

1 Automatic identification of players in the flavonoid 2 biosynthesis with application on the biomedical plant 3 *Croton tiglium*

4 Boas Pucker^{1,2}, Franziska Reiher¹ and Hanna Marie Schilbert^{1,*}

5 ¹ Genetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, 33615 Bielefeld, Germany;
6 bpucker@cebitec.uni-bielefeld.de (B.P.); freiher@cebitec.uni-bielefeld.de (F.R.); hschilbe@cebitec.uni-bielefeld.de
7 (H.M.S.)

8 ² Evolution and Diversity, Department of Plant Sciences, University of Cambridge, Cambridge United Kingdom;
9 bpucker@cebitec.uni-bielefeld.de (B.P.)

10 * Correspondence: hschilbe@cebitec.uni-bielefeld.de (H.M.S.)

11

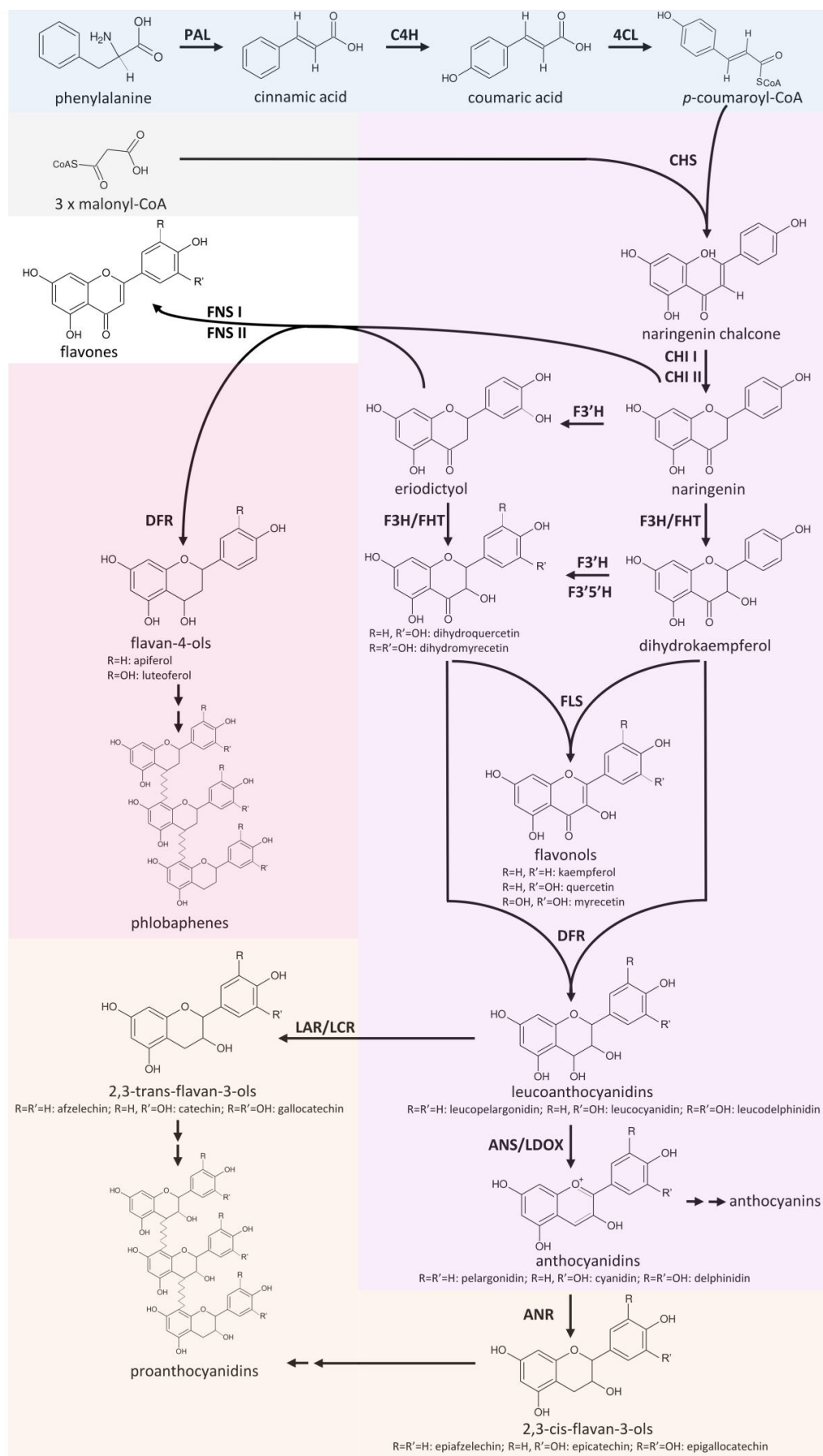
12 **Abstract:** The flavonoid biosynthesis is a well characterised model system for specialised metabolism and
13 transcriptional regulation in plants. Flavonoids have numerous biological functions like UV protection and
14 pollinator attraction, but also biotechnological potential. Here, we present Knowledge-based Identification of
15 Pathway Enzymes (KIPEs) as an automatic approach for the identification of players in the flavonoid
16 biosynthesis. KIPEs combines comprehensive sequence similarity analyses with the inspection of functionally
17 relevant amino acid residues and domains in subjected peptide sequences. Comprehensive sequence sets of
18 flavonoid biosynthesis enzymes and knowledge about functionally relevant amino acids were collected. As a
19 proof of concept, KIPEs was applied to investigate the flavonoid biosynthesis of the medicinal plant *Croton*
20 *tiglium* based on a transcriptome assembly. Enzyme candidates for all steps in the biosynthesis network were
21 identified and matched to previous reports of corresponding metabolites in *Croton* species.

22

23 **Keywords:** anthocyanins, flavonols, proanthocyanidins, general phenylpropanoid pathway, transcriptional
24 regulation, plant pigments, cross-species transcriptomics, specialised metabolism, functional annotation
25

26 1. Introduction

27 Flavonoids are a group of specialised plant metabolites comprising more than 9,000 identified compounds
28 [1] with numerous biological functions [2]. Flavonoids are derived from the aromatic amino acid phenylalanine
29 in a branch of the phenylpropanoid pathway namely the flavonoid biosynthesis (Figure 1). Generally,
30 flavonoids consist of two aromatic C6-rings and one heterocyclic pyran ring [3]. Products of the flavonoid
31 biosynthesis can be assigned to different subgroups, including chalcones, flavones, flavonols, flavandiols,
32 anthocyanins, proanthocyanidins (PA), and aurones [4]. These subclasses are characterised by different
33 oxidation states [5]. In plants, these aglycons are often modified through the addition of various sugars leading
34 to a huge diversity [6].



35 **Figure 1:** Simplified illustration of the general phenylpropanoid pathway and the core flavonoid aglycon
36 biosynthesis network.

37 Flavonoids have important developmental and ecological roles in plants including the control of auxin
38 transport [7], the attraction of pollinators [8], protection of plants against UV light [9], and defense against
39 pathogens and herbivores [10]. Different types of flavonoids can take up these roles. Anthocyanins appear as
40 violet, blue, orange, or red pigments in plants recruiting pollinators and seed dispersers [8]. PAs accumulate in
41 the seed coat leading to the characteristic dark colour of seeds in many species [8]. Flavonols are stored in their
42 glycosylated form in the vacuole of epidermal cells or on occasion in epicuticular waxes [4]. They possess
43 several physiological functions including antimicrobial defense, scavenging of reactive oxygen species (ROS),
44 UV protection, signaling, and colouration of flower pigmentation together with anthocyanins [9].
45 Consequently, the activity of different branches of the flavonoid biosynthesis needs to be adjusted in response
46 to developmental stages and environmental conditions. While the biosynthesis of anthocyanins can be triggered
47 by abiotic factors such as light, temperature, dryness or salts [11], PAs are formed independently of external
48 stimuli in the course of seed development leading to a brown seed colour [11].

49 As the accumulation of flavonoids in fruits and vegetables [12] leads to colouration desired by customers,
50 this pigment pathways is of biotechnological relevance. Therefore, the flavonoid biosynthesis was previously
51 modified by genetic engineering in multiple species (as reviewed in [13]). Flavonoids are not just interesting
52 colourants, but have been reported to have nutritional benefits [14] and even potential in medical applications
53 [15]. Reported anti-oxidative, anti-inflammatory, anti-mutagenic, and anti-carcinogenic properties of
54 flavonoids provide health benefits to humans [16]. For example, kaempferols are assumed to inhibit cancer cell
55 growth and induce cancer cell apoptosis [17]. Heterologous production of flavonoids in plants is considered a
56 promising option to meet customers' demands. Studies already demonstrated that the production of
57 anthocyanins in plant cell cultures is possible [18,19].

58 The flavonoid biosynthesis is one of the best-studied pathways in plants thus serving as a model system for
59 the investigation of specialised metabolism [9]. Academic interest in the synthesis of flavonoids spans multiple
60 fields including molecular genetics, chemical ecology, biochemistry, and health sciences [9,20]. Especially the
61 three subgroups flavonols, anthocyanins, and PAs are well studied in the model organism *Arabidopsis thaliana*
62 [21]. Since a partial lack of flavonoids is not lethal under most conditions, there are large numbers of mutants
63 with visible phenotypes caused by the knockout of various genes in the pathway [22]. For example, seeds
64 lacking PAs show a yellow phenotype due to the absence of brown pigments in the seed coat which inspired the
65 name of mutants in this pathway: *transparent testa* [23]. While the early steps of the flavonoid aglycon
66 biosynthesis are very well known, some later steps require further investigation. Especially the transfer of
67 sugars to PAs and anthocyanidins offer potential for future discoveries [24].
68

69 The core pathway of the flavonoid aglycon biosynthesis comprises several key steps which allow effective
70 channeling of substrates in specific branches (Figure 1). A type III polyketide synthase, the chalcone synthase
71 (CHS), catalyses the initial step of the flavonoid biosynthesis which is the conversion of *p*-coumaroyl-CoA and
72 three malonyl-CoA into naringenin chalcone [25]. Since a knock-out or down-regulation of this step influences
73 all branches of the flavonoid biosynthesis, CHS is well studied in a broad range of species. Flower colour
74 engineering with CHS resulted in the identification of mechanisms for the suppression of gene expression [26].
75 *A. thaliana* CHS can be distinguished from very similar stilbene synthases (STS) based on two diagnostic
76 amino acid residues Q166 and Q167, while a STS would show Q166 H167 or H166 Q167 [27]. The chalcone
77 isomerase (CHI) catalyses the conversion of bicyclic chalcones into tricyclic (S)-flavanones [28]. CHI I
78 converts 6'-tetrahydroxychalcone to 5-hydroxyflavanone, while CHI II additionally converts 6'-deoxychalcone
79 to 5-dexoyflavanone [29]. An investigation of CHI in early land plants revealed the presence of CHI II, which is
80 in contrast to the initial assumption that CHI II activity would be restricted to legumes [30]. A detailed theory
81 about the evolution of functional CHIs from non-enzymatic fatty acid binding proteins and the origin of
82 CHI-like proteins was developed based on evolution experiments [31]. The CHI product naringenin can be
83 processed by different enzymes broadening the flavonoid biosynthesis pathway to a metabolic network.

84 Flavanone 3 β -hydroxylase (F3H/FHT) catalyses 3-hydroxylation of naringenin to dihydroflavonols [32].
85 As a member of the 2-oxoglutarate-dependent dioxygenase (2-ODD) family, F3H utilises the same cofactors
86 and cosubstrate as the two other 2-ODD enzymes in the flavonoid biosynthesis: flavonol synthase (FLS) and
87 leucoanthocyanidin dioxygenase (LDOX) / anthocyanidin synthase (ANS) [33]. The 2-ODD enzymes share
88 overlapping substrate and product selectivities [34]. FLS was identified to be a bifunctional enzyme showing
89 F3H activity in some species including *A. thaliana* [35], *Oryza sativa* [36], and *Ginkgo biloba* [37]. ANS, an

90 enzyme of a late step in the flavonoid biosynthesis pathway, can have both, FLS and F3H activity [38–41]. Due
91 to its FLS side-activity, ANS has to be considered as an additional candidate for the synthesis of flavonols. The
92 flavonoid 3'-hydroxylase (F3'H) catalyses the conversion of naringenin to eriodictyol and the conversion of
93 dihydrokaempferol to dihydroquercetin [42]. Expression and activity of flavonoid 3'5'-hydroxylase (F3'5'H) is
94 essential for the formation of 5'-hydroxylated anthocyanins which cause the blue colour of flowers [13,43].
95 F3'5'H competes with FLS for dihydroflavonols thus it is possible that F3'5'H processes only the excess of
96 these substrates that surpass the FLS capacity [44]. Functionality of enzymes like F3'5'H or F3'H is determined
97 by only a few amino acids. A T487S mutation converted a *Gerbera hybrida* F3'H into a F3'5'H and the reverse
98 mutation in an *Osteospermum hybrida* F3'5'H deleted the F3'5'H activity almost completely while F3'H
99 activity remained [45]. The central enzyme in the flavonol biosynthesis is FLS, which converts a
100 dihydroflavonol into the corresponding flavonol by introducing a double bond between C-2 and C-3 of the
101 heterocyclic pyran ring (Figure 1)[46,47]. FLS activity was first identified in irradiated parsley cells [48] and has
102 then been characterised in several species including *Petunia hybrida* [46], *A. thaliana* [49], and *Zea mays* [24],
103 revealing species-specific substrate specificities and affinities.

104 Another branching pathway channels naringenin into the flavone synthesis. Together with flavonols,
105 flavones occur as primary pigments in white flowers and function as co-pigments with anthocyanins in blue
106 flowers [50]. Flavanones can be oxidized to flavones by flavanone synthase I (FNS I) [51] and FNS II [52].
107 Hence, FNS I and FNS II compete with F3H for flavanones and present a branching reaction in the flavonoid
108 biosynthesis [53]. Being a 2-ODD, FNS I shows only minor differences in its catalytic mechanism compared to
109 F3H, which are determined by only seven amino acid residues [53]. The exchange of all seven residues in
110 parsley F3H resulted in a complete change to FNS I activity [53].

112 Colourful pigments are generated in the anthocyanin and proanthocyanidin biosynthesis. The
113 NADPH-dependent reduction of dihydroflavonols to leucoanthocyanidins by dihydroflavonol-4-reductase
114 (DFR) is the first committed step of the anthocyanin and proanthocyanidin biosynthesis. There is a competition
115 between FLS and DFR for dihydroflavonols [54]. DFR enzymes have different preferences for various
116 dihydroflavonols (dihydrokaempferol, dihydroquercetin, and dihydromyricetin). The molecular basis of these
117 preferences are probably due to differences in a 26-amino acid substrate binding domain of these enzymes [55].
118 N at position 3 of the substrate determining domain was associated with recognition of all three
119 dihydroflavonols [55]. D at position 3 prevented the acceptance of dihydrokaempferols [55], while a L or A lead
120 to a preference for dihydrokaempferol and substantially reduced the processing of dihydromyricetin [55,56].
121 Although this position is central for the substrate specificity, other positions contribute to the substrate
122 specificity [57]. ANS catalyses the last step in the anthocyanin aglycon biosynthesis, the conversion of
123 leucoanthocyanidins into anthocyanidins. The NADPH/NADH-dependent isoflavone-like reductases,
124 leucoanthocyanidin reductase (LAR) / leucocyanidin reductase (LCR), and anthocyanidin reductase (ANR,
125 encoded by *BANYULS* (*BAN*)) are members of the reductase epimerase dehydrogenase superfamily [58]. LAR
126 channels leucoanthocyanidins into the proanthocyanidin biosynthesis which is in competition with the
127 anthocyanidin formation catalysed by ANS. There is also a competition between 3-glucosyltransferases (3GT)
128 and ANR for anthocyanidins [59]. While 3GT generates stable anthocyanins through the addition of a sugar
129 group to anthocyanidins, ANR channels anthocyanidins into the proanthocyanidin biosynthesis.
130 Anthocyanidins are unstable in aqueous solution and fade rapidly unless the pH is extremely low [60].
131 Suppression of *ANR1* and *ANR2* in *Glycine max* caused the formation of red seeds through a reduction in
132 proanthocyanidin biosynthesis and an increased anthocyanin biosynthesis [61]. Substrate preferences of ANR
133 can differ between species as demonstrated for *A. thaliana* and *M. truncatula* [62].

135 As a complex metabolic network with many branches, the flavonoid biosynthesis requires sophisticated
136 regulation. Activity of different branches is mainly regulated at the transcriptional level [63]. In *A. thaliana* as in
137 many other plants, R2R3-MYBs [64,65] and basic helix-loop-helix proteins (bHLH) [66] are two main
138 transcription factor families involved in the regulation of the flavonoid biosynthesis. The WD40 protein TTG1
139 facilitates the interaction of R2R3-MYBs and bHLHs in the regulation of the anthocyanin and proanthocyanidin
140 biosynthesis in *A. thaliana* [67]. Due to its components, this trimeric complex is also referred to as MBW
141 complex [67]. Examples of MBW complexes are MYB123 / bHLH42 / TTG1 and MYB75 / bHLH2 / TTG1,
142 which are involved in anthocyanin biosynthesis regulation in a tissue-specific manner [68]. However, the
143 bHLH-independent R2R3-MYBs like MYB12, MYB11, and MYB111 can activate as single transcriptional
144 activators early genes of the flavonoid biosynthesis including *CHS*, *CHI*, *F3H*, and *FLS* [69].

145
146 Many previous studies performed a systematic investigation of the flavonoid biosynthesis in plant species
147 including *Fragaria x ananassa* [70], *Musa acuminata* [71], *Tricyrtis* spp. [72], and multiple *Brassica* species
148 [73]. In addition to these systematic investigations, genes of the flavonoid biosynthesis are often detected as
149 differentially expressed in transcriptomic studies without particular focus on this pathway [74–76]. In depth
150 investigation of the flavonoid biosynthesis starts with the identification of candidate genes for all steps. This
151 identification of candidates is often relying on an existing annotation or requires tedious manual inspection of
152 sequence alignments. As plant genome sequences and their structural annotations become available with an
153 increasing pace [77], the timely addition of functional annotations is an ever increasing challenge. Therefore,
154 we developed a pipeline for the automatic identification of flavonoid biosynthesis players in any given set of
155 peptide, transcript, or genomic sequences. As a proof of concept, we validate the predictions made by
156 Knowledge-based Identification of Pathway Enzymes (KIPes) with a manual annotation of the flavonoid
157 biosynthesis in the medicinal plant *Croton tiglium*. *C. tiglium* is a member of the family Euphorbiaceae [78] and
158 was first mentioned over 2,200 years ago in China as a medicinal plant probably because of the huge variety of
159 specialised metabolites [79]. Oil of *C. tiglium* was traditionally used to treat gastrointestinal disorders and may
160 have abortifacient and counterirritant effects [80]. Additionally, *C. tiglium* produces phorbol esters and a
161 ribonucleoside analog of guanosine with antitumor activity [81,82]. Characterization of the specialised
162 metabolism of *C. tiglium* will facilitate the unlocking of its potential in agronomical, biotechnological, and
163 medical applications. The flavonoid biosynthesis of *C. tiglium* is largely unexplored. To the best of our
164 knowledge, previous studies only showed the presence of flavonoids through analysis of extracts [83–85].
165 However, transcriptomic resources are available [86] and provide the basis for a systematic investigation of the
166 flavonoid biosynthesis in *C. tiglium*.

167
168 A huge number of publicly available genome and transcriptome assemblies of numerous plant species
169 provide a valuable resource for comparative analysis of the flavonoid biosynthesis. Here, we present an
170 automatic workflow for the identification of flavonoid biosynthesis genes applicable to any plant species and
171 demonstrate the functionality by analyzing a *de novo* transcriptome assembly of *C. tiglium*.

172

173 2. Results

174 We developed a tool for the automatic identification of enzyme sequences in a set of peptide sequences, a
175 transcriptome assembly, or a genome sequence. Knowledge-based Identification of Pathway Enzymes (KIPes)
176 identifies candidate sequences based on overall sequence similarity, functionally relevant amino acid residues,
177 and functionally relevant domains (Figure 2). As a proof of concept, the transcriptome assembly of *Croton*
178 *tiglium* was screened with KIPes to identify the flavonoid aglycon biosynthesis network. Results of the
179 automatic annotation are validated by a manually curated annotation.

180

181 2.1. Concept and components of Knowledge-based Identification of Pathway Enzymes (KIPes)

182 2.1.1. General concept

183 The automatic detection of sequences encoding enzymes of the flavonoid biosynthesis network requires
184 (1) a set of bait sequences covering a broad taxonomic range and (2) information about functionally relevant
185 amino acid residues and domains. Bait sequences were selected to encode enzymes with evidence of
186 functionality i.e. mutant complementation studies or *in vitro* assays. Additional bait sequences were included
187 which were previously studied in comparative analyses of the particular enzyme family. Positions of amino
188 acids and domains with functional relevance need to refer to a reference sequence included in the bait sequence
189 set. All bait sequences and one reference sequence related to one reaction in the network are supplied in one
190 FASTA file. However, many FASTA files can be provided to cover all reactions of a complete metabolic
191 network. Positions of functionally relevant residues and domains are specified in an additional text file based on
192 the reference sequence (see manual for details, <https://github.com/bpucker/KIPes>). Collections of bait
193 sequences and detailed information about the relevant amino acid residues in flavonoid biosynthesis enzymes
194 are provided along with KIPes. However, these collections can be customized by users to reflect updated

195 knowledge and specific research questions. KIPes was developed to have a minimal amount of dependencies.
196 Only the frequently used alignment tools BLAST and MAFFT are required. Both tools are freely available as
197 precompiled binaries without the need for installation.

198

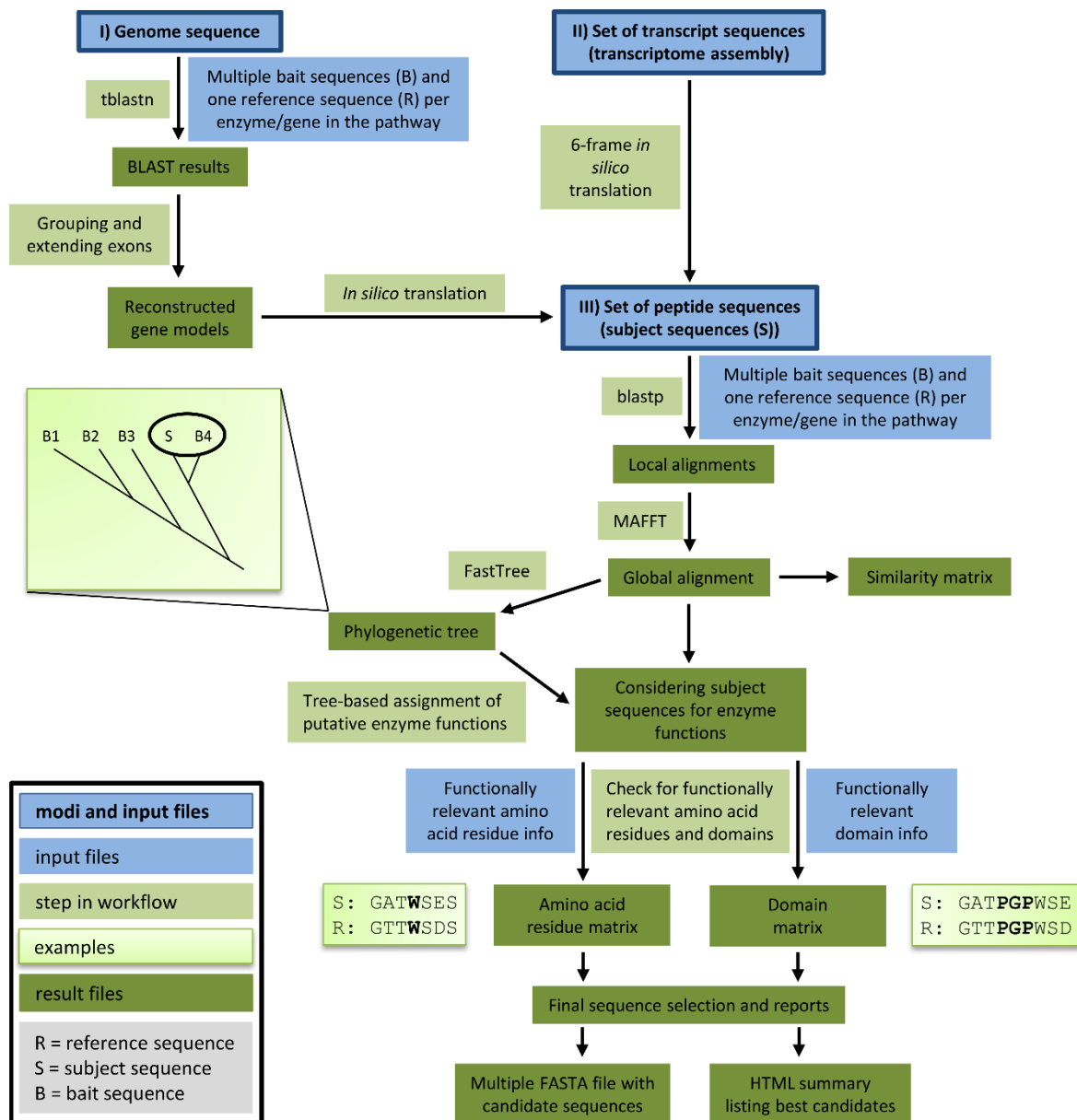
199 2.1.2. Three modes

200 A user can choose between three different analysis modes depending on the available input sequences:
201 peptide sequences, transcript sequences, or a genome sequence. If a reliable peptide sequence annotation is
202 available, these peptide sequences should be subjected to the analysis. Costs in terms of time and computational
203 efforts are substantially lower for the analysis of peptide sequences than for the analysis of genome sequences.
204 The provided peptide sequences are screened via blastp for similarity to previously characterised bait
205 sequences. If default criteria are applied, BLAST hits are considered if the sequence similarity is above 40%
206 and if the score is above 30% of the score resulting from an alignment of the query sequence against itself.
207 These lenient filter criteria are applied to collect a comprehensive set of candidate sequences which is
208 subsequently refined through the construction of global alignments via MAFFT. Next, phylogenetic trees are
209 generated to identify best candidates based on their position in a tree. Candidates are classified based on the
210 closest distance to a bait sequence. Multiple closely related bait sequences can be considered if specified. When
211 transcript sequences are supplied to KIPes, *in silico* translation in all six possible frames generates a set of
212 peptide sequences which are subsequently analysed as described above. Supplied DNA sequences are screened
213 for similarity to the bait peptide sequences via tblastn. Hits reported by tblastn are considered exons or exon
214 fragments and therefore assigned to groups which might represent candidate genes. The connection of these hits
215 is attempted in a way that canonical GT-AG splice site combinations emerge. One isoform per locus is
216 constructed and subsequently analysed as described above.

217

218 2.1.3. Final filtering

219 After identification of initial candidates through overall sequence similarity, a detailed comparison against
220 a well characterised reference sequence with described functionally relevant amino acid residues is performed.
221 All candidates are screened for matching amino acid residues at functionally relevant positions. Sequences
222 encoding functional enzymes are expected to display a matching amino acid residue at all checked positions.
223 Additionally, the conservation of relevant domains is analysed. A prediction about the
224 functionality/non-functionality of the encoded enzyme of all candidate sequences is performed at this step.
225 Results of intermediate steps are stored to allow in depth inspection if necessary.



226

227

228

229

230

231

232

233 2.2. Technical validation of KIPeS

234

235

236

237

238

239

240

241

Figure 2. This overview illustrates the components and steps of Knowledge-based Identification of Pathway Enzymes (KIPeS). Three different modes allow the screening of peptide, transcript, or genome sequences for candidate sequences. Bait sequences and information about functionally relevant features (blue) are supplied by the user. Different modules of KIPeS (light green) are executed consecutively depending on the type of input data. Intermediate results and the final output (dark green) are stored to keep the process transparent.

A first technical validation of KIPeS was performed based on sequence data sets of plant species with previously characterized flavonoid biosyntheses namely *A. lyrata*, *A. thaliana*, *Cicer arietinum*, *Fragaria vesca*, *Glycine max*, *Malus domestica*, *Medicago truncatula*, *Musa acuminata*, *Populus trichocarpa*, *Solanum lycopersicum*, *Solanum tuberosum*, *Theobroma cacao*, and *Vitis vinifera*. The flavonoid biosynthesis of these species was previously characterised thus providing an opportunity for validation. KIPeS identified candidate sequences with conservation of all functionally relevant amino acid residues for the expected enzymes in all species (File S1).

242 2.3. The flavonoid biosynthesis enzymes in *Croton tiglium*

243 Genes in the flavonoid biosynthesis of *C. tiglium* were identified based on bait sequences of over 200 plant
244 species and well characterised reference sequences of *A. thaliana*, *G. max*, *M. sativa*, *Osteospermum* spec.,
245 *Petroselinum crispum*, *P. tomentosa*, and *V. vinifera*. The transcriptome assembly of *C. tiglium* revealed
246 sequences encoding enzymes for all steps in the flavonoid biosynthesis (Table 1). Phylogenetic analyses placed
247 the *C. tiglium* sequences of enzymes in the flavonoid biosynthesis close to the corresponding sequences of
248 related *Malpighiales* species like *Populus tomentosa* (File S2). Conservation of functionally relevant amino
249 acid residues was inspected in an alignment with sequences of characterised enzymes of the respective step
250 (File S3).

251 The general phenylpropanoid biosynthesis is represented by ten phenylalanine ammonia lyase (PAL)
252 candidates, two cinnamate 4-hydroxylases (C4H) candidates, and one 4-coumarate-CoA ligase (4CL) candidate
253 (Table 1, File S4, File S5). Many PAL sequences show a high overall sequence similarity indicating that
254 multiple alleles or isoforms could contribute to the high number. A phylogenetic analysis supports the
255 hypothesis that many PAL candidates might be alleles or alternative transcript variants of the same genes (File
256 S2). Very low transcript abundances indicate that at least three of the PAL candidates can be neglected (Table
257 1).

258 Although multiple CHS candidates were identified based on overall sequence similarity to the *A. thaliana*
259 CHS sequence, only CtCHSa showed all functionally relevant amino acid residues (File S3). Five other
260 candidates were discarded due to the lack of Q166 and Q167, which differentiate CHS from other polyketide
261 synthases like STS or LAP5. Additionally, a CHS signature sequence at the C-terminal end and the
262 malonyl-CoA binding motif at position 313 to 329 in the *A. thaliana* sequence are conserved in CtCHSa. A
263 phylogenetic analysis supported these findings by placing CtCHSa in a clade with *bona fide* chalcone synthases
264 (File S2). There is only one CHI candidate, CtCHI Ia, which contains all functionally relevant amino acid
265 residues (File S3). No CHI II candidate was detected. *C. tiglium* has one F3H candidate, one F3'H candidate,
266 and two F3'5'H candidates. CtF3Ha, CtF3'Ha, CtF3'5'Ha, and CtF3'5'Hb show conservation of the respective
267 functionally relevant amino acid residues (File S3). CtF3'Ha contains the N-terminal proline rich domain and a
268 perfectly conserved oxygen binding pocket at position 302 to 307 in the *A. thaliana* reference sequence. Both,
269 CtF3'5'Ha and CtF3'5'Hb, were also considered as F3'H candidates, but show overall a higher similarity to the
270 F3'5'H bait sequences than to the F3'H bait sequences. The flavone biosynthesis capacities of *C. tiglium*
271 remained elusive. No FNS I candidates with conservation of all functionally relevant amino acids were
272 detected. However, there are four FNS II candidates which show only one substitution of an amino acid residue
273 in the oxygen binding pocket (T313F). The committed step of the flavonol biosynthesis is represented by
274 CtFLSa and CtFLSb which show all functionally relevant residues (File S3).

275 *C. tiglium* contains excellent candidates for all steps of the anthocyanidin and proanthocyanidin
276 biosynthesis. CtDFR shows conservation of the functionally relevant amino acid residues (File S3). We
277 investigated the substrate specificity domain to understand the enzymatic potential of the DFR in *C. tiglium*.
278 Position 3 of this substrate specificity domain shows a D which is associated with low acceptance of
279 dihydrokaempferols. CtLAR is the only LAR candidate with conservation of the functionally relevant amino
280 acid residues (File S3). CtANS is the only ANS candidate with conservation of the functionally relevant amino
281 acid residues (File S3). There are two ANR candidates in *C. tiglium*. CtANRa and CtANRb show conservation
282 of all functionally relevant amino acid residues (File S3). CtANRa shows 74% identical amino acid residues
283 when compared to the reference sequence, which exceeds the 49% of CtANRb substantially.

284 The identification of candidates in a transcriptome assembly already shows transcriptional activity of the
285 respective gene. To resolve the transcriptional activity of genes in greater detail, we quantified the presence of
286 candidate transcripts in different tissues of *C. tiglium* and compared it to *C. draco* through cross-species
287 transcriptomics (File S6). High transcript abundance of almost all flavonoid biosynthesis candidates was
288 observed in seeds, while only a few candidate transcripts were observed in other investigated tissues (Table 1).
289 Transcripts involved in the proanthocyanidin biosynthesis show an exceptionally high abundance in seeds of *C.*
290 *tiglium* and inflorescence of *C. draco*. Overall, the tissue specific abundance of many transcripts is similar
291 between *C. tiglium* and *C. draco*. LAR and ANR show substantially higher transcript abundances in

292 inflorescences of *C. draco* compared to *C. tiglium*. CHS and ANS show the highest transcript abundance in pink
 293 flowers of *C. draco* (File S6).

294

295 **Table 1.** Candidates in the flavonoid biosynthesis of *Croton tiglium*. ‘TRINITY’ prefix of all sequence names was omitted
 296 for brevity. Candidates are sorted by their position in the respective pathway and decreasing similarity to bait sequences.

297 Transcripts per million (TPM) values of the candidates in different tissues are shown: leaf (SRR6239848), stem
 298 (SRR6239849), inflorescence (SRR6239850), root (SRR6239851), and seed (SRR6239852). Displayed values are rounded
 299 to the closest integer thus extremely low abundances are listed as 0. A full table with all available RNA-Seq samples and
 300 transcript abundance values for all candidates is available in the supplements (File S6).

Sequence ID	Function	Leaf	Stem	Inflorescence	Root	Seed
DN23351_c0_g1_i2	<i>CtPALa</i>	18	10	0	1	29
DN32981_c5_g1_i1	<i>CtPALb</i>	0	0	0	0	0
DN32981_c5_g1_i16	<i>CtPALc</i>	4	3	0	0	0
DN32981_c5_g1_i9	<i>CtPALd</i>	0	0	0	0	0
DN32981_c5_g1_i17	<i>CtPALe</i>	0	25	0	5	10
DN32981_c5_g1_i12	<i>CtPALf</i>	5	233	0	18	48
DN32981_c5_g1_i5	<i>CtPALg</i>	0	318	0	3	3
DN32981_c5_g1_i13	<i>CtPALh</i>	0	3	0	0	0
DN32981_c5_g1_i14	<i>CtPALi</i>	0	0	0	0	0
DN23351_c0_g2_i1	<i>CtPALj</i>	18	1	0	9	12
DN32464_c6_g3_i2	<i>CtC4Ha</i>	122	110	2	77	233
DN15593_c0_g1_i1	<i>CtC4Hb</i>	0	0	0	0	3
DN32164_c5_g1_i2	<i>Ct4CLa</i>	46	19	1	2	113
DN50385_c0_g1_i1	<i>CtCHSa</i>	3	6	0	1	588
DN27125_c0_g1_i1	<i>CtCHI Ia</i>	11	2	19	3	88
DN33424_c3_g3_i1	<i>CtF3Ha</i>	4	21	1	2	342
DN33407_c7_g7_i2 ¹	<i>CtFNS IIa</i>	1	7	0	95	1
DN33407_c7_g7_i1 ¹	<i>CtFNS IIb</i>	0	3	1	4	2
DN27999_c0_g1_i2 ¹	<i>CtFNS IIc</i>	0	2	0	75	0
DN33407_c7_g6_i4 ¹	<i>CtFNS IId</i>	0	0	0	22	0
DN252_c0_g1_i1	<i>CtF3'Ha</i>	111	62	0	9	165
DN32466_c16_g7_i1	<i>CtF3'5'Ha</i>	0	0	0	7	266
DN32466_c16_g7_i3	<i>CtF3'5'Hb</i>	0	0	0	0	3
DN25915_c0_g1_i3	<i>CtFLSa</i>	18	19	0	0	84
DN25915_c0_g2_i1	<i>CtFLSb</i>	0	1	0	1	2
DN27402_c0_g1_i3	<i>CtDFRa</i>	0	0	0	0	51
DN32893_c8_g1_i1	<i>CtANSa</i>	0	0	0	0	25
DN33042_c3_g1_i3	<i>CtLARa</i>	3	2	0	1	101
DN30161_c9_g1_i2	<i>CtANRa</i>	0	1	0	1	375
DN30161_c9_g1_i3	<i>CtANRb</i>	0	0	0	0	3

301

302

303

¹ These sequences might encode non-functional enzymes or enzymes with a different function (see results and discussion for details), but represent the best FNS II candidates.

304

305 *2.4. Transcriptional regulators of the flavonoid biosynthesis in Croton tiglium*

306

307 To demonstrate the applicability of KIPes for the investigation of non-enzyme sequences like
 308 transcription factor gene families, we screened the transcriptome assembly of *C. tiglium* for members of the
 309 MYB, bHLH, and WD40 family. This analysis revealed candidates for some key regulators of the flavonoid
 310 biosynthesis namely MYB11/MYB12/MYB111 (subgroup7), MYB123 (subgroup5),
 311 MYB75/MYB90/MYB113/MYB114 (subgroup6), bHLH2/bHLH42, and TTG1 according to the nomenclature
 312 in *A. thaliana* (Table 2, File S7). The MYB subgroups 6 and 7 have multiple members in *A. thaliana* and *C.*
 313 *tiglium*. Therefore, *C. tiglium* candidates are only assigned to an orthogroup (Table 2). The reliable
 314 identification of MYB orthologs between both species was not feasible (File S7). There are five homologous
 315 sequences of MYB123 in *C. tiglium* with one of them probably originating from the same gene. The R2R3
 316 MYB domain was detected in the MYB candidates except for DN21046_c0_g1_i3, DN21046_c0_g1_i3,
 317 DN30455_c10_g1_i1, and DN33314_c5_g2_i4. With the exception of DN33314_c5_g2_i4 (truncated protein)
 318 all CtMYB candidates of subgroup6 have a conserved bHLH interaction domain, while the CtMYB candidates
 319 of the bHLH-independent subgroup7 do not show this conserved domain. There are seven *C. tiglium* sequences
 320 in a clade with the *A. thaliana* bHLH42 (File S7), but these might be alternative isoforms originating from the
 321 same gene. The same is true for the seven isoforms detected as homologous sequences of *A. thaliana* bHLH2
 322 (File S7). Three TTG1 candidates exist in the *C. tiglium* transcriptome assembly, but two of them might be
 323 isoforms belonging to the same gene. The MYB, bHLH, and TTG1 transcription factor candidates show
 324 generally lower transcript abundances than the enzyme candidates (Table 1, Table 2). The highest transcript
 325 abundance of all three MBW complex components was observed in seeds.

326

327 **Table 2.** Transcriptional regulator candidates of the flavonoid biosynthesis. MYB11/MYB12/MYB111 candidates are
 328 summarised as subgroup7 MYBs. MYB75/MYB90/MYB113/MYB114 are summarised as subgroup6. Transcripts per
 329 million (TPM) values of the candidates in different tissues are shown: leaf (SRR6239848), stem (SRR6239849),
 330 inflorescence (SRR6239850), root (SRR6239851), and seed (SRR6239852). Displayed values are rounded to the closest
 331 integer thus extremely low abundances are listed as 0.

Sequence ID	Group	Leaf	Stem	Inflorescence	Root	Seed
DN30455_c10_g1_i1	Subgroup7	0	1	0	0	4
DN21046_c0_g1_i3	Subgroup7	0	0	0	0	0
DN21046_c0_g1_i2	Subgroup7	0	1	0	0	9
DN28041_c1_g1_i4	Subgroup6	0	0	0	0	7
DN28041_c1_g1_i2	Subgroup6	0	0	0	0	0
DN33356_c3_g1_i2	Subgroup6	0	0	0	0	4
DN31144_c5_g1_i2	MYB123	0	0	0	10	29
DN33314_c5_g2_i2	MYB123	3	8	0	1	14
DN33314_c5_g2_i3	MYB123	0	0	0	0	1
DN33314_c5_g2_i4	MYB123	1	2	0	0	6
DN31260_c4_g2_i2	MYB123	0	0	0	0	5
DN30681_c1_g1_i1	bHLH2	0	0	0	0	0
DN30681_c1_g1_i2	bHLH2	0	2	0	2	4
DN30681_c1_g1_i3	bHLH2	0	0	0	0	0
DN30681_c1_g1_i6	bHLH2	0	0	0	0	0
DN30681_c1_g1_i7	bHLH2	2	13	0	16	16

DN30681_c1_g1_i8	bHLH2	2	9	0	20	18
DN30681_c1_g1_i9	bHLH2	0	0	0	0	0
DN32219_c4_g2_i2	bHLH42	0	0	0	0	0
DN32219_c4_g2_i5	bHLH42	0	0	0	0	2
DN32219_c4_g2_i12	bHLH42	0	0	0	0	2
DN32219_c4_g2_i11	bHLH42	0	0	0	0	3
DN32219_c4_g2_i4	bHLH42	0	1	0	0	25
DN32219_c4_g2_i7	bHLH42	0	0	0	0	4
DN32219_c4_g2_i8	bHLH42	0	0	0	0	5
DN32272_c1_g1_i1	TTG1	0	0	0	0	0
DN32272_c1_g2_i2	TTG1	12	8	4	10	9
DN32604_c4_g1_i2	TTG1	0	1	0	1	1

332

333 3. Discussion

334 As previous studies of extracts from *Croton tiglium* and various other *Croton* species revealed the
 335 presence of flavonoids [84,85,87–92], steps in the central flavonoid aglycon biosynthesis network should be
 336 represented by at least one functional enzyme each. However, this is the first identification of candidates
 337 involved in the biosynthesis. Previous reports [84,85,87–92] about flavonoids align well with our observation
 338 (Table 1) that at least one predicted peptide contains all previously described functionally relevant amino acid
 339 residues of the respective enzyme. The only exception is the flavone synthase step. While FNS I is frequently
 340 absent in flavonoid producing species outside the *Apiaceae*, FNS II is more broadly distributed across plants
 341 [53]. *C. tiglium* is not a member of the *Apiaceae* thus the absence of FNS I and the presence of FNS II
 342 candidates are expected.

343 All candidate sequences of presumably functional enzymes belong to actively transcribed genes as
 344 indicated by the presence of these sequences in a transcriptome assembly. Since the flavonoid biosynthesis is
 345 mainly regulated at the transcriptional level [63] and previously reported blocks in the pathway are expected to
 346 be due to transcriptional down-regulation [93,94], we expect most branches of the flavonoid biosynthesis in *C.*
 347 *tiglium* to be functional. No CHI II candidate was detected thus *C. tiglium* probably lacks a 6'-deoxychalcone to
 348 5-dexoyflavanone catalytic activity like most non-leguminous plants [30,95].

349 A domination of proanthocyanidins has been reported for *Croton* species [88]. This high proanthocyanidin
 350 content correlates well with high transcript abundance of proanthocyanidin biosynthesis genes (*CtLAR*,
 351 *CtANR*). PAs have been reported to account for up to 90% of the dried weight of red sap of *Croton lechleri* [96].
 352 Expression of *CtFLSa* in the leaves matches previous reports about flavonol extraction from leaves [90,97].
 353 Interestingly, almost all analysed *Croton* species showed very high amounts of quercetin derivates compared to
 354 kaempferol derivates in their leaf extracts, which significantly correlated with antioxidant potential [97]. This
 355 high quercetin concentration might be due to a high expression level of *CtF3'Ha* in leaves. Since F3'H converts
 356 dihydrokaempferol (DHK) to dihydroquercetin (DHQ), a high gene expression might result in high amounts of
 357 DHQ which can be used from FLS to produce quercetin. At the same time, the production of kaempferols from
 358 DHK is reduced.

359 Flavonols have been extracted from several *Croton* species and several important functions have been
 360 attributed to these flavonols. Quercetin 3,7-dimethyl ether was extracted from *Croton schiedeanus* and elicits
 361 vasorelaxation in isolated aorta [91]. Casticin a methoxylated flavonol from *Croton betulaster* modulates
 362 cerebral cortical progenitors in rats by directly decreasing neuronal death, and indirectly via astrocytes [98].
 363 Besides the anticancer activity of flavonol rich extracts from *Croton celtidifolius* in mice [99], flavonols
 364 extracted from *Croton menyharthii* leaves possess antimicrobial activity [100]. Kaempferol
 365 7-O-β-D-(6"-O-cumaroyl)-glucopyranoside isolated from *Croton piauhiensis* leaves enhanced the effect of
 366 antibiotics and showed antibacterial activity on its own [101]. Flavonols extracted from *Croton cajucara*
 367 showed anti-inflammatory activities [102].

368 The investigating of the CtDFR substrate specificity revealed aspartate at the third position of the substrate
369 specificity domain which was previously reported to reduce the acceptance of dihydrokaempferol [55].
370 Although the substrate specificity of DFR is not completely resolved, a high DHQ affinity would fit to the high
371 transcript abundance of *CtF3'Fs* which encode putative DHQ producing enzymes. Further investigations are
372 needed to reveal how effectively *C. tiglium* produces anthocyanidins and proanthocyanidins based on different
373 dihydroflavonols. As *C. tiglium* is known to produce various proanthocyanidins [83], a functional biosynthetic
374 network must be present. Phlobatannine have been reported in leaves of *C. tiglium* [83] which aligns well with
375 our identification of a probably functional CtDFRa.

376 Our automatic approach for the identification of flavonoid biosynthesis genes could be applied to identify
377 target genes for an experimental validation in a species with a newly sequenced transcriptome or genome. Due
378 to multiple refinement steps, the predictions of KIPes have a substantially higher fidelity than frequently used
379 BLAST results. Especially the distinction of different enzymes with very similar sequences (e.g. CHS, STS,
380 LAP5) was substantially improved by KIPes. Additionally, the automatic identification of flavonoid
381 biosynthesis enzymes/genes across a large number of plant species facilitates comparative analyses which could
382 be a valuable addition to functional studies or might even replace some studies. As functionally relevant amino
383 acid residues are well described for many of the enzymes, an automatic classification of candidate sequences as
384 functional or non-functional is feasible in many cases. It has not escaped our notice that 'non-functionality' only
385 holds with respect to the initially expected enzyme function. Sub- and neofunctionalisation, especially
386 following gene duplications, are likely. Results produced by KIPes could be used to identify species-specific
387 modifications of the general flavonoid biosynthesis. Bi- or even multifunctionality has been described for some
388 members of the 2-ODDs (FLS [36,103,104], F3H, FNS I, and ANS [38–41]). Experimental characterization of
389 these enzymes will still be required to determine the degree of the possible multifunctionalities in one enzyme.
390 However, enzyme characterization experiments could be informed by the results produced by KIPes. As KIPes
391 has a particular focus on high impact amino acid substitutions, it would also be possible to screen sequence data
392 sets of phenotypically interesting plants to identify blocks in pathways. Another potential application is the
393 assessment of the functional impact of amino acid substitutions e.g. in re-sequencing studies. There are
394 established tools like SnpEff [105] for the annotation of sequence variants in re-sequencing studies.
395 Additionally, KIPes could operate on the set of modified peptide sequences to analyse the functional relevance
396 of sequence variants. If functionally relevant amino acids are effected, KIPes could predict that the variant
397 might cause non-functionality.

398
399 Although KIPes can be applied to screen a genome sequence, we recommend to supply peptide or
400 transcript sequences as input whenever possible. Well established gene prediction tools like AUGUSTUS [106]
401 and GeMoMa [107] generate gene models of superior quality in most cases. KIPes is restricted to the
402 identification of canonical GT-AG splice sites. The very low frequency of non-canonical splice sites in plant
403 genomes [108] would cause extreme computational costs and could lead to a substantial numbers of
404 mis-annotations. To the best of our knowledge, non-canonical splice sites have not been reported for genes in
405 the flavonoid biosynthesis. Nevertheless, dedicated gene prediction tools can incorporate additional hints to
406 predict non-canonical introns with high fidelity.

407
408 During the identification of amino acid residues which were previously reported to be relevant for the
409 enzyme function, we observed additional patterns. Certain positions showed not perfect conservation, but
410 multiple amino acids with similar biochemical properties occurred at the respective position. Low relevance of
411 the amino acid at these positions for the enzymatic activity could be one explanation. However, these patterns
412 could also point to lineage specific specializations of various enzymes. A previous study reported the evolution
413 of different F3'H classes in monocots [109]. Subtle differences between isoforms might cause different enzyme
414 properties e.g. altered substrate specificities which could explain the presence of multiple isoforms of the same
415 enzyme in some species. For example, a single amino acid has substantial influence on the enzymatic
416 functionality of F3'H and F3'5'H [45]. This report matches our observation of both F3'5'H candidates being
417 initially also considered as F3'H candidates. A higher overall similarity to the F3'5'H bait sequences than to the
418 F3'H bait sequences allowed an accurate classification. This example showcases the challenges when assigning
419 enzyme functions to peptide sequences.

420
421 We developed KIPes for the automatic identification and annotation of core flavonoid biosynthesis
422 enzymes, because this pathway is well characterised in numerous plant species. Additionally, we demonstrate

423 the applicability for the identification of gene families by screening the transcriptome assembly for MYB,
424 bHLH, and WD40 candidates. Quality and fidelity of the KIPes results depend on the quality of the bait
425 sequence set and the knowledge about functionally relevant amino acid residues. Nevertheless, the
426 implementation of KIPes allows the analysis of additional steps of the flavonoid biosynthesis (e.g. the
427 glycosylation of flavonoids) and even the analysis of other pathways. Here, we presented the identification of
428 enzyme candidates based on single amino acid residues with functional relevance. Functionally characterized
429 domains were subordinate in this enzyme detection process. However, KIPes can also assess the conservation
430 of domains. This function is not only relevant for the analysis of enzymes, but could be applied to the analysis of
431 other proteins like transcription factors with specific binding domains.
432

433 **4. Materials and Methods**

434 4.1. Retrieval of bait and reference sequences

435 The NCBI protein database was screened for sequences of the respective enzyme for all steps in the core
436 flavonoid biosynthesis by searching for the common names. Listed sequences were screened for associated
437 publications about functionality of the respective sequence. Only peptide sequences with evidence for enzyme
438 functionality were retrieved (File S8). To generate a comprehensive set of bait sequences, we also considered
439 sequences with indirect evidence like clear differential expression associated with a phenotype and sequences
440 which were previously included in analyses of the respective enzyme family. The set of bait and reference
441 sequences used for the analyses described in this manuscript is designated FlavonoidBioSynBaits_v1.0.
442

443 4.2. Collection of information about important amino acid residues

444 All bait sequences and one reference sequence per step in the flavonoid biosynthesis were subjected to a
445 global alignment via MAFFT v7 [111]. Highly conserved positions, which were also reported in the literature to
446 be functionally relevant, are referred to as ‘functionally relevant amino acid residues’ in this manuscript (File
447 S9). The amino acid residues and their positions in a designated reference sequence are provided in one table
448 per reaction in the network (<https://github.com/bpucker/KIPes>). A customized Python script was applied to
449 identify contrasting residues between two sequence sets e.g. chalcone and stilbene synthases
450 (<https://github.com/bpucker/KIPes>).
451

452 4.3. Implementation and availability of KIPes

453 KIPes is implemented in Python 2.7. The script is freely available at github:
454 <https://github.com/bpucker/KIPes>. Details about the usage are described in the manual provided along with the
455 Python script. Collections of bait and reference sequences as well as data tables about functionally relevant
456 amino acid residues are included. In summary, these data sets allow the automatic identification of flavonoid
457 biosynthesis genes in other plant species via KIPes. Customization of all data sets is possible to enable the
458 analysis of other pathways. Mandatory dependencies of KIPes are blastp [112], tblastn [112], and MAFFT
459 [111]. FastTree2 [113] is an optional dependency which substantially improves the fidelity of the candidate
460 identification and classification. Positions of candidate sequences in a phylogenetic tree are used to identify the
461 closest bait sequences. The function of the closest bait sequence is then transferred to the candidate. However, it
462 is possible to consider a candidate sequence for multiple different functions. If the construction of phylogenetic
463 trees is not possible, the highest similarity to a bait sequence in a global alignment is used instead to predict a
464 function. An analysis of functionally relevant amino acid residues in the candidate sequences is finally used to
465 assign a function.
466

467 4.4 Phylogenetic analysis

468 Alignments were generated with MAFFT v7 [111] and cleaned with pxclsq [114] to remove alignment
469 columns with very low occupancy. Phylogenetic trees were constructed with FastTree v2.1.10 [113] using the
470 WAG+CAT model. FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>) was used to visualize the phylogenetic
471 trees. Alignments were visualized online at <http://esprict.ibcp.fr/ESPrict/ESPrict/index.php> v3.0 [115] using
472 3D structures of reference enzymes derived from the Protein Data Bank (PDB) [116] (File S10). If no PDB
473 entry was available, the amino acid sequence of the respective reference enzyme was subjected to I-TASSER

474 [117] for protein structure prediction and modelling (File S10, File S11). Functionally relevant amino acid
475 residues in the *C. tiglium* sequences were subsequently highlighted in the generated PDFs (File S3).

476

477 4.5 Transcript abundance quantification

478 All available RNA-Seq data sets of *C. tiglium* [86,118] and *C. draco* [119] were retrieved from the
479 Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>) via fastq-dump v2.9.6
480 (<https://github.com/ncbi/sra-tools>). Kallisto v0.44 [120] was applied with default parameters to quantify the
481 abundance of transcripts based on the *C. tiglium* transcriptome assembly [86].

482

483

484 4.6 Application of KIPEs for the identification of transcription factors

485 KIPEs was run with sets of MYB, bHLH, and WD40 peptide sequences (MYB_bHLH_WD40_v1.0) to
486 identify corresponding candidates in the *C. tiglium* transcriptome assembly. MYB sequences of *A. thaliana*
487 [64], *Vitis vinifera* [121], *Beta vulgaris* [122], and *Musa acuminata* [123] were subject to KIPEs as baits. bHLH
488 bait sequences were collected from *A. thaliana* [124], *V. vinifera* [125], *Nelumbo nucifera* [126], *Citrus grandis*
489 [127], *M. acuminata* [128], and *Solanum melongena* [129]. WD40 sequences of *A. thaliana* [130], *Triticum*
490 *aestivum* [131], and *Setaria italica* [132] were collected as bait sequences for the identification of the WD40
491 protein TTG1. Phylogenetic trees with the candidates reported by KIPEs, the sets of bait sequences derived
492 from the genome-wide studies, and selected sequences retrieved from the NCBI were generated with FastTree
493 v2.1.10 [113] based on alignments constructed with MAFFT v7 [111]. MYB domain and bHLH-interaction
494 domain were identified with a Python script (<https://github.com/bpucker/bananaMYB>) based on previously
495 defined patterns [123].

496

497 5. Conclusions

498 KIPEs enables the automatic identification of enzymes involved in flavonoid biosynthesis in
499 uninvestigated sequence data sets of plants, thus paving the way for comparative studies and the identification
500 of lineage specific differences. While we demonstrate the applicability of KIPEs for the identification and
501 sequence-based characterization of players in the core flavonoid biosynthesis, we envision applications beyond
502 this pathway. Various enzymes of entire metabolic networks can be identified if sufficient knowledge about
503 functionally relevant amino acids is available.

504

505 **Supplementary Materials:** The following are available online: File S1: KIPEs evaluation results, File S2: Phylogenetic
506 trees of candidates, File S3: Multiple sequence alignments of candidates (yellow highlighting is used for functional relevant
507 residues in the *C. tiglium* sequences, acc=relative accessibility, black background indicates perfect conservation across all
508 sequences), File S4: Coding sequences of *C. tiglium* flavonoid biosynthesis genes, File S5: Peptide sequences of *C. tiglium*
509 flavonoid biosynthesis genes, File S6: Gene expression heatmap of all candidate genes, File S7: Unrooted phylogenetic trees
510 of MYB, bHLH, and WD40 candidates in *C. tiglium* and corresponding bait sequences; File S8: List of bait and reference
511 sequences, File S9: Functionally relevant amino acid residues considered for analysis of flavonoid biosynthesis enzymes,
512 File S10: Information about used crystal structures of previously characterized enzymes and protein models produced in this
513 study, File S11: 3D models of flavonoid biosynthesis enzyme structures generated by I-TASSER.

514

515 **Author Contributions:** B.P. and H.M.S. conceived the project. B.P., F.R., and H.M.S. conducted data analysis. B.P., F.R.,
516 and H.M.S. wrote the manuscript. B.P. supervised the project. All authors have read and agreed to the final version of this
517 manuscript.

518

519 **Funding:** This research received no external funding.

520 **Acknowledgments:** We are extremely grateful to all researchers who characterised enzymes in the flavonoid biosynthesis,
521 submitted the underlying sequences to the appropriate databases, and published their experimental findings.

522 **Conflicts of Interest:** The authors declare no conflict of interest.

523 References

524 1. Williams, C.A.; Grayer, R.J. Anthocyanins and other flavonoids. *Nat Prod Rep* **2004**, *21*, 539–573, doi:10.1039/b311404j.

- 525 2. Jaakola, L.; Hohtola, A. Effect of latitude on flavonoid biosynthesis in plants. *Plant, Cell & Environment* **2010**, *33*, 1239–1247,
526 doi:10.1111/j.1365-3040.2010.02154.x.
- 527 3. Kandaswami, C.; Kanadaswami, C.; Lee, L.-T.; Lee, P.-P.H.; Hwang, J.-J.; Ke, F.-C.; Huang, Y.-T.; Lee, M.-T. The antitumor
528 activities of flavonoids. *In Vivo* **2005**, *19*, 895–909.
- 529 4. Winkel-Shirley, B. Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. *Plant*
530 *Physiology* **2001**, *126*, 485–493, doi:10.1104/pp.126.2.485.
- 531 5. Marais, J.P.J.; Deavours, B.; Dixon, R.A.; Ferreira, D. The Stereochemistry of Flavonoids. In: Grotewold, E., Ed.; Springer New
532 York, 2006; pp. 47–69.
- 533 6. Pourcel, L.; Routaboul, J.-M.; Cheyrier, V.; Lepiniec, L.; Debeaujon, I. Flavonoid oxidation in plants: from biochemical properties
534 to physiological functions. *Trends Plant Sci.* **2007**, *12*, 29–36, doi:10.1016/j.tplants.2006.11.006.
- 535 7. Murphy, A.; Peer, W.A.; Taiz, L. Regulation of auxin transport by aminopeptidases and endogenous flavonoids. *Planta* **2000**, *211*,
536 315–324, doi:10.1007/s004250000300.
- 537 8. Mol, J.; Grotewold, E.; Koes, R. How genes paint flowers and seeds. *Trends in Plant Science* **1998**, *3*, 212–217,
538 doi:10.1016/S1360-1385(98)01242-4.
- 539 9. Harborne, J.B.; Williams, C.A. Advances in flavonoid research since 1992. *Phytochemistry* **2000**, *55*, 481–504,
540 doi:10.1016/s0031-9422(00)00235-1.
- 541 10. Harborne, J.B. Recent advances in chemical ecology. *Nat Prod Rep* **1999**, *16*, 509–523, doi:10.1039/a804621b.
- 542 11. Appelhagen, I.; Jahns, O.; Bartelniewoehner, L.; Sagasser, M.; Weisshaar, B.; Stracke, R. Leucoanthocyanidin Dioxygenase in
543 *Arabidopsis thaliana*: characterization of mutant alleles and regulation by MYB-BHLH-TTG1 transcription factor complexes. *Gene*
544 **2011**, *484*, 61–68, doi:10.1016/j.gene.2011.05.031.
- 545 12. Panche, A.N.; Diwan, A.D.; Chandra, S.R. Flavonoids: an overview. *J Nutr Sci* **2016**, *5*, doi:10.1017/jns.2016.41.
- 546 13. Nishihara, M.; Nakatsuka, T. Genetic engineering of flavonoid pigments to modify flower color in floricultural plants. *Biotechnol*
547 *Lett* **2011**, *33*, 433–441, doi:10.1007/s10529-010-0461-z.
- 548 14. Kozłowska, A.; Szostak-Wegierek, D. Flavonoids--food sources and health benefits. *Rocz Panstw Zakl Hig* **2014**, *65*, 79–85.
549 doi:10.1016/s0163-7258(02)00298-x.
- 550 15. Havsteen, B.H. The biochemistry and medical significance of the flavonoids. *Pharmacol. Ther.* **2002**, *96*, 67–202,
551 doi:10.1016/s0163-7258(02)00298-x.
- 552 16. Rice-Evans, C.; Miller, N.; Paganga, G. Antioxidant properties of phenolic compounds. *Trends in Plant Science* **1997**, *2*, 152–159,
553 doi:10.1016/S1360-1385(97)01018-2.
- 554 17. Chen, A.Y.; Chen, Y.C. A review of the dietary flavonoid, kaempferol on human health and cancer chemoprevention. *Food*
555 *Chemistry* **2013**, *138*, 2099–2107, doi:10.1016/j.foodchem.2012.11.139.
- 556 18. Zhang, W.; Furusaki, S. Production of anthocyanins by plant cell cultures. *Biotechnol. Bioprocess Eng.* **1999**, *4*, 231–252,
557 doi:10.1007/BF02933747.
- 558 19. Appelhagen, I.; Wulff-Vester, A.K.; Wendell, M.; Hvoslef-Eide, A.-K.; Russell, J.; Oertel, A.; Martens, S.; Mock, H.-P.; Martin,
559 C.; Matros, A. Colour bio-factories: Towards scale-up production of anthocyanins in plant cell cultures. *Metab Eng* **2018**, *48*,
218–232, doi:10.1016/j.ymben.2018.06.004.
- 560 20. Forkmann, G. Flavonoids as Flower Pigments: The Formation of the Natural Spectrum and its Extension by Genetic Engineering.
561 *Plant Breeding* **1991**, *106*, 1–26, doi:10.1111/j.1439-0523.1991.tb00474.x.
- 562 21. Saito, K.; Yonekura-Sakakibara, K.; Nakabayashi, R.; Higashi, Y.; Yamazaki, M.; Tohge, T.; Fernie, A.R. The flavonoid
563 biosynthetic pathway in *Arabidopsis*: structural and genetic diversity. *Plant Physiol. Biochem.* **2013**, *72*, 21–34,
564 doi:10.1016/j.plaphy.2013.02.001.
- 565 22. Koornneef, M. Mutations affecting the testa color in *Arabidopsis*. *Arabidopsis Inf. Serv.* **1990**, *28*, 1–4.
- 566 23. Shirley, B.W.; Hanley, S.; Goodman, H.M. Effects of ionizing radiation on a plant genome: analysis of two *Arabidopsis* transparent
567 testa mutations. *Plant Cell* **1992**, *4*, 333–347, doi:10.1105/tpc.4.3.333.

- 568 24. Falcone Ferreyra, M.L.; Casas, M.I.; Questa, J.I.; Herrera, A.L.; Deblasio, S.; Wang, J.; Jackson, D.; Grotewold, E.; Casati, P.
569 Evolution and expression of tandem duplicated maize flavonol synthase genes. *Front Plant Sci* **2012**, *3*, 101,
570 doi:10.3389/fpls.2012.00101.
- 571 25. Ferrer, J.-L.; Jez, J.M.; Bowman, M.E.; Dixon, R.A.; Noel, J.P. Structure of chalcone synthase and the molecular basis of plant
572 polyketide biosynthesis. *Nature Structural Biology* **1999**, *6*, 775–784, doi:10.1038/11553.
- 573 26. Krol, A.R. van der; Mur, L.A.; Beld, M.; Mol, J.N.; Stuitje, A.R. Flavonoid genes in petunia: addition of a limited number of gene
574 copies may lead to a suppression of gene expression. *The Plant Cell* **1990**, *2*, 291–299, doi:10.1105/tpc.2.4.291.
- 575 27. Schröder, G.; Schröder, J. A single change of histidine to glutamine alters the substrate preference of a stilbene synthase. *J. Biol.*
576 *Chem.* **1992**, *267*, 20558–20560.
- 577 28. Jez, J.M.; Bowman, M.E.; Dixon, R.A.; Noel, J.P. Structure and mechanism of the evolutionarily unique plant enzyme chalcone
578 isomerase. *Nature Structural Biology* **2000**, *7*, 786–791, doi:10.1038/79025.
- 579 29. Jez, J.M.; Noel, J.P. Reaction mechanism of chalcone isomerase. pH dependence, diffusion control, and product binding
580 differences. *J. Biol. Chem.* **2002**, *277*, 1361–1369, doi:10.1074/jbc.M109224200.
- 581 30. Cheng, A.-X.; Zhang, X.; Han, X.-J.; Zhang, Y.-Y.; Gao, S.; Liu, C.-J.; Lou, H.-X. Identification of chalcone isomerase in the basal
582 land plants reveals an ancient evolution of enzymatic cyclization activity for synthesis of flavonoids. *New Phytologist* **2018**, *217*,
583 909–924, doi:10.1111/nph.14852.
- 584 31. Kaltenbach, M.; Burke, J.R.; Dindo, M.; Pabis, A.; Munsberg, F.S.; Rabin, A.; Kamerlin, S.C.L.; Noel, J.P.; Tawfik, D.S. Evolution
585 of chalcone isomerase from a noncatalytic ancestor. *Nature Chemical Biology* **2018**, *14*, 548–555, doi:10.1038/s41589-018-0042-3.
- 586 32. Forkmann, G.; Heller, W.; Grisebach, H. Anthocyanin Biosynthesis in Flowers of *Matthiola incana* Flavanone 3- and Flavonoid
587 3'-Hydroxylases. *Zeitschrift für Naturforschung C* **1980**, *35*, 691–695, doi:10.1515/znc-1980-9-1004.
- 588 33. Prescott, A.G.; John, P. DIOXYGENASES: Molecular Structure and Role in Plant Metabolism. *Annu. Rev. Plant Physiol. Plant*
589 *Mol. Biol.* **1996**, *47*, 245–271, doi:10.1146/annurev.arplant.47.1.245.
- 590 34. Cheng, A.-X.; Han, X.-J.; Wu, Y.-F.; Lou, H.-X. The Function and Catalysis of 2-Oxoglutarate-Dependent Oxygenases Involved in
591 Plant Flavonoid Biosynthesis. *Int J Mol Sci* **2014**, *15*, 1080–1095, doi:10.3390/ijms15011080.
- 592 35. Owens, D.K.; Alerding, A.B.; Crosby, K.C.; Bandara, A.B.; Westwood, J.H.; Winkel, B.S.J. Functional analysis of a predicted
593 flavonol synthase gene family in Arabidopsis. *Plant Physiol.* **2008**, *147*, 1046–1061, doi:10.1104/pp.108.117457.
- 594 36. Park, S.; Kim, D.-H.; Park, B.-R.; Lee, J.-Y.; Lim, S.-H. Molecular and Functional Characterization of *Oryza sativa* Flavonol
595 Synthase (OsFLS), a Bifunctional Dioxygenase. *J. Agric. Food Chem.* **2019**, *67*, 7399–7409, doi:10.1021/acs.jafc.9b02142.
- 596 37. Xu, F.; Li, L.; Zhang, W.; Cheng, H.; Sun, N.; Cheng, S.; Wang, Y. Isolation, characterization, and function analysis of a flavonol
597 synthase gene from *Ginkgo biloba*. *Mol. Biol. Rep.* **2012**, *39*, 2285–2296, doi:10.1007/s11033-011-0978-9.
- 598 38. Turnbull, J.J.; Nagle, M.J.; Seibel, J.F.; Welford, R.W.D.; Grant, G.H.; Schofield, C.J. The C-4 stereochemistry of leucocyanidin
599 substrates for anthocyanidin synthase affects product selectivity. *Bioorganic & Medicinal Chemistry Letters* **2003**, *13*, 3853–3857,
600 doi:10.1016/S0960-894X(03)00711-X.
- 601 39. Welford, R.W.D.; Turnbull, J.J.; Claridge, T.D.W.; Prescott, A.G.; Schofield, C.J. Evidence for oxidation at C-3 of the flavonoid
602 C-ring during anthocyanin biosynthesis. *Chem. Commun.* **2001**, 1828–1829, doi:10.1039/B105576N.
- 603 40. Yan, Y.; Chemler, J.; Huang, L.; Martens, S.; Koffas, M.A.G. Metabolic Engineering of Anthocyanin Biosynthesis in *Escherichia*
604 *coli*. *Appl Environ Microbiol* **2005**, *71*, 3617–3623, doi:10.1128/AEM.71.7.3617-3623.2005.
- 605 41. Almeida, J.R.M.; D'Amico, E.; Preuss, A.; Carbone, F.; de Vos, C.H.R.; Deiml, B.; Mourgues, F.; Perrotta, G.; Fischer, T.C.; Bovy,
606 A.G.; et al. Characterization of major enzymes and genes involved in flavonoid and proanthocyanidin biosynthesis during fruit
607 development in strawberry (*Fragaria xananassa*). *Arch. Biochem. Biophys.* **2007**, *465*, 61–71, doi:10.1016/j.abb.2007.04.040.
- 608 42. Brugliera, F.; Barri-Rewell, G.; Holton, T.A.; Mason, J.G. Isolation and characterization of a flavonoid 3'-hydroxylase cDNA clone
609 corresponding to the Ht1 locus of *Petunia hybrida*. *Plant J.* **1999**, *19*, 441–451.

- 610 43. Vetten, N. de; Horst, J. ter; Schaik, H.-P. van; Boer, A. de; Mol, J.; Koes, R. A cytochrome b5 is required for full activity of
611 flavonoid 3',5'-hydroxylase, a cytochrome P450 involved in the formation of blue flower colors. *PNAS* **1999**, *96*, 778–783,
612 doi:10.1073/pnas.96.2.778.
- 613 44. Olsen, K.M.; Hehn, A.; Jugdé, H.; Slimestad, R.; Larbat, R.; Bourgaud, F.; Lillo, C. Identification and characterisation of
614 CYP75A31, a new flavonoid 3',5'-hydroxylase, isolated from *Solanum lycopersicum*. *BMC Plant Biology* **2010**, *10*, 21,
615 doi:10.1186/1471-2229-10-21.
- 616 45. Seitz, C.; Ameres, S.; Forkmann, G. Identification of the molecular basis for the functional difference between flavonoid
617 3'-hydroxylase and flavonoid 3',5'-hydroxylase. *FEBS Lett.* **2007**, *581*, 3429–3434, doi:10.1016/j.febslet.2007.06.045.
- 618 46. Holton, T.A.; Brugliera, F.; Tanaka, Y. Cloning and expression of flavonol synthase from *Petunia hybrida*. *Plant J.* **1993**, *4*,
619 1003–1010, doi:10.1046/j.1365-313x.1993.04061003.x.
- 620 47. Forkmann, G.; Vlaming, P. de; Spribille, R.; Wiering, H.; Schram, A.W. Genetic and Biochemical Studies on the Conversion of
621 Dihydroflavonols to Flavonols in Flowers of *Petunia hybrida*. *Zeitschrift für Naturforschung C* **1985**, *41*, 179–186,
622 doi:10.1515/znc-1986-1-227.
- 623 48. Britsch, L.; Heller, W.; Grisebach, H. Conversion of Flavanone to Flavone, Dihydroflavonol and Flavonol with an Enzyme System
624 from Cell Cultures of Parsley. *Zeitschrift für Naturforschung C* **1981**, *36*, 742–750, doi:10.1515/znc-1981-9-1009.
- 625 49. Pelletier, M.K.; Murrell, J.R.; Shirley, B.W. Characterization of Flavonol Synthase and Leucoanthocyanidin Dioxygenase Genes in
626 *Arabidopsis* (Further Evidence for Differential Regulation of “Early” and “Late” Genes). *Plant Physiology* **1997**, *113*, 1437–1445,
627 doi:10.1104/pp.113.4.1437.
- 628 50. Hostetler, G.L.; Ralston, R.A.; Schwartz, S.J. Flavones: Food Sources, Bioavailability, Metabolism, and Bioactivity. *Adv Nutr*
629 **2017**, *8*, 423–435, doi:10.3945/an.116.012948.
- 630 51. Martens, S.; Forkmann, G.; Britsch, L.; Wellmann, F.; Matern, U.; Lukacin, R. Divergent evolution of flavonoid
631 2-oxoglutarate-dependent dioxygenases in parsley. *FEBS Lett.* **2003**, *544*, 93–98, doi:10.1016/s0014-5793(03)00479-4.
- 632 52. Martens, S.; Forkmann, G. Cloning and expression of flavone synthase II from *Gerbera* hybrids. *Plant J.* **1999**, *20*, 611–618,
633 doi:10.1046/j.1365-313x.1999.00636.x.
- 634 53. Gebhardt, Y.H.; Witte, S.; Steuber, H.; Matern, U.; Martens, S. Evolution of Flavone Synthase I from Parsley Flavanone
635 3 β -Hydroxylase by Site-Directed Mutagenesis. *Plant Physiol* **2007**, *144*, 1442–1454, doi:10.1104/pp.107.098392.
- 636 54. Davies, K.M.; Schwinn, K.E.; Deroles, S.C.; Manson, D.G.; Lewis, D.H.; Bloor, S.J.; Bradley, J.M. Enhancing anthocyanin
637 production by altering competition for substrate between flavonol synthase and dihydroflavonol 4-reductase. *Euphytica* **2003**, *131*,
638 259–268, doi:10.1023/A:1024018729349.
- 639 55. Johnson, E.T.; Ryu, S.; Yi, H.; Shin, B.; Cheong, H.; Choi, G. Alteration of a single amino acid changes the substrate specificity of
640 dihydroflavonol 4-reductase. *Plant J.* **2001**, *25*, 325–333, doi:10.1046/j.1365-313x.2001.00962.x.
- 641 56. Miosic, S.; Thill, J.; Milosevic, M.; Gosch, C.; Pober, S.; Molitor, C.; Ejaz, S.; Rompel, A.; Stich, K.; Halbwirth, H.
642 Dihydroflavonol 4-Reductase Genes Encode Enzymes with Contrasting Substrate Specificity and Show Divergent Gene
643 Expression Profiles in *Fragaria* Species. *PLOS ONE* **2014**, *9*, e112707, doi:10.1371/journal.pone.0112707.
- 644 57. Katsu, K.; Suzuki, R.; Tsuchiya, W.; Inagaki, N.; Yamazaki, T.; Hisano, T.; Yasui, Y.; Komori, T.; Koshio, M.; Kubota, S.; et al. A
645 new buckwheat dihydroflavonol 4-reductase (DFR), with a unique substrate binding structure, has altered substrate specificity.
646 *BMC Plant Biol* **2017**, *17*, doi:10.1186/s12870-017-1200-6.
- 647 58. Gang, D.R.; Kasahara, H.; Xia, Z.Q.; Vander Mijnsbrugge, K.; Bauw, G.; Boerjan, W.; Van Montagu, M.; Davin, L.B.; Lewis, N.G.
648 Evolution of plant defense mechanisms. Relationships of phenylcoumaran benzylic ether reductases to pinorensinol-lariciresinol and
649 isoflavone reductases. *J. Biol. Chem.* **1999**, *274*, 7516–7527, doi:10.1074/jbc.274.11.7516.
- 650 59. Gao, J.; Shen, L.; Yuan, J.; Zheng, H.; Su, Q.; Yang, W.; Zhang, L.; Nnaemeka, V.E.; Sun, J.; Ke, L.; et al. Functional analysis of
651 GhCHS, GhANR and GhLAR in colored fiber formation of *Gossypium hirsutum* L. *BMC Plant Biol* **2019**, *19*, 455,
652 doi:10.1186/s12870-019-2065-7.

- 653 60. Timberlake, C.F.; Bridle, P. Spectral Studies of Anthocyanin and Anthocyanidin Equilibria in Aqueous Solution. *Nature* **1966**, *212*,
654 158–159, doi:10.1038/212158a0.
- 655 61. Kovinich, N.; Saleem, A.; Rintoul, T.L.; Brown, D.C.W.; Arnason, J.T.; Miki, B. Coloring genetically modified soybean grains
656 with anthocyanins by suppression of the proanthocyanidin genes ANR1 and ANR2. *Transgenic Res.* **2012**, *21*, 757–771,
657 doi:10.1007/s11248-011-9566-y.
- 658 62. Xie, D.-Y.; Sharma, S.B.; Dixon, R.A. Anthocyanidin reductases from *Medicago truncatula* and *Arabidopsis thaliana*. *Arch.*
659 *Biochem. Biophys.* **2004**, *422*, 91–102, doi:10.1016/j.abb.2003.12.011.
- 660 63. Weisshaar, B.; Jenkins, G.I. Phenylpropanoid biosynthesis and its regulation. *Curr. Opin. Plant Biol.* **1998**, *1*, 251–257,
661 doi:10.1016/s1369-5266(98)80113-1.
- 662 64. Stracke, R.; Werber, M.; Weisshaar, B. The R2R3-MYB gene family in *Arabidopsis thaliana*. *Current Opinion in Plant Biology*
663 **2001**, *4*, 447–456, doi:10.1016/S1369-5266(00)00199-0.
- 664 65. Du, H.; Liang, Z.; Zhao, S.; Nan, M.-G.; Tran, L.-S.P.; Lu, K.; Huang, Y.-B.; Li, J.-N. The Evolutionary History of R2R3-MYB
665 Proteins Across 50 Eukaryotes: New Insights Into Subfamily Classification and Expansion. *Sci Rep* **2015**, *5*, 1–16,
666 doi:10.1038/srep11037.
- 667 66. Hichri, I.; Barrieu, F.; Bogs, J.; Kappel, C.; Delrot, S.; Lauvergeat, V. Recent advances in the transcriptional regulation of the
668 flavonoid biosynthetic pathway. *J. Exp. Bot.* **2011**, *62*, 2465–2483, doi:10.1093/jxb/erq442.
- 669 67. Ramsay, N.A.; Glover, B.J. MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. *Trends Plant Sci.* **2005**, *10*,
670 63–70, doi:10.1016/j.tplants.2004.12.011.
- 671 68. Carretero-Paulet, L.; Galstyan, A.; Roig-Villanova, I.; Martínez-García, J.F.; Bilbao-Castro, J.R.; Robertson, D.L. Genome-wide
672 classification and evolutionary analysis of the bHLH family of transcription factors in *Arabidopsis*, poplar, rice, moss, and algae.
673 *Plant Physiol.* **2010**, *153*, 1398–1412, doi:10.1104/pp.110.153593.
- 674 69. Stracke, R.; Ishihara, H.; Huep, G.; Barsch, A.; Mehrtens, F.; Niehaus, K.; Weisshaar, B. Differential regulation of closely related
675 R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J.*
676 **2007**, *50*, 660–677, doi:10.1111/j.1365-313X.2007.03078.x.
- 677 70. Pillet, J.; Yu, H.-W.; Chambers, A.H.; Whitaker, V.M.; Folta, K.M. Identification of candidate flavonoid pathway genes using
678 transcriptome correlation network analysis in ripe strawberry (*Fragaria × ananassa*) fruits. *J Exp Bot* **2015**, *66*, 4455–4467,
679 doi:10.1093/jxb/erv205.
- 680 71. Pandey, A.; Alok, A.; Lakhwani, D.; Singh, J.; Asif, M.H.; Trivedi, P.K. Genome-wide Expression Analysis and Metabolite
681 Profiling Elucidate Transcriptional Regulation of Flavonoid Biosynthesis and Modulation under Abiotic Stresses in Banana. *Sci*
682 *Rep* **2016**, *6*, 1–13, doi:10.1038/srep31361.
- 683 72. Otani, M.; Kanemaki, Y.; Oba, F.; Shibuya, M.; Funayama, Y.; Nakano, M. Comprehensive isolation and expression analysis of the
684 flavonoid biosynthesis-related genes in *Tricyrtis* spp. *Biol Plant* **2018**, *62*, 684–692, doi:10.1007/s10535-018-0802-7.
- 685 73. Qu, C.; Zhao, H.; Fu, F.; Wang, Z.; Zhang, K.; Zhou, Y.; Wang, X.; Wang, R.; Xu, X.; Tang, Z.; et al. Genome-Wide Survey of
686 Flavonoid Biosynthesis Genes and Gene Expression Analysis between Black- and Yellow-Seeded *Brassica napus*. *Front. Plant Sci.*
687 **2016**, *7*, doi:10.3389/fpls.2016.01755.
- 688 74. Wang, J.; Zhang, Q.; Cui, F.; Hou, L.; Zhao, S.; Xia, H.; Qiu, J.; Li, T.; Zhang, Y.; Wang, X.; et al. Genome-Wide Analysis of Gene
689 Expression Provides New Insights into Cold Responses in *Thellungiella salsuginea*. *Front Plant Sci* **2017**, *8*,
690 doi:10.3389/fpls.2017.00713.
- 691 75. Amirbakhtiar, N.; Ismaili, A.; Ghaffari, M.R.; Firouzabadi, F.N.; Shobbar, Z.-S. Transcriptome response of roots to salt stress in a
692 salinity-tolerant bread wheat cultivar. *PLOS ONE* **2019**, *14*, e0213305, doi:10.1371/journal.pone.0213305.
- 693 76. Sicilia, A.; Testa, G.; Santoro, D.F.; Cosentino, S.L.; Lo Piero, A.R. RNASeq analysis of giant cane reveals the leaf transcriptome
694 dynamics under long-term salt stress. *BMC Plant Biol* **2019**, *19*, 355, doi:10.1186/s12870-019-1964-y.

- 695 77. Pucker, B.; Schilbert, H.M. Genomics and Transcriptomics Advance in Plant Sciences. In *Molecular Approaches in Plant Biology*
696 *and Environmental Challenges*; Singh, S.P., Upadhyay, S.K., Pandey, A., Kumar, S., Eds.; Energy, Environment, and
697 Sustainability; Springer: Singapore, 2019; pp. 419–448 ISBN 9789811506901.
- 698 78. Kalwij, J.M. Review of ‘The Plant List, a working list of all plant species.’ *Journal of Vegetation Science* **2012**, *23*, 998–1002,
699 doi:10.1111/j.1654-1103.2012.01407.x.
- 700 79. Pope, J. On a New Preparation of Croton Tiglium. *Med Chir Trans* **1827**, *13*, 97–102, doi:10.1177/09595287270130p111.
- 701 80. Gläser, S.; Sorg, B.; Hecker, E. A Method for Quantitative Determination of Polyfunctional Diterpene Esters of the Tiglane Type
702 in Croton tiglium. *Planta Med.* **1988**, *54*, 580, doi:10.1055/s-2006-962595.
- 703 81. Kim, J.H.; Lee, S.J.; Han, Y.B.; Moon, J.J.; Kim, J.B. Isolation of isoguanosine from Croton tiglium and its antitumor activity.
704 *Arch. Pharm. Res.* **1994**, *17*, 115–118, doi:10.1007/bf02974234.
- 705 82. El-Mekawy, S.; Meselhy, M.R.; Nakamura, N.; Hattori, M.; Kawahata, T.; Otake, T. Anti-HIV-1 phorbol esters from the seeds of
706 Croton tiglium. *Phytochemistry* **2000**, *53*, 457–464, doi:10.1016/s0031-9422(99)00556-7.
- 707 83. Abbas, M.; Shahid, M.; Sheikh, M.A.; Muhammad, G. Phytochemical Screening of Plants Used in Folkloric Medicine: Effect of
708 Extraction Method and Solvent. *Asian Journal of Chemistry* **2014**, *26*, 6194–6198, doi:10.14233/ajchem.2014.17125.
- 709 84. Palmeira Jr., S.F.; Conserva, L.M.; Silveira, E.R. Two clerodane diterpenes and flavonoids from Croton brasiliensis. *Journal of the*
710 *Brazilian Chemical Society* **2005**, *16*, 1420–1424, doi:10.1590/S0103-50532005000800021.
- 711 85. Kostova, I.; Iossifova, T.; Rostan, J.; Vogler, B.; Kraus, W.; Navas, H. Chemical and biological studies on Croton panamensis latex
712 (Dragon’s Blood). *Pharmaceutical and Pharmacological Letters* **1999**, *9*, 34–36.
- 713 86. Haak, M.; Vinke, S.; Keller, W.; Droste, J.; Rückert, C.; Kalinowski, J.; Pucker, B. High Quality de Novo Transcriptome Assembly
714 of Croton tiglium. *Front. Mol. Biosci.* **2018**, *5*, doi:10.3389/fmolb.2018.00062.
- 715 87. Li, C.; Wu, X.; Sun, R.; Zhao, P.; Liu, F.; Zhang, C. Croton Tiglium Extract Induces Apoptosis via Bax/Bcl-2 Pathways in Human
716 Lung Cancer A549 Cells. *Asian Pac J Cancer Prev* **2016**, *17*, 4893–4898, doi:10.22034/APJCP.2016.17.11.4893.
- 717 88. Salatino, A.; Salatino, M.L.F.; Negri, G. Traditional uses, chemistry and pharmacology of Croton species (Euphorbiaceae). *Journal*
718 *of the Brazilian Chemical Society* **2007**, *18*, 11–33, doi:10.1590/S0103-50532007000100002.
- 719 89. Tsacheva, I.; Rostan, J.; Iossifova, T.; Vogler, B.; Odjakova, M.; Navas, H.; Kostova, I.; Kojouharova, M.; Kraus, W. Complement
720 inhibiting properties of dragon’s blood from Croton draco. *Z. Naturforsch., C, J. Biosci.* **2004**, *59*, 528–532,
721 doi:10.1515/znc-2004-7-814.
- 722 90. Maciel, M.A.; Pinto, A.C.; Arruda, A.C.; Pamplona, S.G.; Vanderlinde, F.A.; Lapa, A.J.; Echevarria, A.; Grynberg, N.F.; Cólus,
723 I.M.; Farias, R.A.; et al. Ethnopharmacology, phytochemistry and pharmacology: a successful combination in the study of Croton
724 cajucara. *J Ethnopharmacol* **2000**, *70*, 41–55, doi:10.1016/s0378-8741(99)00159-2.
- 725 91. Guerrero, M.F.; Puebla, P.; Carrón, R.; Martín, M.L.; Román, L.S. Quercetin 3,7-dimethyl ether: a vasorelaxant flavonoid isolated
726 from Croton schiedeanus Schlecht. *Journal of Pharmacy and Pharmacology* **2002**, *54*, 1373–1378,
727 doi:10.1211/002235702760345455.
- 728 92. Krebs, H.C.; Ramiarantsoa, H. Clerodane diterpenes of Croton hoverum. *Phytochemistry* **1997**, *45*, 379–381,
729 doi:10.1016/S0031-9422(96)00815-1.
- 730 93. Shimada, S.; Otsuki, H.; Sakuta, M. Transcriptional control of anthocyanin biosynthetic genes in the Caryophyllales. *J Exp Bot*
731 **2007**, *58*, 957–967, doi:10.1093/jxb/erl256.
- 732 94. Nesi, N.; Debeaujon, I.; Jond, C.; Pelletier, G.; Caboche, M.; Lepiniec, L. The TT8 Gene Encodes a Basic Helix-Loop-Helix
733 Domain Protein Required for Expression of DFR and BAN Genes in Arabidopsis Siliques. *The Plant Cell* **2000**, *12*, 1863–1878,
734 doi:10.1105/tpc.12.10.1863.
- 735 95. Ni, R.; Zhu, T.-T.; Zhang, X.-S.; Wang, P.-Y.; Sun, C.-J.; Qiao, Y.-N.; Lou, H.-X.; Cheng, A.-X. Identification and evolutionary
736 analysis of chalcone isomerase-fold proteins in ferns. *J Exp Bot* **2020**, *71*, 290–304, doi:10.1093/jxb/erz425.
- 737 96. Cai, Y.; Evans, F.J.; Roberts, M.F.; Phillipson, J.D.; Zenk, M.H.; Gleba, Y.Y. Polyphenolic compounds from Croton lechleri.
738 *Phytochemistry* **1991**, *30*, 2033–2040, doi:10.1016/0031-9422(91)85063-6.

- 739 97. Furlan, C.M.; Santos, K.P.; Sedano-Partida, M.D.; Motta, L.B. da; Santos, D.Y.A.C.; Salatino, M.L.F.; Negri, G.; Berry, P.E.; van
740 Ee, B.W.; Salatino, A. Flavonoids and antioxidant potential of nine Argentinian species of Croton (Euphorbiaceae). *Braz. J. Bot*
741 **2015**, *38*, 693–702, doi:10.1007/s40415-014-0115-9.
- 742 98. de Sampaio e Spohr, T.C.L.; Stipursky, J.; Sasaki, A.C.; Barbosa, P.R.; Martins, V.; Benjamim, C.F.; Roque, N.F.; Costa, S.L.;
743 Gomes, F.C.A. Effects of the flavonoid casticin from Brazilian Croton betulaster in cerebral cortical progenitors in vitro: direct and
744 indirect action through astrocytes. *J. Neurosci. Res.* **2010**, *88*, 530–541, doi:10.1002/jnr.22218.
- 745 99. Biscaro, F.; Parisotto, E.B.; Zanette, V.C.; Günther, T.M.F.; Ferreira, E.A.; Gris, E.F.; Correia, J.F.G.; Pich, C.T.; Mattivi, F.; Filho,
746 D.W.; et al. Anticancer activity of flavonol and flavan-3-ol rich extracts from Croton celtidifolius latex. *Pharmaceutical Biology*
747 **2013**, *51*, 737–743, doi:10.3109/13880209.2013.764331.
- 748 100. Aderogba, M.A.; Ndhkala, A.R.; Rengasamy, K.R.R.; Van Staden, J. Antimicrobial and selected in vitro enzyme inhibitory effects
749 of leaf extracts, flavonols and indole alkaloids isolated from Croton menyharthii. *Molecules* **2013**, *18*, 12633–12644,
750 doi:10.3390/molecules181012633.
- 751 101. Cruz, B.G.; Dos Santos, H.S.; Bandeira, P.N.; Rodrigues, T.H.S.; Matos, M.G.C.; Nascimento, M.F.; de Carvalho, G.G.C.;
752 Braz-Filho, R.; Teixeira, A.M.R.; Tintino, S.R.; et al. Evaluation of antibacterial and enhancement of antibiotic action by the
753 flavonoid kaempferol 7-O-β-D-(6"-O-cumaroyl)-glucopyranoside isolated from Croton piauhiensis müll. *Microb. Pathog.* **2020**,
754 *143*, 104144, doi:10.1016/j.micpath.2020.104144.
- 755 102. Nascimento, A.M.; Maria-Ferreira, D.; Dal Lin, F.T.; Kimura, A.; de Santana-Filho, A.P.; Werner, M.F. de P.; Iacomini, M.;
756 Sasaki, G.L.; Cipriani, T.R.; de Souza, L.M. Phytochemical analysis and anti-inflammatory evaluation of compounds from an
757 aqueous extract of Croton cajucara Benth. *J Pharm Biomed Anal* **2017**, *145*, 821–830, doi:10.1016/j.jpba.2017.07.032.
- 758 103. Prescott, A.G.; Stamford, N.P.J.; Wheeler, G.; Firmin, J.L. In vitro properties of a recombinant flavonol synthase from Arabidopsis
759 thaliana. *Phytochemistry* **2002**, *60*, 589–593, doi:10.1016/s0031-9422(02)00155-3.
- 760 104. Lukacin, R.; Wellmann, F.; Britsch, L.; Martens, S.; Matern, U. Flavonol synthase from Citrus unshiu is a bifunctional dioxygenase.
761 *Phytochemistry* **2003**, *62*, 287–292, doi:10.1016/s0031-9422(02)00567-8.
- 762 105. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating
763 and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain
764 w1118; iso-2; iso-3. *Fly (Austin)* **2012**, *6*, 80–92, doi:10.4161/fly.19695.
- 765 106. Stanke, M.; Steinkamp, R.; Waack, S.; Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids*
766 *Res* **2004**, *32*, W309–W312, doi:10.1093/nar/gkh379.
- 767 107. Keilwagen, J.; Hartung, F.; Paulini, M.; Twardziok, S.O.; Grau, J. Combining RNA-seq data and homology-based gene prediction
768 for plants, animals and fungi. *BMC Bioinformatics* **2018**, *19*, 189, doi:10.1186/s12859-018-2203-5.
- 769 108. Pucker, B.; Brockington, S.F. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes.
770 *BMC Genomics* **2018**, *19*, 980, doi:10.1186/s12864-018-5360-z.
- 771 109. Jia, Y.; Li, B.; Zhang, Y.; Zhang, X.; Xu, Y.; Li, C. Evolutionary dynamic analyses on monocot flavonoid 3'-hydroxylase gene
772 family reveal evidence of plant-environment interaction. *BMC Plant Biology* **2019**, *19*, 347, doi:10.1186/s12870-019-1947-z.
- 773 110. Nielsen, K.; Deroles, S.C.; Markham, K.R.; Bradley, M.J.; Podivinsky, E.; Manson, D. Antisense flavonol synthase alters
774 copigmentation and flower color in lisianthus. *Molecular Breeding* **2002**, *9*, 217–229, doi:10.1023/A:1020320809654.
- 775 111. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and
776 Usability. *Molecular Biology and Evolution* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.
- 777 112. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology*
778 **1990**, *215*, 403–410, doi:10.1016/S0022-2836(05)80360-2.
- 779 113. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*
780 **2010**, *5*, e9490, doi:10.1371/journal.pone.0009490.
- 781 114. Brown, J.W.; Walker, J.F.; Smith, S.A. Phyx: phylogenetic tools for unix. *Bioinformatics* **2017**, *33*, 1886–1888,
782 doi:10.1093/bioinformatics/btx063.

- 783 115. Robert, X.; Gouet, P. Deciphering key features in protein structures with the new ENDScript server. *Nucleic Acids Res* **2014**, *42*,
784 W320–W324, doi:10.1093/nar/gku316.
- 785 116. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data
786 Bank. *Nucleic Acids Res* **2000**, *28*, 235–242, doi:10.1093/nar/28.1.235.
- 787 117. Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat*
788 *Protoc* **2010**, *5*, 725–738, doi:10.1038/nprot.2010.5.
- 789 118. Leebens-Mack, J.H.; Barker, M.S.; Carpenter, E.J.; Deyholos, M.K.; Gitzendanner, M.A.; Graham, S.W.; Grosse, I.; Li, Z.;
790 Melkonian, M.; Mirarab, S.; et al. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **2019**, *574*,
791 679–685, doi:10.1038/s41586-019-1693-2.
- 792 119. Canedo-Téxon, A.; Ramón-Farías, F.; Monribot-Villanueva, J.L.; Villafán, E.; Alonso-Sánchez, A.; Pérez-Torres, C.A.; Ángeles,
793 G.; Guerrero-Analco, J.A.; Ibarra-Laclette, E. Novel findings to the biosynthetic pathway of magnoflorine and taspine through
794 transcriptomic and metabolomic analysis of *Croton draco* (Euphorbiaceae). *BMC Plant Biology* **2019**, *19*, 560,
795 doi:10.1186/s12870-019-2195-y.
- 796 120. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **2016**,
797 *34*, 525–527, doi:10.1038/nbt.3519.
- 798 121. Matus, J.T.; Aquea, F.; Arce-Johnson, P. Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades
799 and conserved gene structure organization across *Vitis* and *Arabidopsis* genomes. *BMC Plant Biology* **2008**, *8*, 83,
800 doi:10.1186/1471-2229-8-83.
- 801 122. Stracke, R.; Holtgräwe, D.; Schneider, J.; Pucker, B.; Sörensen, T.R.; Weisshaar, B. Genome-wide identification and
802 characterisation of R2R3-MYB genes in sugar beet (*Beta vulgaris*). *BMC Plant Biol.* **2014**, *14*, 249,
803 doi:10.1186/s12870-014-0249-8.
- 804 123. Pucker, B.; Pandey, A.; Weisshaar, B.; Stracke, R. The R2R3-MYB gene family in banana (*Musa acuminata*): genome-wide
805 identification, classification and expression patterns. *bioRxiv* **2020**, 2020.02.03.932046, doi:10.1101/2020.02.03.932046.
- 806 124. Heim, M.A.; Jakoby, M.; Werber, M.; Martin, C.; Weisshaar, B.; Bailey, P.C. The Basic Helix–Loop–Helix Transcription Factor
807 Family in Plants: A Genome-Wide Study of Protein Structure and Functional Diversity. *Mol Biol Evol* **2003**, *20*, 735–747,
808 doi:10.1093/molbev/msg088.
- 809 125. Wang, P.; Su, L.; Gao, H.; Jiang, X.; Wu, X.; Li, Y.; Zhang, Q.; Wang, Y.; Ren, F. Genome-Wide Characterization of bHLH Genes
810 in Grape and Analysis of their Potential Relevance to Abiotic Stress Tolerance and Secondary Metabolite Biosynthesis. *Front.*
811 *Plant Sci.* **2018**, *9*, doi:10.3389/fpls.2018.00064.
- 812 126. Mao, T.-Y.; Liu, Y.-Y.; Zhu, H.-H.; Zhang, J.; Yang, J.-X.; Fu, Q.; Wang, N.; Wang, Z. Genome-wide analyses of the bHLH gene
813 family reveals structural and functional characteristics in the aquatic plant *Nelumbo nucifera*. *PeerJ* **2019**, *7*, e7153,
814 doi:10.7717/peerj.7153.
- 815 127. Zhang, X.-Y.; Qiu, J.-Y.; Hui, Q.-L.; Xu, Y.-Y.; He, Y.-Z.; Peng, L.-Z.; Fu, X.-Z. Systematic analysis of the basic/helix-loop-helix
816 (bHLH) transcription factor family in pummelo (*Citrus grandis*) and identification of the key members involved in the response to
817 iron deficiency. *BMC Genomics* **2020**, *21*, 233, doi:10.1186/s12864-020-6644-7.
- 818 128. Wang, Z.; Jia, C.; Wang, J.-Y.; Miao, H.-X.; Liu, J.-H.; Chen, C.; Yang, H.-X.; Xu, B.; Jin, Z. Genome-Wide Analysis of Basic
819 Helix-Loop-Helix Transcription Factors to Elucidate Candidate Genes Related to Fruit Ripening and Stress in Banana (*Musa*
820 *acuminata* L. AAA Group, cv. Cavendish). *Front. Plant Sci.* **2020**, *11*, doi:10.3389/fpls.2020.00650.
- 821 129. Tian, S.; Li, L.; Wei, M.; Yang, F. Genome-wide analysis of basic helix–loop–helix superfamily members related to anthocyanin
822 biosynthesis in eggplant (*Solanum melongena* L.). *PeerJ* **2019**, *7*, e7768, doi:10.7717/peerj.7768.
- 823 130. van Nocker, S.; Ludwig, P. The WD-repeat protein superfamily in *Arabidopsis*: conservation and divergence in structure and
824 function. *BMC Genomics* **2003**, *4*, 50, doi:10.1186/1471-2164-4-50.
- 825 131. Hu, R.; Xiao, J.; Gu, T.; Yu, X.; Zhang, Y.; Chang, J.; Yang, G.; He, G. Genome-wide identification and analysis of WD40 proteins
826 in wheat (*Triticum aestivum* L.). *BMC Genomics* **2018**, *19*, 803, doi:10.1186/s12864-018-5157-0.

- 827 132. Mishra, A.K.; Muthamilarasan, M.; Khan, Y.; Parida, S.K.; Prasad, M. Genome-Wide Investigation and Expression Analyses of
828 WD40 Protein Family in the Model Plant Foxtail Millet (*Setaria italica* L.). *PLOS ONE* **2014**, *9*, e86852,
829 doi:10.1371/journal.pone.0086852.
830