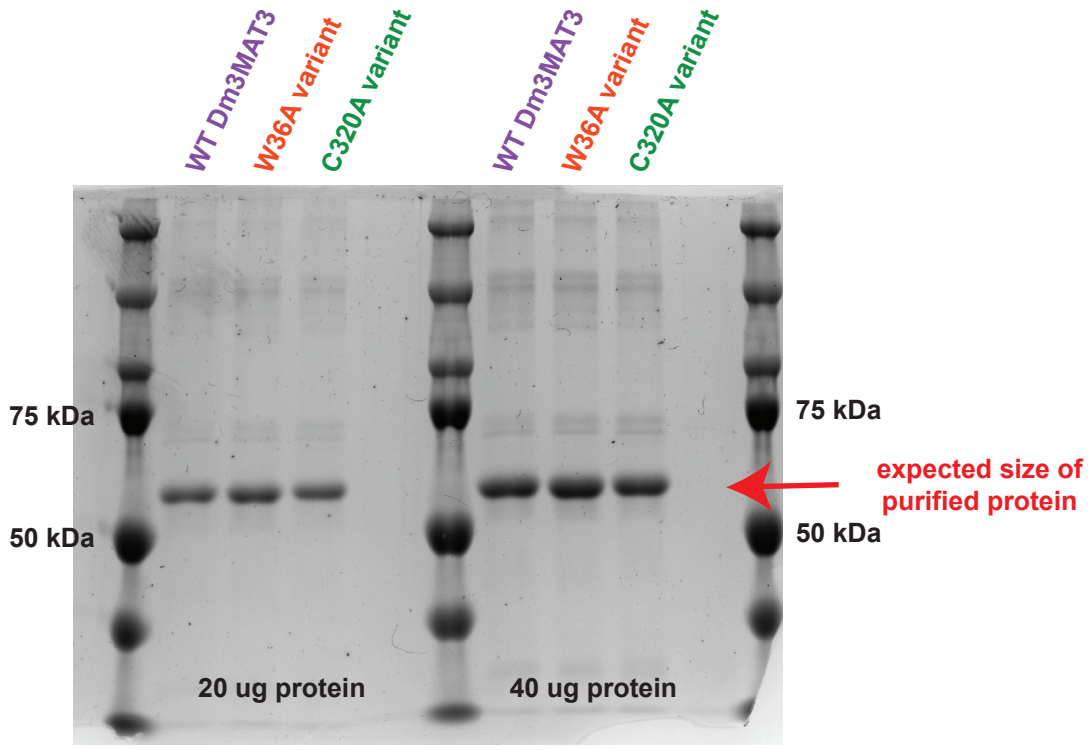


Figure S10. Substrate docking. Cyanidin 3-O-glucoside (C3G) docking into the Dm3MaT3 acceptor site for **(A)** WT, **(B)** W36A, and **(C)** C320A variants. WT is represented as purple, W36A is represented as orange-red, and C320A is represented as green.

Figure S11

A



B

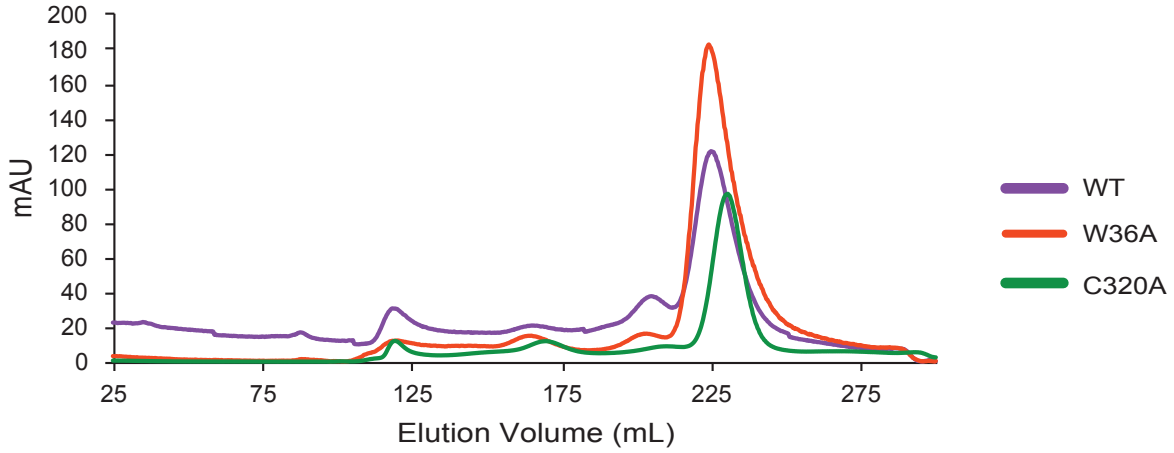


Figure S11. SDS-PAGE and Size Exclusion Chromatography of Dm3MAT3 variants. (A) Coomassie-stained SDS-PAGE of Dm3MAT3 variants purified to homogeneity using Ni-NTA-agarose. 20 and 40 ug protein were separated on a 12% gel. Similar migration pattern and purity for all variants was observed. **(B)** Size exclusion chromatography traces of the different Dm3MAT3 variants show that mutant proteins retain to similar elution volumes as wild-type, indicating proper size and shape of the protein.