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[8,9], 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, Hamburg 22607, 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

[10] Department of Biological Sciences, Western Michigan University, Kalamazoo, MI 49008, USA

[11] Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[12] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

#Equal contributions

*Correspondence to Adrian G. Turjanski (adrian@qb.fcen.uba.ar) or Todd J. Barkman (todd.barkman@wmich.edu)

Abstract

Yerba mate (*Ilex paraguariensis*) is an economically important crop marketed for the elaboration of mate, the third-most widely consumed caffeine-containing infusion worldwide. Here we report the first genome assembly of this species, which has a total length of 1.06 Gb and contains 53,390 protein-coding genes. Comparative analyses revealed that the large yerba mate genome size is partly due to a whole-genome duplication (Ip-α) during the early evolutionary history of *Ilex*, in addition to the hexaploidization event (γ) shared by core eudicots. Characterization of the genome allowed us to clone the genes encoding methyltransferase enzymes that catalyse multiple reactions required for caffeine production. To our surprise, this species has converged upon a different biochemical pathway compared to that of its relatives, coffee and tea. In order to gain insight into the structural basis for the convergent enzyme activities, we obtained a crystal structure for the terminal enzyme in the

pathway that forms caffeine. The structure reveals that convergent solutions have evolved for substrate positioning because different amino acid residues facilitate a different substrate orientation such that efficient methylation occurs in the independently evolved enzymes in yerba mate and coffee. While our results show phylogenomic constraint limits the genes coopted for convergence of caffeine biosynthesis, the x-ray diffraction data suggests structural constraints are minimal for the convergent evolution of individual reactions.

Introduction

In the genomic era, hundreds of plant species have had their nucleotide sequences determined to provide unprecedented insight into the genetic basis of many traits. One of the few species of economic importance for which no genomic data exists is *Ilex paraguariensis* var. *paraguariensis* A. St. Hilaire (Aquifoliaceae), colloquially known as yerba mate (YM), which is a caffeinated diploid tree-species ($2n=2x=40$) endemic to the subtropical rainforests of South America[1]. The dried leaves and twigs of this dioecious evergreen are used to prepare a traditional infusion named mate, or chimarrão, widely consumed around the world. Approximately 300,000 ha are cultivated with this tree crop, with Argentina responsible for 80% of world-wide production[2]. The mate infusion has been shown to have numerous beneficial effects in humans including as an antioxidant[3,4,5], antidiabetic[6,7], as well as central nervous system stimulant[8], among others. Several bioactive compounds have been identified in YM that might be responsible for its effects, including terpenes, flavonoids, phenolics and methylxanthines[2]. Although its stimulant properties are best known and mostly related to caffeine content, little is known about the genetic and biochemical mechanisms of how YM synthesizes this, or any, of its important metabolites. Despite the recent release of three other *Ilex* genome sequences[9–11], none of the species produce caffeine, making the genetic basis 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 **Results and Discussion**
111
112 **Yerba mate genome sequencing, assembly and annotation**
113
114 The YM genome was sequenced combining Illumina and PacBio sequencing
115 technologies. With Illumina sequencing, we generated ~263.2 Gb of short reads from various
116 DNA fragment sizes (350 bp, 550 bp, 3 kbp, 8 kbp and 12 kbp), while with PacBio sequencing,
117 we generated ~77.5 Gb of long reads. These reads represent ~157.5- and ~49.3-fold base-
118 pair coverage of the genome, respectively (Supplementary Table 1). The total assembly length
119 was ~1.06 Gb and consisted of 10,611 scaffolds (≥1 kb) with an N50 length of ~510.8 Kb
120 (Supplementary Table 2). To assess the completeness of the assembly we aligned the
121 available YM transcriptome reads[25,26,27] and the YM genome short reads generated in this
122 study with the genome assembly, establishing that 99.3% of the YM transcriptome reads and
123 99.5% of the YM genome short reads mapped to the assembly. This result suggests that we
124 have assembled almost all available sequence data. The GC content of the genome assembly
125 was 36.33% (Supplementary Table 2), similar to that of other eudicots (33.70-38.20 GC%)[28]
126 and almost identical to that of *Ilex polyneura* (36.08 GC%)[9], *Ilex asprella* (36.25 GC%)[10], and
127 *Ilex latifolia* (36.44 GC%)[11], the only three *Ilex* species with sequenced genomes. About
128 64.63% of the genome assembly was composed of repetitive sequences, of which ~36.22%
129 were retrotransposons, ~1.80% were DNA transposons, ~0.74% were simple repeats and

130 ~0.15% were low complexity regions. Long terminal-repeat (LTR) retrotransposons of the
131 Gypsy and Copia families were the most abundant transposable elements, as observed in
132 many sequenced plant genomes[29], followed by long interspersed nuclear elements (LINEs)
133 and hobo-Activator transposons, among others (Supplementary Table 3).

134 A total of 53,390 protein-coding genes were predicted in the genome, with a mean
135 coding sequence length of 3,062 bp and 4.23 exons per gene. Of these, 41,483 (~77.63%)
136 could be annotated with GO terms, EC numbers or Pfam domains. In addition, we identified
137 4,530 non-coding RNA genes, including 2,670 small nucleolar RNAs, 815 transfer RNAs, 471
138 ribosomal RNAs, 348 small nuclear RNAs, and 226 micro RNAs (Supplementary Notes S1
139 and Supplementary Tables 4-6). To further assess the completeness of the assembly, we
140 aligned the scaffolds with the KOG[30] and DEG[31] databases, determining that 98% of the core
141 gene families from the KOG database and 97.5% of the *Arabidopsis thaliana* DEG subset
142 were present in the assembly. Then, we performed a Benchmarking Universal Single-Copy
143 Orthologs (BUSCO)[32] assessment using the eudicot ODB10 database. Among 2,326
144 conserved single-copy genes, ~77.30% (1,798) of the gene set was completely retrieved,
145 ~12.12% (282) was partially retrieved and only ~10.57% (100) was missing. These results
146 suggest that the coding region of the assembly is nearly complete. The number of estimated
147 genes for YM is higher than the ca. 39,000 reported from the genome sequences of other *Ilex*
148 species[9–11]. This could be at least partly due to the larger genome size of YM estimated from
149 flow cytometry relative to the other species[33].

150

151 **Evolutionary analysis of yerba mate genome provides evidence of whole-genome**
152 **duplication in an early *Ilex* ancestor**

153

154 Most plant lineages have experienced ancient polyploidization events followed by
155 massive duplicate gene losses and genome rearrangements, which may have contributed to

156 the evolution of developmental and metabolic complexity[34,35]. Recent transcriptome-based
157 analyses[36,37] reported an ancient polyploidization event in the *Ilex* lineage around 60 Ma
158 (Cretaceous–Paleogene boundary), based on phylogenomic and synonymous substitution
159 rate (K_s) evidence. Evolutionary analyses of *Ilex polyneura*[9] and *Ilex latifolia*[11] genomes also
160 provided evidence of a shared *Ilex*-specific whole genome duplication (WGD). As YM is the
161 first American holly to have its genome sequenced, we performed synteny-based analyses of
162 its genome to deepen our understanding of Aquifoliales evolution (Fig. 2 and Supplementary
163 Fig. 1). The K_s distribution of YM paralogues (Fig. 2b) revealed a significant peak with a
164 median K_s value of ~0.37, not shared with the rest of the eudicot genomes analysed (Fig. 2b
165 and Supplementary Fig. 1). This confirms the lineage-specific polyploidization event (Ip-α)
166 previously reported in *Ilex*[36,37], in addition to the shared ancestral WGT-γ which is indicated by
167 a median K_s value of ~1.4 (Fig. 2b). A WGD in the common ancestor of *Ilex* species is further
168 supported by 2:1 syntenic depth ratios between the YM genome and the coffee and grape
169 genomes, which did not experience additional duplication events after the ancestral WGT-γ
170 (Fig. 2c). In order to determine the age of Ip-α, we used two different phylogenies (Fig. 2a).
171 The plastid genome phylogeny supports the monophyly of Aquifoliales as the first diverging
172 clade of campanulids[38]; the alternative nuclear genome phylogeny supports *Ilex* in Aquifoliales
173 I as an early branching lineage of lamiids[36]. With the former phylogeny, we estimated the age
174 of the WGD event between 48.75-69.63 Ma while, with the latter, divergence was estimated
175 at 49.43-70.62 Ma (Fig. 2a). Both estimates are consistent with that of Zhang et al. 2020 and
176 validate the age of Ip-α near the origin of *Ilex*, which is estimated between 43-89 Ma[39].
177
178 **Convergent evolution of caffeine biosynthesis in yerba mate**
179
180 In order to determine the genes and biochemical pathway responsible for caffeine
181 biosynthesis in YM, we used bioinformatic analyses to identify SABATH enzyme family

182 members in the genome[19,22,23]. There appear to be 28 full-length SABATH genes in YM that
183 encode members of the functionally diverse clades of the family, including SAMT[40] and JMT[41],
184 among others (Fig. 3a). Our phylogenetic analysis showed that although the YM genome does
185 not appear to encode XMT-type caffeine-producing enzymes like *Coffea* and *Citrus*, it does
186 contain three recently and tandemly duplicated genes that encode CS-type enzymes, IpCS1-
187 3 (Fig. 3a, c). The duplicated IpCS1-3 are 86-91% identical at the amino acid level and are
188 expressed at highest levels in caffeine-accumulating tissue (Fig. 3b)[19]. IpCS1-3 also appear
189 to be of recent origin, since non-caffeine accumulating *Ilex* species only have a single gene or
190 gene fragment in the syntenic region (Fig. 3c). In *Camellia*, *Theobroma* and *Paullinia*, recent
191 duplications of the CS-type enzymes responsible for the successive steps of xanthine alkaloid
192 methylation have also independently occurred (Fig. 3a)[19,24]. Two other YM genes encode
193 IpCS4 & 5, but these are not syntenic with IpCS1-3 and are not highly expressed in any tissues
194 studied (Fig. 3b); therefore, we did not characterize them further.

195 To investigate the biochemical activities of the enzymes encoded by the three CS-type
196 genes, we cloned them into bacterial expression vectors and determined heterologous protein
197 functions. One enzyme, IpCS1, appears to primarily methylate X to catalyse the formation of
198 3X (Fig. 4). A second enzyme, IpCS2, shows activity only with 3X to produce TB, while a third
199 enzyme, IpCS3, exhibits a preference to methylate TB to form CF (Fig. 4). Thus, collectively,
200 these three enzymes appear capable of catalysing a complete pathway from xanthine to
201 caffeine. The apparent K_M for the preferred substrates of all three enzymes ranges from 85-
202 197 μM, and the k_{cat} / K_M estimates are comparable to those determined for other caffeine
203 biosynthetic enzymes[24] (Supplementary Table 7). Further evidence for this biosynthetic
204 pathway has been reported by ^{14}C xanthine tracer studies in young leaf segments of *I.*
205 *paraguariensis* that showed radioactivity in 3X and TB in addition to CF[42]. A pathway from
206 X→3X→TB→CF has also been reported for *Theobroma* and *Paullinia* using CS-type SABATH
207 enzymes[19]. Like Huang et al. 2016, this represents another departure from the long-assumed

208 pathway to caffeine biosynthesis (XR→7X→TB→CF) as reported in coffee and tea (Fig. 1).
209 This instance in *Ilex* is particularly notable since YM is an Asterid, like coffee and tea. The fact
210 that *Ilex*, *Theobroma* and *Paullinia* convergently recruited CS genes that independently
211 duplicated and evolved to encode enzymes with similar substrate preferences to catalyse a
212 common pathway to caffeine, in spite of their divergence more than 100 Ma[43], is remarkable
213 and suggests a high degree of genetic constraint governing the repeated origin of this trait.
214 While the substrate preferences shown in Fig. 4 suggest pathway flux from
215 X→3X→TB→CF, IpCS1 also shows secondary activity with 7X to produce TB and IpCS3 can
216 catalyse the formation of CF from paraxanthine (PX) (Fig. 4a). Thus, flux through other
217 branches of the xanthine alkaloid biosynthetic network (Fig. 1) cannot be excluded. However,
218 it is not clear how 7X or PX would be produced *in planta* since none of the three enzymes
219 studied here is capable of their formation; therefore, these secondary activities may not be
220 physiologically relevant. In addition, it has been proposed that TP may also be a precursor to
221 caffeine biosynthesis in *I. paraguariensis* based on radioisotopic feeding studies[42], although
222 its levels in plant tissues are 30-160 times lower than TB[44]. Our *in vitro* enzyme assays provide
223 no experimental evidence for that biosynthetic route; however, it is possible that additional MT
224 enzymes from the SABATH (or other) gene family not characterized in this study may perform
225 such reactions. Alternatively, if the exogenously supplied TP was first catabolized to 3X in YM
226 tissues, then the caffeine detected previously[42] could have been synthesized via the route
227 described above for IpCS2+IpCS3 (Fig. 4).
228
229 **The caffeine biosynthetic pathway in YM evolved from ancestral networks with different**
230 **inferred flux**
231
232 The fact that caffeine is produced within only one small lineage of *Ilex* that diverged
233 and experienced CS gene duplication (Fig. 3) within the last 11 million years[39,44] indicates that

234 the pathway has only recently evolved. The nature by which novel multistep biochemical
235 pathways evolve is a central question in biology[45]. To investigate the caffeine pathway origin
236 in YM, we used Ancestral Sequence Resurrection[46,47] to study AncIpCS1 & AncIpCS2, the
237 ancestors of the three modern-day enzymes implicated in caffeine biosynthesis in YM (Fig. 5
238 and Supplementary Fig. 2). The ancestral enzyme, AncIpCS1, which gave rise to all three
239 modern-day YM enzymes, exhibits highest relative activity with X, 3X and 7X (Fig. 5a).
240 Methylation of 7X by AncIpCS1 occurred at the N3 position resulting in TB synthesis, whereas
241 xanthine methylation occurred at either the N1 or N3 position to form 1X and 3X, respectively
242 (Fig. 5b and Supplementary Fig. 3). AncIpCS1 was capable of methylation of 3X at N1 to
243 produce TP, while methylation at the N7 position led to TB formation (Fig. 5b). These data
244 demonstrate that, although AncIpCS1 could not produce caffeine, it could methylate xanthine
245 alkaloids at 3 different positions of the planar heterocyclic ring structures and this combination
246 of activities would have allowed for the ancestor of YM to produce a cocktail of 1X, 3X, TP and
247 TB by flux through multiple branches of the xanthine alkaloid biosynthetic network with a single
248 enzyme (Fig. 5b).

249 After gene duplication of AncIpCS1, one daughter enzyme ultimately gave rise to
250 IpCS1, which exhibits preference to methylate xanthine to produce 3X (Fig. 5a). The other
251 daughter enzyme, AncIpCS2, appears to have maintained highest activity with X, 3X and 7X
252 like AncIpCS1 (Fig. 5a). However, unlike its ancestor, AncIpCS1, AncIpCS2 evolved high
253 relative activity with 7X to produce not just TB, but also PX by methylation at the N1 position
254 (Fig. 5a and Supplementary Fig. 3b). AncIpCS2 retained the ancestral activity of AncIpCS1
255 with xanthine to produce 1X, but also evolved the ability to methylate X at the N7 position (Fig.
256 5b and Supplementary Fig. 3). This enzyme also retained ancestral activity with 3X to produce
257 only TB by N7 methylation but lost the ability to methylate 3X at the N1 position to form TP.
258 These activities of AncIpCS2 would have allowed for ancestral flux to produce 1X, 7X, TB and
259 PX but not caffeine. Because a YM ancestor could have possessed both AncIpCS2 and a

260 descendant of AncIpCS1, AncIpCS1' (Fig. 5b), additional pathway flux is possible. If
261 AncIpCS1' retained activities of its ancestor, AncIpCS1, then the ancestral *Ilex* species could
262 have also produced 3X and TP making for an even more diverse array of xanthine alkaloids
263 in its tissues (Fig. 5b). It has been shown that the xanthine alkaloids, 1X, 3X, and TP, can bind
264 to modern-day rat adenosine receptors[48]. Therefore, if these molecules were to accumulate in
265 ancestral *Ilex* tissues, they could have conferred a selective advantage which would likely
266 result in retention of the ancient genes. Ultimately, once gene duplication led to the generation
267 of the three modern-day CS-type enzymes in YM, pathway flux could be channelled away
268 from intermediates like 1X and TP such that the modern-day pathway to caffeine evolved (Fig.
269 5b). Not only did the modern-day CS enzymes of YM evolve to catalyse a pathway from
270 X>3X>TB>CF from ancestral biosynthetic networks of different products, *Theobroma* and
271 *Paullinia* also independently evolved enzymes with similar properties[19]. And, they did so from
272 ancestral pathways that, like YM, had alternative ancestral fluxes[24]. While it could be due to
273 chance alone that all three lineages converged to catalyse a similar pathway from differing
274 ancestral networks, it is also possible that it was advantageous to specialize for flux to TB via
275 X and 3X because either it is more enzymatically favourable or these intermediates have
276 greater adaptive value than other structural isomers.
277
278 **Protein crystal structure of IpCS3 reveals convergent structural basis for methylation**
279 **of theobromine to form caffeine**
280
281 We successfully crystallised and determined the 2.7 Å resolution structure of IpCS3,
282 that converts TB into CF (Fig. 6a, Supplementary Table 8). As is typical for the SABATH family
283 of methyltransferases, IpCS3 exhibits a Rossman-like fold which accounts for the SAM binding
284 portion of the enzyme as well as an alpha-helical cap where methyl-acceptor binding occurs[49–
285 52]. This structural information of the enzyme bound to both of its products, S-

286 adenosylhomocysteine (SAH) and caffeine, facilitates an in-depth comparison of the active
287 site structures of the caffeine-producing CS-type enzyme found in *Ilex* and the XMT-type
288 enzyme in *Coffea*[49] to determine the extent to which convergence of physicochemical
289 properties of the active site has allowed for independent specialisation for TB methylation by
290 the paralogous SABATH enzymes. Although the IpCS3 structure was obtained in complex
291 with its product, caffeine, it can be assumed that the binding mode is conserved for its
292 precursor, theobromine. Indeed, our computational modelling of TB in the active site of IpCS3
293 predicts it to be oriented as we have discerned from the diffraction data (Supplementary Fig.
294 4). Thus, in the following comparisons, the atomic numbering for the theobromine precursor
295 will be used to facilitate comparison to the DXMT structure.

296 In both DXMT and IpCS3, there are several conserved residues, shared by nearly all
297 SABATH enzymes, that form part of the active site pocket and appear to play important roles
298 in binding many different substrates[50,51]. His160 and Trp161 in DXMT are in the same relative
299 positions as His155 and Trp156 in IpCS3 (Fig. 6b, c). These residues are ca. 3 Å from TB and
300 participate in H-bonding but to different atoms of the substrate. In DXMT, the NE1 of Trp161
301 and NE2 of His160 form hydrogen bonds to carbonyl O2 of TB when positioned for N1
302 methylation; yet, in the structure of IpCS3 these corresponding side chain groups form
303 hydrogen bonds to O6 of TB. Despite these two residues being conserved for H-bonding, the
304 substrates are rotated 180 degrees along an axis going through N1 and C4. Thus, the
305 conserved His and Trp residues interact with opposing carbonyls in TB/CF but still position
306 the substrate for N1 methylation (Fig. 6b, c).

307 On the other hand, there are residues that differ between the two enzymes but appear
308 to provide for important substrate interactions. Specifically, in the structure of DXMT, the
309 hydroxyl group of Ser237 allows specific hydrogen bonding with N9 to position Theobromine
310 for N1 methylation (Fig. 6c). In IpCS3, His236 is found at the homologous position in the
311 structure. Nevertheless, its involvement in H-bonding with N9 is uncertain as the distance

312 between heavy atoms is ca. 4Å. Tyr368 of DXMT is found to participate in π-π interactions with
313 the ring structure of TB[53]; yet in IpCS3, Asn353 is found in the homologous position and the
314 amine forms a hydrogen bond with N9 due to its proximity within 3.2 Å, which is additionally
315 stabilised by Asn221 and His236 (Fig. 6b). Because the residues in these positions of IpCS3
316 and DXMT differ yet contribute to TB binding, these independent replacements represent
317 convergent structural solutions for N1 methylation of the substrate.
318
319 **Computational modelling of IpCS1 and IpCS2 active sites predict convergent substrate**
320 **binding residues for xanthine and 3-methylxanthine methylation**
321
322 Previous studies used site-directed mutagenesis of two sequence regions in CS-type
323 caffeine biosynthetic enzymes from *Theobroma* and *Paullinia* to uncover the mutational basis
324 for the convergent evolution of substrate preference switches towards their preferred
325 substrates, X and 3X[24]. In order to determine whether the same regions were convergently
326 mutated in IpCS1 and IpCS2 and provide binding interactions with X and 3X, respectively,
327 AlphaFold2[54] models and subsequent docking studies were performed (Supplementary Fig.
328 5). Modelling of substrate binding in the predicted active sites of IpCS1 and IpCS2 (Fig. 7a, b)
329 shows that the preferred substrates have optimal binding orientations that would result in
330 methylation to form the products that were experimentally detected in our assays shown in
331 Fig. 4. From our docking simulations, IpCS1 residues W156, N221 and Y265 are positioned
332 for hydrogen bonding with the carbonyl moieties of xanthine to position N3 for methyl transfer
333 from SAM (Fig. 7a). Although *Theobroma* and *Paullinia* CS1 enzymes specialized for xanthine
334 methylation also possess W156 and Y265 in the homologous positions (Supplementary Fig.
335 6), these residues are highly conserved amongst nearly all angiosperm SABATH enzymes.
336 On the other hand, the homologous position to N221 did not change concomitantly with the
337 evolution of X preference in either *Theobroma* and *Paullinia* (Supplementary Fig. 6); instead,

338 when their "region III" was mutated, activity with X improved in both[24]. Because IpCS1 was
339 not mutated in the homologous region III, there appear to be convergent solutions allowing for
340 efficient positioning of X for 3X biosynthesis amongst these enzymes. In the case of IpCS2,
341 two hydrogen bond donors S24 and T25 appear to contribute to the positioning of 3X in the
342 active site (Fig. 7b). This homologous region was experimentally mutated in *Theobroma* and
343 *Paullinia* CS2 enzymes and improved specialization for 3X methylation in both, although the
344 actual substitutions are different in each case[24]. Thus, this may represent an additional
345 instance where convergent mutations of the same region lead to specialization for 3X
346 methylation.
347
348 **Comparative phylogenomic analyses of caffeine biosynthetic genes reveal historical**
349 **constraints to convergent gene co-option**
350
351 Many nearly ubiquitous specialized metabolites involved in defence, development and
352 floral scent are produced by SABATH enzyme family members that appear to be conserved
353 across diverse angiosperm lineages, such as SAMT that methylates salicylic acid[55] and IAMT
354 that methylates Indole-3-acetic acid[52]. However, caffeine is sporadically distributed amongst
355 disparate angiosperm lineages and seems to have only recently evolved by convergence in a
356 few distantly related orders[19]. Our comparative evolutionary genomic analysis of the CS and
357 XMT syntenic regions across angiosperm (Fig. 8) indicates that predicting which SABATH
358 locus a given lineage might co-opt for caffeine biosynthesis is more dependent upon the
359 idiosyncratic history of gene loss than phylogenetic relatedness. For example, in the case of
360 the CS syntenic region used for caffeine biosynthesis in YM and *Theobroma*, *Coffea* lacks a
361 CS ortholog and none can be detected from its genome (Supplementary Fig. 7a). Thus, only
362 XMT was historically available for recruitment in *Coffea*. Conversely, YM appears to have lost
363 any vestiges of XMT orthologues known to be responsible for caffeine biosynthesis in *Coffea*

364 and *Citrus* (Supplementary Fig. 7b-d). This lack of genomic potential may be seen as an
365 evolutionary constraint to gene recruitment for caffeine biosynthesis in *Coffea* and YM.
366 A broader phylogenetic perspective on the XMT and CS syntenic regions provides
367 further insight into genomic canalization and allows for predictions about the underlying
368 genetic basis for caffeine biosynthesis in as-of-yet characterized lineages. As shown in Fig. 8,
369 several angiosperm lineages have neither XMT nor CS and this may explain why caffeine has
370 apparently never evolved in the large and diverse orders Brassicales, Asterales, Solanales
371 and Lamiales even though it has been shown to be advantageous in transgenic plants[56,57]. In
372 the case of *Cola*, a caffeine-producing genus from Africa[58], it is predicted to have co-opted CS
373 genes for xanthine alkaloid methylation because the order Malvales to which it belongs
374 appears to have lost XMT orthologues prior to its origin (Fig. 8). Tests of this hypothesis await
375 genomic sequences and functional studies of *Cola* enzymes. However, even with a functional
376 XMT or CS enzyme, duplication and diversification appears to be required to assemble a
377 complete pathway to caffeine as shown here for YM.
378
379 **Methods**
380
381 **Plant materials.** All plant materials used in this study were collected from *Ilex paraguariensis*
382 breeding line CA 8/74, kindly provided by the Banco de Germoplasma de Yerba Mate y Té at
383 the Estación Experimental Agropecuaria INTA Cerro Azul (EEA-INTA-CA), Misiones,
384 Argentina. This accession has one of the highest productivities, allows macro-propagation and
385 shows increased resistance to drought conditions (https://inta.gob.ar/).
386
387 **DNA extraction and sequencing.** Young leaves, preserved in silica-gel, were used for the
388 isolation of total genomic DNA with the DNeasy® Plant Mini Kit (QIAGEN®), following
389 manufacturer's instructions. Two approaches were combined to sequence the resulting DNA

390 and improve the accuracy of genome assembly. First, paired-end libraries (with insert sizes of
391 350 bp and 550 bp) and mate-pair libraries (with insert sizes of 3 kbp, 8 kbp and 12 kbp) were
392 constructed using the Illumina TruSeq® DNA Sample Preparation Kit and Illumina Nextera®
393 Mate Pair Library Preparation Kit following the kit's instructions, respectively. The obtained
394 libraries were sequenced on an Illumina HiSeq 1500 platform, generating ~263.2 Gb of raw
395 data. Second, long-reads libraries were prepared using Sequel Binding Kit 1.0 (Pacific
396 Biosciences), following manufacturer's instructions. The obtained libraries were sequenced on
397 PacBio Sequel I (Pacific Biosciences) using Sequel Sequencing Kit 1.0 (Pacific Biosciences)
398 and SMRT Cell 1M (Pacific Biosciences), generating ~77.5 Gb of additional raw data.
399
400 **Genome assembly and quality assessment.** We opted for a pipeline that could take
401 advantage of both short and long sequencing technologies. For the short-reads, we applied
402 Trimmomatic[59] v.0.39 to remove adaptor contaminations and filter low-quality reads (reads
403 with mean quality scores ≤ 25, reads where the quality of the bases at the head or tail was ≤
404 10 and reads with a length ≤ 30 bp). The resulting clean reads were then corrected using
405 Quake[60] v.0.3. Contig assembly and scaffolding was done using the assembler
406 SOAPdenovo[61] v.2 (55-mer size), with the mate-pair reads being used to link contigs into
407 scaffolds. After the assembly, DeconSeq[62] v.0.4.3 was used to detect and remove sequence
408 contaminants. Contigs and scaffolds clearly belonging to the chloroplast and mitochondria
409 genomes were also discarded. YM transcriptome sequences[25,26,27] and public databases
410 (KOG[30] and DEG[31]) were used to validate the genome assembly. Canu[63] v.2.2 was used to
411 perform the self-correction and assembly of the long reads, using the default parameters and
412 stopOnLowCoverage=20. For both short and long assemblies, we separated the assembly
413 haplotypes (haplotigs) using PurgeHaplotigs[64] with the recommended parameter values.
414 Then, we merged both SOAPdenovo v.2 and Canu v.2.2 curated assemblies using
415 Quickmerge[65] v.03, where only contigs with minimum overlap of 5000 bp (-ml 5000) were

416 merged and only the contigs greater than 1000 bp (-l 1000) were retained. The resulting
417 scaffolds and contigs were refined further with the gap-filling module in SOAPdenovo v.2
418 (GapCloser) and SSPACE[66] v.2.1.1.
419
420 **Gene prediction and annotation.** First, we masked the genome assembly with
421 RepeatMasker (http://repeatmasker.org/). Then, we predicted the protein coding and non-
422 coding genes using Funannotate[67] v.1.8.13 previously training it with the available *Ilex*
423 *paraguariensis* RNA-Seq experiments (NCBI projects PRJNA315513, PRJNA375923 and
424 PRJNA251985). Then, Infernal[68] v.1.1.4 was employed to improve the prediction of small
425 RNAs and microRNAs, while tRNAScan-SE[69] v.2.0 was used to improve the prediction of
426 transfer RNAs. Ribosomal RNAs were predicted using RNAmmer[70] v.1.2. The TAPIR web
427 server[71] (http://bioinformatics.psb.ugent.be/webtools/tapir) and the TargetFinder software[72]
428 v.1.7 were used to identify miRNA targets. InterProScan[73] v.5.55-88.0 and eggNOG-mapper[74]
429 v.2.1.7 were employed for the functional assignment of the predicted genes.
430
431 **Repeat content estimation.** The repeat content was estimated employing Dfam TE Tools
432 v.1.5 (https://github.com/Dfam-consortium/TETools). First, we used RepeatModeler v.2.0.3[75]
433 to build a database with *Ilex* repeat families. Then, we merged that database with Dfam v3.6[76]
434 and GIRI Repbase ver 20181026[77]. Finally, we ran RepeatMasker on the assembly using the
435 merged database to look for repeat sequences.
436
437 **Genome duplication analysis.** Rates of synonymous substitution (K_s) between paralogous
438 genes and orthologous genes in *Lactuca sativa*, *Daucus carota*, *Ilex paraguariensis*, *Coffea*
439 *canephora* and *Vitis vinifera* were determined using CoGe's tool SynMap
440 (https://genomevolution.org/). Gaussian mixture models (GMMs) were fitted to the resulting K_s
441 distributions with the mclust R package[78] v.5.0, and significant peaks were identified using the

442 SiZer R package[79] v.0.1-7. To estimate the age of the lineage-specific polyploidization event
443 (Ip-α) in *Ilex*, we considered two different phylogenies (a multiple nuclear genome phylogeny
444 and a plastid genome phylogeny). With the median K_s value of yerba mate-grape orthologues
445 (~0.89) and the divergence date of the two species (125.64 Ma for the multiple nuclear
446 genome phylogeny and 123.7 Ma for the plastid genome phylogeny), we calculated the
447 number of substitution per synonymous site per year (r) for YM (divergence date = $K_s/(2 \times r)$).
448 Conforming to the multiple nuclear genome phylogeny, the YM r value is 3.54E-9; while for
449 the plastid genome phylogeny, the YM r value is 3.59E-9. These r values and the SiZer K_s
450 range of YM paralogues (~0.35-0.5) were then applied to estimate the age of Ip-α. Finally, to
451 determine the syntenic depth ratio between *I. paraguariensis* and *C. canephora* and *V.*
452 *vinifera*, we employed CoGe's tool SynFind (https://genomevolution.org/), using a distance
453 cutoff of 10 genes and requiring at least 5 gene pairs per synteny block.
454
455 **Gene expression quantitation.** First, YM transcriptome reads (PRJNA315513) were mapped
456 to the IpCS1-5 transcripts using BWA[80]. Then, with the number of mapped reads, the
457 abundance of each transcript was calculated, normalized by transcript length and
458 transcriptome size (quantification in RPKM, reads per kilobase per million mapped reads).
459
460 **Cloning, heterologous expression and purification of enzymes.** Two different approaches
461 were used to clone IpCS genes: RT-PCR from leaf tissue and custom gene synthesis. For RT-
462 PCR of IpCS2, cDNA was obtained from 1 μg of RNA from fresh YM leaves using standard
463 procedures and cycling conditions with the following two primers: IpCS2F 5'-
464 ATGGACGTGAAGGAAGCAC-3' and IpCS2R 5'-CTATCCCATGGTCCTGCTAAG-3'.
465 Following amplification, cDNA was cloned using the pTrcHis TOPO® TA Expression Kit
466 (Invitrogen, Carlsbad, CA). Ligation of cDNA into the pTrcHis vector and subsequent
467 transformation into Top10 *E. coli* cells was carried out according to the manufacturer's

468 protocol. The transformation mixture was incubated overnight at 37°C on LB plates containing
469 50 μg/mL ampicillin. Colonies were screened by PCR to obtain full-length inserts that were
470 subsequently verified for insert orientation by DNA sequencing. For IpCS1 & 3 and AncIpCS1
471 & 2, gene sequences were synthesized by GenScript with codons optimized for expression in
472 *E. coli*. Synthesized genes were subcloned from the pUC57 cloning vector into the pET-15b
473 (Novagen) expression vector using 1.5 µg of DNA and NdeI and BamHI in 30 µl reactions.
474 Linear fragments corresponding to the expected sizes were gel purified using the QIAEX II
475 Gel Extraction Kit (Qiagen Corp.) according to the manufacturer's instructions. Purified DNA
476 fragments were ligated into pET-15b using T4 DNA ligase from New England Biolabs. Ligation
477 products were transformed into Top10 *E. coli* cells using 2 µl of the ligation reaction. Minipreps
478 of positive transformants were obtained using a QIAprep Spin Miniprep Kit (Qiagen Corp.) and
479 10 ng of each plasmid was used to transform BL21 *E. coli* cells using standard plating and
480 incubation methods.

481 Induction of His_6-protein was achieved in 50 ml cell cultures of BL21 (DE3) with IpCS1
482 & 3 and AncIpCS1 & 2 in pET-15b or Top10 with IpCS2 in pTrcHis with the addition of 1mM
483 IPTG at 23° C for 6 hours. Purification of the His_6-tagged protein was achieved by TALON®
484 spin columns (Takara Bio) following the manufacturer's instructions. Bradford assays were
485 used to determine purified protein concentration, and recombinant protein purity was
486 evaluated on SDS-PAGE gels.

487

488 **Enzyme assays.** All enzymes were tested for activity with the eight xanthine alkaloid
489 substrates shown in Fig. 1. Radiochemical assays were performed at 24° C for 60 minutes in
490 50 μl reactions that included 50 mM Tris-HCl buffer, 0.01 μCi (0.5 μl) ^{14}C-labelled SAM, 10-20
491 μl purified protein and 1 mM methyl acceptor substrate dissolved in 0.5 M NaOH. Negative
492 controls were composed of the same reagents, except that the methyl acceptor substrate was
493 omitted and 1 μl of 0.5 M NaOH was added instead. Methylated products were extracted in

494 200 μl ethyl acetate and quantified using a liquid scintillation counter. The highest enzyme
495 activity reached with a specific substrate was set to 1.0 and relative activities with remaining
496 substrates were calculated. Each assay was run at least twice so that mean, plus standard
497 deviation, could be calculated.
498
499 **High performance liquid chromatography (HPLC).** Product identity of enzyme assays was
500 determined using HPLC on 500 µl scaled-up reactions utilizing all the same reagents as
501 described above except that non-radioactive SAM was used as the methyl donor and reactions
502 were allowed to progress for 4 hours. Whole reactions were filtered through Vivaspin®
503 columns (Sartorius) to remove all protein prior to injection in the HPLC. Mixtures were
504 separated by HPLC using a two-solvent system with a 250 mm x 4.6 mm Kinetex® 5 µM EVO
505 C18 column (Phenomenex). Solvent A was 99.9% water with 0.1% TFA and Solvent B was
506 80% acetonitrile, 19.9% water and 0.1% TFA and a 0-20% gradient was generated over 16
507 minutes with a flow rate of 1.0 ml/minute. Product identity was determined by comparing
508 retention times and absorbance at 254 and 272 nm of authentic standards. Reactions were
509 compared to negative controls in which no methyl acceptor substrates were added.
510
511 **Phylogenetic analyses.** In order to accurately determine the orthology of YM SABATH
512 sequences encoded in the genome, we compared them to all previously characterized gene
513 family members in other species. We also included CS and XMT orthologues from relatives
514 from the orders of caffeine-producing species (Malvales, Ericales, Gentianales, Sapindales)
515 available in public databases (GenBank, OneKP) as shown in Fig. 3. Accession numbers for
516 all sequences are provided in Supplementary Table 9. Alignment of amino acid sequences
517 was achieved using MAFFT[81] v.7.0 and employing the auto search strategy to maximize
518 accuracy and speed. A phylogenetic estimate was obtained using FastTree[82] v.2 assuming
519 the Jones-Taylor-Thorton model of amino acid substitution with a CAT approximation using

520 20 rate categories. Reliability of individual nodes was estimated from local support values
521 using the Shimodaira-Hasegawa test as implemented in FastTree.
522
523 **Ancestral sequence resurrection.** In order to obtain accurate ancestral CS protein
524 estimates, we assembled a dataset that included 154 sequences including all CS-type
525 enzymes we could retrieve from GenBank and China National Gene Bank as well as
526 representatives of all other functionally characterized clades of SABATH enzymes
527 (Supplementary Fig. 2). IQTree[83] was used to obtain a tree describing the relationships
528 amongst the sequences assuming the Jones, Taylor, and Thorton (JTT) matrix model for
529 amino acid substitution and the Free rate model for among-site rate heterogeneity[84].
530 Confidence in ancestral sequences was high as indicated by the average site-specific
531 posterior probability for AncIpCS1 (0.99) and AncIpCS2 (0.995). In order to determine
532 ancestral protein lengths in regions with alignment gaps, we coded each gap for the number
533 of amino acids possessed and used parsimony to determine ancestral residue numbers as in
534 our previous studies[19]. The estimated sequences were synthesized by Genscript Corp. and
535 had codons chosen for optimal protein expression in *Escherichia coli*.
536
537 **Crystallization, data collection, phasing and refinement of IpCS3.** Crystallization
538 screening was performed using the IpCS3 methyltransferase at a concentration of 30 mg/mL
539 incubated with 2 mM Theobromine and 2 mM S-Adenosyl methionine. Sitting-drop for
540 crystallization screening was set up by equal volume of precipitant and protein solution
541 (0.25:0.25 μL) using a Crystal Gryphon robot (Art Robbins Instruments) and a reservoir volume
542 of 45 μL. Trays were incubated at 9 °C. Initial hits were further optimized using the hanging-
543 drop method at 9°C, with 150 μL reservoir solution and 1:1 ratio of precipitant to protein and
544 ligand solution in a 2 μL drop. The latter was composed of 33 mg/mL IpCS3 protein
545 concentration and 4 mM Theobromine and 2 mM S-Adenosyl methionine. The crystallization

546 condition was optimized to 25% PEG 3350, 0.2 M NH_4SO_4, 0.1 M Bis-Tris methane pH 5.5.
547 Square crystals grew over 10 days and were subsequently soaked for 3h at 9°C in the
548 precipitant solution supplemented with 10 mM S-Adenosyl homocysteine and 10 mM
549 Theobromine. Crystals were transiently soaked in the precipitant solution supplemented with
550 20% Ethylene Glycol immediately prior to vitrification by direct immersion into liquid nitrogen.
551 Diffraction data were collected at the Advanced Photon Source (APS) at Argonne National
552 Laboratory Sector-21 via the Life Sciences-Collaborative Access Team (LS-CAT) at beamline
553 21-ID-G. Diffraction data were indexed, integrated, and scaled using the autoPROC software
554 package[85]. The structure was solved by molecular replacement using Phaser-MR included in
555 the Phenix software package[86], using PDBID 6LYH structure as the replacement model. The
556 model was subject to rounds of manual building followed by refinement using REFMAC5[87],
557 and was manually built in COOT[88] v.0.9.8.3. Crystallographic statistics are listed in
558 Supplementary Table 8.
559
560 **Structure prediction and molecular docking.** Protein structures of IpCS enzymes were
561 predicted using the ColabFold implementation of AlphaFold2[54] with no template. Diagnostic
562 plots depicting the MSA coverage, alignment error and LDDT are shown in the supplementary
563 information (Supplementary Fig. 5). Structures of xanthine alkaloid ligands (X, 3X, and TB)
564 were downloaded from the ChEMBL database[89]; protonation states were checked by
565 Chemicalize[90] and optimized using the VMD Molefacture plugin[91]. The receptor structures
566 were prepared following the standard AutoDock protocol[92] using the prepare_receptor4.py
567 script from AutoDock Tools. All non-polar hydrogens were merged, and Gasteiger charges
568 and atom types were added. The ligand PDBQT was prepared using the prepare_ligand4.py
569 script. The grid size and position were chosen to contain the whole ligand-binding site
570 (including all protein atoms closer than 5 Å from all ligands). For each system, 10 different
571 docking runs were performed. Docking was performed using AutoDock Vina v.1.2.0[93]. The

572 docking results were further analysed by visual inspection. Images of the molecules were
573 prepared using the PyMOL molecular graphics system[94].
574
575 **Synteny comparisons and phylogenetic distribution of CS and XMT.** The presence or
576 absence of CS and XMT genes was determined for orders of plants for which at least one
577 genomic sequence exists, as shown in Fig. 8. For those species which do not yet have an
578 available assembly, we used BLAST[95] analyses of GenBank (nr and TSA databases),
579 Phytozome[96] as well as the OneKP dataset[37] in China National GeneBank. BLAST combined
580 with subsequent phylogenetic analyses were also used to verify presence/absence of CS- or
581 XMT-type sequences in cases where the syntenic regions did not appear to encode one or
582 the other gene. Comparisons of the CS and XMT syntenic regions were performed using
583 CoGe's tool GEvo (https://genomevolution.org/).
584
585 **Data availability**
586
587 The raw sequence data, assembled genome and annotation were deposited in the
588 European Nucleotide Archive (ENA) under BioProject No. PRJEB65927. The plasmids used
589 to produce proteins are freely available upon request. The atomic coordinates and structure
590 factors have been deposited in the Protein Data Bank, Research Collaboratory for Structural
591 Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rscb.org) with the
592 accession code 8T2G for the IPCS3 structure bound to caffeine and SAH.
593
594 **Acknowledgments**
595
596 This work was supported by Consejo Nacional de Investigaciones Científicas y
597 Técnicas de Argentina (CONICET); PRO.MATE.AR project, funded by Secretaría de Políticas

598 Universitarias del Ministerio de Educación de la Nación Argentina; CABANA project, funded
599 by UKRI-BBSRC on behalf of the Global Challenges Research Fund (BB/P027849/1), and
600 National Science Foundation (NSF) grants (MCB-1120624 and MCB-2325341 to T.J.B.). We
601 thank Centro de Cómputos de Alto Rendimiento (CeCAR) and UBA-FCEN-QB-Cluster for
602 granting use of computational resources which allowed us to perform most of the
603 computational analyses included in this work. Kevin Blair and the Department of Chemistry is
604 thanked for facilitating our HPLC analyses.

605

606 **Contributions**

607

608 F.A.V., C.P.M., T.J.B. and A.G.T. designed the project and coordinated the research activities.
609 F.A.V., T.J.B. and A.G.T. wrote the manuscript with input from all co-authors. F.A.V. was
610 involved in all analyses. R.M.A. extracted the genomic DNA. E.J.S., G.F.B., R.R.M.O., G.L.N.,
611 and G.O. performed the assembly and annotation of the genome. F.A.V. conducted the
612 evolutionary and syntenic-gene analyses. A.H.G. and T.J.B. carried out the expression,
613 purification and activity assays of CS enzymes. L.A.D made the AlphaFold2 models. L.A.D.
614 and C.P.M. performed the molecular dockings. A.H.G. and S.N. obtained the crystal structure
615 of IpCS3-CF complex. T.J.B., M.R., and P.D.Z. performed the comparative genomic analyses
616 of caffeine-producing enzymes across angiosperms. P.D.Z., D.A.M., G.O., S.N., T.J.B., and
617 A.G.T. obtained funding to carry out the project. All authors read and approved the final article.

618

619 **Competing interests**

620

621 The authors declare no competing interests.

622

623 **References**

1. Rey, H. Y. *et al.* Colchicine, trifluralin, and oryzalin promoted development of somatic embryos in *Ilex paraguariensis* (Aquifoliaceae). *Euphytica* **123**, 49–56 (2002).
2. Heck, C. I. & de Mejia, E. G. Yerba Mate tea (*Ilex paraguariensis*): a comprehensive review on chemistry, health implications, and technological considerations. *Journal of Food Science* **72**, R138–R151 (2007).
3. Gugliucci, A. Antioxidant effects of *Ilex paraguariensis*: induction of decreased oxidability of human LDL *in vivo*. *Biochem. Biophys. Res. Commun.* **224**, 338–344 (1996).
4. Vieira, M. A. *et al.* Phenolic acids and methylxanthines composition and antioxidant properties of mate (*Ilex paraguariensis*) residue. *Journal of Food Science* **75**, C280–C285 (2010).
5. Boado, L. S., Fretes, R. M. & Brumovsky, L. A. Bioavailability and antioxidant effect of the *Ilex paraguariensis* polyphenols. *Nutrition & Food Science* **45**, 326–335 (2015).
6. Kang, Y.-R. *et al.* Anti-obesity and anti-diabetic effects of Yerba Mate (*Ilex paraguariensis*) in C57BL/6J mice fed a high-fat diet. *Laboratory Animal Research* **28**, 23 (2012).
7. Ríos, J., Francini, F. & Schinella, G. Natural products for the treatment of type 2 diabetes mellitus. *Planta Medica* **81**, 975–994 (2015).
8. Santos, E. C. S. *et al.* Anxiolytic-like, stimulant and neuroprotective effects of *Ilex paraguariensis* extracts in mice. *Neuroscience* **292**, 13–21 (2015).
9. Yao, X., Lu, Z., Song, Y., Hu, X. & Corlett, R. T. A chromosome-scale genome assembly for the holly (*Ilex polyneura*) provides insights into genomic adaptations to elevation in Southwest China. *Hortic Res* **9**, (2022).
10. Kong, B. L.-H. *et al.* Chromosomal level genome of *Ilex asprella* and insight into antiviral triterpenoid pathway. *Genomics* **114**, 110366 (2022).
11. Xu, K.-W. *et al.* The chromosome-level holly (*Ilex latifolia*) genome reveals key enzymes in triterpenoid saponin biosynthesis and fruit color change. *Front. Plant Sci.* **13**, (2022).

650 12. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: new insights and
651 directions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190102 (2019).
652 13. Thorogood, C. J., Bauer, U. & Hiscock, S. J. Convergent and divergent evolution in
653 carnivorous pitcher plant traps. *New Phytol.* **217**, 1035–1041 (2018).
654 14. Wan, T. *et al.* A genome for gnetophytes and early evolution of seed plants. *Nat Plants* **4**,
655 82–89 (2018).
656 15. Yang, X. *et al.* The Kalanchoë genome provides insights into convergent evolution and
657 building blocks of crassulacean acid metabolism. *Nat. Commun.* **8**, 1899 (2017).
658 16. Pichersky, E. & Lewinsohn, E. Convergent evolution in plant specialized metabolism.
659 *Annu. Rev. Plant Biol.* **62**, 549–566 (2011).
660 17. Anaya, A. L., Cruz-Ortega, R. & Waller, G. R. Metabolism and ecology of purine alkaloids.
661 *Front. Biosci.* **11**, 2354–2370 (2006).
662 18. Stevenson, P. C., Nicolson, S. W. & Wright, G. A. Plant secondary metabolites in nectar:
663 impacts on pollinators and ecological functions. *Funct. Ecol.* **31**, 65–75 (2017).
664 19. Huang, R., O'Donnell, A. J., Barboline, J. J. & Barkman, T. J. Convergent evolution of
665 caffeine in plants by co-option of exapted ancestral enzymes. *Proc. Natl. Acad. Sci. U. S.*
666 *A.* **113**, 10613–10618 (2016).
667 20. Suzuki, T. & Takahashi, E. Caffeine biosynthesis in *Camellia sinensis*. *Phytochemistry*
668 **15**, 1235–1239 (1976).
669 21. Ashihara, H., Monteiro, A. M., Gillies, F. M. & Crozier, A. Biosynthesis of caffeine in leaves
670 of coffee. *Plant Physiol.* **111**, 747–753 (1996).
671 22. Kato, M. *et al.* Caffeine biosynthesis in young leaves of *Camellia sinensis*: in vitro studies
672 on N-methyltransferase activity involved in the conversion of xanthosine to caffeine.
673 *Physiol. Plant.* **98**, 629–636 (1996).
674 23. Uefuji, H., Ogita, S., Yamaguchi, Y., Koizumi, N. & Sano, H. Molecular cloning and
675 functional characterization of three distinct N-methyltransferases involved in the caffeine

676 biosynthetic pathway in coffee plants. *Plant Physiol.* **132**, 372–380 (2003).

677 24. O'Donnell, A. J., Huang, R., Barboline, J. J. & Barkman, T. J. Convergent biochemical
678 pathways for xanthine alkaloid production in plants evolved from ancestral enzymes with
679 different catalytic properties. *Mol. Biol. Evol.* **38**, 2704–2714 (2021).

680 25. Debat, H. J. *et al.* Exploring the genes of yerba mate (*Ilex paraguariensis* A. St.-Hil.) by
681 NGS and de novo transcriptome assembly. *PLoS One* **9**, e109835 (2014).

682 26. Fay, J. V. *et al.* Yerba mate (*Ilex paraguariensis*, A. St.-Hil.) de novo transcriptome
683 assembly based on tissue specific genomic expression profiles. *BMC Genomics* **19**, 891
684 (2018).

685 27. Acevedo, R. M. *et al.* Transcript and metabolic adjustments triggered by drought in *Ilex*
686 *paraguariensis* leaves. *Planta* **250**, 445–462 (2019).

687 28. Singh, R., Ming, R. & Yu, Q. Comparative analysis of GC content variations in plant
688 genomes. *Tropical Plant Biology* vol. 9 136–149 (2016).

689 29. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M.-A. LTR-
690 retrotransposons in plants: engines of evolution. *Gene* **626**, 14–25 (2017).

691 30. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC*
692 *Bioinformatics* **4**, 41 (2003).

693 31. Luo, H., Lin, Y., Gao, F., Zhang, C.-T. & Zhang, R. DEG 10, an update of the database of
694 essential genes that includes both protein-coding genes and noncoding genomic
695 elements. *Nucleic Acids Res.* **42**, D574–80 (2014).

696 32. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update:
697 novel and streamlined workflows along with broader and deeper phylogenetic coverage
698 for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654
699 (2021).

700 33. Gottlieb, A. M. & Poggio, L. Quantitative and qualitative genomic characterization of
701 cultivated *Ilex* L. species. *Plant Genet. Resour.* **13**, 142–152 (2015).

702 34. Sankoff, D. & Zheng, C. Whole genome duplication in plants: implications for evolutionary
703 analysis. *Methods Mol. Biol.* **1704**, 291–315 (2018).
704 35. Landis, J. B. *et al.* Impact of whole-genome duplication events on diversification rates in
705 angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
706 36. Zhang, C. *et al.* Asterid phylogenomics/phylotranscriptomics uncover morphological
707 evolutionary histories and support phylogenetic placement for numerous whole-genome
708 duplications. *Mol. Biol. Evol.* **37**, 3188–3210 (2020).
709 37. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and
710 the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
711 38. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A
712 metacalibrated time-tree documents the early rise of flowering plant phylogenetic
713 diversity. *New Phytol.* **207**, 437–453 (2015).
714 39. Yao, X., Song, Y., Yang, J.-B., Tan, Y.-H. & Corlett, R. T. Phylogeny and biogeography
715 of the hollies (*Ilex* L., Aquifoliaceae). *J. Syst. Evol.* **59**, 73–82 (2021).
716 40. Ross, J. R., Nam, K. H., D'Auria, J. C. & Pichersky, E. S-Adenosyl-L-methionine:salicylic
717 acid carboxyl methyltransferase, an enzyme involved in floral scent production and plant
718 defense, represents a new class of plant methyltransferases. *Arch. Biochem. Biophys.*
719 **367**, 9–16 (1999).
720 41. Seo, H. S. *et al.* Jasmonic acid carboxyl methyltransferase: a key enzyme for jasmonate-
721 regulated plant responses. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4788–4793 (2001).
722 42. Yin, Y., Katahira, R. & Ashihara, H. Metabolism of purine alkaloids and xanthine in leaves
723 of maté (*Ilex paraguariensis*). *Nat. Prod. Commun.* **10**, 707–712 (2015).
724 43. Yang, L. *et al.* Phylogenomic Insights into deep phylogeny of angiosperms based on
725 broad nuclear gene sampling. *Plant Commun* **1**, 100027 (2020).
726 44. Negrin, A., Long, C., Motley, T. J. & Kennelly, E. J. LC-MS metabolomics and
727 chemotaxonomy of caffeine-containing holly (*Ilex*) species and related taxa in the

728 Aquifoliaceae. *J. Agric. Food Chem.* **67**, 5687–5699 (2019).

729 45. Noda-Garcia, L., Liebermeister, W. & Tawfik, D. S. Metabolite-enzyme coevolution: from
730 single enzymes to metabolic pathways and networks. *Annu. Rev. Biochem.* **87**, 187–216
731 (2018).

732 46. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules.
733 *Nat. Rev. Genet.* **5**, 366–375 (2004).

734 47. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the
735 functional synthesis. *Nat. Rev. Genet.* **8**, 675–688 (2007).

736 48. Daly, J. W., Butts-Lamb, P. & Padgett, W. Subclasses of adenosine receptors in the
737 central nervous system: interaction with caffeine and related methylxanthines. *Cell. Mol.*
738 *Neurobiol.* **3**, 69–80 (1983).

739 49. McCarthy, A. A. & McCarthy, J. G. The structure of two N-methyltransferases from the
740 caffeine biosynthetic pathway. *Plant Physiol.* **144**, 879–889 (2007).

741 50. Zubieta, C. *et al.* Structural basis for substrate recognition in the salicylic acid carboxyl
742 methyltransferase family. *Plant Cell* **15**, 1704–1716 (2003).

743 51. Petronikolou, N., Hollatz, A. J., Schuler, M. A. & Nair, S. K. Loganic acid
744 methyltransferase: insights into the specificity of methylation on an iridoid glycoside.
745 *Chembiochem* **19**, 784–788 (2018).

746 52. Zhao, N. *et al.* Structural, biochemical, and phylogenetic analyses suggest that indole-3-
747 acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family.
748 *Plant Physiol.* **146**, 455–467 (2008).

749 53. Lanzarotti, E., Defelipe, L. A., Marti, M. A. & Turjanski, A. G. Aromatic clusters in protein–
750 protein and protein–drug complexes. *J. Cheminform.* **12**, 1–9 (2020).

751 54. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**,
752 679–682 (2022).

753 55. Dubs, N. M. *et al.* A collaborative classroom investigation of the evolution of SABATH

754 methyltransferase substrate preference shifts over 120 my of flowering plant history. *Mol.*
755 *Biol. Evol.* **39**, (2022).
756 56. Kim, Y.-S., Uefuji, H., Ogita, S. & Sano, H. Transgenic tobacco plants producing caffeine:
757 a potential new strategy for insect pest control. *Transgenic Res.* **15**, 667–672 (2006).
758 57. Kim, Y.-S. *et al.* Resistance against beet armyworms and cotton aphids in caffeine-
759 producing transgenic chrysanthemum. *Plant Biotechnol.* **28**, 393–395 (2011).
760 58. Niemenak, N., Onomo, P. E., Fotso, Lieberei, R. & Ndoumou, D. O. Purine alkaloids and
761 phenolic compounds in three *Cola* species and *Garcinia kola* grown in Cameroon. *S. Afr.*
762 *J. Bot.* **74**, 629–638 (2008).
763 59. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
764 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
765 60. Kelley, D. R., Schatz, M. C. & Salzberg, S. L. Quake: quality-aware detection and
766 correction of sequencing errors. *Genome Biol.* **11**, R116 (2010).
767 61. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de*
768 *novo* assembler. *Gigascience* **1**, 18 (2012).
769 62. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination
770 from genomic and metagenomic datasets. *PLoS One* **6**, e17288 (2011).
771 63. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive -mer
772 weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
773 64. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig
774 reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460
775 (2018).
776 65. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and
777 accurate *de novo* assembly of metazoan genomes with modest long read coverage.
778 *Nucleic Acids Res.* **44**, e147 (2016).
779 66. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-

780 assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).

781 67. Palmer, J. & Stajich, J. *nextgenusfs/funannotate: funannotate v1.5.3*. (2019).
782 doi:10.5281/zenodo.2604804.

783 68. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches.
784 *Bioinformatics* **29**, 2933–2935 (2013).

785 69. Chan, P. P. & Lowe, T. M. tRNAscan-SE: searching for tRNA genes in genomic
786 sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).

787 70. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
788 *Nucleic Acids Res.* **35**, 3100–3108 (2007).

789 71. Bonnet, E., He, Y., Billiau, K. & Van de Peer, Y. TAPIR, a web server for the prediction of
790 plant microRNA targets, including target mimics. *Bioinformatics* **26**, 1566–1568 (2010).

791 72. Fahlgren, N. & Carrington, J. C. miRNA target prediction in plants. *Methods in Molecular*
792 *Biology* 51–57 (2010).

793 73. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
794 *Bioinformatics* **30**, 1236–1240 (2014).

795 74. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
796 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
797 *Res.* **47**, D309–D314 (2019).

798 75. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable
799 element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).

800 76. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**,
801 D81–9 (2016).

802 77. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet.*
803 *Genome Res.* **110**, 462–467 (2005).

804 78. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. Mclust 5: clustering, classification and
805 density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016).

806 79. Chaudhuri, P. & Marron, J. S. Scale space view of curve estimation. *Ann. Stat.* **28**, 408–
807 428 (2000).
808 80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
809 transform. *Bioinformatics* **25**, 1754–1760 (2009).
810 81. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
811 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
812 82. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
813 trees for large alignments. *PLoS One* **5**, e9490 (2010).
814 83. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online
815 phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–5
816 (2016).
817 84. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**,
818 993–1005 (1995).
819 85. Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta*
820 *Crystallogr. D Biol. Crystallogr.* **67**, 293–302 (2011).
821 86. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular
822 structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
823 87. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal
824 structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).
825 88. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot.
826 *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
827 89. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids*
828 *Res.* **47**, D930–D940 (2019).
829 90. Swain, M. chemicalize.org. *J. Chem. Inf. Model.* **52**, 613–615 (2012).
830 91. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.*
831 **14**, 33–8, 27–8 (1996).

832 92. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: automated docking with selective
833 receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
834 93. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: new
835 docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**,
836 3891–3898 (2021).
837 94. Schrodinger, L. L. C. The PyMOL molecular graphics system. *Version 1.8* (2015).
838 95. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
839 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
840 96. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics.
841 *Nucleic Acids Res.* **40**, D1178–86 (2012).
842 97. Iorizzo, M. *et al.* A high-quality carrot genome assembly provides new insights into
843 carotenoid accumulation and asterid genome evolution. *Nat. Genet.* **48**, 657–666 (2016).
844 98. Reyes-Chin-Wo, S. *et al.* Genome assembly with in vitro proximity ligation data and
845 whole-genome triplication in lettuce. *Nat. Commun.* **8**, 14953 (2017).

846

847

848

849

850

851

852

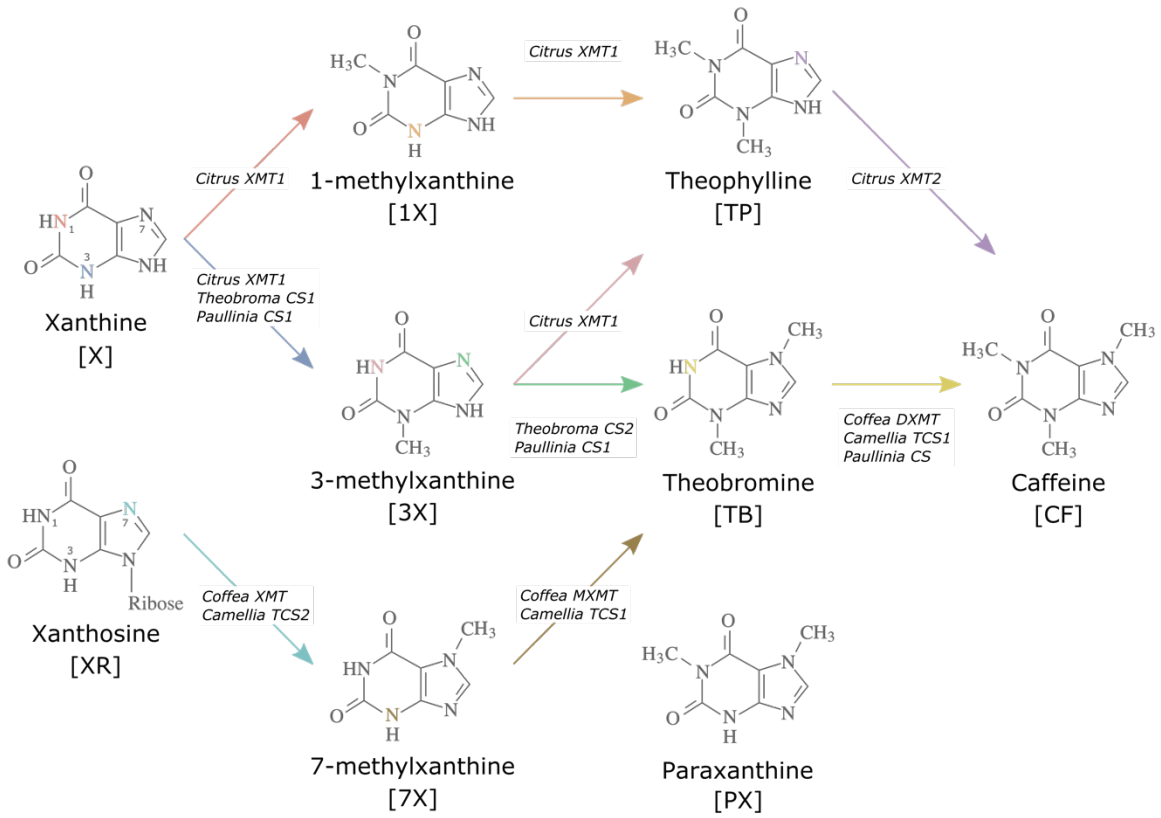


Figure 1. Biosynthetic routes to caffeine within the xanthine alkaloid network. CF, caffeine; PX, paraxanthine; TB, theobromine; TP, theophylline; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; XR, xanthosine; X, xanthine. Nitrogen atoms are coloured to match the arrows corresponding to the enzymes that methylate them.

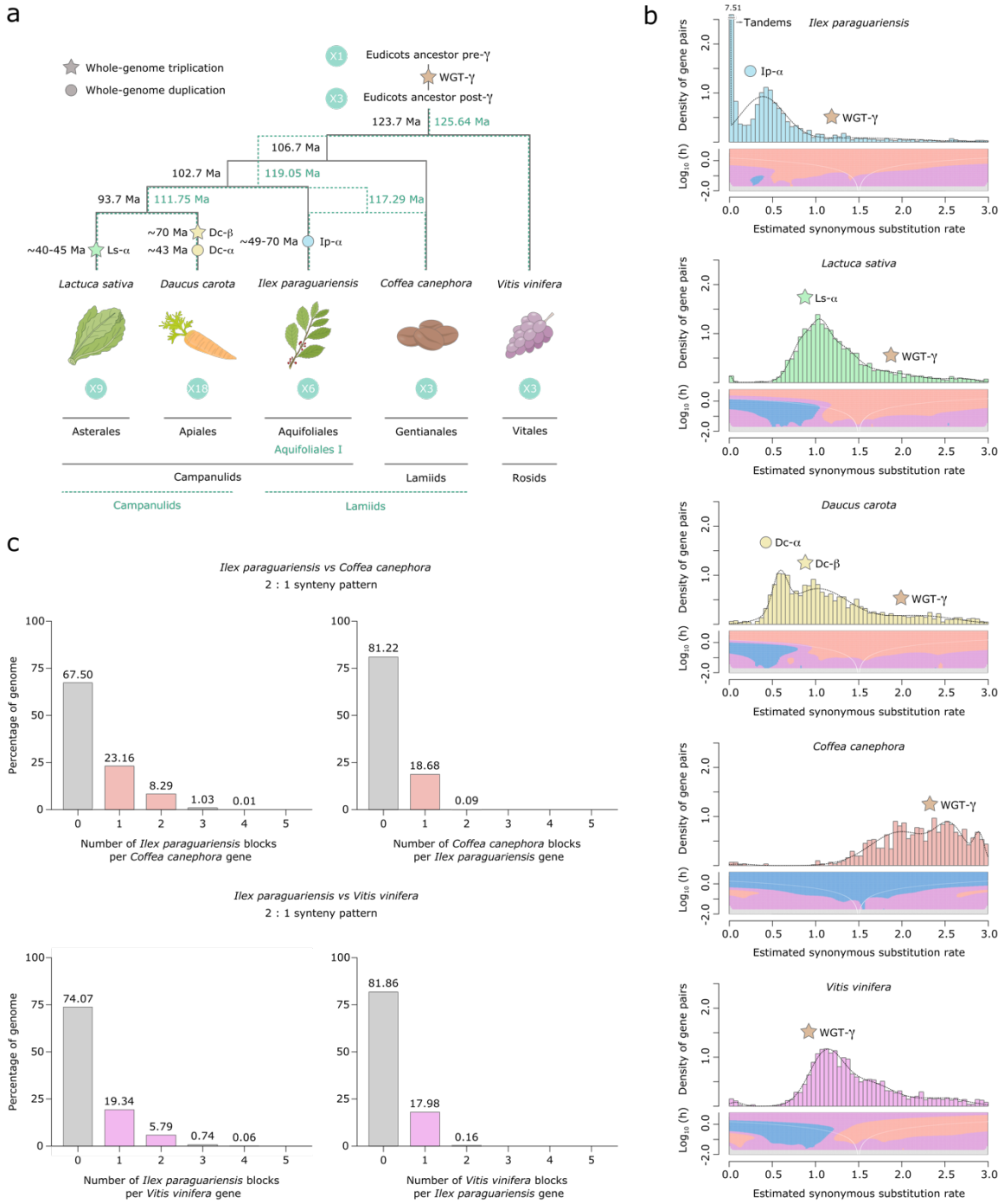


Figure 2. Yerba mate genome duplication history. **a**, Evolutionary scenario of the eudicot genomes of *Lactuca sativa*, *Daucus carota*, *Ilex paraguariensis*, *Coffea canephora* and *Vitis vinifera*, from their ancestor pre-γ. The plastid genome phylogeny is represented with solid black lines, while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

871 (triplications). Divergence time estimates for the lineages, as well as age estimates for the *L.*
872 *sativa* and *D. carota* paleopolyploidizations were obtained from the literature[36,38,97,98]. Ma,
873 million years ago. **b**, K_s distributions with Gaussian mixture model and SiZer analyses of *I.*
874 *paraguariensis* (blue), *L. sativa* (green), *D. carota* (yellow), *C. canephora* (red) and *V. vinifera*
875 (purple) paralogues. SiZer maps below histograms identify significant peaks at corresponding
876 K_s values. Blue represents significant increases in slope, red indicates significant decreases,
877 purple represents no significant slope change, and grey indicates not enough data for the test.
878 **c**, Comparative genomic synteny analyses of *I. paraguariensis* with *C. canephora* and *V.*
879 *vinifera*.
880
881
882
883
884
885
886
887
888
889
890
891

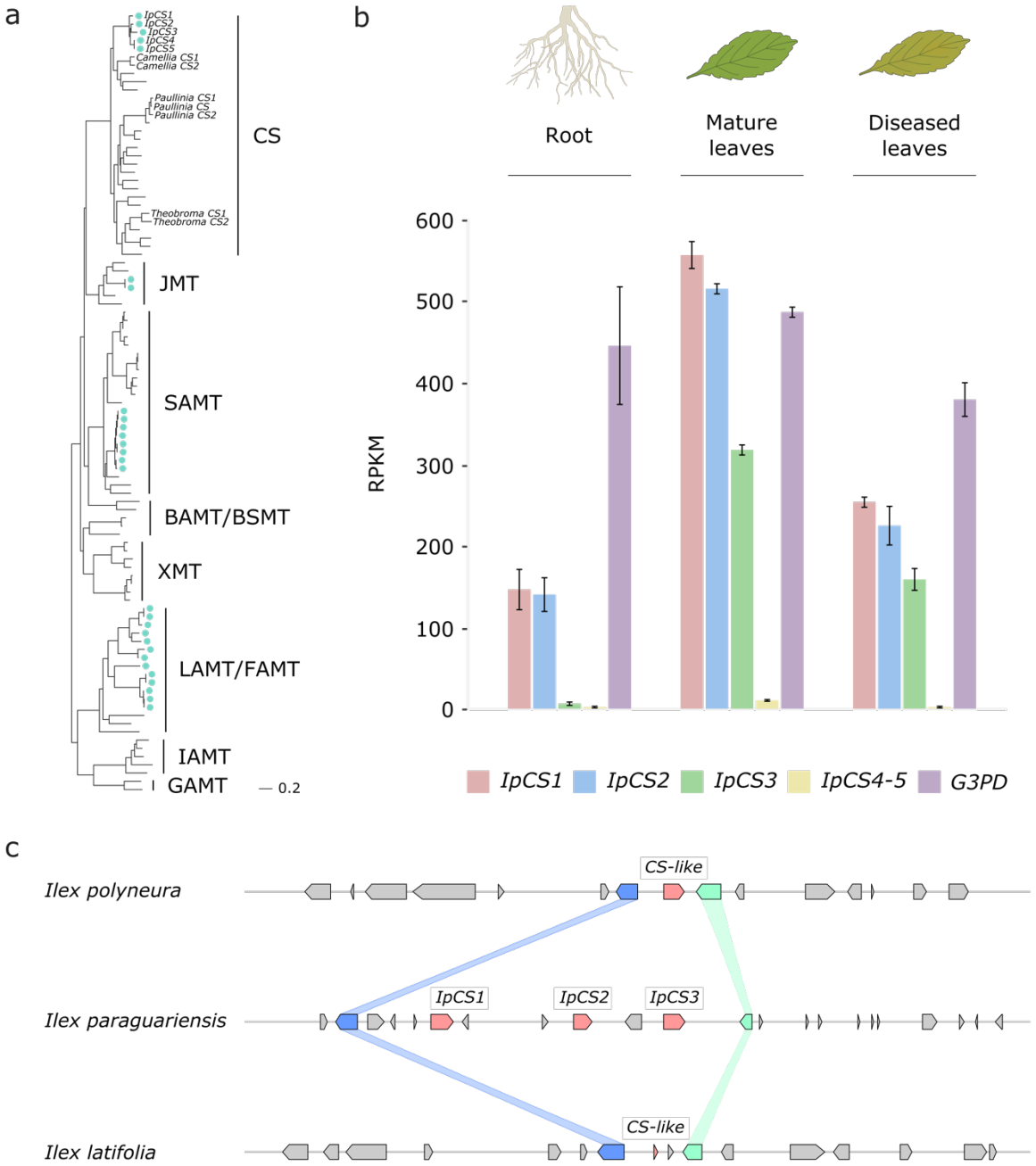
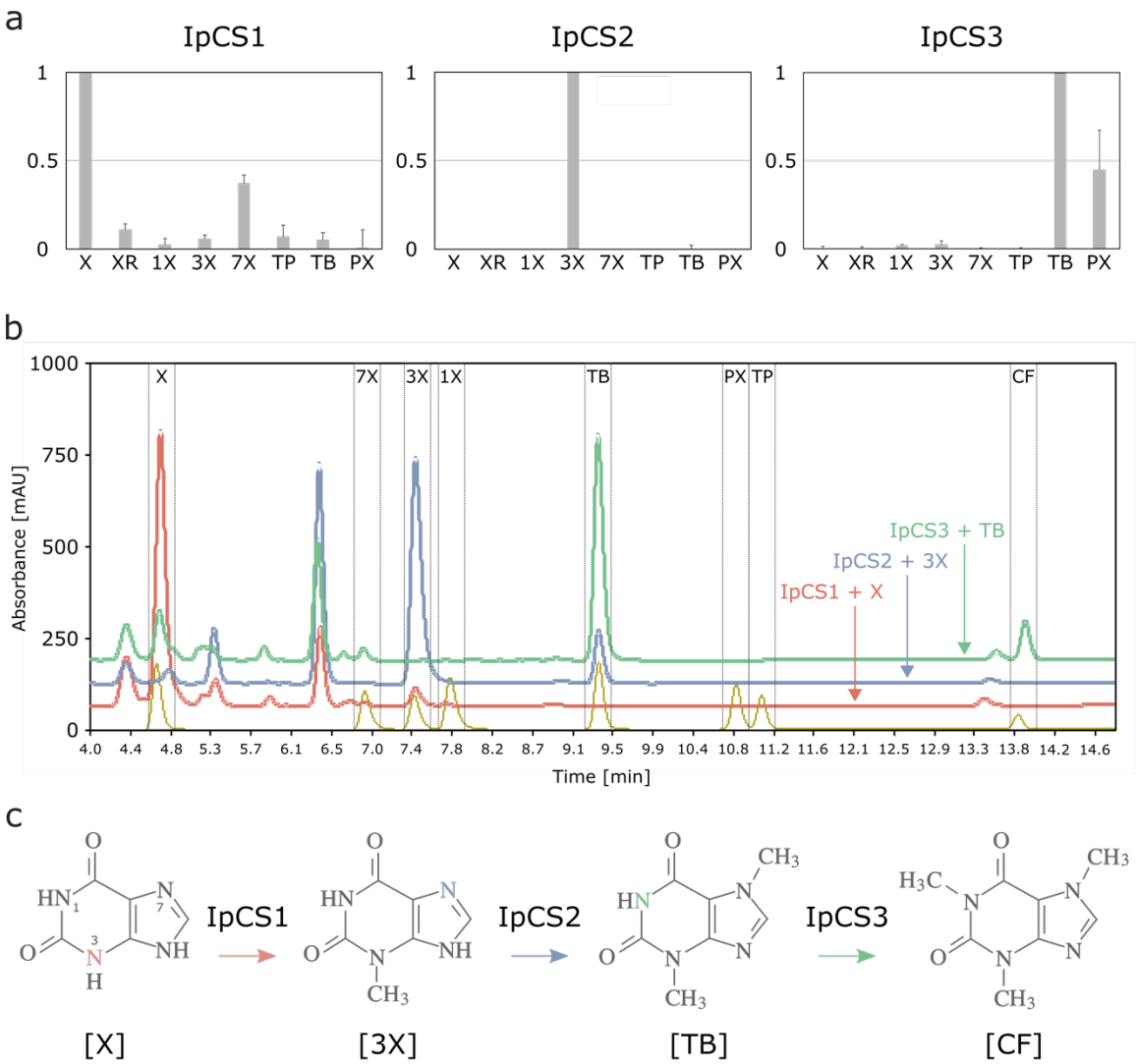


Figure 3. The yerba mate genome encodes three recently duplicated CS-type SABATH proteins that are expressed in caffeine-producing tissues. a, SABATH gene tree estimate (LnL = -34265.473) shows the placement of full length YM proteins (marked by blue-green dots) within clades that have published functions. 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

899 acid MT; JMT, jasmonic acid MT; CS, caffeine synthase in *Theobroma*, *Camellia* and *Paullinia*.
900 **b**, Gene expression analysis of IpCS1-5 as indicated by the relative abundance of YM
901 transcriptome reads mapped to the IpCS1-5 transcripts. RPKM, reads per kilobase per million
902 mapped reads. Error bars indicate standard deviation from the mean. Housekeeping gene:
903 *G3PD*, Glyceraldehyde-3-phosphate dehydrogenase. **c**, Synteny-based analysis of the CS
904 genomic region for *I. paraguariensis*, *I. polyneura* and *I. latifolia*.
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919

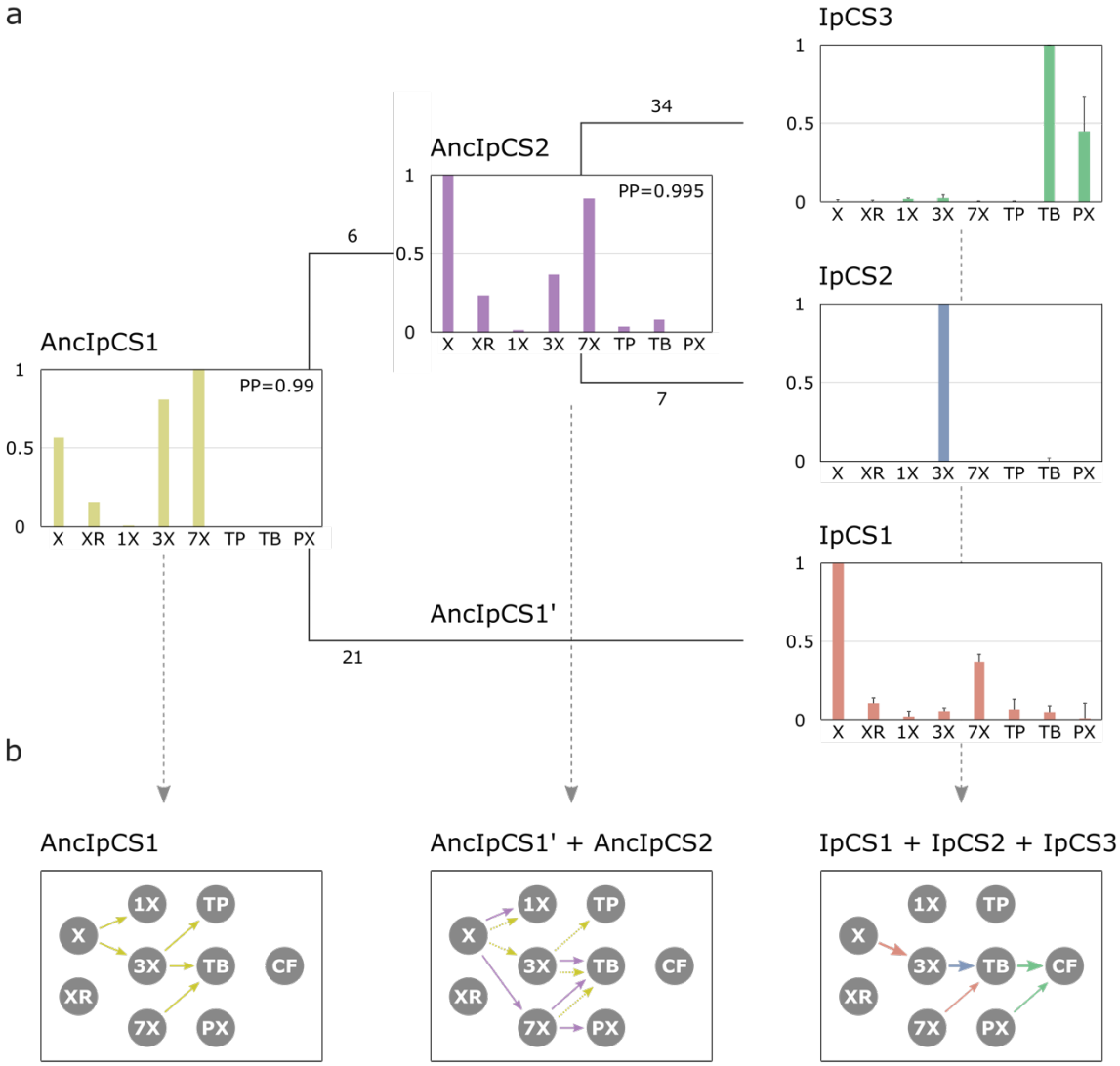


920

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

928

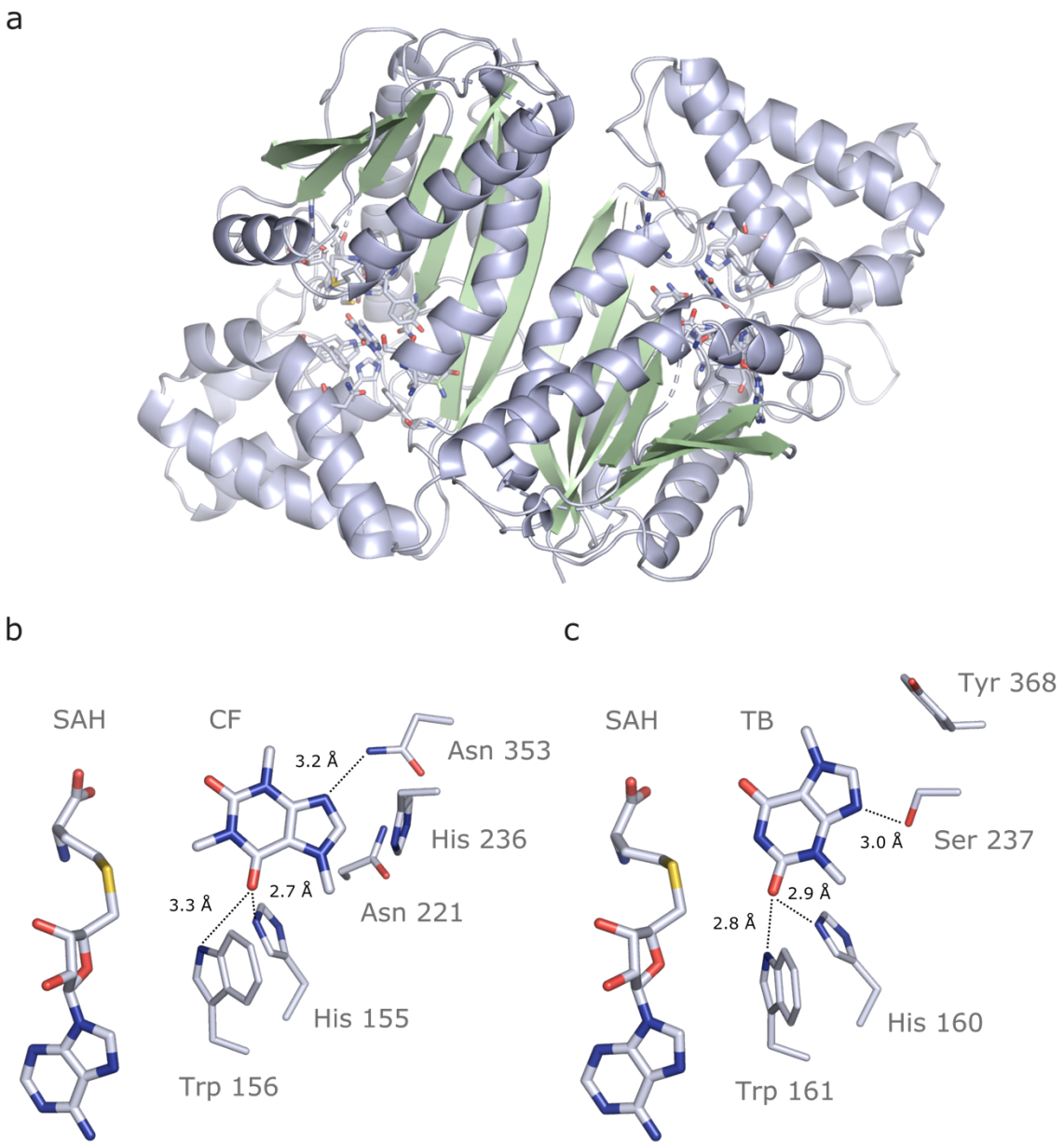
929



930

931 **Figure 5. Ancestral sequence resurrection reveals ancestral xanthine alkaloid pathway**
932 **flux. a**, Simplified evolutionary history of three YM xanthine alkaloid-methylating enzymes and
933 their two ancestors, AncIpCS1 and AncIPCS2. Average site-specific posterior probabilities for
934 each ancestral enzyme estimate are provided. Numbers above each branch of the phylogeny
935 represents the number of amino acid replacements between each enzyme shown. **b**, Inferred
936 pathway flux is shown for the antecedent pathways that could have been catalysed by the
937 ancestral or modern-day combinations of enzymes that would have existed at three time points
938 in the history of the enzyme lineage. Arrows linking metabolites are coloured according to the
939 activities detected from each enzyme shown in panel a. Dotted arrows are shown for

940 AncIpCS1' because it is unknown what characteristics it would possess; it is assumed that it
941 would have at least catalysed the formation of 3X from X since both its ancestor and
942 descendant enzyme do so. X, xanthine; XR, xanthosine; 1X, 1-methylxanthine; 3X, 3-
943 methylxanthine; 7X, 7-methylxanthine; TP, theophylline; TB, theobromine; PX, paraxanthine.
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959



960

961 **Figure 6. Crystal structure of IpCS3 in complex with Caffeine (CF) and S-adenosyl-**
962 **homocysteine (SAH) and comparison with the active site of *Coffea* DXMT. a**, Overview
963 of the crystal structure of IpCS3 (PDBID 8T2G) depicting the active site of the enzyme in
964 complex with CF and SAH. **b**, Relevant residues in IpCS3 for ligand recognition. **c**, Relevant
965 residues in DMXT (PDBID 2EFJ) for ligand recognition. Protein residues are displayed as lines
966 with carbon atoms coloured in bluewhite while small molecules - caffeine (CF), theobromine
967 (TB), and S-adenosyl-homocysteine (SAH) - are drawn as sticks. Colour code for the rest of

968 the atoms: nitrogen (blue), oxygen (red) and sulphur (yellow). Hydrogen bond interactions are
969 indicated as black dotted lines.
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986

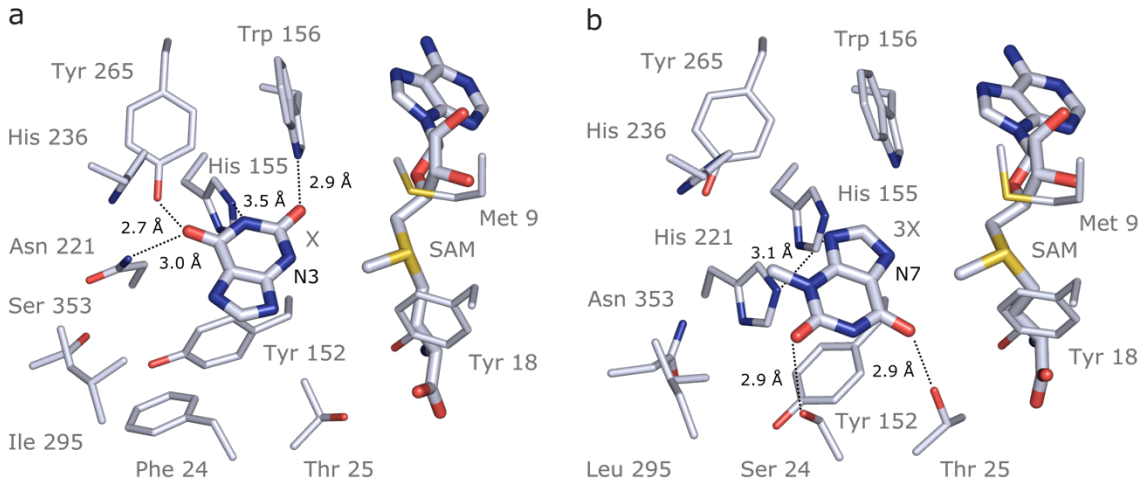


Figure 7. Docking models of xanthine alkaloids in IpCS1 and IpCS2 active sites. **a**, IpCS1-X complex. **b**, IpCS2-3X complex. Protein residues are displayed as lines with carbon atoms coloured in bluewhite while small molecules - xanthine (X), 3-methylxanthine (3X), caffeine (CF), paraxanthine (PX), 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.

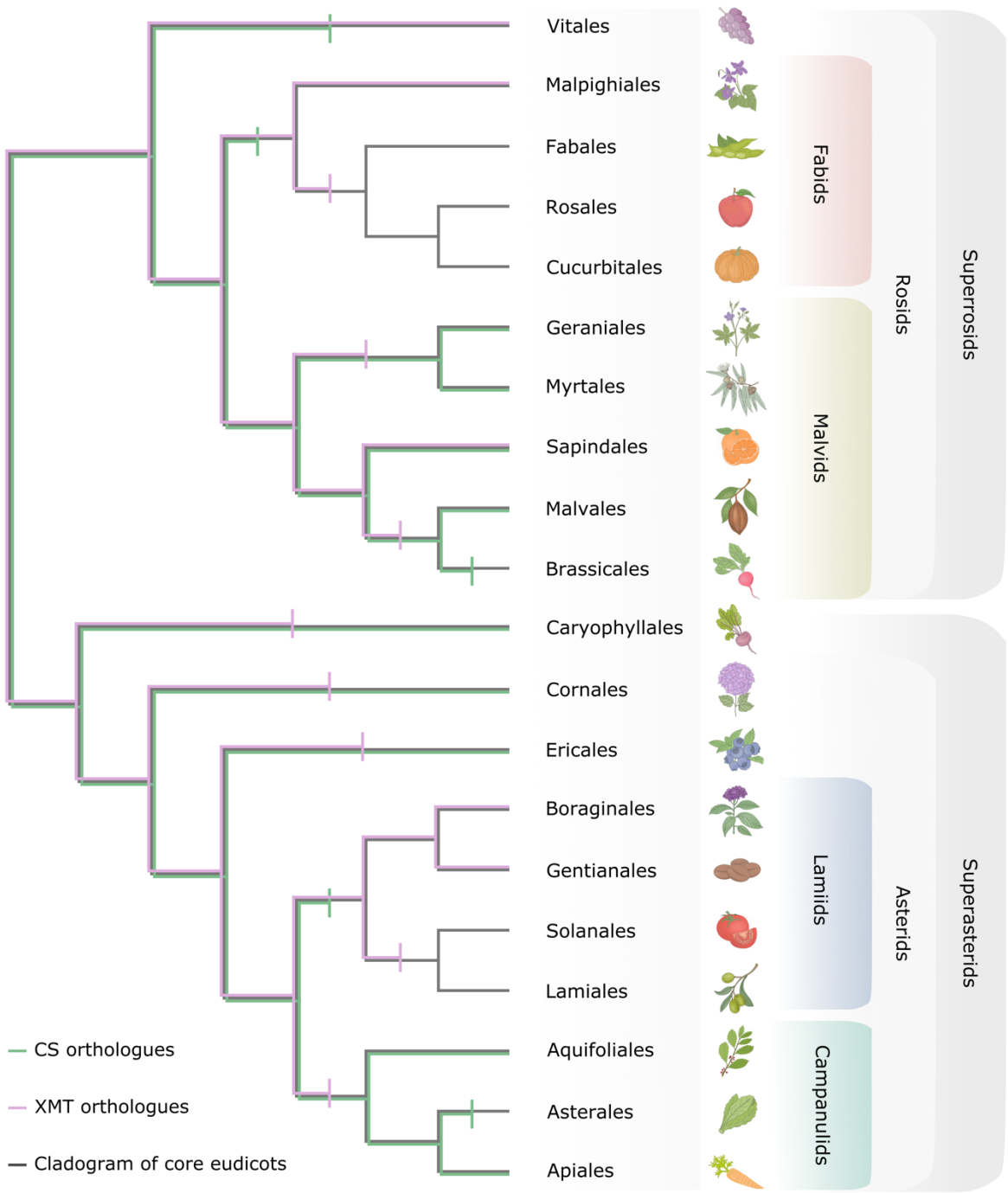


Figure 8. Only CS genes were available for co-option and utilization for xanthine alkaloid biosynthesis in yerba mate whereas coffee only had XMT genes. Both CS- and XMT-type caffeine biosynthetic enzymes were present in the ancestor of core eudicots but numerous apparent losses of one or the other or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