Yerba mate (*Ilex paraguariensis*) genome provides new insights into convergent evolution of caffeine biosynthesis

Federico A. Vignale[1#], Andrea Hernandez Garcia[2#], Carlos P. Modenutti[3,4], Ezequiel J. Sosa[3,4], Lucas A. Defelipe[1], Renato R.M. Oliveira[5], Gisele L. Nunes[5], Raúl M. Acevedo[6], German F. Burguener[7], Maximiliano Rossi[8], Pedro D. Zapata[8], Dardo A. Marti[9], Pedro A. Sansberro[6], Guilherme Oliveira[5], Madeline N. Smith[10], Nicole M. Dubs[10], Satish Nair[2,11,12], Todd J. Barkman[10#*], and Adrian G. Turjanski[3,4#*]

[1] European Molecular Biology Laboratory - Hamburg Unit, Notkestrasse 85, 22607 Hamburg, Germany

[2] Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[3] IQUIBICEN-CONICET, Ciudad Universitaria, Pabellón 2, C1428EHA Ciudad Autónoma de Buenos Aires, Argentina

[4] Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, C1428EHA Ciudad Autónoma de Buenos Aires, Argentina

[5] Instituto Tecnológico Vale, Belém, Pará, Brazil

[6] Laboratorio de Biotecnología Aplicada y Genómica Funcional, Instituto de Botánica del Nordeste (IBONE-CONICET), Facultad de Ciencias Agrarias, Universidad Nacional del Nordeste, W3402BKG, Corrientes, Argentina

[7] Department of Plant Sciences, University of California, Davis, CA 95616, USA

[8] Instituto de Biotecnología de Misiones, Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones (INBIOMIS-FCEQyN-UNaM), Misiones, Argentina

[9] Instituto de Biología Subtropical, Universidad Nacional de Misiones (IBS-UNaM-CONICET), Posadas, Misiones, Argentina

28 [10] Department of Biological Sciences, Western Michigan University, Kalamazoo, MI 49008,
29 USA
30 [11] Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign,
31 Urbana, IL 61801, USA
32 [12] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign,
33 Urbana, IL 61801, USA
34
35 #Equal contributions
36 *Correspondence to Adrian G. Turjanski (adrian@qb.fcen.uba.ar) or Todd J. Barkman
37 (todd.barkman@wmich.edu)
38
39 **Abstract**
40
41 Yerba mate (*Ilex paraguariensis*) is an economically important crop marketed for the
42 elaboration of mate, the third-most widely consumed caffeine-containing infusion worldwide.
43 Here we report the first genome assembly of this species, which has a total length of 1.06 Gb
44 and contains 53,390 protein-coding genes. Comparative analyses revealed that the large
45 yerba mate genome size is partly due to a whole-genome duplication (Ip-α) during the early
46 evolutionary history of *Ilex*, in addition to the hexaploidization event (γ) shared by core
47 eudicots. Characterization of the genome allowed us to clone the genes encoding
48 methyltransferase enzymes that catalyse multiple reactions required for caffeine production.
49 To our surprise, this species has converged upon a different biochemical pathway compared
50 to that of its relatives, coffee and tea. In order to gain insight into the structural basis for the
51 convergent enzyme activities, we obtained a crystal structure for the terminal enzyme in the
52 pathway that forms caffeine. The structure reveals that convergent solutions have evolved for
53 substrate positioning because different amino acid residues facilitate a different substrate
54 orientation such that efficient methylation occurs in the independently evolved enzymes in
55 yerba mate and coffee. While our results show phylogenomic constraint limits the genes

56 coopted for convergence of caffeine biosynthesis, the x-ray diffraction data suggests structural
57 constraints are minimal for the convergent evolution of individual reactions.
58
59 **Introduction**
60
61 In the genomic era, hundreds of plant species have had their nucleotide sequences
62 determined to provide unprecedented insight into the genetic basis of many traits. One of the
63 few species of economic importance for which no genomic data exists is *Ilex paraguariensis*
64 var. *paraguariensis* A. St. Hilaire (Aquifoliaceae), colloquially known as yerba mate (YM),
65 which is a caffeinated diploid tree-species ($2n=2x=40$) endemic to the subtropical rainforests
66 of South America[1]. The dried leaves and twigs of this dioecious evergreen are used to prepare
67 a traditional infusion named mate, or chimarrão, widely consumed around the world.
68 Approximately 300,000 ha are cultivated with this tree crop, with Argentina responsible for
69 80% of world-wide production[2]. The mate infusion has been shown to have numerous
70 beneficial effects in humans including as an antioxidant[3,4,5], antidiabetic[6,7], as well as central
71 nervous system stimulant[8], among others. Several bioactive compounds have been identified
72 in YM that might be responsible for its effects, including terpenes, flavonoids, phenolics and
73 methylxanthines[2]. Although its stimulant properties are best known and mostly related to
74 caffeine content, little is known about the genetic and biochemical mechanisms of how YM
75 synthesizes this, or any, of its important metabolites. Despite the recent release of three other
76 *Ilex* genome sequences[9–11], none of the species produce caffeine, making the genetic basis
77 for convergent evolution of this trait in YM unclear.
78 Convergent evolution has occurred throughout the tree of life and is particularly
79 rampant in plants[12] where examples of repeated origins of morphological[13], anatomical[14],
80 physiological[15], and biochemical[16] traits have been documented. Caffeine (CF) is a xanthine
81 alkaloid that has independently evolved no less than six times across angiosperms and has
82 implications for pollination, insect defence and allelopathy[17,18]. There are multiple biosynthetic
83 routes to caffeine possible within the xanthine alkaloid network (Fig. 1). Within the Rosid

84 genera *Theobroma*, *Paullinia*, and *Citrus*, sequential methylation of xanthine (X), 1- and/or 3-
85 methylxanthine (1X, 3X) and, finally, either theophylline (TP) or theobromine (TB) leads to
86 caffeine[19]. In contrast, the Asterids, *Coffea* and *Camellia*, appear to sequentially methylate
87 xanthosine (XR), 7-methylxanthine (7X), and theobromine (TB) to yield caffeine[20,21] (Fig. 1).
88 Regardless of which pathway is utilized, species differ in terms of which SABATH enzyme
89 family members were convergently recruited to synthesize caffeine: Xanthine
90 Methyltransferase (XMT) is used by *Citrus* and *Coffea* while the paralogous Caffeine Synthase
91 (CS) is used by *Camellia*, *Theobroma* and *Paullinia*[19,22,23] (Fig. 1). Convergence appears to
92 also extend to the mutational level, since different amino acid replacements to homologous
93 regions of CS and XMT enzymes appear to govern the evolution of substrate preference
94 switches[24]. However, it remains unclear whether mutations lead to convergent three-
95 dimensional protein structures to confer convergent substrate interactions and catalysis by the
96 enzymes. Because XMT- and CS-type enzymes have been convergently recruited in both
97 Rosids and Asterids to catalyse the same or different pathways, it suggests considerable
98 evolutionary lability underlying caffeine production in plants[19]. As a result, it is difficult to predict
99 what sets of genes, biochemical reactions, and structural properties might lead to the evolution
100 of caffeine biosynthesis in YM, or any plant, *a priori*.
101 Although some transcriptomic resources have been generated for YM[25,26,27], a
102 complete genome sequence has the potential to advance our understanding of the metabolic
103 potential of this important crop and facilitate improvement. Here, we describe the first draft
104 genome of YM and report on its composition, organization and evolution. The genomic
105 sequence enabled us to uncover the genetic, biochemical and structural bases for convergent
106 evolution of caffeine in YM. Our comparative analyses of caffeine-producing enzymes across
107 angiosperms reveals how convergence may be the result of constrained evolutionary genomic
108 potential but relatively unconstrained structural potential.
109
110
111

Results and Discussion

Yerba mate genome sequencing, assembly and annotation

The YM genome was sequenced combining Illumina and PacBio sequencing technologies. With Illumina sequencing, we generated ~263.2 Gb of short reads from various DNA fragment sizes (350 bp, 550 bp, 3 kbp, 8 kbp and 12 kbp), while with PacBio sequencing, we generated ~77.5 Gb of long reads. These reads represent ~158.5- and ~49.3-fold base-pair coverage of the genome, respectively (Supplementary Table 1). The total assembly length was ~1.06 Gb and consisted of 10,611 scaffolds (≥1 kb) with an N50 length of ~510.8 Kb (Supplementary Table 2). To assess the completeness of the assembly we aligned the available YM transcriptome reads[25,26,27] and the YM genome short reads generated in this study with the genome assembly, establishing that 99.3% of the YM transcriptome reads and 99.5% of the YM genome short reads mapped to the assembly. This result suggests that we have assembled almost all available sequence data. The GC content of the genome assembly was 36.33% (Supplementary Table 2), similar to that of other eudicots (33.70-38.20 GC%)[28] and almost identical to that of *Ilex polyneura* (36.08 GC%)[9], *Ilex asprella* (36.25 GC%)[10], and *Ilex latifolia* (36.44 GC%)[11], the only three *Ilex* species with sequenced genomes. About 64.63% of the genome assembly was composed of repetitive sequences, of which ~36.22% were retrotransposons, ~1.80% were DNA transposons, ~0.74% were simple repeats and ~0.15% were low complexity regions. Long terminal-repeat (LTR) retrotransposons of the Gypsy and Copia families were the most abundant transposable elements, as observed in many sequenced plant genomes[29], followed by long interspersed nuclear elements (LINEs) and hobo-Activator transposons, among others (Supplementary Table 3).

A total of 53,390 protein-coding genes were predicted in the genome, with a mean coding sequence length of 3,062 bp and 4.23 exons per gene. Of these, 41,483 (~77.63%) could be annotated with GO terms, EC numbers or Pfam domains. In addition, we identified 4,530 non-coding RNA genes, including 2,670 small nucleolar RNAs, 815 transfer RNAs, 471

140 ribosomal RNAs, 348 small nuclear RNAs, and 226 micro RNAs (Supplementary Notes S1
141 and Supplementary Tables 4-6). To further assess the completeness of the assembly, we
142 aligned the scaffolds with the KOG[30] and DEG[31] databases, determining that 98% of the core
143 gene families from the KOG database and 97.5% of the *Arabidopsis thaliana* DEG subset
144 were present in the assembly. Then, we performed a Benchmarking Universal Single-Copy
145 Orthologs (BUSCO)[32] assessment using the eudicot ODB10 database. Among 2,326
146 conserved single-copy genes, ~96.20% were retrieved, of which ~78.80% were complete and
147 single copies, ~17.40% were complete and in duplicates, ~3.10% were fragmented and only
148 ~0.70% were missing. These results suggest that the coding region of the assembly is nearly
149 complete. The number of estimated genes for YM is higher than the ca. 39,000 reported from
150 the genome sequences of other *Ilex* species[9–11]. This could be at least partly due to the larger
151 genome size of YM estimated from flow cytometry relative to the other species[33].
152
153 **Evolutionary analysis of yerba mate genome provides evidence of whole-genome**
154 **duplication in an early *Ilex* ancestor**
155
156 Most plant lineages have experienced ancient polyploidization events followed by
157 massive duplicate gene losses and genome rearrangements, which may have contributed to
158 the evolution of developmental and metabolic complexity[34,35]. Recent transcriptome-based
159 analyses[36,37] reported an ancient polyploidization event in the *Ilex* lineage around 60 Ma
160 (Cretaceous–Paleogene boundary), based on phylogenomic and synonymous substitution
161 rate (K_s) evidence. Evolutionary analyses of *Ilex polyneura*[9] and *Ilex latifolia*[11] genomes also
162 provided evidence of a shared *Ilex*-specific whole genome duplication (WGD). As YM is the
163 first American holly to have its genome sequenced, we performed synteny-based analyses of
164 its genome to deepen our understanding of Aquifoliales evolution (Fig. 2 and Supplementary
165 Fig. 1). The K_s distribution of YM paralogues (Fig. 2b) revealed a significant peak with a
166 median K_s value of ~0.37, not shared with the rest of the eudicot genomes analysed (Fig. 2b
167 and Supplementary Fig. 1). This confirms the lineage-specific polyploidization event (Ip-α)

168 previously reported in *Ilex*[9,11,36,37], in addition to the shared ancestral WGT-γ which is indicated
169 by a median K_s value of ~1.4 (Fig. 2b). A WGD in the common ancestor of *Ilex* species is
170 further supported by 2:1 syntenic depth ratios between the YM genome and the coffee and
171 grape genomes, which did not experience additional duplication events after the ancestral
172 WGT-γ (Fig. 2c). In order to determine the age of Ip-α, we used two different phylogenies (Fig.
173 2a). The plastid genome phylogeny supports the monophyly of Aquifoliales as the first
174 diverging clade of campanulids[38]; the alternative nuclear genome phylogeny supports *Ilex* in
175 Aquifoliales I as an early branching lineage of lamiids[36]. With the former phylogeny, we
176 estimated the age of the WGD event between 48.75-69.63 Ma while, with the latter,
177 divergence was estimated at 49.43-70.62 Ma (Fig. 2a). Both estimates are consistent with that
178 of Zhang et al. 2020 and validate the age of Ip-α near the origin of *Ilex*, which is estimated
179 between 43-89 Ma[39].
180
181 **Convergent evolution of caffeine biosynthesis in yerba mate**
182
183 In order to determine the genes and biochemical pathway responsible for caffeine
184 biosynthesis in YM, we used bioinformatic analyses to identify SABATH enzyme family
185 members in the genome[19,22,23]. There appear to be 28 full-length SABATH genes in YM that
186 encode members of the functionally diverse clades of the family, including SAMT[40] and JMT[41],
187 among others (Fig. 3a). Our phylogenetic analysis showed that although the YM genome does
188 not appear to encode XMT-type caffeine-producing enzymes like *Coffea* and *Citrus*, it does
189 contain three recently and tandemly duplicated genes that encode CS-type enzymes, IpCS1-
190 3 (Fig. 3a, c). The duplicated IpCS1-3 are 86-91% identical at the amino acid level and are
191 expressed at highest levels in caffeine-accumulating tissue (Fig. 3b)[19]. IpCS1-3 also appear
192 to be of recent origin, since non-caffeine accumulating *Ilex* species only have a single gene or
193 gene fragment in the syntenic region (Fig. 3c). In *Camellia*, *Theobroma* and *Paullinia*, recent
194 duplications of the CS-type enzymes responsible for the successive steps of xanthine alkaloid
195 methylation have also independently occurred (Fig. 3a)[19,24]. Two other YM genes encode

196 IpCS4 & 5, but these are not syntenic with IpCS1-3 and are not highly expressed in any tissues
197 studied (Fig. 3b); therefore, we did not characterize them further.
198 To investigate the biochemical activities of the enzymes encoded by the three CS-type
199 genes, we cloned them into bacterial expression vectors and determined heterologous protein
200 functions. One enzyme, IpCS1, appears to primarily methylate X to catalyse the formation of
201 3X (Fig. 4). A second enzyme, IpCS2, shows activity only with 3X to produce TB, while a third
202 enzyme, IpCS3, exhibits a preference to methylate TB to form CF (Fig. 4). Thus, collectively,
203 these three enzymes appear capable of catalysing a complete pathway from xanthine to
204 caffeine. The apparent K_M for the preferred substrates of all three enzymes ranges from 85-
205 197 μM, and the k_{cat} / K_M estimates are comparable to those determined for other caffeine
206 biosynthetic enzymes[24] (Supplementary Table 7). Further evidence for this biosynthetic
207 pathway has been reported by ^{14}C xanthine tracer studies in young leaf segments of *I.*
208 *paraguariensis* that showed radioactivity in 3X and TB in addition to CF[42]. A pathway from
209 X→3X→TB→CF has also been reported for *Theobroma* and *Paullinia* using CS-type SABATH
210 enzymes[19]. Like Huang et al. 2016, this represents another departure from the long-assumed
211 pathway to caffeine biosynthesis (XR→7X→TB→CF) as reported in coffee and tea (Fig. 1).
212 This instance in *Ilex* is particularly notable since YM is an Asterid, like coffee and tea. The fact
213 that *Ilex*, *Theobroma* and *Paullinia* convergently recruited CS genes that independently
214 duplicated and evolved to encode enzymes with similar substrate preferences to catalyse a
215 common pathway to caffeine, in spite of their divergence more than 100 Ma[43], is remarkable
216 and suggests a high degree of genetic constraint governing the repeated origin of this trait.
217 While the substrate preferences shown in Fig. 4 suggest pathway flux from
218 X→3X→TB→CF, IpCS1 also shows secondary activity with 7X to produce TB and IpCS3 can
219 catalyse the formation of CF from paraxanthine (PX) (Fig. 4a). Thus, flux through other
220 branches of the xanthine alkaloid biosynthetic network (Fig. 1) cannot be excluded. However,
221 it is not clear how 7X or PX would be produced *in planta* since none of the three enzymes
222 studied here is capable of their formation; therefore, these secondary activities may not be
223 physiologically relevant. In addition, it has been proposed that TP may also be a precursor to

224 caffeine biosynthesis in *I. paraguariensis* based on radioisotopic feeding studies[42], although
225 its levels in plant tissues are 30-160 times lower than TB[44]. Our *in vitro* enzyme assays provide
226 no experimental evidence for that biosynthetic route; however, it is possible that additional MT
227 enzymes from the SABATH (or other) gene family not characterized in this study may perform
228 such reactions. Alternatively, if the exogenously supplied TP was first catabolized to 3X in YM
229 tissues, then the caffeine detected previously[42] could have been synthesized via the route
230 described above for IpCS2+IpCS3 (Fig. 4).
231
232 **The caffeine biosynthetic pathway in YM evolved from ancestral networks with different**
233 **inferred flux**
234
235 The fact that caffeine is produced within only one small lineage of *Ilex* that diverged
236 and experienced CS gene duplication (Fig. 3) within the last 11 million years[39,44] indicates that
237 the pathway has only recently evolved. The nature by which novel multistep biochemical
238 pathways evolve is a central question in biology[45]. To investigate the caffeine pathway origin
239 in YM, we used Ancestral Sequence Resurrection[46,47] to study AncIpCS1 & AncIpCS2, the
240 ancestors of the three modern-day enzymes implicated in caffeine biosynthesis in YM (Fig. 5
241 and Supplementary Fig. 2-5). The ancestral enzyme, AncIpCS1, which gave rise to all three
242 modern-day YM enzymes, exhibits highest relative activity with X, 3X and 7X (Fig. 5a).
243 Methylation of 7X by AncIpCS1 occurred at the N3 position resulting in TB synthesis, whereas
244 xanthine methylation occurred at either the N1 or N3 position to form 1X and 3X, respectively
245 (Fig. 5b and Supplementary Fig. 6). AncIpCS1 was capable of methylation of 3X at N1 to
246 produce TP, while methylation at the N7 position led to TB formation (Fig. 5b). These data
247 demonstrate that, although AncIpCS1 could not produce caffeine, it could methylate xanthine
248 alkaloids at 3 different positions of the planar heterocyclic ring structures and this combination
249 of activities would have allowed for the ancestor of YM to produce a cocktail of 1X, 3X, TP and
250 TB by flux through multiple branches of the xanthine alkaloid biosynthetic network with a single
251 enzyme (Fig. 5b).

252 After gene duplication of AncIpCS1, one daughter enzyme ultimately gave rise to
253 IpCS1, which exhibits preference to methylate xanthine to produce 3X (Fig. 5a). The other
254 daughter enzyme, AncIpCS2, appears to have maintained highest activity with X, 3X and 7X
255 like AncIpCS1 (Fig. 5a). However, unlike its ancestor, AncIpCS1, AncIpCS2 evolved high
256 relative activity with 7X to produce not just TB, but also PX by methylation at the N1 position
257 (Fig. 5a and Supplementary Fig. 6b). AncIpCS2 retained the ancestral activity of AncIpCS1
258 with xanthine to produce 1X, but also evolved the ability to methylate X at the N7 position (Fig.
259 5b and Supplementary Fig. 6). This enzyme also retained ancestral activity with 3X to produce
260 only TB by N7 methylation but lost the ability to methylate 3X at the N1 position to form TP.
261 These activities of AncIpCS2 would have allowed for ancestral flux to produce 1X, 7X, TB and
262 PX but not caffeine. Because a YM ancestor could have possessed both AncIpCS2 and a
263 descendant of AncIpCS1, AncIpCS1' (Fig. 5b), additional pathway flux is possible. If
264 AncIpCS1' retained activities of its ancestor, AncIpCS1, then the ancestral *Ilex* species could
265 have also produced 3X and TP making for an even more diverse array of xanthine alkaloids
266 in its tissues (Fig. 5b). It has been shown that the xanthine alkaloids, 1X, 3X, and TP, can bind
267 to modern-day rat adenosine receptors[48]. Therefore, if these molecules were to accumulate in
268 ancestral *Ilex* tissues, they could have conferred a selective advantage which would likely
269 result in retention of the ancient genes. Ultimately, once gene duplication led to the generation
270 of the three modern-day CS-type enzymes in YM, pathway flux could be channelled away
271 from intermediates like 1X and TP such that the modern-day pathway to caffeine evolved (Fig.
272 5b). Not only did the modern-day CS enzymes of YM evolve to catalyse a pathway from
273 X>3X>TB>CF from ancestral biosynthetic networks of different products, *Theobroma* and
274 *Paullinia* also independently evolved enzymes with similar properties[19]. And, they did so from
275 ancestral pathways that, like YM, had alternative ancestral fluxes[24]. While it could be due to
276 chance alone that all three lineages converged to catalyse a similar pathway from differing
277 ancestral networks, it is also possible that it was advantageous to specialize for flux to TB via
278 X and 3X because either it is more enzymatically favourable or these intermediates have
279 greater adaptive value than other structural isomers.

280 **Protein crystal structure of IpCS3 reveals convergent structural basis for methylation**
281 **of theobromine to form caffeine**
282
283 We successfully crystallized and determined the 2.7 Å resolution structure of IpCS3
284 (PDB ID: 8UZD), that converts TB into CF. This enzyme crystallizes as a holo-homodimer,
285 bound to both of its reaction products: S-adenosyl-homocysteine (SAH) and caffeine (Fig. 6a,
286 Supplementary Table 8). As is typical for the SABATH family of methyltransferases, IpCS3
287 exhibits a Rossman-like fold composed of 7 β-strands surrounded by 5 α-helices which bind
288 the methyl-donor S-adenosyl-L-methionine (SAM), as well as an alpha-helical cap which bind
289 the methyl-acceptor theobromine[49–52]. This structural information of the enzyme bound to both
290 of its products, SAH and caffeine, facilitates an in-depth comparison of the active site
291 structures of the caffeine-producing CS-type enzyme found in *Ilex* and the XMT-type enzyme
292 in *Coffea*[49] to determine the extent to which convergence of physicochemical properties of the
293 active site has allowed for independent specialization for theobromine methylation by the
294 paralogous SABATH enzymes. Although the IpCS3 structure was obtained in complex with its
295 product, caffeine, it can be assumed that the binding mode is conserved for its precursor,
296 theobromine (Supplementary Fig. 7). Indeed, our computational modelling of theobromine in
297 the active site of IpCS3 predicts it to be oriented as we have discerned from the diffraction
298 data (Supplementary Fig. 8). Thus, in the following comparisons, the atomic numbering for the
299 theobromine precursor will be used to facilitate comparison to the DXMT structure.
300 In both DXMT and IpCS3, there are several conserved residues, shared by nearly all
301 SABATH enzymes, that form part of the active site pocket and appear to play important roles
302 in binding many different substrates[50,51]. His160 and Trp161 in DXMT are in the same relative
303 positions as His155 and Trp156 in IpCS3 (Fig. 6b, c). These residues are ca. 3 Å from TB and
304 participate in H-bonding but to different atoms of the substrate. In DXMT, the NE1 of Trp161
305 and NE2 of His160 form hydrogen bonds to carbonyl O2 of TB when positioned for N1
306 methylation; yet, in the structure of IpCS3 these corresponding side chain groups form
307 hydrogen bonds to O6 of TB. Despite these two residues being conserved for H-bonding, the

308 substrates are rotated 180 degrees along an axis going through N1 and C4. Thus, the
309 conserved His and Trp residues interact with opposing carbonyls in TB/CF but still position
310 the substrate for N1 methylation (Fig. 6b, c).

311 On the other hand, there are residues that differ between the two enzymes but appear
312 to provide for important substrate interactions. Specifically, in the structure of DXMT, the
313 hydroxyl group of Ser237 allows specific hydrogen bonding with N9 to position TB for N1
314 methylation (Fig. 6c). In IpCS3, His236 is found at the homologous position in the structure.
315 Nevertheless, its involvement in H-bonding with N9 is uncertain as the distance between
316 nitrogen atoms is ca. 4Å. Tyr368 of DXMT is found to participate in π-π interactions with the
317 ring structure of TB[53]; yet in IpCS3, Asn353 is found in the homologous position and the amine
318 forms a hydrogen bond with N9 due to its proximity within 3.2 Å, which is additionally stabilized
319 by Asn221 and His236 (Fig. 6b). Because the residues in these positions of IpCS3 and DXMT
320 differ yet contribute to TB binding, these independent replacements represent convergent
321 structural solutions for N1 methylation of the substrate.

322

323 **Computational modelling of IpCS1 and IpCS2 active sites predict convergent substrate**
324 **binding residues for xanthine and 3-methylxanthine methylation**

325

326 Previous studies used site-directed mutagenesis of two sequence regions in CS-type
327 caffeine biosynthetic enzymes from *Theobroma* and *Paullinia* to uncover the mutational basis
328 for the convergent evolution of substrate preference switches towards their preferred
329 substrates, X and 3X[24]. In order to determine whether the same regions were convergently
330 mutated in IpCS1 and IpCS2 and provide binding interactions with X and 3X, respectively,
331 AlphaFold2[54] models and subsequent docking studies were performed (Supplementary Fig.
332 9). Modelling of substrate binding in the predicted active sites of IpCS1 and IpCS2 (Fig. 7a, b)
333 shows that the preferred substrates have optimal binding orientations that would result in
334 methylation to form the products that were experimentally detected in our assays shown in
335 Fig. 4. From our docking simulations, IpCS1 residues W156, N221 and Y265 are positioned

336 for hydrogen bonding with the carbonyl moieties of xanthine to position N3 for methyl transfer
337 from SAM (Fig. 7a). Although *Theobroma* and *Paullinia* CS1 enzymes specialized for xanthine
338 methylation also possess W156 and Y265 in the homologous positions (Supplementary Fig.
339 10), these residues are highly conserved amongst nearly all angiosperm SABATH enzymes.
340 On the other hand, the homologous position to N221 did not change concomitantly with the
341 evolution of X preference in either *Theobroma* and *Paullinia* (Supplementary Fig. 10); instead,
342 when their "region III" was mutated, activity with X improved in both[24]. Because IpCS1 was
343 not mutated in the homologous region III, there appear to be convergent solutions allowing for
344 efficient positioning of X for 3X biosynthesis amongst these enzymes. In the case of IpCS2,
345 two hydrogen bond donors S24 and T25 appear to contribute to the positioning of 3X in the
346 active site (Fig. 7b). This homologous region was experimentally mutated in *Theobroma* and
347 *Paullinia* CS2 enzymes and improved specialization for 3X methylation in both, although the
348 actual substitutions are different in each case[24]. Thus, this may represent an additional
349 instance where convergent mutations of the same region lead to specialization for 3X
350 methylation.
351
352 **Comparative phylogenomic analyses of caffeine biosynthetic genes reveal historical**
353 **constraints to convergent gene co-option**
354
355 Many nearly ubiquitous specialized metabolites involved in defence, development and
356 floral scent are produced by SABATH enzyme family members that appear to be conserved
357 across diverse angiosperm lineages, such as SAMT that methylates salicylic acid[55] and IAMT
358 that methylates Indole-3-acetic acid[52]. However, caffeine is sporadically distributed amongst
359 disparate angiosperm lineages and seems to have only recently evolved by convergence in a
360 few distantly related orders[19]. Our comparative evolutionary genomic analysis of the CS and
361 XMT syntenic regions across angiosperm (Fig. 8) indicates that predicting which SABATH
362 locus a given lineage might co-opt for caffeine biosynthesis is more dependent upon the
363 idiosyncratic history of gene loss than phylogenetic relatedness. For example, in the case of

364 the CS syntenic region used for caffeine biosynthesis in YM and *Theobroma*, *Coffea* lacks a
365 CS ortholog and none can be detected from its genome (Supplementary Fig. 11a). Thus, only
366 XMT was historically available for recruitment in *Coffea*. Conversely, YM appears to have lost
367 any vestiges of XMT orthologues known to be responsible for caffeine biosynthesis in *Coffea*
368 and *Citrus* (Supplementary Fig. 11b-d). This lack of genomic potential may be seen as an
369 evolutionary constraint to gene recruitment for caffeine biosynthesis in *Coffea* and YM.

370 A broader phylogenetic perspective on the XMT and CS syntenic regions provides
371 further insight into genomic canalization and allows for predictions about the underlying
372 genetic basis for caffeine biosynthesis in as-of-yet characterized lineages. As shown in Fig. 8,
373 several angiosperm lineages have neither XMT nor CS and this may explain why caffeine has
374 apparently never evolved in the large and diverse orders Brassicales, Asterales, Solanales
375 and Lamiales even though it has been shown to be advantageous in transgenic plants[56,57]. In
376 the case of *Cola*, a caffeine-producing genus from Africa[58], it is predicted to have co-opted CS
377 genes for xanthine alkaloid methylation because the order Malvales to which it belongs
378 appears to have lost XMT orthologues prior to its origin (Fig. 8). Tests of this hypothesis await
379 genomic sequences and functional studies of *Cola* enzymes. However, even with a functional
380 XMT or CS enzyme, duplication and diversification appears to be required to assemble a
381 complete pathway to caffeine as shown here for YM.

382

383 **Methods**

384

385 **Plant materials.** *Ilex paraguariensis* A. St.-Hil. var. *paraguariensis*, cv CA 8/74 (INTA-EEA
386 Cerro Azul, Misiones, Argentina) and cv SI-49 (Establecimiento Las Marías S.A.C.I.F.A.,
387 Corrientes, Argentina) were used in this study. High productivity, increased tolerance to
388 drought, and ease of propagation with stem cuttings characterize these genotypes[27,59,60].

389

390 **DNA extraction and sequencing.** Two DNA extraction and sequencing approaches were
391 combined to improve the accuracy of genome assembly. First, young leaves of cv CA 8/74,

392 preserved in silica-gel, were used to isolate total genomic DNA with the DNeasy® Plant Mini
393 Kit (QIAGEN®), following manufacturer's instructions. Paired-end libraries (with insert sizes of
394 350 bp and 550 bp) and mate-pair libraries (with insert sizes of 3 kbp, 8 kbp and 12 kbp) were
395 constructed using the Illumina TruSeq® DNA Sample Preparation Kit and Illumina Nextera®
396 Mate Pair Library Preparation Kit following the kit's instructions, respectively. The obtained
397 libraries were sequenced on an Illumina HiSeq 1500 platform, generating ~263.2 Gb of raw
398 data. Second, young leaves of cv SI-49, preserved in silica-gel, were used to purify high
399 molecular weight DNA with the Quick-DNA™ HMW MagBead Kit (Zymo Research), according
400 to the manufacturer's instructions. Long-reads libraries were prepared using Sequel Binding
401 Kit 1.0 (Pacific Biosciences), following manufacturer's instructions. The obtained libraries were
402 subsequently sequenced on PacBio Sequel I (Pacific Biosciences) using Sequel Sequencing
403 Kit 1.0 (Pacific Biosciences) and SMRT Cell 1M (Pacific Biosciences), generating ~77.5 Gb of
404 additional raw data.
405
406 **Genome assembly and quality assessment.** We opted for a pipeline that could take
407 advantage of both short and long sequencing technologies. For the short-reads, we applied
408 Trimmomatic[61] v.0.39 to remove adaptor contaminations and filter low-quality reads (reads
409 with mean quality scores ≤ 25, reads where the quality of the bases at the head or tail was ≤
410 10 and reads with a length ≤ 30 bp). The resulting clean reads were then corrected using
411 Quake[62] v.0.3. Contig assembly and scaffolding was done using the assembler
412 SOAPdenovo[63] v.2 (55-mer size), with the mate-pair reads being used to link contigs into
413 scaffolds. After the assembly, DeconSeq[64] v.0.4.3 was used to detect and remove sequence
414 contaminants. Contigs and scaffolds clearly belonging to the chloroplast and mitochondria
415 genomes were also discarded. YM transcriptome sequences[25,26,27] and public databases
416 (KOG[30] and DEG[31]) were used to validate the genome assembly. Canu[65] v.2.2 was used to
417 perform the self-correction and assembly of the long reads, using the default parameters and
418 stopOnLowCoverage=20. For both short and long assemblies, we separated the assembly
419 haplotypes (haplotigs) using PurgeHaplotigs[66] with the recommended parameter values.

420 Then, we merged both SOAPdenovo v.2 and Canu v.2.2 curated assemblies using
421 Quickmerge[67] v.03, where only contigs with minimum overlap of 5000 bp (-ml 5000) were
422 merged and only the contigs greater than 1000 bp (-l 1000) were retained. The resulting
423 scaffolds and contigs were refined further with the gap-filling module in SOAPdenovo v.2
424 (GapCloser) and SSPACE[68] v.2.1.1.
425
426 **Gene prediction and annotation.** First, we masked the genome assembly with
427 RepeatMasker (http://repeatmasker.org/). Then, we predicted the protein coding and non-
428 coding genes using Funannotate[69] v.1.8.13 previously training it with the available *Ilex*
429 *paraguariensis* RNA-Seq experiments (NCBI projects PRJNA315513, PRJNA375923 and
430 PRJNA251985). Then, Infernal[70] v.1.1.4 was employed to improve the prediction of small
431 RNAs and microRNAs, while tRNAScan-SE[71] v.2.0 was used to improve the prediction of
432 transfer RNAs. Ribosomal RNAs were predicted using RNAmmer[72] v.1.2. The TAPIR web
433 server[73] (http://bioinformatics.psb.ugent.be/webtools/tapir) and the TargetFinder software[74]
434 v.1.7 were used to identify miRNA targets. InterProScan[75] v.5.55-88.0 and eggNOG-mapper[76]
435 v.2.1.7 were employed for the functional assignment of the predicted genes.
436
437 **Repeat content estimation.** The repeat content was estimated employing Dfam TE Tools
438 v.1.5 (https://github.com/Dfam-consortium/TETools). First, we used RepeatModeler v.2.0.3[77]
439 to build a database with *Ilex* repeat families. Then, we merged that database with Dfam v3.6[78]
440 and GIRI Repbase ver 20181026[79]. Finally, we ran RepeatMasker on the assembly using the
441 merged database to look for repeat sequences.
442
443 **Genome duplication analysis.** Rates of synonymous substitution (K_s) between paralogous
444 genes and orthologous genes in *Lactuca sativa*, *Daucus carota*, *Ilex paraguariensis*, *Coffea*
445 *canephora* and *Vitis vinifera* were determined using CoGe's tool SynMap
446 (https://genomevolution.org/). Gaussian mixture models (GMMs) were fitted to the resulting K_s
447 distributions with the mclust R package[80] v.5.0, and significant peaks were identified using the

448 SiZer R package[81] v.0.1-7. To estimate the age of the lineage-specific polyploidization event
449 (Ip-α) in *Ilex*, we considered two different phylogenies (a multiple nuclear genome phylogeny
450 and a plastid genome phylogeny). With the median K_s value of yerba mate-grape orthologues
451 (~0.89) and the divergence date of the two species (125.64 Ma for the multiple nuclear
452 genome phylogeny and 123.7 Ma for the plastid genome phylogeny), we calculated the
453 number of substitution per synonymous site per year (*r*) for YM (divergence date = $K_s/(2 \times r)$).
454 Conforming to the multiple nuclear genome phylogeny, the YM *r* value is 3.54E-9; while for
455 the plastid genome phylogeny, the YM *r* value is 3.59E-9. These *r* values and the SiZer K_s
456 range of YM paralogues (~0.35-0.5) were then applied to estimate the age of Ip-α. Finally, to
457 determine the syntenic depth ratio between *I. paraguariensis* and *C. canephora* and *V.*
458 *vinifera*, we employed CoGe's tool SynFind (https://genomevolution.org/), using a distance
459 cutoff of 10 genes and requiring at least 5 gene pairs per synteny block.
460
461 **Gene expression quantitation.** First, YM transcriptome reads (PRJNA315513) were mapped
462 to IpCS1-5 transcripts, obtained from the *de novo* transcriptome assembly and annotation,
463 using BWA[82]. Then, with the number of mapped reads, the abundance of each transcript was
464 calculated, normalized by transcript length and transcriptome size (quantification in RPKM,
465 reads per kilobase per million mapped reads).
466
467 **Cloning, heterologous expression and purification of enzymes.** Two different approaches
468 were used to clone IpCS genes: RT-PCR from leaf tissue and custom gene synthesis. For RT-
469 PCR of IpCS2, cDNA was obtained from 1 µg of RNA from fresh YM leaves using standard
470 procedures and cycling conditions with the following two primers: IpCS2F 5'-
471 ATGGACGTGAAGGAAGCAC-3' and IpCS2R 5'-CTATCCCATGGTCCTGCTAAG-3'.
472 Following amplification, cDNA was cloned using the pTrcHis TOPO® TA Expression Kit
473 (Invitrogen, Carlsbad, CA). Ligation of cDNA into the pTrcHis vector and subsequent
474 transformation into Top10 *E. coli* cells was carried out according to the manufacturer's
475 protocol. The transformation mixture was incubated overnight at 37°C on LB plates containing

476 50 µg/mL ampicillin. Colonies were screened by PCR to obtain full-length inserts that were
477 subsequently verified for insert orientation by DNA sequencing. For IpCS1 & 3 and AncIpCS1
478 & 2, gene sequences were synthesized by GenScript with codons optimized for expression in
479 *E. coli*. Synthesized genes were subcloned from the pUC57 cloning vector into the pET-15b
480 (Novagen) expression vector using 1.5 µg of DNA and NdeI and BamHI in 30 µl reactions.
481 Linear fragments corresponding to the expected sizes were gel purified using the QIAEX II
482 Gel Extraction Kit (Qiagen Corp.) according to the manufacturer's instructions. Purified DNA
483 fragments were ligated into pET-15b using T4 DNA ligase from New England Biolabs. Ligation
484 products were transformed into Top10 *E. coli* cells using 2 µl of the ligation reaction. Minipreps
485 of positive transformants were obtained using a QIAprep Spin Miniprep Kit (Qiagen Corp.) and
486 10 ng of each plasmid was used to transform BL21 *E. coli* cells using standard plating and
487 incubation methods.

488 Induction of His_6-protein was achieved in 50 ml cell cultures of BL21 (DE3) with IpCS1
489 & 3 and AncIpCS1 & 2 in pET-15b or Top10 with IpCS2 in pTrcHis with the addition of 1mM
490 IPTG at 23° C for 6 hours. Purification of the His_6-tagged protein was achieved by TALON®
491 spin columns (Takara Bio) following the manufacturer's instructions. Bradford assays were
492 used to determine purified protein concentration, and recombinant protein purity was
493 evaluated on SDS-PAGE gels.

494

495 **Enzyme assays.** All enzymes were tested for activity with the eight xanthine alkaloid
496 substrates shown in Fig. 1. Radiochemical assays were performed at 24° C for 60 minutes in
497 50 µl reactions that included 50 mM Tris-HCl buffer, 0.01 µCi (0.5 µl) ^{14}C-labelled SAM, 10-20
498 µl purified protein and 1 mM methyl acceptor substrate dissolved in 0.5 M NaOH. Negative
499 controls were composed of the same reagents, except that the methyl acceptor substrate was
500 omitted and 1 µl of 0.5 M NaOH was added instead. Methylated products were extracted in
501 200 µl ethyl acetate and quantified using a liquid scintillation counter. The highest enzyme
502 activity reached with a specific substrate was set to 1.0 and relative activities with remaining

503 substrates were calculated. Each assay was run at least twice so that mean, plus standard
504 deviation, could be calculated.
505
506 **High performance liquid chromatography (HPLC).** Product identity of enzyme assays was
507 determined using HPLC on 500 µl scaled-up reactions utilizing all the same reagents as
508 described above except that non-radioactive SAM was used as the methyl donor and reactions
509 were allowed to progress for 4 hours. Whole reactions were filtered through Vivaspin®
510 columns (Sartorius) to remove all protein prior to injection in the HPLC. Mixtures were
511 separated by HPLC using a two-solvent system with a 250 mm x 4.6 mm Kinetex® 5 µM EVO
512 C18 column (Phenomenex). Solvent A was 99.9% water with 0.1% TFA and Solvent B was
513 80% acetonitrile, 19.9% water and 0.1% TFA and a 0-20% gradient was generated over 16
514 minutes with a flow rate of 1.0 ml/minute. Product identity was determined by comparing
515 retention times and absorbance at 254 and 272 nm of authentic standards. Reactions were
516 compared to negative controls in which no methyl acceptor substrates were added.
517
518 **Phylogenetic analyses.** In order to accurately determine the orthology of YM SABATH
519 sequences encoded in the genome, we compared them to all previously characterized gene
520 family members in other species. We also included CS and XMT orthologues from relatives
521 from the orders of caffeine-producing species (Malvales, Ericales, Gentianales, Sapindales)
522 available in public databases (GenBank, OneKP) as shown in Fig. 3. Accession numbers for
523 all sequences are provided in Supplementary Table 9. Alignment of amino acid sequences
524 was achieved using MAFFT[83] v.7.0 and employing the auto search strategy to maximize
525 accuracy and speed. A phylogenetic estimate was obtained using FastTree[84] v.2 assuming
526 the Jones-Taylor-Thorton model of amino acid substitution with a CAT approximation using
527 20 rate categories. Reliability of individual nodes was estimated from local support values
528 using the Shimodaira-Hasegawa test as implemented in FastTree.
529

530 **Ancestral sequence resurrection.** In order to obtain accurate ancestral CS protein
531 estimates, we assembled two datasets to assess variation in terms of sampling. The first
532 dataset included 154 sequences including all CS-type enzymes we could retrieve from
533 GenBank and China National Gene Bank as well as representatives of all other functionally
534 characterized clades of SABATH enzymes (Supplementary Fig. 2). This set of ancestral
535 enzymes was obtained using only the sequences from YM which were available during our
536 initial analyses. Subsequently, once additional *Ilex* genomes became available, we estimated
537 a second set of ancestral sequences using 29 CS-type enzymes from asterids to assess
538 uncertainty in our estimates (Supplementary Fig. 3). IQTree[85] was used to obtain trees
539 describing the relationships amongst the sequences for both datasets. For the first, the Jones,
540 Taylor, and Thorton (JTT) matrix model for amino acid substitution and the Free rate model
541 for among-site rate heterogeneity[86] was determined to be the best fit. For the second dataset,
542 the Q matrix as estimated for plants[87] with a gamma model for rate heterogeneity was the
543 preferred model. Confidence in ancestral sequences was high in both analyses as indicated
544 by the average site-specific posterior probability for AncIpCS1 (0.99) and AncIpCS2 (0.99). In
545 order to determine ancestral protein lengths in regions with alignment gaps, we coded each
546 gap for the number of amino acids possessed and used parsimony to determine ancestral
547 residue numbers as in our previous studies[19]. The estimated sequences were synthesized by
548 Genscript Corp. and had codons chosen for optimal protein expression in *Escherichia coli*.
549 Although the two separate estimates are highly similar to one another (>95% in both cases),
550 the two AncIpCS1 proteins differ at 10 positions and those for AncIpCS2 differ at 7 positions
551 (Supplementary Fig. 4 & 5).
552
553 **Crystallization, data collection, phasing and refinement of IpCS3.** Initial crystallization
554 screening was performed using the IpCS3 methyltransferase at a concentration of 30 mg/mL
555 incubated with 2 mM TB and 2 mM SAM. Sitting-drop for crystallization screening was set up
556 by equal volume of precipitant and protein solution (0.25:0.25 µL) using a Crystal Gryphon
557 robot (Art Robbins Instruments) and a reservoir volume of 45 µL. Trays were incubated at 9

558 ℃. Initial hits were further optimized using the hanging-drop method at 9°C, with 150 µL
559 reservoir solution and 1:1 ratio of precipitant to protein and ligand solution in a 2 µL drop.
560 Attempts to crystallize with SAH or uncleavable SAM analogs and TB to attain a pre
561 methylation structure were unsuccessful given the poor diffraction of these crystals. Therefore,
562 the latter was composed of 33 mg/mL IpCS3 protein concentration, 4 mM TB and 2 mM SAM,
563 and the crystallization condition was optimized to 25% PEG 3350, 0.2 M NH_4SO_4, 0.1 M Bis-
564 Tris methane pH 5.5. Square crystals grew over 10 days, but initial X-ray crystallography data
565 revealed a poor electron density for SAH and an electron density in the active site for CF, the
566 product, rather than for TB. Consequently, crystals were grown in the aforementioned
567 condition and subsequently soaked for 4h at 9°C in the precipitant solution supplemented with
568 10 mM SAH and 10 mM TB. The idea was to supply an excess of the expended methyl source
569 and additional TB to convert any existing SAM as we did not have access to caffeine as a
570 reagent. Crystals were transiently soaked in the precipitant solution supplemented with 20%
571 Ethylene Glycol immediately prior to vitrification by direct immersion into liquid nitrogen.
572 Diffraction data were collected at the Advanced Photon Source (APS) at Argonne National
573 Laboratory Sector-21 via the Life Sciences-Collaborative Access Team (LS-CAT) at beamline
574 21-ID-G. Diffraction data were indexed, integrated, and scaled using the autoPROC software
575 package[88]. The structure was solved by molecular replacement using Phaser-MR included in
576 the Phenix software package[89], using PDB ID 6LYH structure as the replacement model. The
577 model was subject to rounds of manual building followed by refinement using REFMAC5[90],
578 and was manually built in COOT[91] v.0.9.8.3. Crystallographic statistics are listed in
579 Supplementary Table 8.
580
581 **Structure prediction and molecular docking.** Protein structures of IpCS enzymes were
582 predicted using the ColabFold implementation of AlphaFold2[54] with no template. Diagnostic
583 plots depicting the MSA coverage, alignment error and LDDT are shown in the supplementary
584 information (Supplementary Fig. 9). Structures of xanthine alkaloid ligands (X, 3X, and TB)
585 were downloaded from the ChEMBL database[92]; protonation states were checked by

586 Chemicalize[93] and optimized using the VMD Molefacture plugin[94]. The receptor structures
587 were prepared following the standard AutoDock protocol[95] using the prepare_receptor4.py
588 script from AutoDock Tools. All non-polar hydrogens were merged, and Gasteiger charges
589 and atom types were added. The ligand PDBQT was prepared using the prepare_ligand4.py
590 script. The grid size and position were chosen to contain the whole ligand-binding site
591 (including all protein atoms closer than 5 Å from all ligands). For each system, 10 different
592 docking runs were performed. Docking was performed using AutoDock Vina v.1.2.0[96]. The
593 docking results were further analysed by visual inspection. Images of the molecules were
594 prepared using the PyMOL molecular graphics system[97].
595
596 **Synteny comparisons and phylogenetic distribution of CS and XMT.** The presence or
597 absence of CS and XMT genes was determined for orders of plants for which at least one
598 genomic sequence exists, as shown in Fig. 8. For those species which do not yet have an
599 available assembly, we used BLAST[98] analyses of GenBank (nr and TSA databases),
600 Phytozome[99] as well as the OneKP dataset[37] in China National GeneBank. BLAST combined
601 with subsequent phylogenetic analyses were also used to verify presence/absence of CS- or
602 XMT-type sequences in cases where the syntenic regions did not appear to encode one or
603 the other gene. Comparisons of the CS and XMT syntenic regions were performed using
604 CoGe's tool GEvo (https://genomevolution.org/).
605
606 **Data availability**
607
608 The Illumina and PacBio raw sequence data, assembly and annotation were deposited
609 in the European Nucleotide Archive (ENA) under BioProject No. PRJEB65927. An assembly
610 obtained only with the Illumina data was also deposited in ENA under BioProject No.
611 PRJEB36685. The plasmids used to produce proteins are freely available upon request. The
612 atomic coordinates and structure factors have been deposited in the Protein Data Bank,
613 Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ

614 (http://www.rscb.org) with the accession code 8UZD for the IpCS3 structure bound to caffeine
615 and SAH.
616
617 **Acknowledgments**
618
619 This work was supported by Consejo Nacional de Investigaciones Científicas y
620 Técnicas de Argentina (CONICET); PRO.MATE.AR project, funded by Secretaría de Políticas
621 Universitarias del Ministerio de Educación de la Nación Argentina; CABANA project, funded
622 by UKRI-BBSRC on behalf of the Global Challenges Research Fund (BB/P027849/1), the U.S.
623 National Science Foundation (NSF) (grants MCB-1120624 and MCB-2325341 to T.J.B.); and
624 the European Molecular Biology Laboratory (EMBL). We would like to express our gratitude
625 to Centro de Cómputos de Alto Rendimiento (CeCAR) and UBA-FCEN-QB-Cluster for
626 providing access to computational resources, which facilitated the majority of computational
627 analyses in this work. Special thanks are extended to Kevin Blair and the Department of
628 Chemistry for facilitating our HPLC analyses.
629
630 **Contributions**
631
632 F.A.V., C.P.M., T.J.B. and A.G.T. designed the project and coordinated the research
633 activities. F.A.V., T.J.B. and A.G.T. wrote the manuscript with input from all co-authors. F.A.V.
634 was involved in all analyses. R.M.A. extracted the genomic DNA. E.J.S., G.F.B., R.R.M.O.,
635 G.L.N., R.M.A., and G.O. performed the assembly and annotation of the genome. F.A.V.
636 conducted the evolutionary and syntenic-gene analyses. A.H.G. and T.J.B. carried out the
637 expression, purification and activity assays of CS enzymes. L.A.D made the AlphaFold2
638 models. L.A.D. and C.P.M. performed the molecular dockings. A.H.G. and S.N. obtained the
639 crystal structure of IpCS3-CF complex. T.J.B., M.R., and P.D.Z. performed the comparative
640 genomic analyses of caffeine-producing enzymes across angiosperms. P.D.Z., D.A.M.,

P.A.S., G.O., S.N., T.J.B., and A.G.T. obtained funding to carry out the project. All authors read and approved the final article.

Competing interests

The authors declare no competing interests.

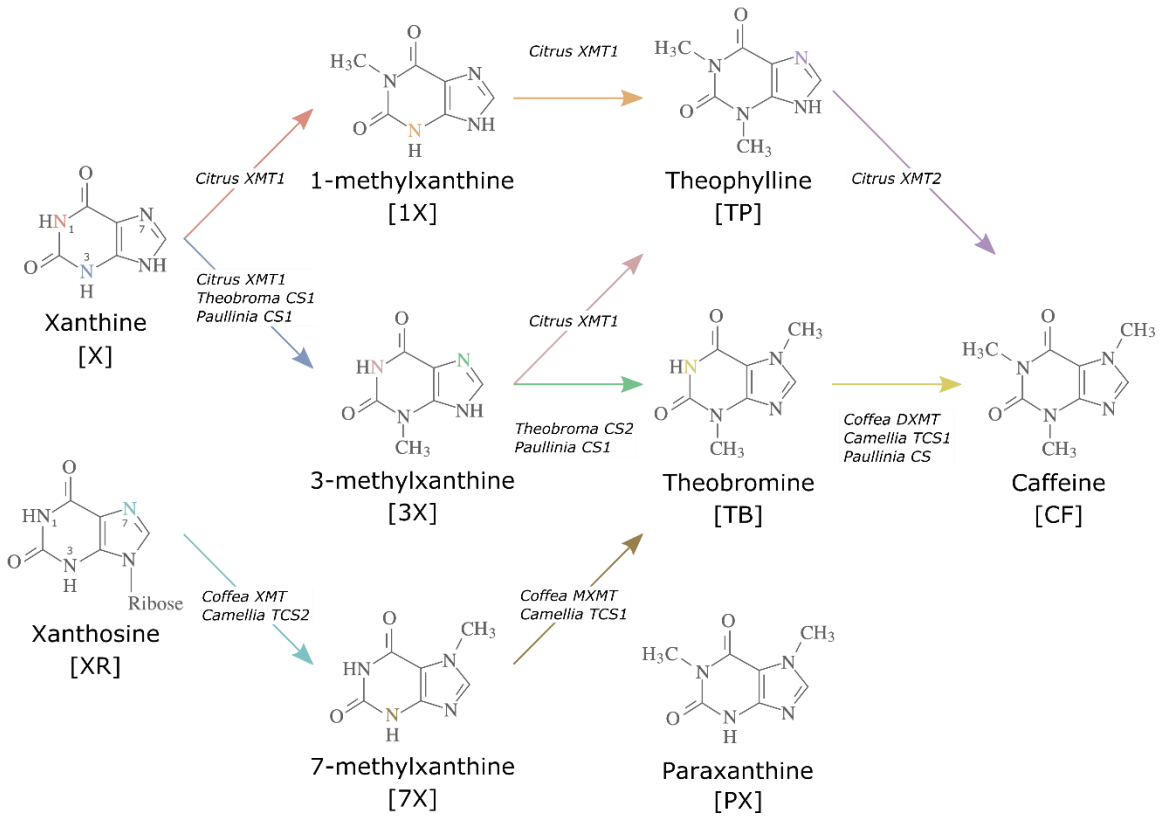


Fig. 1. Biosynthetic routes to caffeine within the xanthine alkaloid network. CF, caffeine; PX, paraxanthine; TB, theobromine; TP, theophylline; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; XR, xanthosine; X, xanthine. Nitrogen atoms are coloured to match the arrows corresponding to the enzymes that methylate them. Adapted with permission from O'Donnell et al.[24]

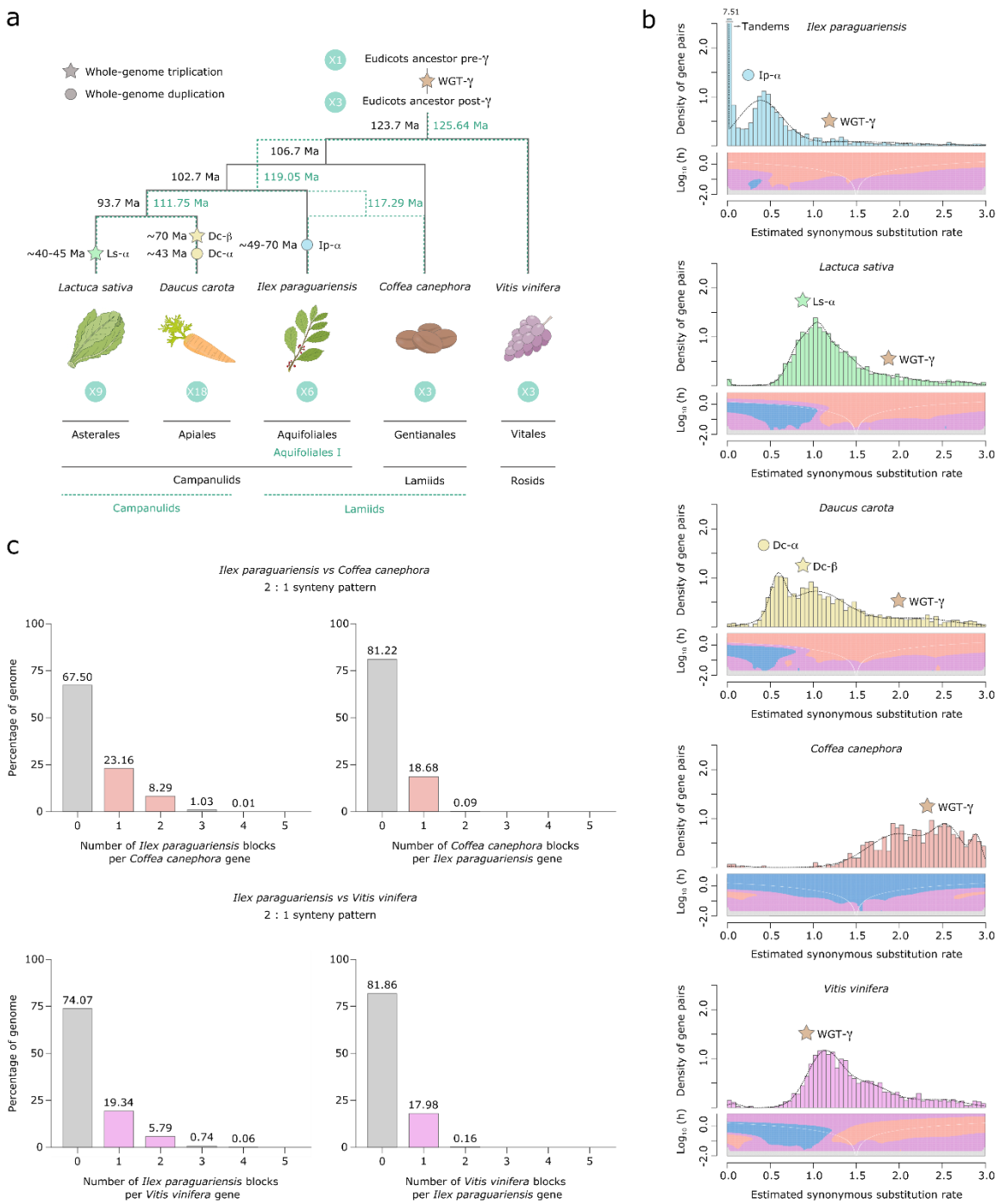


Fig. 2. Yerba mate genome duplication history. **a**, Evolutionary scenario of the eudicot genomes of *Lactuca sativa*, *Daucus carota*, *Ilex paraguariensis*, *Coffea canephora* and *Vitis vinifera*, from their ancestor pre-γ. The plastid genome phylogeny is represented with solid black lines, while the multiple nuclear genome phylogeny is represented with green dashed lines. Paleopolyploidizations are shown with coloured dots (duplications) and stars (triplications). Divergence time estimates for the lineages, as well as age estimates for the *L.*

670 *sativa* and *D. carota* paleopolyploidizations were obtained from the literature[36,38,100,101]. Ma,
671 million years ago. **b**, K_s distributions with Gaussian mixture model and SiZer analyses of *I.*
672 *paraguariensis* (blue), *L. sativa* (green), *D. carota* (yellow), *C. canephora* (red) and *V. vinifera*
673 (purple) paralogues. SiZer maps below histograms identify significant peaks at corresponding
674 K_s values. Blue represents significant increases in slope, red indicates significant decreases,
675 purple represents no significant slope change, and grey indicates not enough data for the test.
676 **c**, Comparative genomic synteny analyses of *I. paraguariensis* with *C. canephora* and *V.*
677 *vinifera*.

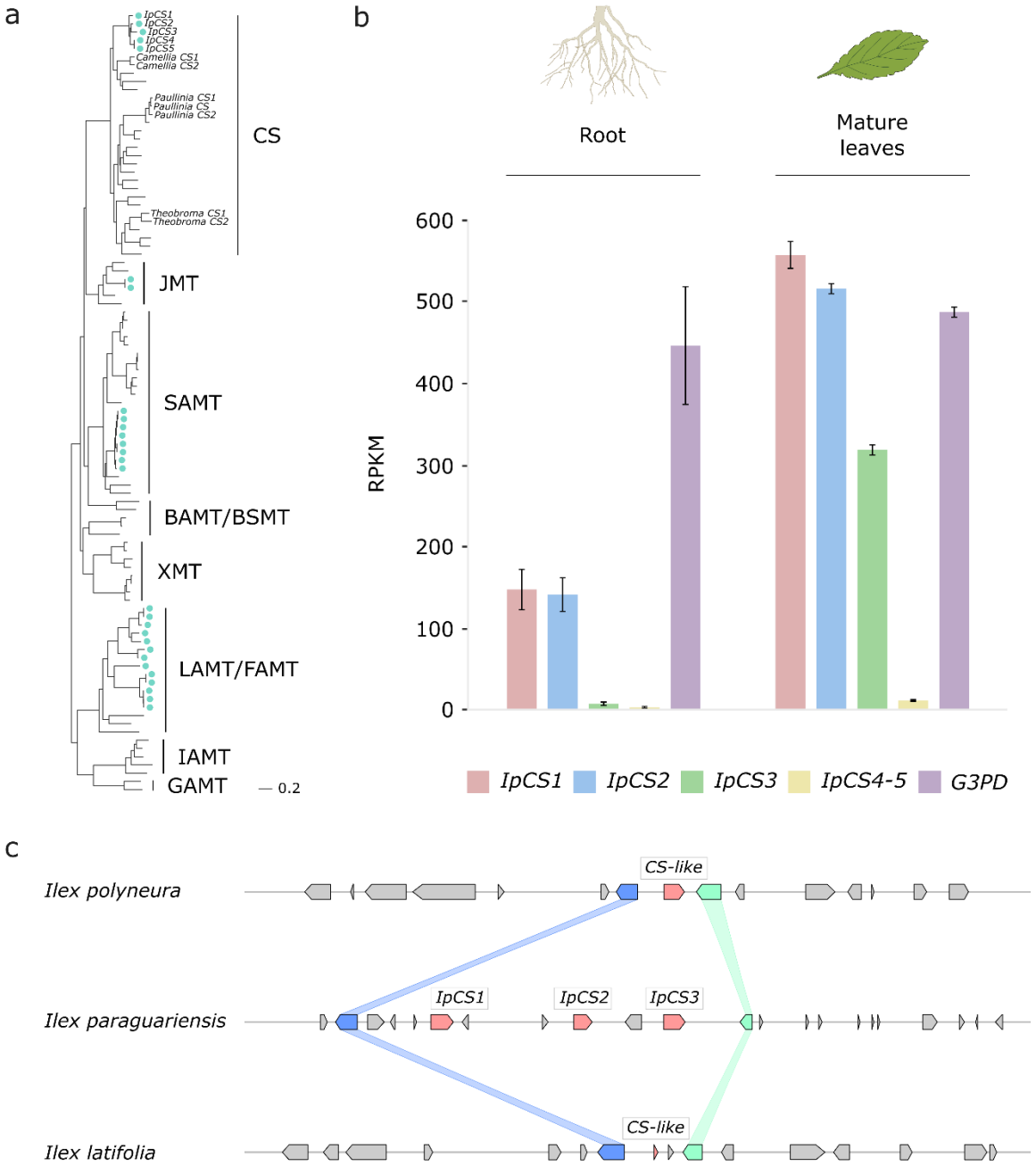


Fig. 3. The yerba mate genome encodes three recently duplicated CS-type SABATH proteins that are expressed in caffeine-producing tissues. a, SABATH gene tree estimate (LnL = -34265.473) shows the placement of full length YM proteins (marked by blue-green dots) within clades that have published functions. GAMT, gibberellin MT; IAMT, indole-3-acetic acid MT; LAMT/FAMT, loganic/farnesoic acid MT; BAMT/BSMT, benzoic/salicylic acid MT; XMT, xanthine alkaloid MT used for caffeine biosynthesis in *Coffea* and *Citrus*; SAMT, salicylic acid MT; JMT, jasmonic acid MT; CS, caffeine synthase in *Theobroma*, *Camellia* and *Paullinia*.

686 **b**, Gene expression analysis of IpCS1-5 as indicated by the relative abundance of YM
687 transcriptome reads mapped to the IpCS1-5 transcripts. RPKM, reads per kilobase per million
688 mapped reads. Error bars indicate standard deviation from the mean. Housekeeping gene:
689 *G3PD*, Glyceraldehyde-3-phosphate dehydrogenase. **c**, Synteny-based analysis of the CS
690 genomic region for *I. paraguariensis*, *I. polyneura* and *I. latifolia*.
691
692
693
694
695
696
697

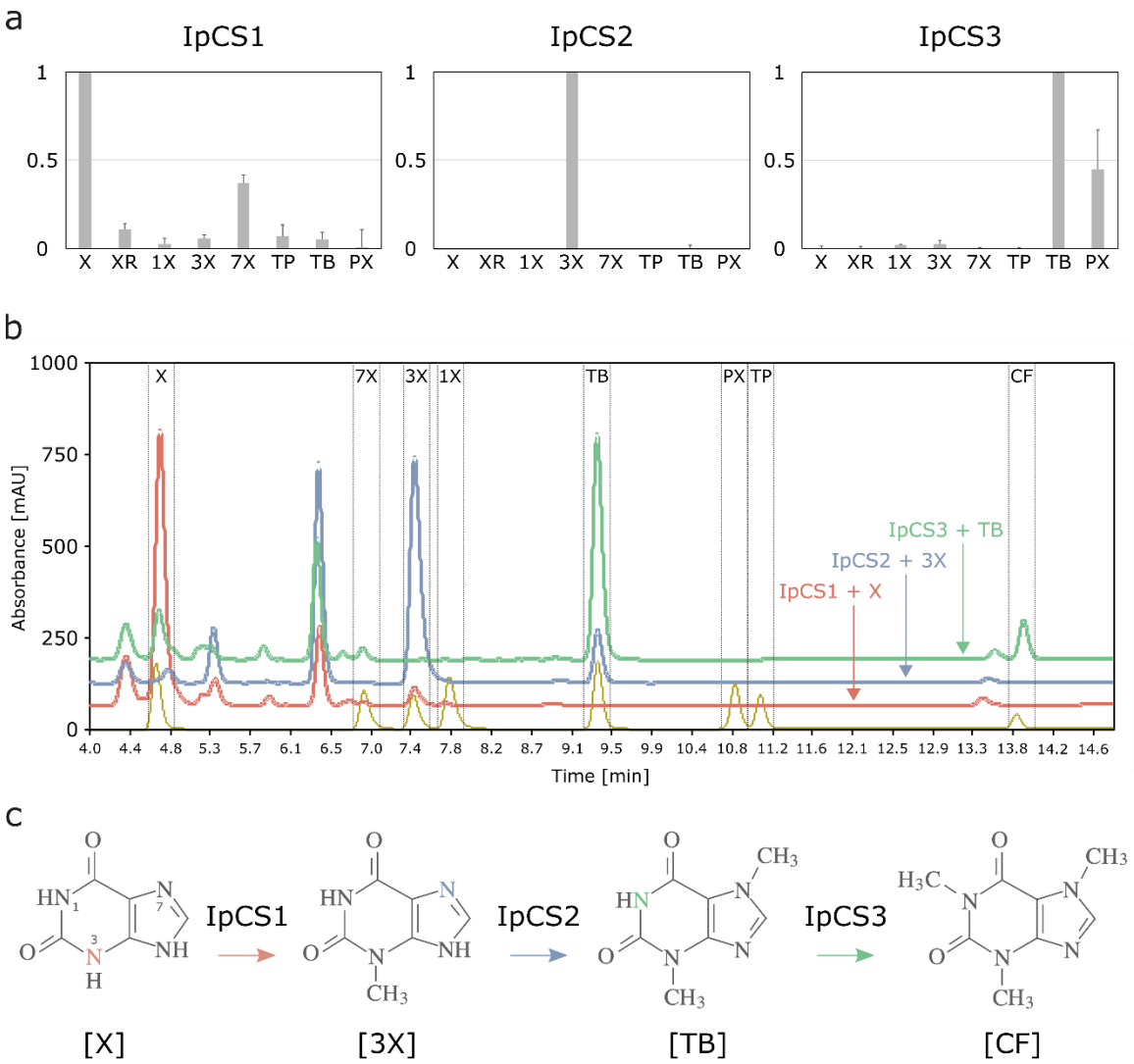


Fig. 4. SABATH enzymes have evolved to catalyse the biosynthesis of caffeine in yerba mate. a, Relative enzyme activities of three CS-type SABATH enzymes with eight xanthine alkaloid substrates. **b**, HPLC traces showing products formed by three encoded CS-type enzymes. **c**, Proposed biosynthetic pathway for caffeine in yerba mate. X, xanthine; XR, xanthosine; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; TP, theophylline; TB, theobromine; PX, paraxanthine. Coloured atoms and arrows indicate atoms that act as methyl acceptors for a given reaction.

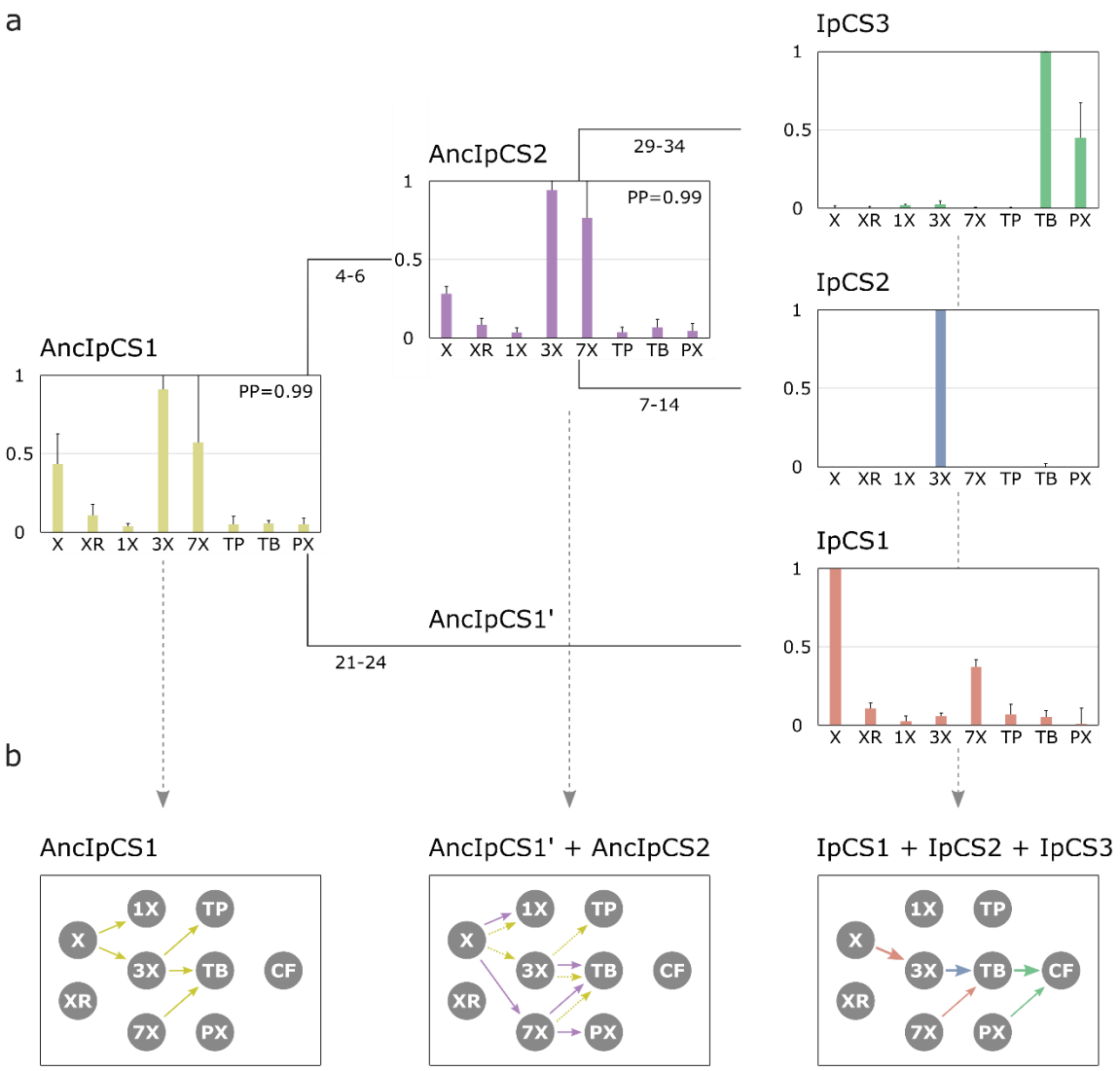
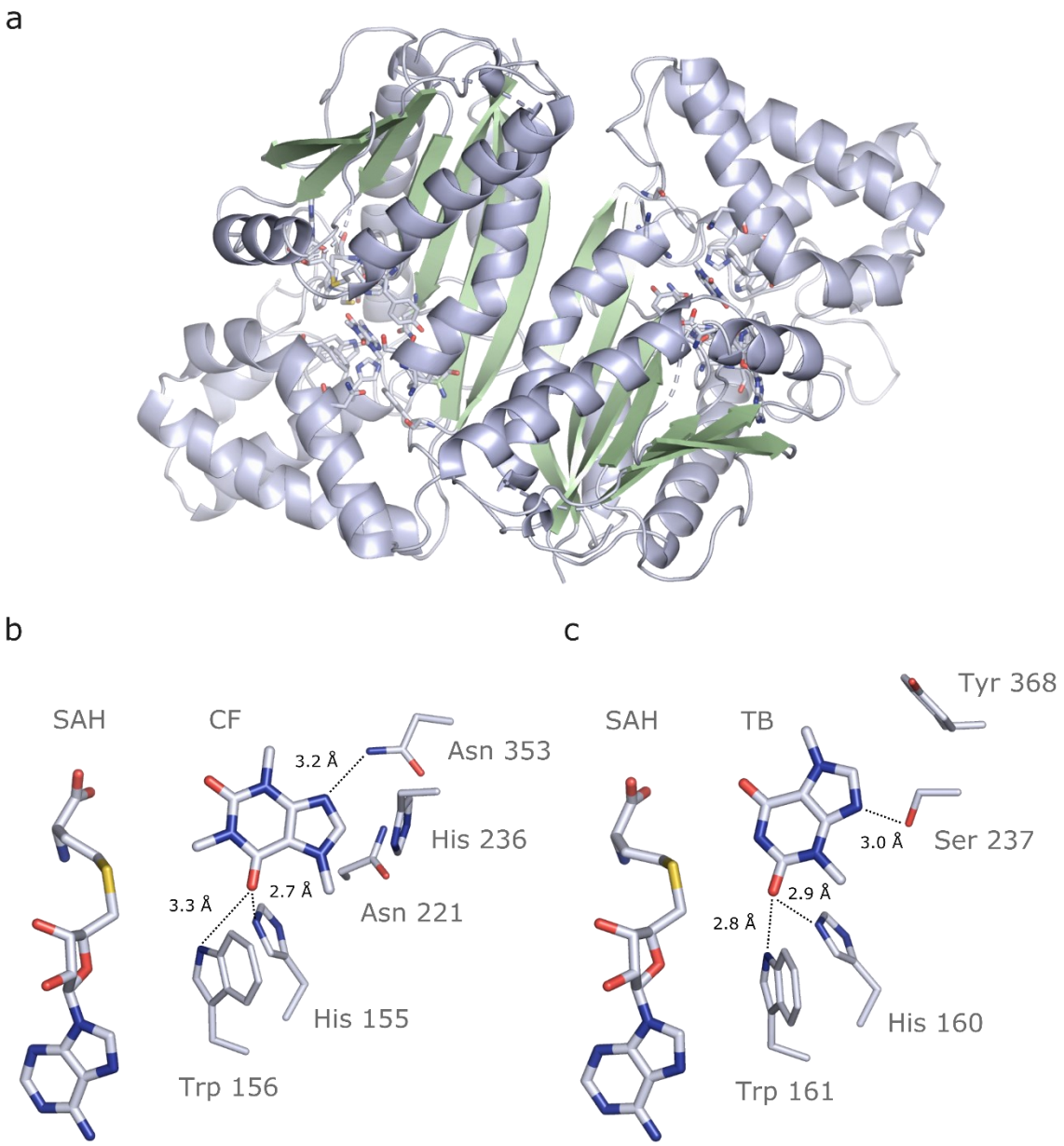


Fig. 5. Ancestral sequence resurrection reveals ancestral xanthine alkaloid pathway flux. a, Simplified evolutionary history of three YM xanthine alkaloid-methylating enzymes and their two ancestors, AncIpCS1 and AncIPCS2. Average site-specific posterior probabilities for each ancestral enzyme estimate are provided. Numbers below each branch of the phylogeny represents the number of amino acid replacements between each enzyme shown. These two ancestral relative activity charts show the averaged activities of two allelic variants of each enzyme. **b**, Inferred pathway flux is shown for the antecedent pathways that could have been catalysed by the ancestral or modern-day combinations of enzymes that would have existed at three time points in the history of the enzyme lineage. Arrows linking metabolites are coloured according to the activities detected from each enzyme shown in panel a. Dotted

717 arrows are shown for AncIpCS1' because it is unknown what characteristics it would possess;
718 it is assumed that it would have at least catalysed the formation of 3X from X since both its
719 ancestor and descendant enzyme do so. X, xanthine; XR, xanthosine; 1X, 1-methylxanthine;
720 3X, 3-methylxanthine; 7X, 7-methylxanthine; TP, theophylline; TB, theobromine; PX,
721 paraxanthine.
722
723
724
725



726

727 **Fig 6. Crystal structure of IpCS3 in complex with Caffeine (CF) and S-adenosyl-**
728 **homocysteine (SAH) and comparison with the active site of *Coffea* DXMT. a**, Overview
729 of the crystal structure of IpCS3 (PDB ID: 8UZD) depicting the active site of the enzyme in
730 complex with CF and SAH. **b**, Relevant residues in IpCS3 for ligand recognition. **c**, Relevant
731 residues in DMXT (PDB ID: 2EFJ) for ligand recognition. Protein residues are displayed as
732 lines with carbon atoms coloured in bluewhite while small molecules - caffeine (CF),
733 theobromine (TB), and S-adenosyl-homocysteine (SAH) - are drawn as sticks. Colour code

for the rest of the atoms: nitrogen (blue), oxygen (red) and sulphur (yellow). Hydrogen bond interactions are indicated as black dotted lines.

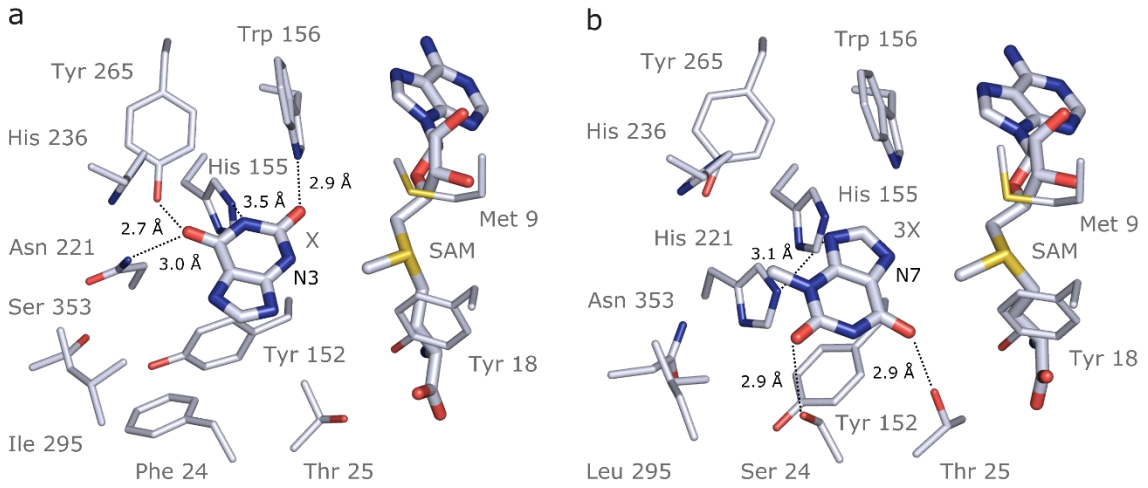


Fig. 7. Docking models of xanthine alkaloids in IpCS1 and IpCS2 active sites. **a**, IpCS1-X complex. **b**, IpCS2-3X complex. Protein residues are displayed as lines with carbon atoms coloured in bluewhite while small molecules - xanthine (X), 3-methylxanthine (3X), caffeine (CF), paraxanthine (PX), S-adenosyl-L-methionine (SAM) and S-adenosyl-homocysteine (SAH) - are drawn as sticks. Colour code for the rest of the atoms: nitrogen (blue), oxygen (red) and sulphur (yellow). Hydrogen bond interactions are indicated as black dotted lines.

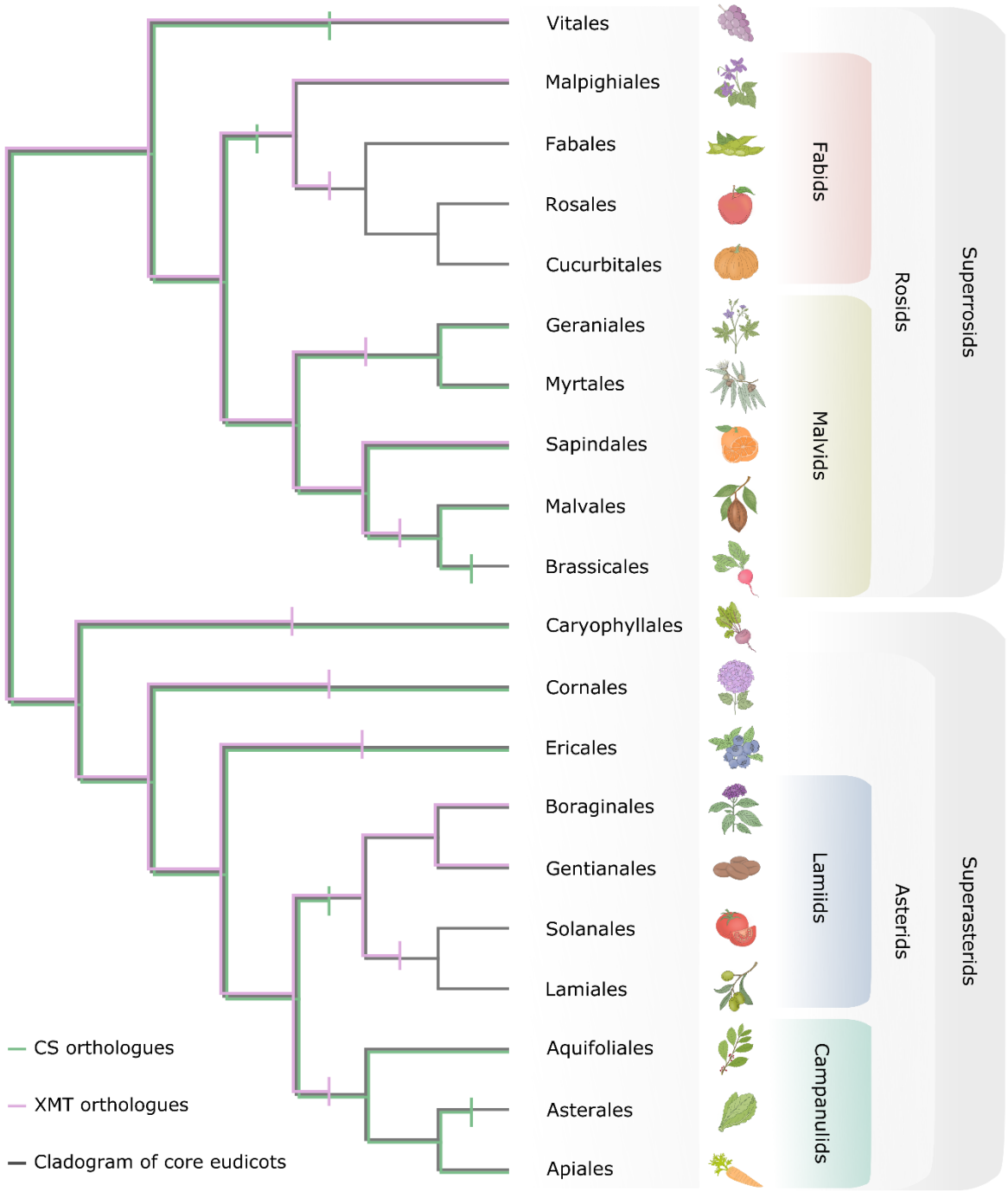


Fig. 8. Only CS genes were available for co-option and utilization for xanthine alkaloid biosynthesis in yerba mate whereas coffee only had XMT genes. Both CS- and XMT-type caffeine biosynthetic enzymes were present in the ancestor of core eudicots but numerous apparent losses of one or the other or both has occurred during lineage diversification. Gene loss is represented by vertical bar on relevant branches of the cladogram.

768 **References**

769 1. Niklas, C. O. Estudios embriológicos y citológicos en la yerba mate Ilex Paraguariensis
770 (Aquifoliaceae). *Bonplandia* **6**, 45 (1987).
771 2. Heck, C. I. & de Mejia, E. G. Yerba Mate tea (Ilex paraguariensis): a comprehensive
772 review on chemistry, health implications, and technological considerations. *Journal of*
773 *Food Science* vol. 72 R138–R151 Preprint at https://doi.org/10.1111/j.1750-
774 3841.2007.00535.x (2007).
775 3. Antioxidant effects of Ilex paraguariensis: induction of decreased oxidability of human
776 LDL in vivo. *Biochem. Biophys. Res. Commun.* **224**, 338–344 (1996).
777 4. Vieira, M. A. *et al.* Phenolic acids and methylxanthines composition and antioxidant
778 properties of mate (Ilex paraguariensis) residue. *Journal of Food Science* vol. 75 C280–
779 C285 Preprint at https://doi.org/10.1111/j.1750-3841.2010.01548.x (2010).
780 5. Boado, L. S., Fretes, R. M. & Brumovsky, L. A. Bioavailability and antioxidant effect of the
781 Ilex Paraguariensis polyphenols. *Nutrition & Food Science* vol. 45 326–335 Preprint at
782 https://doi.org/10.1108/nfs-08-2014-0079 (2015).
783 6. Kang, Y.-R. *et al.* Anti-obesity and anti-diabetic effects of Yerba Mate (Ilex paraguariensis)
784 in C57BL/6J mice fed a high-fat diet. *Laboratory Animal Research* vol. 28 23 Preprint at
785 https://doi.org/10.5625/lar.2012.28.1.23 (2012).
786 7. Ríos, J., Francini, F. & Schinella, G. Natural products for the treatment of type 2 diabetes
787 mellitus. *Planta Medica* vol. 81 975–994 Preprint at https://doi.org/10.1055/s-0035-
788 1546131 (2015).
789 8. Santos, E. C. S. *et al.* Anxiolytic-like, stimulant and neuroprotective effects of Ilex
790 paraguariensis extracts in mice. *Neuroscience* vol. 292 13–21 Preprint at
791 https://doi.org/10.1016/j.neuroscience.2015.02.004 (2015).
792 9. Yao, X., Lu, Z., Song, Y., Hu, X. & Corlett, R. T. A chromosome-scale genome assembly
793 for the holly (Ilex polyneura) provides insights into genomic adaptations to elevation in
794 Southwest China. *Hortic Res* **9**, (2022).

795 10. Kong, B. L.-H. *et al.* Chromosomal level genome of Ilex asprella and insight into antiviral
796 triterpenoid pathway. *Genomics* **114**, 110366 (2022).
797 11. Xu, K.-W. *et al.* The chromosome-level holly (Ilex latifolia) genome reveals key enzymes
798 in triterpenoid saponin biosynthesis and fruit color change. *Front. Plant Sci.* **13**, (2022).
799 12. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: new insights and
800 directions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190102 (2019).
801 13. Thorogood, C. J., Bauer, U. & Hiscock, S. J. Convergent and divergent evolution in
802 carnivorous pitcher plant traps. *New Phytol.* **217**, 1035–1041 (2018).
803 14. Wan, T. *et al.* A genome for gnetophytes and early evolution of seed plants. *Nat Plants* **4**,
804 82–89 (2018).
805 15. Yang, X. *et al.* The Kalanchoë genome provides insights into convergent evolution and
806 building blocks of crassulacean acid metabolism. *Nat. Commun.* **8**, 1899 (2017).
807 16. Pichersky, E. & Lewinsohn, E. Convergent evolution in plant specialized metabolism.
808 *Annu. Rev. Plant Biol.* **62**, 549–566 (2011).
809 17. Anaya, A. L., Cruz-Ortega, R. & Waller, G. R. Metabolism and ecology of purine alkaloids.
810 *Front. Biosci.* **11**, 2354–2370 (2006).
811 18. Stevenson, P. C., Nicolson, S. W. & Wright, G. A. Plant secondary metabolites in nectar:
812 impacts on pollinators and ecological functions. *Funct. Ecol.* **31**, 65–75 (2017).
813 19. Huang, R., O'Donnell, A. J., Barboline, J. J. & Barkman, T. J. Convergent evolution of
814 caffeine in plants by co-option of exapted ancestral enzymes. *Proc. Natl. Acad. Sci. U. S.*
815 *A.* **113**, 10613–10618 (2016).
816 20. Suzuki, T. & Takahashi, E. Caffeine biosynthesis in Camellia sinensis. *Phytochemistry*
817 **15**, 1235–1239 (1976).
818 21. Ashihara, H., Monteiro, A. M., Gillies, F. M. & Crozier, A. Biosynthesis of caffeine in leaves
819 of coffee. *Plant Physiol.* **111**, 747–753 (1996).
820 22. Kato, M. *et al.* Caffeine biosynthesis in young leaves of Camellia sinensis: In vitro studies
821 on N-methyltransferase activity involved in the conversion of xanthosine to caffeine.
822 *Physiol. Plant.* **98**, 629–636 (1996).

823 23. Uefuji, H., Ogita, S., Yamaguchi, Y., Koizumi, N. & Sano, H. Molecular cloning and
824 functional characterization of three distinct N-methyltransferases involved in the caffeine
825 biosynthetic pathway in coffee plants. *Plant Physiol.* **132**, 372–380 (2003).
826 24. O'Donnell, A. J., Huang, R., Barboline, J. J. & Barkman, T. J. Convergent biochemical
827 pathways for xanthine alkaloid production in plants evolved from ancestral enzymes with
828 different catalytic properties. *Mol. Biol. Evol.* **38**, 2704–2714 (2021).
829 25. Debat, H. J. *et al.* Exploring the genes of yerba mate (Ilex paraguariensis A. St.-Hil.) by
830 NGS and de novo transcriptome assembly. *PLoS One* **9**, e109835 (2014).
831 26. Fay, J. V. *et al.* Yerba mate (Ilex paraguariensis, A. St.-Hil.) de novo transcriptome
832 assembly based on tissue specific genomic expression profiles. *BMC Genomics* **19**, 891
833 (2018).
834 27. Acevedo, R. M. *et al.* Transcript and metabolic adjustments triggered by drought in Ilex
835 paraguariensis leaves. *Planta* **250**, 445–462 (2019).
836 28. Singh, R., Ming, R. & Yu, Q. Comparative analysis of GC content variations in plant
837 genomes. *Tropical Plant Biology* vol. 9 136–149 Preprint at
838 https://doi.org/10.1007/s12042-016-9165-4 (2016).
839 29. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M.-A. LTR-
840 retrotransposons in plants: Engines of evolution. *Gene* **626**, 14–25 (2017).
841 30. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC*
842 *Bioinformatics* **4**, 41 (2003).
843 31. Luo, H., Lin, Y., Gao, F., Zhang, C.-T. & Zhang, R. DEG 10, an update of the database of
844 essential genes that includes both protein-coding genes and noncoding genomic
845 elements. *Nucleic Acids Res.* **42**, D574–80 (2014).
846 32. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update:
847 novel and streamlined workflows along with broader and deeper phylogenetic coverage
848 for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654
849 (2021).
850 33. Gottlieb, A. M. & Poggio, L. Quantitative and qualitative genomic characterization of

851 cultivated Ilex L. species. *Plant Genet. Resour.* **13**, 142–152 (2015).
852 34. Sankoff, D. & Zheng, C. Whole genome duplication in plants: implications for evolutionary
853 analysis. *Methods Mol. Biol.* **1704**, 291–315 (2018).
854 35. Landis, J. B. *et al.* Impact of whole-genome duplication events on diversification rates in
855 angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
856 36. Zhang, C. *et al.* Asterid phylogenomics/phylotranscriptomics uncover morphological
857 evolutionary histories and support phylogenetic placement for numerous whole-genome
858 duplications. *Mol. Biol. Evol.* **37**, 3188–3210 (2020).
859 37. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and
860 the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
861 38. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A
862 metacalibrated time-tree documents the early rise of flowering plant phylogenetic
863 diversity. *New Phytol.* **207**, 437–453 (2015).
864 39. Yao, X., Song, Y., Yang, J.-B., Tan, Y.-H. & Corlett, R. T. Phylogeny and biogeography
865 of the hollies (Ilex L., Aquifoliaceae). *J. Syst. Evol.* **59**, 73–82 (2021).
866 40. Ross, J. R., Nam, K. H., D'Auria, J. C. & Pichersky, E. S-Adenosyl-L-methionine:salicylic
867 acid carboxyl methyltransferase, an enzyme involved in floral scent production and plant
868 defense, represents a new class of plant methyltransferases. *Arch. Biochem. Biophys.*
869 **367**, 9–16 (1999).
870 41. Seo, H. S. *et al.* Jasmonic acid carboxyl methyltransferase: a key enzyme for jasmonate-
871 regulated plant responses. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4788–4793 (2001).
872 42. Yin, Y., Katahira, R. & Ashihara, H. Metabolism of purine alkaloids and xanthine in leaves
873 of maté (Ilex paraguariensis). *Nat. Prod. Commun.* **10**, 707–712 (2015).
874 43. Yang, L. *et al.* Phylogenomic Insights into deep phylogeny of angiosperms based on
875 broad nuclear gene sampling. *Plant Commun* **1**, 100027 (2020).
876 44. Negrin, A., Long, C., Motley, T. J. & Kennelly, E. J. LC-MS metabolomics and
877 chemotaxonomy of caffeine-containing holly (Ilex) species and related taxa in the
878 Aquifoliaceae. *J. Agric. Food Chem.* **67**, 5687–5699 (2019).

879 45. Noda-Garcia, L., Liebermeister, W. & Tawfik, D. S. Metabolite-enzyme coevolution: from
880 single enzymes to metabolic pathways and networks. *Annu. Rev. Biochem.* **87**, 187–216
881 (2018).
882 46. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules.
883 *Nat. Rev. Genet.* **5**, 366–375 (2004).
884 47. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the
885 functional synthesis. *Nat. Rev. Genet.* **8**, 675–688 (2007).
886 48. Daly, J. W., Butts-Lamb, P. & Padgett, W. Subclasses of adenosine receptors in the
887 central nervous system: interaction with caffeine and related methylxanthines. *Cell. Mol.*
888 *Neurobiol.* **3**, 69–80 (1983).
889 49. McCarthy, A. A. & McCarthy, J. G. The structure of two N-methyltransferases from the
890 caffeine biosynthetic pathway. *Plant Physiol.* **144**, 879–889 (2007).
891 50. Zubieta, C. *et al.* Structural basis for substrate recognition in the salicylic acid carboxyl
892 methyltransferase family. *Plant Cell* **15**, 1704–1716 (2003).
893 51. Petronikolou, N., Hollatz, A. J., Schuler, M. A. & Nair, S. K. Loganic acid
894 methyltransferase: insights into the specificity of methylation on an iridoid glycoside.
895 *Chembiochem* **19**, 784–788 (2018).
896 52. Zhao, N. *et al.* Structural, biochemical, and phylogenetic analyses suggest that indole-3-
897 acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family.
898 *Plant Physiol.* **146**, 455–467 (2008).
899 53. Lanzarotti, E., Defelipe, L. A., Marti, M. A. & Turjanski, A. G. Aromatic clusters in protein–
900 protein and protein–drug complexes. *J. Cheminform.* **12**, 1–9 (2020).
901 54. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**,
902 679–682 (2022).
903 55. Dubs, N. M. *et al.* A collaborative classroom investigation of the evolution of SABATH
904 methyltransferase substrate preference shifts over 120 my of flowering plant history. *Mol.*
905 *Biol. Evol.* **39**, (2022).
906 56. Kim, Y.-S., Uefuji, H., Ogita, S. & Sano, H. Transgenic tobacco plants producing caffeine:

907 a potential new strategy for insect pest control. *Transgenic Res.* **15**, 667–672 (2006).
908 57. Kim, Y.-S. *et al.* Resistance against beet armyworms and cotton aphids in caffeine-
909 producing transgenic chrysanthemum. *Plant Biotechnol.* **28**, 393–395 (2011).
910 58. Niemenak, N., Onomo, P. E., Fotso, Lieberei, R. & Ndoumou, D. O. Purine alkaloids and
911 phenolic compounds in three Cola species and Garcinia kola grown in Cameroon. *S. Afr.*
912 *J. Bot.* **74**, 629–638 (2008).
913 59. Avico, E. H. *et al.* Integrating Transcriptional, Metabolic, and Physiological Responses to
914 Drought Stress in Roots. *Plants* **12**, (2023).
915 60. Tarragó, J., Filip, R., Mroginski, L. & Sansberro, P. Influence of the irradiance on phenols
916 content and rooting of Ilex paraguariensis cuttings collected from adult plants. *Acta*
917 *Physiol. Plant* **34**, 2419–2424 (2012).
918 61. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
919 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
920 62. Kelley, D. R., Schatz, M. C. & Salzberg, S. L. Quake: quality-aware detection and
921 correction of sequencing errors. *Genome Biol.* **11**, R116 (2010).
922 63. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de
923 novo assembler. *Gigascience* **1**, 18 (2012).
924 64. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination
925 from genomic and metagenomic datasets. *PLoS One* **6**, e17288 (2011).
926 65. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive -mer
927 weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
928 66. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig
929 reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460
930 (2018).
931 67. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and
932 accurate *de novo* assembly of metazoan genomes with modest long read coverage.
933 Preprint at https://doi.org/10.1101/029306.
934 68. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-

935 assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
936 69. Palmer, J. & Stajich, J. *nextgenusfs/funannotate: funannotate v1.5.3*. (2019).
937 doi:10.5281/zenodo.2604804.
938 70. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches.
939 *Bioinformatics* **29**, 2933–2935 (2013).
940 71. Chan, P. P. & Lowe, T. M. tRNAscan-SE: searching for tRNA genes in genomic
941 sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
942 72. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
943 *Nucleic Acids Res.* **35**, 3100–3108 (2007).
944 73. Bonnet, E., He, Y., Billiau, K. & Van de Peer, Y. TAPIR, a web server for the prediction of
945 plant microRNA targets, including target mimics. *Bioinformatics* **26**, 1566–1568 (2010).
946 74. Fahlgren, N. & Carrington, J. C. miRNA target prediction in plants. *Methods in Molecular*
947 *Biology* 51–57 Preprint at https://doi.org/10.1007/978-1-60327-005-2_4 (2010).
948 75. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
949 *Bioinformatics* **30**, 1236–1240 (2014).
950 76. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
951 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
952 *Res.* **47**, D309–D314 (2019).
953 77. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable
954 element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
955 78. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**,
956 D81–9 (2016).
957 79. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet.*
958 *Genome Res.* **110**, 462–467 (2005).
959 80. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. Mclust 5: Clustering, classification
960 and density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016).
961 81. Chaudhuri, P. & Marron, J. S. Scale space view of curve estimation. *Ann. Stat.* **28**, 408–
962 428 (2000).

963 82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
964 transform. *Bioinformatics* **25**, 1754–1760 (2009).
965 83. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
966 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
967 84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
968 trees for large alignments. *PLoS One* **5**, e9490 (2010).
969 85. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online
970 phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–5
971 (2016).
972 86. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**,
973 993–1005 (1995).
974 87. Ran, J.-H., Shen, T.-T., Wang, M.-M. & Wang, X.-Q. Phylogenomics resolves the deep
975 phylogeny of seed plants and indicates partial convergent or homoplastic evolution
976 between Gnetales and angiosperms. *Proc. Biol. Sci.* **285**, (2018).
977 88. Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta*
978 *Crystallogr. D Biol. Crystallogr.* **67**, 293–302 (2011).
979 89. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular
980 structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
981 90. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal
982 structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).
983 91. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot.
984 *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
985 92. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids*
986 *Res.* **47**, D930–D940 (2019).
987 93. Swain, M. chemicalize.org. *J. Chem. Inf. Model.* **52**, 613–615 (2012).
988 94. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.*
989 **14**, 33–8, 27–8 (1996).
990 95. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective

991 receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
992 96. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: new
993 docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**,
994 3891–3898 (2021).
995 97. Schrodinger, L. L. C. The PyMOL molecular graphics system. *Version 1.8* (2015).
996 98. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
997 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
998 99. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics.
999 *Nucleic Acids Res.* **40**, D1178–86 (2012).
1000 100. Iorizzo, M. *et al.* A high-quality carrot genome assembly provides new insights into
1001 carotenoid accumulation and asterid genome evolution. *Nat. Genet.* **48**, 657–666 (2016).
1002 101. Reyes-Chin-Wo, S. *et al.* Genome assembly with in vitro proximity ligation data and
1003 whole-genome triplication in lettuce. *Nat. Commun.* **8**, 14953 (2017).

1004