6-2015

Evolutionary Convergence of the Caffeine Biosynthetic Pathway in Chocolate Followed Duplication of a Constrained Ancestral Enzyme

Andrew J. O'Donnell
Western Michigan University

EVOLUTIONARY CONVERGENCE OF THE CAFFEINE BIOSYNTHETIC PATHWAY IN CHOCOLATE FOLLOWED DUPLICATION OF A CONSTRAINED ANCESTRAL ENZYME

by

Andrew J. O'Donnell

A thesis submitted to the faculty of the Graduate College
in partial fulfillment of the requirements
for the degree of Master of Science
Biological Sciences
Western Michigan University
June 2015

Thesis Committee:

Todd Barkman, Ph.D., Chair
Yan Lu, Ph.D.
Andre Venter, Ph.D.

EVOLUTIONARY CONVERGENCE OF THE CAFFEINE BIOSYNTHETIC PATHWAY IN CHOCOLATE FOLLOWED DUPLICATION OF A CONSTRAINED ANCESTRAL ENZYME

Andrew J. O'Donnell, M.S.

Western Michigan University, 2015

Caffeine biosynthesis is widely distributed in flowering plants and requires three consecutive methylation steps of xanthine alkaloids. Genes that have previously been reported to participate in the multi-step pathway in *Coffea sp.* (coffee) and *Camellia sinensis* (tea) encode members of the SABATH family of methyltransferases. Two genes highly expressed in fruits of *Theobroma cacao* (cacao) are orthologous to the caffeine genes in tea and appear to have diversified following gene duplication. Biochemical characterization of the enzymes (XMTs) encoded by these genes strongly suggest an unprecedented major pathway to theobromine, a precursor to caffeine. These findings imply that caffeine biosynthesis evolved convergently in plants and raise two major questions about the evolution of the caffeine pathway in cocoa: 1) What ancestral conditions led to duplication, evolution, and maintenance of paralogous XMTs in the evolution of theobromine accumulation in cacao fruits? and 2) How has gene duplication played a role in the independent evolution of flux through the caffeine pathway? To answer these questions, two ca. 50-million-year-old enzyme ancestors of the cocoa XMTs were resurrected in vitro and biochemically characterized with xanthine alkaloids as substrates.

ACKNOWLEDGMENTS

I would like to first and foremost acknowledge my thesis advisor, Dr. Todd Barkman, for the role he has played in teaching me the skills necessary for a career as a scientist. Through my interactions with him in the past years, I have learned that many surprising breakthroughs come only by challenging conventional thinking. He has inspired me to pursue a career in research, and to tell the evolutionary story contained in this thesis.

I also owe great thanks to my committee members, Dr. Andre Venter and Dr. Yan Lu, for providing strong theoretical training in analytical chemistry and plant physiology that will serve me well for the rest of my career, and for providing thoughtful feedback on my project over the duration of my master's training.

I would also like to thank my fellow students, Ruiqi Huang and Jessica Barboline, for making their molecular and cellular expertise available to me, and Celene Jackson, Chris Jackson, Dr. John Geiser, and Dr. Wendy Beane for providing equipment and materials in support of my training. I also owe thanks to Dr. Pamela Hoppe, Dr. Don Kane, Dr. Rachel Warga, and Dr. Doug Coulter for providing discussion and career advice.

Thanks also to my family for supporting me in all my career decisions.

Andrew J. O'Donnell

TABLE OF CONTENTS

Table of Contents—continued

Table of Contents—continued

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Caffeine is arguably the most economically important plant alkaloid due to its natural presence in a variety of commercial crops such as coffee, tea, and cacao. The world produced 6.7 million tons of coffee, 2.15 million tons of black tea, and 3.7 million tons of cocoa per year in 2000 (1). At one point during the 20th century, coffee comprised more than half of the total exports of Brazil and is still an important crop today (2). In beans of coffee, caffeine is thought to contribute to flavor properties such as bitterness (3). Similarly, high concentrations of caffeine precursors in cocoa beans correlate significantly with bitterness (4). Thus, understanding the link between genes and the biochemical mechanisms of caffeine biosynthesis in plants has garnered considerable attention, especially in the past 20 years (Ashihara 2013, book chapter). The following introductory sections describe what is known about the biosynthesis, ecology, and evolution of caffeine in flowering plants, and set the stage for hypothesis tests of evolutionary convergence on caffeine biosynthesis in *Theobroma cacao*.

Ecological roles for caffeine

Various flowering plants are reported to accumulate caffeine (1,3,7 - trimethylxanthine) and its precursors in at least one tissue type (Table 1). It was first shown over 100 years ago that caffeine plays an allelopathic role in perturbing the

germination of neighboring seeds in soil (5). Since then, considerable evidence has emerged that caffeine functions as a plant defensive compound in a wide host of ecological interactions. For instance, caffeine applied to the leaves of *Coffea arabica* increased the movement of herbivores in comparison to control groups so that they spent more time moving and less time feeding (6). Various Lepidoptera that were fed caffeine-treated leaves laid fewer eggs and produced eggs with reduced protein and energy content (7). In the leaves of *Theobroma cacao*, caffeine may play a defensive role against wounding by herbivores and pathogen attack (8). Caffeine production is induced when normal leaves are either excised or treated with hormones that elicit physiological defenses in response to pathogen infection (8). Importantly, Aneja *et al.* (2001) demonstrated that treated leaves also significantly perturbed the growth of the pathogenic fungus *Crinipellis perniciosa*.

In addition to its functions in plant defense, a surprising role for caffeine in pollination is supported. Caffeine was shown to be present in both the nectar of flowers of *Tilia cordata* and in the guts of bees following pollination (9). Years later, it was shown that caffeine in *Citrus* nectar enhances the memory of reward in honeybees by triggering physiological mechanisms of memory formation in their brains, thus improving the likelihood that each flower will be

Table 1 Caffeine production in well-studied species of flowering plants

Latin Name	Major Xanthine Alkaloid	Known Tissue(s)	Reference
Camellia sinensis	caffeine	leaves[‡], flowers	(10, 11)
Citrus sp.	caffeine[†]	flowers[‡], leaves	(12)
Coffea arabica	caffeine	seeds[‡], leaves, seedlings, roots	(13-15)
Cola nitidia	caffeine	seeds	(16)
Ilex paraguariensis	caffeine	leaves	(17)
Tilia cordata	caffeine[†]	flowers	(9)
Theobroma cacao	theobromine	seeds[‡], leaves, shoots	(18-20)
Paullinia cupana	caffeine	seeds[‡], leaves, stems	(21)

[†] *abundance relative to at least one other xanthine alkaloid not quantified*
[‡] *putative primary site of xanthine alkaloid biosynthesis*

revisited by an affected bee (22). Not surprisingly, a study in 2015 by Thomson *et al.* demonstrated that artificial flowers with caffeinated nectar received more pollination service than those with no caffeine present (23).

Biosynthesis of caffeine and other xanthine alkaloids in Coffea and Camellia

Caffeine biosynthesis requires the sequential N-methylation of xanthine alkaloids (Figure 1). In the pathway, xanthosine, and xanthine bases, are methylated by enzymes using the universal methyl-donor S-adenosyl-L-methionine (SAM) (24). In leaves and seeds of *Coffea arabica*, caffeine is the major purine alkaloid (14, 25). In young *C. arabica* leaves incubated with radiolabeled adenine, a significant proportion of radioactivity was incorporated into theobromine and caffeine (14). In the study by

Ashihara *et al.* (1996), the conversion of adenine to caffeine appeared to slow down as leaves aged. At all leaf stages, xanthine and 7-methylxanthine with incorporated radioactivity were present at low concentrations (14). Theophylline was not detected at any life stage when leaves were incubated with labeled precursor. Similar results were obtained when leaves of each stage were incubated with radiolabeled guanine in that much of the provided precursor was incorporated into theobromine and caffeine in the younger tissues, with the conversion slowing down as leaves matured (14). In the same study, pulse-chase using labeled xanthosine resulted in significant incorporation into 7-methylxanthine, theobromine, caffeine, and xanthine. As reviewed by the authors, xanthine is not likely a key precursor to caffeine based on results from previous studies that show either negligible or no conversion of this molecule to caffeine. From these data the authors argue that caffeine biosynthesis is initiated by the degradation of purine nucleotides, and

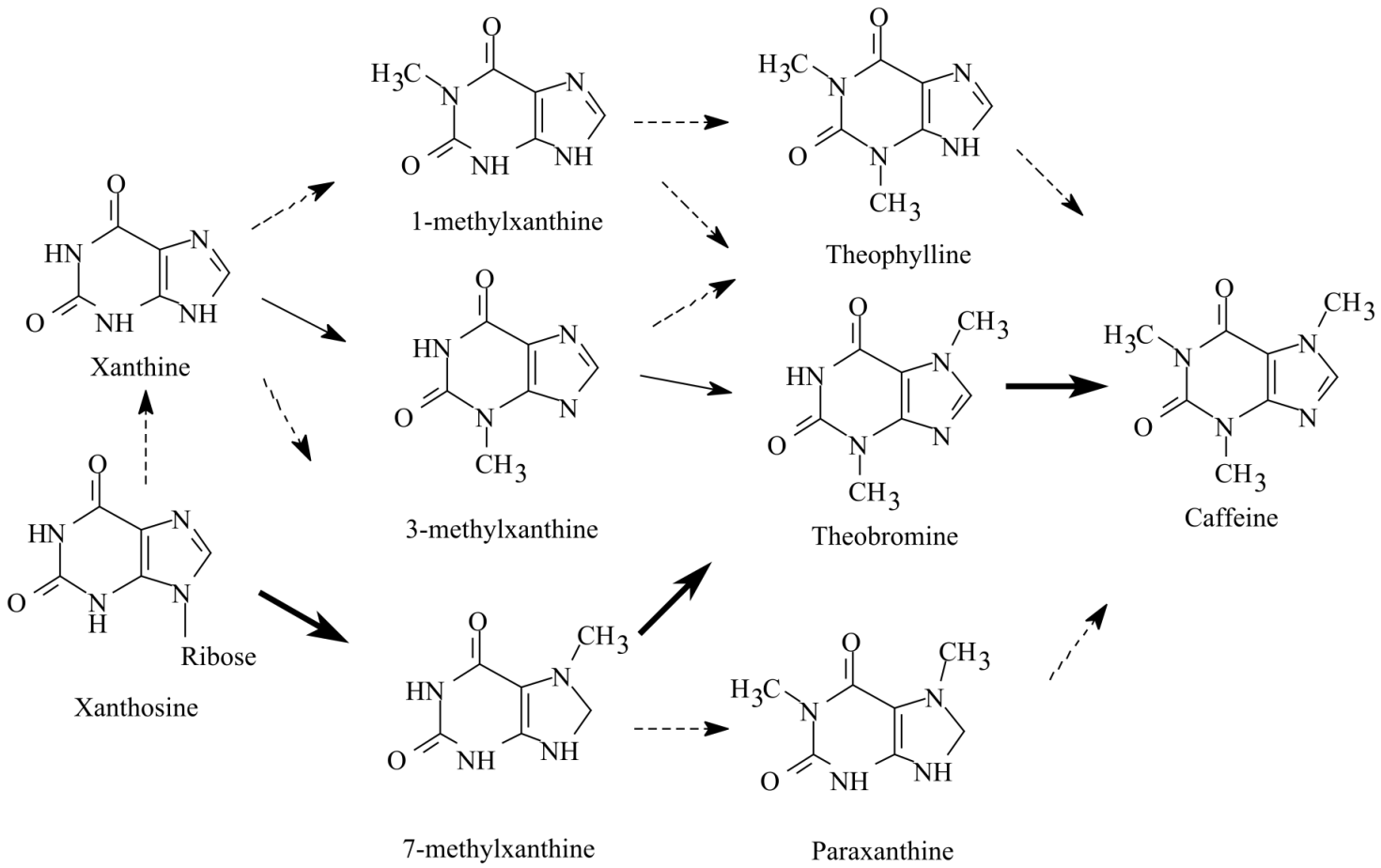


Figure 1. Potential pathways for caffeine biosynthesis in plants. Arrows represent a methyl transfer reaction from SAM to a ring nitrogen atom. The putative major pathway in plants is indicated by bold arros. Thin black arrows indicate a minor route suggestesed by Koyama et al. (2003). All other possible routes are represented by dashed arros. Conversion of xanthosine to the 7-methylxanthosine intermediate is not shown.

that the conversion of xanthosine to caffeine via 7-methylxanthine and theobromine comprises the major biosynthetic route (Figure 1) (14).

The first published SAM-dependent N-metyhyltransferase from *C. arabica* capable of methylating purine alkaloids was cloned from leaves (26). This enzyme, named by the authors CaMXMT, preferentially methylated 7-methylxanthine in the 3-N position, thus forming theobromine. CaMXMT appeared to be a highly specific enzyme, as no secondary activity was detected when assayed with all of the purine precursors shown in figure 1 (26). Moreover, the same study used transgenic onion to determine subcellular localization patterns of CaMXMT. The results of these latter experiments suggested that CaMXMT naturally localizes to the cytoplasm of plant cells (26).

Uefuji *et al.* (2003) later discovered the identities of three distinct N-methyltransferases from *C. arabica* fruits that together reconstitute the conversion of xanthosine to caffeine (27). The three enzymes (named by the authors CaXMT1, CaMXMT2, and CaDXMT1) were each found to be specialized enzymes. CaXMT1 was shown to have activity only with xanthosine, methylating it in the 7-N position to produce 7-methylxanthine (27). This result is consistent with the purine alkaloid metabolism study by Ashihara *et al.* (1996), which suggests that 5' nucleosidase activity occurs during or immediately after transmethylation (14). CaMXMT2 was found to have an activity profile highly similar to that of CaMXMT1 in that it preferentially methylates 7-methylxanthine to produce theobromine, with minor secondary activity with paraxanthine. CaDXMT1 appeared to prefer paraxanthine with lesser activity with 7-

methylxanthine and theobromine and very high activity with paraxanthine. However, because none of the other enzymes studied were capable of producing paraxanthine, the authors conclude that this is simply a fortuitous secondary activity and that theobromine is the genuine substrate for CaDXMT1. Indeed, when all three enzymes are present *in vitro* with SAM and xanthosine as substrates, 7-methylxanthine, theobromine, and caffeine were detected as products (Figure 1) (27).

Considerably less is known about the identities of N-methyltransferases participating in caffeine biosynthesis in *Camellia* (Figure 1). Kato *et al.* (1999) partially purified a caffeine synthase (CS) from *Camellia sinensis* (28). The CSs they isolated appeared to be somewhat promiscuous, with its highest activity being 3-N methylation of paraxanthine, and lesser activity in decreasing order with 7-methylxanthine, theobromine, 3-methylxanthine, 1-methylxanthine, and only trace levels of activity with theophylline and xanthine. However, due to the previous finding by Kato *et al.* (1996) that paraxanthine biosynthesis is limited in *C.* sinensis leaves, the authors argue that 3-N methylation to produce theobromine is the primary function of this enzyme (29). Moreover, due to their finding of secondary activity with theobromine to produce caffeine, they demonstrate a likely dual-function role for this enzyme in catalyzing the last two steps of caffeine biosynthesis in *C. sinensis* leaves (Figure) (28). Kato *et al.* (2000) later cloned CS out of *C. sinensis* leaves, sequenced the gene, and renamed it TCS1. They also found that it shared only 40% sequence identity with non-N-methyltransferase SABATH members (30).

Assuming that purine alkaloid biosynthesis genes in other species are similar in sequence to TCS1, Yokoyama *et al.*(2006) used the coding sequence of TCS1 to clone homologous sequences expressed in young leaves of *C. sinensis, C irrawadiensis, C. ptilophylla,* and *T. cacao*. With an interest in characterizing CS enzymes from other theobromine and caffeine-producing species, and to understand the biochemical basis for diversity in purine alkaloid metabolism, they expressed and characterized the homologous enzymes *in vitro* (31). They found that TCS1 has a close ortholog in the closely-related species *C. irrawadiensis* (ICS1) and *C. ptilophylla* (PCS1). After characterization of the encoded enzymes, they found that the activity profiles of ICS1 and PCS1 were similar to TCS1 in that they preferred 7-methylxanthine and were capable of converting it to theobromine (Figure 1). Interestingly, they also found a second sequence (TCS2) expressed in *C. sinensis* tissues with a predicted protein product that shares 89% sequence identity with TCS1. TCS2 from *C. sinensis* also had orthologs in *C. irrawadiensis* (ICS2) and *C. ptilophylla* (PCS2) (31).

Although they were unable to detect any activity from TCS2, ICS2, or PCS2, they were able to clone and characterize an N-methyltransferase with sequence homology to XMTs from *Camellia* from the leaves of *T. cacao* (31). An interesting result was that this enzyme, named BTS1, was found to be 55% identical to the CS enzymes from *Camellia* and only 40% identical to those from *Coffea*. Characterization of BTS1 showed that it had exclusive activity with 7-methylxanthine, converting it to theobromine (Figure 1) (31). The authors do, however, report a *K*m value of 2.4 mM of BTS1 with 7-

methylxanthine, which is substantially higher than those reported for ICS1 (51 μM) and PCS1 (85 μM) with the same substrate. Due to the ability of these enzymes to function as xanthine alkaloid methyltransferases, the class of enzymes capable of participating in caffeine biosynthesis will hereafter be referred to as XMTs.

Purine alkaloid metabolism in Theobroma cacao

Theobromine is the principal purine alkaloid in *T. cacao* leaves and fruits in most tissues stage (18, 19). Koyama et al. (2003) showed that young leaves had the highest purine alkaloid concentration of any other leaf stage, 75% of which was converted to salvage products or CO_2 in older stages (19). Young leaves that were supplied with the radiolabeled purine bases and nucleosides adenine, adenosine, guanine, guanosine, hypoxanthine, inosine incorporated each into theobromine, 7-methylxanthine, 7-methylxanthosine, and 3-methylxanthine (19). Xanthine and xanthosine were also supplied to young leaves and was incorporated into 3-methylxanthine and theobromine. However, their incorporation into 7-methylxanthine or 7-methylxanthosine could not be detected (19). Of the purine alkaloids, theobromine invariably received the most incorporated radioactive material regardless of the precursor.

When radiolabeled theobromine was supplied to young leaves, about 12% was converted to caffeine after incubation (19). Not surprisingly, no conversion of theobromine to caffeine could be detected in older leaf tissues. As tissue stage advanced, increasing proportions of theobromine were found to be converted to 3-methylxanthine,

while conversion to 7-methylxanthine could not be detected at any stage (19). Similarly, supplied caffeine was converted to 3-methylxanthine in some stages, while none was found to be incorporated into 7-methylxanthine. In all leaf stages, caffeine was converted to theobromine and theophylline (19). These results suggest that xanthine alkaloids that are produced in young leaves begin to degrade in older stages following a pathway of caffeine -> theobromine -> 3-methylxanthine -> xanthine -> salvage products + CO_2.

Taken together with evidence from *Coffea* and *Camellia*, the authors argue that a major route to theobromine biosynthesis in *T. cacao* leaves involves the conversion of xanthosine to 7-methylxanthosine and 7-methylxanthine. Due to the metabolism of xanthine, they also suggest that a minor pathway is the conversion of xanthine to 3-methylxanthine, and 3-methylxanthine to theobromine (19). There is additional support for this minor pathway in plants, as it appears that a similar one is utilized in *Camellia ptilophylla* (32).

Endogenous levels of purine alkaloids appear to be similar in *T. cacao* fruits in that theobromine is higher than caffeine in nearly all tissues and stages (18). Zheng et al (2004) show, however, that a major difference in fruits is that theobromine production appears to accelerate in older tissues, as the highest concentrations of theobromine per gram of fresh tissue was found in ”stage c” fruits. Concentrations of theobromine were most extreme in the cotyledon (18). Coinciding with the increase in theobromine concentration is an increase in all mono- and di-methylated purine nucleosides and bases with the exception of theophylline, which appeared to decrease slightly (18).

As was the case in *T. cacao* leaves (19), supplied adenine was converted to theobromine in young fruits. However, this effect is pronounced in stage c cotyledons, consistent with the observation that theobromine production increases with tissue age in fruits. It is important to note that adenine was also incorporated into 3- and 7-methylxanthine in young fruits, as was the case in leaves (18, 19). As might be expected by the high ratio of endogenous theobromine to caffeine in fruits, conversion of supplied theobromine to caffeine was not detected. As was the case in leaves, supplied theobromine and caffeine was incorporated into 3-methylxanthine and other degradation products (18). The results by Zheng *et al.* (2004) also indicate a pathway of caffeine degradation via theobromine to 3-methylxanthine, and finally xanthine.

The SABATH family of methyltransferases

One of the results by Yoneyama *et al* (2006) was that XMTs from *Coffea*, *Camellia*, and *T. cacao are* phylogenetically related to members of the SABATH family of methyltransferases, as will be discussed in more detail below (31). To date, all characterized SABATH members transfer a methyl group from the SAM to oxygen, sulfur, or nitrogen atoms on their substrates (33). Structural studies have shown that SABATH members (SAMT from *Clarkia breweri* , XMT and DXMT from *Coffea canephora,* and IAMT from *Arabidopsis thaliana*) exist as homodimers in solution and possess an α-helical capping domain and a highly conserved SAM-binding domain (34-36). Like SAMT and IAMT, each monomer of XMT and DXMT homodimer complexes

has its own independent active site. The dimerization interface itself is composed mostly of hydrophobic residues and is highly conserved in sequence between XMT, DXMT, and SAMT (35). All three structural studies revealed that substrates are orientated proximally to the SAM methyl donor with receptor atoms facing towards the reactive methyl group within the active site in preparation for transmethylation.

Because this study focuses on the evolution of XMTs, only the structures of XMT and DXMT will be reviewed in detail. Interestingly, the overall structures of XMT and DXMT are nearly identical to each other (35), which might suggest that enzymes within this family achieved their specificities through mutations that subtly alter binding preference. In fact, the majority of differences between the two enzymes occur in loop regions within or near the xanthine alkaloid binding pocket. The SAM binding region occurs near the purine alkaloid pocket and is highly conserved in sequence, shape and position between resolved SABATH enzymes. SAM appears to be held in position through hydrogen-bonding, and through van der Waals forces such as π-stacking between SAM's adenine ring and nearby amino acid residues. Substrates in all characterized members of this family are also held in place through hydrogen bonding, and through van der Waals interactions such as hydrophobic interactions between the purine ring and active site residues.

Evolution of xanthine methyltransferases

Across the angiosperm phylogeny, the presence of caffeine is polyphyletic and appears to be thinly distributed (Figure 2). For example, in the genus *Camellia* it is

known that *C. irrawadiensis* and *C. sinensis* produce theobromine and caffeine in leaves, respectively. However, *C. japonica* is not known to make either theobromine or caffeine (37). It is possible that plants producing xanthine alkaloids have not yet been discovered to have this trait. However, due to the fact that only a few members of each order are reported to produce caffeine, the most parsimonious explanation for this distribution is that caffeine biosynthesis is a relatively recent innovation and has evolved independently multiple times.

The gene phylogeny behind the species distribution of caffeine corroborates the convergence hypothesis in a striking way; the phylogenetic tree by Yoneyama et al (2006) shows that XMTs in *Coffea* and *Camellia* appear to have duplicated sometime after the divergence of these two lineages. Moreover, their finding that the *Camellia* XMTs are more similar to the SAMT family of SABATH methyltransferases than they are to *Coffea* XMTs strongly suggests the independent recruitments of different SABATH members for caffeine biosynthesis in caffeine-producing lineages. A recent study by Denoeud *et al*. (2014) also highlighted the prevalence of lineage-independent tandem gene duplications of XMT genes in *Coffea*, *Camellia,* and *T. cacao,* lending credence to the paradigm that caffeine biosynthetic machinery has recently diversified in each lineage.

Although it is readily apparent that duplicate XMTs arose and were maintained following the divergence of caffeine-producing lineages, it is not yet known to what extent convergence explains the independent acquisitions of flux through this pathway.

Due to the observation that some caffeine synthase genes have putative dual roles, as is the case with TCS1, it is possible that an ancestral MT in the predecessors of *Coffea*, *Camellia*, and *T. cacao*, potentially dating back to the ancestor of all core eudicots, was able to perform either some or all steps of the caffeine pathway. However, the scenario of an ancient caffeine-producing ancestor is less favorable given the tenuous distribution of the caffeine phenotype in flowering plants. If, however, the ancestral MTs were always methylating xanthine alkaloids, perhaps functioning in the first step of the pathway, we would conclude that convergence on theobromine biosynthesis occurred in each lineage through the *de novo* acquisition of the second methylation reaction following gene duplication. A second possibility is that ancestors of the enzymes performing the modern-day pathway had primary functions outside of methylating xanthine alkaloids. This would imply that gene duplications somehow permitted the recruitment of an ancient SABATH gene towards its new function in the caffeine pathway. This second scenario would implicate convergence of the entire biosynthetic

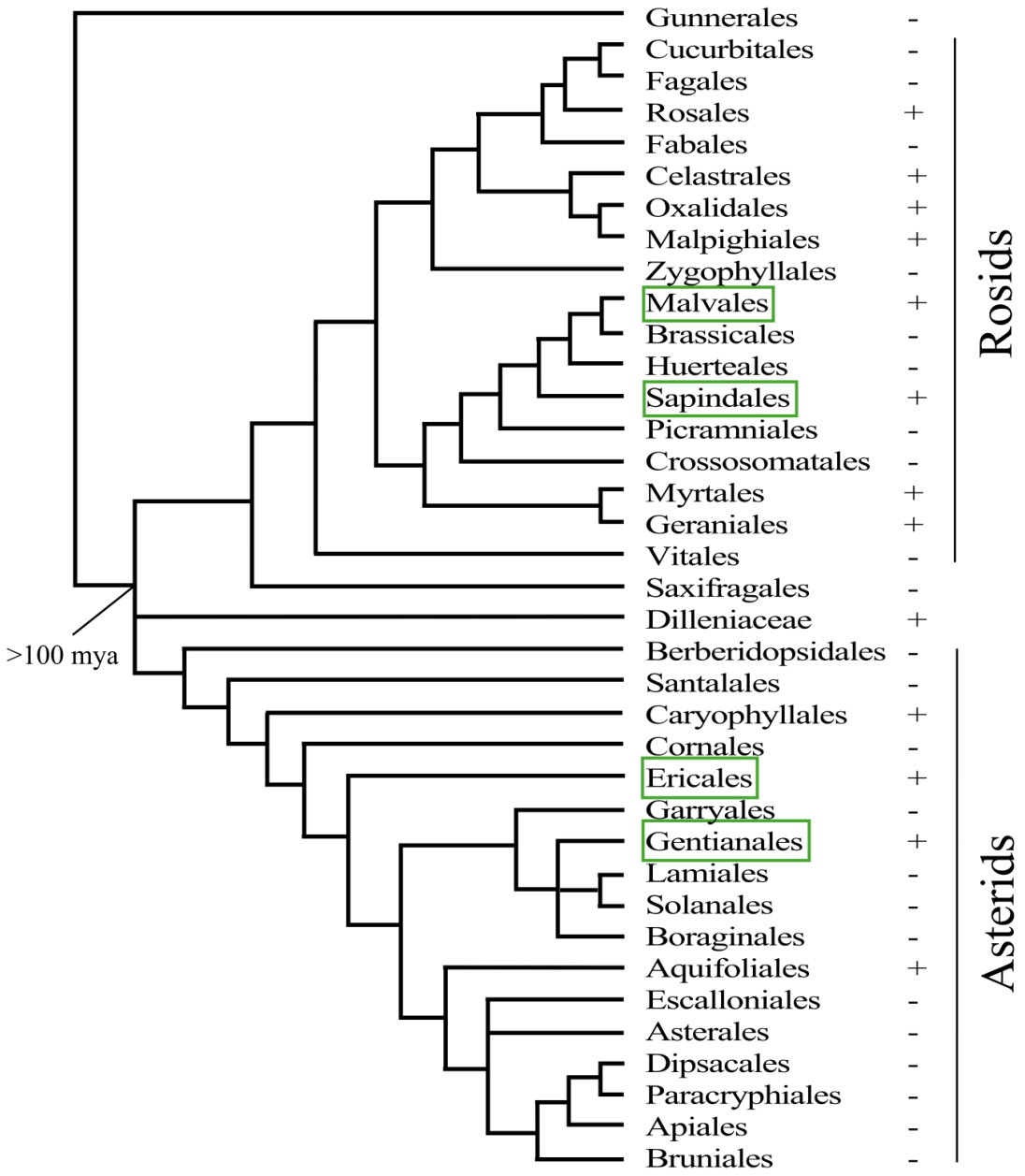


Figure 2. Phylogeny of the core eudicots (APG 2007) and distribution of caffeine biosynthesis. Lineages known to produce caffeine based on literature reports are shown with a "+". Lineages that have been sampled but show no evidence for caffeine accumulation or remain unsampled are marked with a "-". Orders with species discussed

in this study are in green boxes - Malvales (T. cacao), Sapindales (P. cupana), Ericales (C. sinensis), and Gentianales (C. arabica).

route in caffeine-producing species through molecular co-option.

An interesting additional layer to the story is that pre-duplicated genes must have been maintained for over 100 million years if indeed BTS1 is orthologous to the CS genes from *Camellia* (Figure 2). **This observation, and the above hypotheses about the nature of convergence, begs the characterization of the ancestral conditions preceding the relatively recent gene duplications. Examining these characteristics would be a robust test for hypotheses of convergence and would generally help us to understand what conditions led to convergence on caffeine production and to the diversity we see in modern-day xanthine alkaloid metabolism.**

Models of Enzyme Evolution

In order to understand the genetic underpinnings of evolution of the caffeine pathway, it is important to review how gene duplications can lead to novel traits. The role of gene duplication in the evolution of new phenotypes has been debated for over 45 years. It was first proposed by Ohno in 1970 that gene duplication frees one gene copy from selection to acquire a new function (neofunctionalize) while foregoing the old function because one paralog is available to continue performing its original ancestral

role (Ohno 1970) (Figure 3a). Due to the rarity of gain-of-function mutations with respect to degenerative ones, an inherent property of Ohno's model is that most duplicate genes are either lost during segregation or become pseudogenes. The fact that gain-of-function mutations are less common than null mutations, and that neutral alleles degenerate or are lost quickly, Ohno's model raised an important question about fates of duplicate copies - How is a duplicate gene maintained for long enough in the absence of selection to acquire neofunctionalizing mutations (38)?

One elegant way around this problem was to invoke a model, known as subfunctionalization, in which all modern-day functions existed in the ancestor as subfunctions and are partitioned between duplicate copies via selectively-neutral loss-of-function mutations (Figure 3b). This process could rapidly lead to the maintenance of paralogous genes because each copy becomes an obligate participant in performing the entire role of the ancestral gene (39). Indeed, it has been shown empirically that subfunctionalization sufficiently explains the fates of at least some duplicated genes in nature (40).

As discussed by Bergthorsson et al., subfunctionalization does not explain the *de novo* origins of adaptive phenotypes (38). They therefore proposed the Innovation, Amplification, and Divergence (IAD) model of gene evolution (Figure 3c). In their model, novel secondary activities evolve fortuitously preceding gene duplication (innovation). The gene duplicates in the presence of positive selection for the secondary function, resulting in elevated levels of gene product and amplification of the beneficial,

secondary trait (amplification). Due to selection for the secondary activity, paralogs could be maintained for long enough for one copy to specialize on the secondary function (divergence). Additionally, the increase in gene copy number also increases the number of mutational targets, thus increasing the likelihood of at least one copy to acquire function-improving mutations. The IAD model would explain how novel

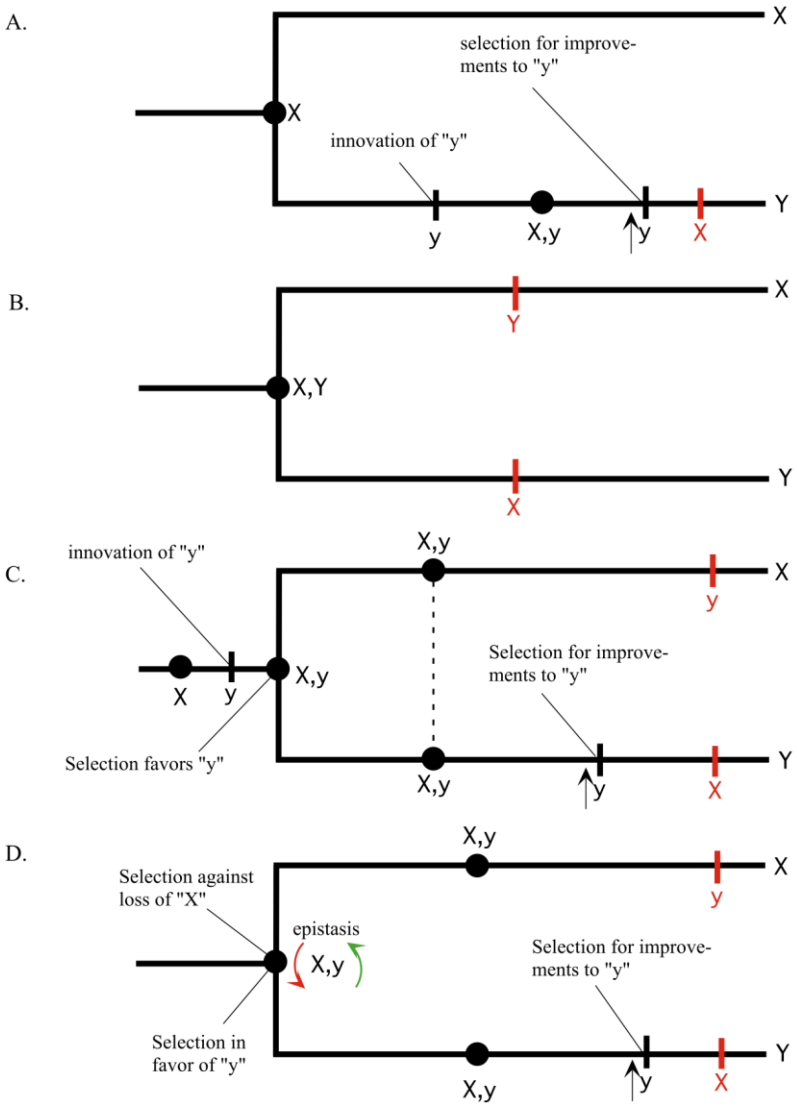


Figure 3 Prominent models of protein evolution following gene duplication. **A.** Classical neofunctionalization **B.** subfunctionalization **C.** Innovation-amplification-divergence, and **D.** escape from adaptive conflict. Black circles and branches represent ancestral proteins and protein evolution via mutation, respectively. Capital and lower-case letters represent primary and secondary functions, respectively. Red letters represent functional loss of an activity. Maintenance of paralogs due to selection for protein dosage is indicated by a dotted line. Upward-pointing black arrows representmutations that cause an improvement to the indicated function. Curved red and green arrows in a functional profile indicate epistatic interactions that would cause a tradeoff of one function for

another, respectively.

activities are able to evolve in descendant genes before they are lost. A study in 2012 showed that the IAD model explains the adaptive evolution of novel metabolic routes in amino acid biosynthesis in bacteria (41).

Evidence has arisen in recent years that some proteins experience an adaptive conflict before duplication (42-44). That is, a protein is constrained from improving simultaneously on multiple functions due to pleiotropic effects of mutation (42). For example, a protein with subfunctions A and B loses efficiency in performing function A by acquiring gain-of-function mutations for function B. This model is known as escape from adaptive conflict (EAC) (Figure 3d). In the presence of selection for both functions, gene duplication could resolve this conflict by allowing each encoded daughter protein to acquire adaptive mutations without the consequence of losing either function. This evolutionary process appears to explain the evolution of antifreeze proteins (AFPs), which perturb the formation of ice crystals in arctic fishes, from an enzyme that performs an unrelated function as a sialic acid synthase (SAS) (42). The study by Deng *et al.* (2010) elegantly demonstrated that the patterns of evolution surrounding AFP neofunctionalization fit the three predictions of EAC: 1) the ancestral enzymes were probably bi-functional, 2) the selected function was constrained from improving, and 3) adaptive changes occurred in the descendant gene copies following duplication (42).

Evolution of biochemical pathways

Protein evolution following gene duplication is often the focus in studies on the evolution of new biochemical pathways (45), and likely explains part of caffeine evolution. Generally, plants are famous for their vast repertoire of specialized chemicals, many of which have applications ranging from human health to use as biofuels. A central goal in plant biochemistry is to achieve greater yields of metabolic products that are of interest to humans. Thus, many studies aim to understand the mechanisms by which flux through a pathway is either controlled or constrained. Perhaps the most common target of such studies is the expression level of rate-limiting enzymes that participate in a pathway of interest, whereby an enzyme suggested to play a controlling role in flux through a pathway is overexpressed *in vivo* and the resultant product concentrations, and in some cases their organismal fitness effects, are assessed (46, 47). Another common pursuit in metabolic engineering is to improve kinetic parameters such as substrate specificity or product turnover rate by introducing mutations to a pathway enzyme that are predicted have an effect on its biochemical properties (48, 49). In more dramatic cases, entire subunits have been incorporated into engineered proteins in order to achieve kinetic improvements, as was the case in engineering Rubisco to have a higher product turnover rate in transgenic rice (50).

The evolutionary process is a metabolic engineer with 3.5 billion years worth of success. In the evolution of plant secondary metabolism, enzymes that participate in core

metabolic pathways are recruited towards a new function (51), and this can result in the production of novel compounds that are visible to selection. There is a growing volume of research that suggests an important role for gene duplication and operon-like gene clusters in the evolution of new metabolic pathways (45, 52). One view with nascent empirical support is that pathway flux control is unevenly distributed between enzymes, with it being most heavily vested in early-pathway enzymes (53, 54). However, in the evolution of new biochemical pathways, the genetic mechanisms that drive new patterns of flux remain unclear. For example, an increase in pathway flux could be achieved through elevated expression of early-pathway enzymes. Conversely, rate-limiting enzymes could acquire mutations that enhance a kinetic parameter such as substrate specificity.

In the subfunctionalization model discussed above, the partitioning of ancestral functions through degenerative mutations into descendent copies of genes following duplication could provides some advantage in that descendants are able to specialize on separate functions (39). However, due to the notion that duplicate genes could confer positive enzyme dosage effects, and that mutations acquired throughout evolution can improve enzyme function (38), it is unclear whether degenerative mutations themselves play a role in improving the flow of metabolites through a pathway. Thus, it is of great interests to biochemists and evolutionary biologists alike how enzymes and their networks ultimately achieve improvements in product accumulation. **In addition to investing the ancestral conditions surrounding the story of convergence on caffeine**

biosynthesis, this work examines how historical mutations that followed gene duplication in the caffeine biosynthetic pathway of *Theobroma cacao* resulted in an improved metabolic network, and why two enzymes are simply better than one.

Hypotheses with associated predictions and experiments in this study

The first question being asked is whether or not methylation of xanthine alkaloids for the biosynthesis of caffeine and its precursors was inherited through the common ancestors of caffeine-producing species. To address this question, the evolutionary history of xanthine alkaloid metabolism in *T. cacao* will be examined and compared to the available data in other caffeine-producing species. Here it is hypothesized that *T. cacao* recently engineered new enzymatic machinery in favor of theobromine accumulation through gene duplication, independently of *Coffea arabica* and *Camellia sinensis*. The major prediction of this hypothesis is that an ancestral gene inherited by an ancestor of *T. cacao* was not operating in the conversion of xanthosine to theobromine, and that mutations acquired, following duplication, promoted pathway flux. Alternatively, it may be the case that an ancestral enzyme involved in theobromine production was inherited by all caffeine-producing lineages, and that convergence exists only on the evolution of the final methylation steps.

The second question being addressed is why gene duplication is so prevalent among known and predicted XMTs in *Coffea, Camellia,* and *T. cacao*. One hypothesis is that that gene duplication provided additional genetic material for the partitioning of

ancestral activities, which would predict that the activities of modern-day XMTs from *T. cacao* were present in some form in an ancestor of *T. cacao*, either as primary or secondary activities. A second hypothesis is that gene duplication was required for the acquisition of novel primary methylating activities, such as the conversion of 3-methylxanthine to theobromine. This hypothesis predicts that one or more activity in modern-day XMTs in *T. cacao* were absent in the ancestor.

An empirical test of these hypotheses requires a method of statistically inferring and characterizing ancestral enzymes known as ancestral sequence resurrection (ASR) (55). As reviewed by Harms and Thornton (2010), ASR is a mature technique in understanding how diversity in protein sequence and structure contributes to their modern-day phenotypes by taking the evolutionary mutational process into account (56). This so-called "vertical approach" contrasts with what the authors refer to as a "horizontal approach," the latter being the more common method of comparing diversity between diverse modern-day enzymes, thereby extracting information on the relationship between sequence and function. As reviewed by the authors, the vertical approach has the distinct advantage of examining mutations as they occurred through evolutionary time, thus allowing researchers to narrow down the subset of mutations contributing to phenotype by focusing on those that were concomitant with evolutionary functional shifts (56). More generally, ASR has led to unprecedented insight into the evolutionary patterns and processes underlying modern-day organismal diversity (57). Using available genomic and transcriptomic data coupled with phylogenetic analyses and biochemical assays,

enzymes catalyzing the modern-day xanthine alkaloid pathway in *T. cacao* will first be identified and characterized. To assess the characteristics of ancestral enzymes in *T. cacao*, ASR will be used following characterization of the modern-day pathway and *in vitro* analysis of enzyme function will be performed.

MATERIALS & METHODS

Retrieval of SABATH-like coding sequences from the T. cacao genome

To obtain expressed sequences from *T. cacao* fruits that share sequence homology with known SABATH members, a tblastn search was performed on all expressed sequence tag (EST) libraries from *T. cacao* tissues on Genbank using the published BTS1 amino acid sequence (Genbank accession no. AB096699). For the general search algorithm parameters, 20,000 max target sequences were allowed, expected threshold was set to 10, word size to 3, and max matches in a query range was set to zero. For the scoring parameters, the BLOSUM62 parameter was used, with existence and extension gap costs set to 11 and 1, respectively, and a conditional compositional score matrix adjustment was allowed. Only low complexity regions were filtered.

To determine the set of SABATH-like genomic loci that could be coding for XMTs and other SABATH enzymes in *T. cacao* fruits, BTS1 was used to query Phytozome using the tblastn search function within the masked genome database for *T. cacao*. Only genomic open reading frames that corresponded to a gene product cutoff

of between 200 and 399 amino acids were considered further. All genomic loci and ESTs retrieved by the two searches were assembled into contigs using Sequencher v4.1 with minimum match and minimum overlap set to 85% and 20%, respectively.

Phylogenetic analyses of SABATH-like sequences from T. cacao and of all characterized SABATH members

Genomic loci with putative gene products which passed the length screening were translated to amino acids and a multiple sequence alignment (MSA) using the local pairwise alignment algorithm was performed using MAFFT version 7.122b (58). Subsequent minor adjustments were performed manually. A maximum likelihood search was performed in PhyML version 3.1 using the aforementioned alignment. The JTT model of amino acid substitution was used with the amino acid frequency parameter set to "model." The proportion of variable sites was fixed with the p-invar parameter set to zero. Gamma was distributed across sites and its distribution was estimated 4 substitution rate categories were allowed. A BioNJ starting tree was used and a fast, approximate tree topology search operation was performed. The optimal tree topology was used to reconstruct the phylogeny of all SABATH like sequences in the *T. cacao* genome.

To reconstruct a phylogeny of all characterized SABATH sequences, published SABATH members were acquired from Genbank using accession numbers from the literature. All published SABATH members, and two sequences from this study

(TcXMT1 and TcXMT2) were aligned using amino acid sequences. The alignment and phylogenetic analysis was repeated exactly as above. Due to the difficulty of aligning highly divergent regions across SABATH enzymes, alternate analyses were performed exactly as above, but by arbitrarily removing poorly aligned segments of the MSA. This allowed for the assessment of any phylogenetic effects of divergent regions.

Inferring ancestral XMTs

To infer the ancestral sequences of XMT genes in modern-day *T. cacao,* various putatively expressed XMT genes from Malvales and Sapindales (MalvSap XMTs) were acquired using the blast function from OneKP (www.bioinfodata.org/Blast4OneKP/blast) in November of 2013. BTS1 was used as a query sequence in the tblastn search function using the pairwise alignment parameter with the E-value parameter set to 0.01. The search was allowed to retrieve a maximum number of sequence hits of 200,000. To limit the number of gaps in the MSA of MalvSap XMTs, sequences were concatenated when multiple short EST reads occurred within a single species and were thought to be part of the same gene. Sequence alignment and phylogenetic analysis using maximum likelihood were performed using the methods in the previous section.

To reconstruct the ancestral states of XMT-B orthologs in *T. cacao*, a MSA of MalvSap XMTs was used to generate a maximum likelihood tree using the methods above. The MSA of the MalvSap XMTs was converted to codons and used along with the generated tree in an analysis using PAML version 4.2 (59). For the PAML search,

default parameters were used with the exception of using an estimated gamma shape parameter (alpha was not fixed). The joint ancestral estimates for the ancestor of BTS1, TcXMT1 and TcXMT2 (Fig X, node A), and for the ancestor of TcXMT1 and TcXMT2 alone (Fig X, node B) had gaps removed using an amino acid alignment, were renamed AncXMT1 and AncXMT2, respectively, and used for further analysis.

Cloning and expression of modern and ancestral XMTs

AncXMT1, AncXMT2, and TcXMT2, were synthesized by Genewiz in Puc57 vectors and digested at 37 °C for 6 hours using 1.5 ng of DNA and NdeI and BamHI in 30 µl reactions. Linear fragments corresponding to the expected sizes were gel purified using the QIAEXII gel extraction kit according to the manufacturer's instructions. 7.5 µl of purified DNA fragments were ligated into pET15b expression vectors using pet15b double-digested with NdeI and BamHI, and T4 DNA ligase and ligase buffer from New England Biolabs. Reactions were incubated at 16 °C overnight and either stored at -20 °C or used immediately in subsequent steps.

TcXMT1, TcXMT2, and BTS1 were then transformed into Top10 *Escherichia coli* cells using 2 µl of ligation reaction. 200 µl of each transformation reaction were plated onto LB-agar plates containing 0.1 g/ml ampicillin and incubated for 16 hours at 37 °C. Incubation of all subcloning reactions was carried out in a PTC-100™ programmable thermal cycler by MJ Research, Inc. Starter cultures were then seeded with colonies of each transformant, incubated overnight with aeration at 37 °C, and

miniprepped using a Qiaprep spin miniprep kit by Qiagen and eluting with milliQ water. Concentrations of purified DNA were obtained using a Nanodrop 2000 by Thermoscientific. All steps requiring incubation at 37 °C with aeration were carried out using a G24 environmental incubator shaker by New Brunswick Scientific. Samples were sequenced by Genewiz, and 10ng of each plasmid was used to transform and grow BL21*E. coli* cells using the same plating and incubation methods above.

Due to solubility and/or expression issues of TcXMT2 in pET15b, TcXMT2 was cloned into T7 SUMO by Life Technologies with an immediate (no spacer sequence) c-terminal 6x polyhistidine tag according the manufacturer's instructions.

50 ml of transformed BL21 cells in LB broth were grown to an OD_{600} of between 0.6 and 0.8 in the presence of 0.1 g/ml ampicillin. OD_{600} was measured using a Spectronic 21 spectrophotometer by Bausch & Lomb. To induce protein expression, cells were aerated for 6 hours at 24 °C in 1mM IPTG using a Gyrotory® water bath shaker G76 by New Brunswick Scientific. Cells were harvested at 8C by centrifugation using an Eppendorf centrifuge model 5804R at 4k RPM for 15 minutes. Cells were then lysed by resuspending pellets in 4ml equilibration-wash buffer (50mM monobasic sodium phosphate, 300 mM NaCl, 10 mM imidazole, 10% glycerol and 10 mM β-mercaptoethanol) containing 3mg of lysozyme and incubated on ice for 30 minutes. Cell lysate was then gently sonicated 3 times for 30 seconds on ice, with 30 second resting intervals between each sonication using a Sonciator® Ultrasonic Processor XL. Total protein including the overexpressed peptides was isolated by centrifuging cell lysate at

13k RPM at 4 °C for 20 minutes using an Eppendorf centrifuge model 5415R. Supernatant was collected after centrifugation and either used in total protein assays or purified. Protein purification was performed on total protein using Talon cobalt columns as per the manufacturer's instructions.

Biochemical characterization of modern and ancestral XMTs

Relative activities and kinetic parameters were obtained by incubating enzymes with ^{14}C-SAM (Sigma) and xanthine alkaloids as substrates. For relative activity assays, most reactions consisted of 10 µl purified enzyme, 10 µl buffer (250 mM Tris-HCl pH = 7.5), 1 µl of 5 mM xanthine alkaloid substrate dissolved in 0.5M NaOH, and 28.5 µl H_2O. In the case of TcXMT1 expressed in SUMO, 38.5 µl of total protein was used and no water was added. All relative activity reactions were initiated with 0.5 µl ^{14}C-SAM and incubated at 22 °C for 20 minutes. 200 µl of 2-butatnone was used to extract methylated products and 120 µl of the organic phase was added to 3 ml of liquid scintillation cocktail. DPMs were counted for 2 minutes using a PerkinElmer liquid scintillation counter.

Analysis of pathway flux

Because activity with 3MX in AncXMT2 is secondary and generally lower than in TcXMT2, the concentration of AncXMT2 was titrated in 20-minute ^{14}C-SAM assays so that the amount of product was greater than that of TcXMT2. This ensured that

reactions containing AncXMT2 were equal to or given the advantage in carrying out both steps of the caffeine pathway in pathway flux analyses. Two 100-ml cultures of AncXMT2 were grown and induced with IPTG and extracted for total protein as above. Total protein from each 100-ml pellet was purified as above, and the first fraction of each purification was combined. AncXMT2 reactions were 770 µl purified fraction 1, 200 µl Tris-HCl buffer (above), 20 µl of 25mM xanthine in 0.5 mM NaOH, 10 µl of 32 mM SAM (Sigma). After reactions were complete, caffeine in 100% EtOH to a final reaction concentration of 740 µM was added as an internal standard. Reactions containing both modern-day enzymes had 670 µl TcXMT1 fraction 1, 200 µl TcXMT2 fraction 2, SAM and xanthine as in the previous reaction and caffeine in EtOH to a final concentration of 500 µM. Samples were allowed to react overnight at room temperature.

HPLC

All samples used in HPLC analyses were adjusted to a pH of 5-6. Samples were filtered through a polyethersulfone spin column with a molecular weight cutoff of 5 kDa. Samples were concentrated eight-fold via speedvac at room temperature using a Centrivap Concentrator by Labconco.

A 150 x 2.1mm C18 reverse-phase column by Phenomene was used for all HPLC. 10 µl of each sample was injected into buffer A (99.9% H2O, 0.1% TFA) at a flow rate of 0.5 ml/min with a gradient of buffer B (80% acetonitrile, 19.9% H2O, and 0.1% TFA) applied over 11 minutes to a final value of 40% of buffer A. Purine alkaloids were

scanned at 254 and 272 nm. For pathway flux analyses, areas under peaks were calculated and product concentrations were derived from the internal caffeine standards.

RESULTS

Phylogeny and expression of SABATH-like sequences in T. cacao fruits suggest a role for two novel enzymes in xanthine alkaloid biosynthesis

The first objective was to identify the enzymes involved in the caffeine pathway in modern-day cacao fruits. The general approach was to target SABATH-like sequences that are highly expressed in the fruits of *T. cacao*, the primary site of xanthine alkaloid biosynthesis in this species. First, a BLAST search was performed using a previously-reported XMT, BTS1, as a query against the *T. cacao* genome published on Phytozome to determine all genomic SABATH-like sequences that could encode the XMTs and other SABATH members. The search retrieved 43 putative genes with open reading frames, 26 of which translated to amino acid sequences that met a lower cutoff value of 300 residues. A phylogenetic analysis of all 26 putative SABATH-like sequences in the *T. cacao* genome was performed (Figure 4a). One sequence, which was found to correspond to BTS1 in the BLAST search, was closely related to two other genomic sequences, which were named TcXMT1 and TcXMT2 (figure 4a). TcXMT1 and TcXMT2 therefore appear to have arisen from gene duplication and are paralogs of BTS1. It was found TcXMT1 and TcXMT2 share 79% identity with each other, while each shares only about 67% and 64% identity to BTS1, respectively. BTS1 and TcXMT1

were each found to be encoded for by a single locus, while TcXMT2 was found to have two loci with identical exonic regions. The fact that both coding regions for TcXMT2 are identical suggests that one copy arose from a very recent gene duplication.

Next, a BLAST search against expressed sequence tag (EST) libraries on NCBI generated from various tissues of cacao was performed. Using a previously-reported theobromine synthase (BTS1) as a query sequence, the tblastn search function was used to retrieve all SABATH-like sequences expressed in sampled *T. cacao* tissues. The search retrieved 256 ESTs, 222 of which were greater than 40% identical to BTS1. All ESTs were then assembled into contiguous sequences (contigs), the 26 genomic coding regions in figure 4a were matched to each contig, and the number of ESTs representing each coding locus was quantified and categorized according to tissue type of origin (Figure 4a and 4b). TcXMT1 and TcXMT2, were clearly represented to a greater extent than all other SABATH-like sequences retrieved by the search (Figure 4b). TcXMT1 and TcXMT2 encode predicted proteins with estimated sizes of 40.9 and 40.8 kDa, respectively, which is consistent with the previously reported BTS1with an estimated weight of 40.9 kDa. In the fruits, five SABATH-like sequences were found to be expressed - TcXMT1, TcXMT2, BTS1, and two sequences that were named JMTb, based on phylogenetic closeness to AtJMT (data not shown), and MT3. MT3 and JMTb were each represented in fruit EST libraries by only one or two ESTs, respectively.

However, figure 4b shows that TcXMT1 was represented in fruits by 32 ESTs while TcXMT2 had 125 ESTs. TcXMT1 and TcXMT2 therefore outnumber each of the

other three sequences by more than 15-fold in fruits. Moreover, the number of ESTs corresponding to TcXMT1 and TcXMT2 are highest in libraries generated from fruits than from other tissues of *T. cacao* (Figure 4b). TcXMT1 and TcXMT2 are also represented to a greater extent in leaves, which are also known to accumulate caffeine as previously discussed, than in stems, fruits and flowers. Taken together, these results strongly suggest a role for TcXMT1 and TcXMT2 in theobromine accumulation in *T. cacao*. Surprisingly, despite the isolation of BTS1 from leaves of *T. cacao* and methylation activity with 7-methylxanthine reported by Yonayama *et al* (2006), only one EST was found to represent BTS1 in leaves, and two in fruits (Figure 4a and 4b). This result, along with the finding that TcXMT1 and TcXMT2 are represented highly in fruits, suggests that BTS1 has less influence over xanthine alkaloid methylation *in vivo* than its two paralogs.

Phylogenetic relationships between XMTs in Coffea, Camellia, and Paullinia, and their relatives in T. cacao, suggest the independent evolution of caffeine biosynthetic machinery across caffeine-producing lineages

To assess the phylogenetic context of TcXMT1 and TcXMT2 among characterized SABATH methyltransferases, especially the XMTs,

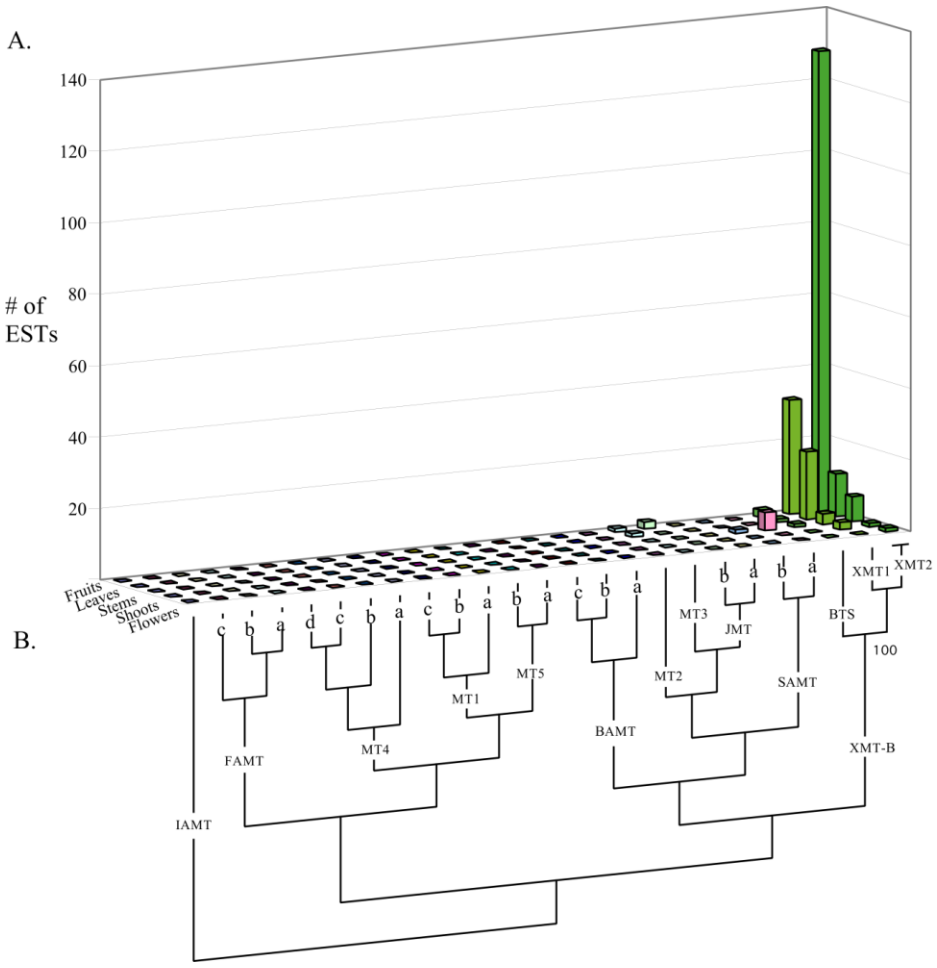


Figure 4. EST numbers and phylogenetic relationships of SABATH-like genomic coding sequences. **A.** EST numbers representing expression of all known SABATH genes in various tissues from the genome of Theobroma cacao. The gene tree shown represents relationships among all known SABATH gene sequences the Theobroma genome. Only the closely related TcXMT1 and TcXMT2 are highly expressed in fruits where theobromine and caffeine accumulate. IAMT (Thecc1EG030787), FAMTc (Thecc1EG019318), FAMTb (Thecc1EG019315), FAMTa (Thecc1EG019314), MT4d (Thecc1EG011287), MT4c (Thecc1EG011286), MT4b (Thecc1EG011290), MT4a (Thecc1EG011291), MT1c (Thecc1EG045368), MT1b (Thecc1EG045372), MT1a (Thecc1EG045370), MT5b (Thecc1EG031006), MT5a(Thecc1EG012604), MT6 (Thecc1EG000168), BAMTc (Thecc1EG000331), BAMTb (Thecc1EG000328), BAMTa (Thecc1EG040854), MT2 (Thecc1EG006850), MT3 (Thecc1EG000336), JMTb (Thecc1EG034091), JMTa (Thecc1EG034089), SAMTb (Thecc1EG000326), SAMTa (Thecc1EG000324), BTS1 (Thecc1EG042576), XMT1 (TcXMT1) (Thecc1EG042578), XMT2 (TcXMT2) (Thecc1EG042587 & Thecc1EG042590). B. Maximum-likelihood

phylogeny of genomic SABATH-like sequences in T. cacao

a separate phylogenetic analysis was performed using these sequences (Figure 5). It was found that TcXMT1, TcXMT2, and BTS1 form a distinct clade that is orthologous to XMTs in *Camellia* and *Paullinia* (Figure 5, green branches). This clade, which was named XMT-B, appears to contain members that underwent gene duplication within *T. cacao* independent of *Camellia* (Figure 5). Specifically, BTS1, TcXMT1, and TcXMT2 are more closely related to each other than they are to PcCS1 from *Paullinia*, which belongs to the order Sapindales (see Figure 2 legend for details). Thus, it appears likely that BTS1, TcXMT1, and TcXMT2 diversified since the divergence of Malvales and Sapindales.

As shown in Figure 5, the XMT-B clade is paralogous to other members of the SABATH family. One clade that is paralogous to the XMT-B lineage, which was named XMT-A for the purposes of this study, is comprised entirely of XMTs in *Coffea* (Figure 5, purple branches). The finding that *T. cacao* expresses sequences related to the XMT-B lineage, and that this lineage forms a clade that is phylogenetically distinct from the XMT-As, was highly unexpected because *Coffea* and *Camellia* are both Asterids, while *T. cacao* is a Rosid. Thus, it appears that *Camellia* sinensis and *T. cacao,* two species estimated to have diverged over 100 million years ago (Figure 2) (60), anciently inherited the same SABATH member that recently diversified independently in each lineage . In contrast, *Coffea* recruited a completely different SABATH member for xanthine alkaloid methylation than did *Camellia* and *Paullinia* (Figure 5).

It is worthy of mention that, despite a high level of confidence (>95%) in relationships near the tips of branches, it is not possible to determine the pattern of relatedness in the more ancient nodes of the tree with high levels of confidence. Therefore, it is unclear which class of SABATH methyltransferase could have been recruited anciently towards xanthine alkaloid biosynthesis. Nevertheless, the XMT-As and XMT-Bs invariably formed separate clades, and XMTs from *Camellia* always appeared orthologous to BTS1, TcXMT1 and TcXMT2 when phylogenetic analysis was performed using multiple alternative MSAs without alignment ambiguous sites (data not shown). Thus, it is clear that xanthine alkaloid methylation has at least two independent evolutionary origins.

Biochemical activities of TcXMT1 and TcXMT2 suggest an unprecedented major route in caffeine biosynthesis

Relative activities and reaction products of all enzymes in this study are summarized in Figure 6. To assess whether TcXMT1 and TcXMT2 are capable of methylating xanthine alkaloids, each coding gene was expressed in *E. coli*, and either the *E. coli* post induction total protein (TcXMT1) or purified protein (TcXMT2) was

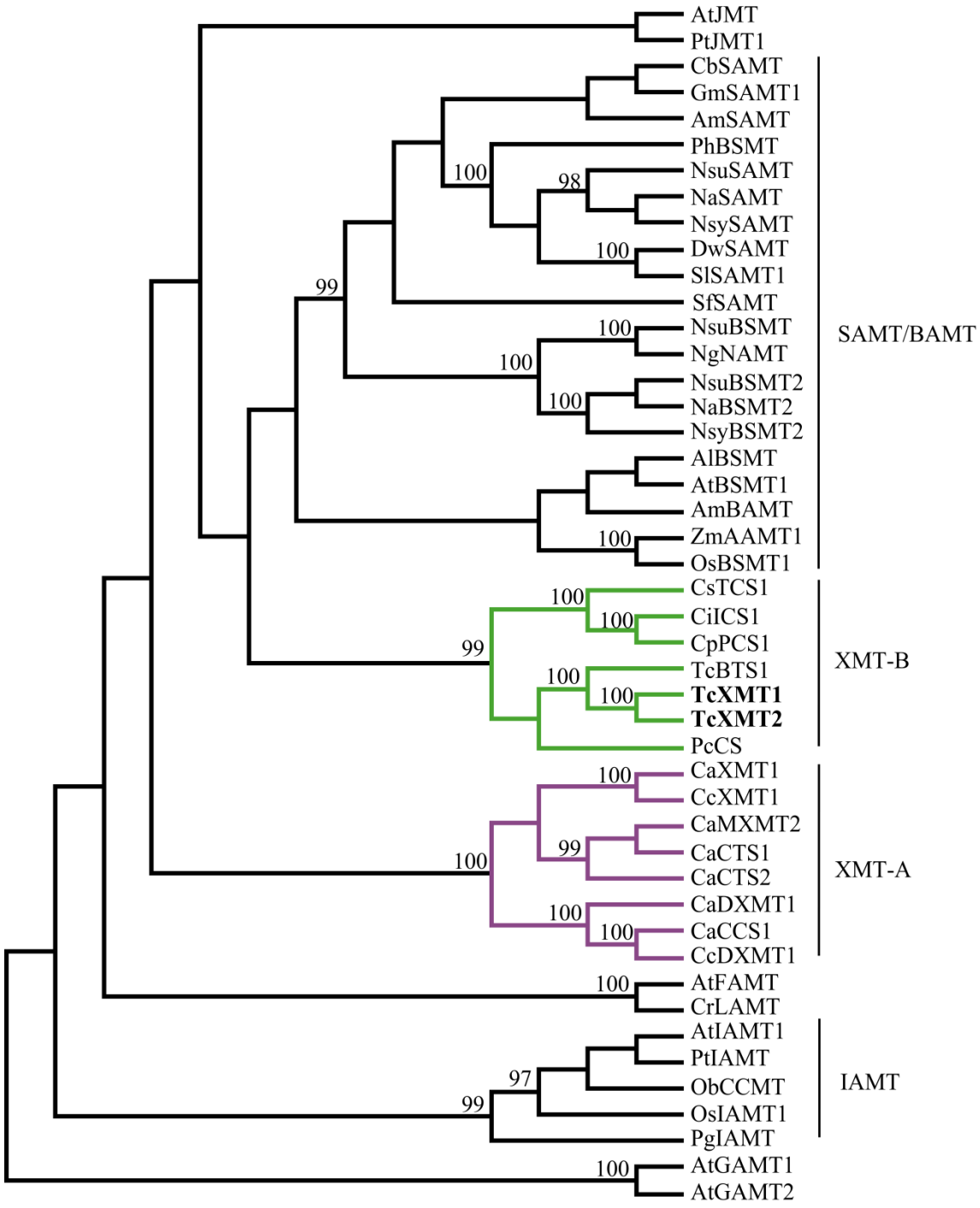

AtJMT
PtJMT1
CbSAMT
GmSAMT1
AmSAMT
PhBSMT
NsuSAMT
NaSAMT
NsySAMT
DwSAMT
SlSAMT1
SfSAMT
NsuBSMT
NgNAMT
NsuBSMT2
NaBSMT2
NsyBSMT2
AlBSMT
AtBSMT1
AmBAMT
ZmAAMT1
OsBSMT1
CsTCS1
CiICS1
CpPCS1
TcBTS1
TcXMT1
TcXMT2
PcCS
CaXMT1
CcXMT1
CaMXMT2
CaCTS1
CaCTS2
CaDXMT1
CaCCS1
CcDXMT1
AtFAMT
CrLAMT
AtIAMT1
PtIAMT
ObCCMT
OsIAMT1
PgIAMT
AtGAMT1
AtGAMT2
100
98
100
99
100
100
100
100
100
99
100
100
100
99
100
100
100
99
100
97
99
100
SAMT/BAMT
XMT-B
XMT-A
IAMT

Figure 5. Maximum-likelihood phylogeny of the characterized SABATH and the two XMTs from this study. TcXMT1 and TcXMT2 are indicated in bold. BTS1 is labeled as "TcBTS1" for naming consistency. Bootstrap values >95% are indicated by each node.

assayed with all xanthine alkaloid precursors in Figure 1 and ^{14}C-SAM. Surprisingly, it was found that TcXMT1 preferentially methylated xanthine, converting it to 3-methylxanthine (Figure 6a). This was highly unexpected because the first reaction in the caffeine pathway in *Camellia* and *Coffea* is the conversion of xanthosine to 7-methylxanthine. There is evidence for some level of substrate promiscuity in this enzyme, as shown in the inset bar chart of Figure 6a. However, activity with the next-best substrate (7-methylxanthine) was much lower than with xanthine. This strongly suggests that conversion of xanthine to 3-methylxanthine is the genuine primary function of this enzyme. Because TcXMT1 is clearly capable of N3 methylation based on these results, the possibility should be considered that the methylation of 7-methylxanthine is occurring at the N3 position, thus functioning *in vivo* in minor production of theobromine. However, the activity of TcXMT1 with 7-methylxanthine is too low to allow for detection on HPLC using the present methods.

Similar to TcXMT1, TcXMT2 also appears to be a specific enzyme (Figure 6b). A striking result of these assays was the apparent preference of TcXMT2 for 3-methylxanthine: the methylated product of TcXMT1 (Figure 6a). Thus, as expected, TcXMT2 methylated 3-methylxanthine in the N7 position, producing theobromine as the major product (Figure 6b). These results provide strong evidence that TcXMT1 and TcXMT2 together reconstitute a novel primary biochemical route towards theobromine

(and eventually caffeine) biosynthesis that has thus far not been implicated in either *Camellia* or *Coffea*. Thus, convergent evolution at the biochemical level explains part of the diversity seen in modern-day caffeine biosynthetic pathways.

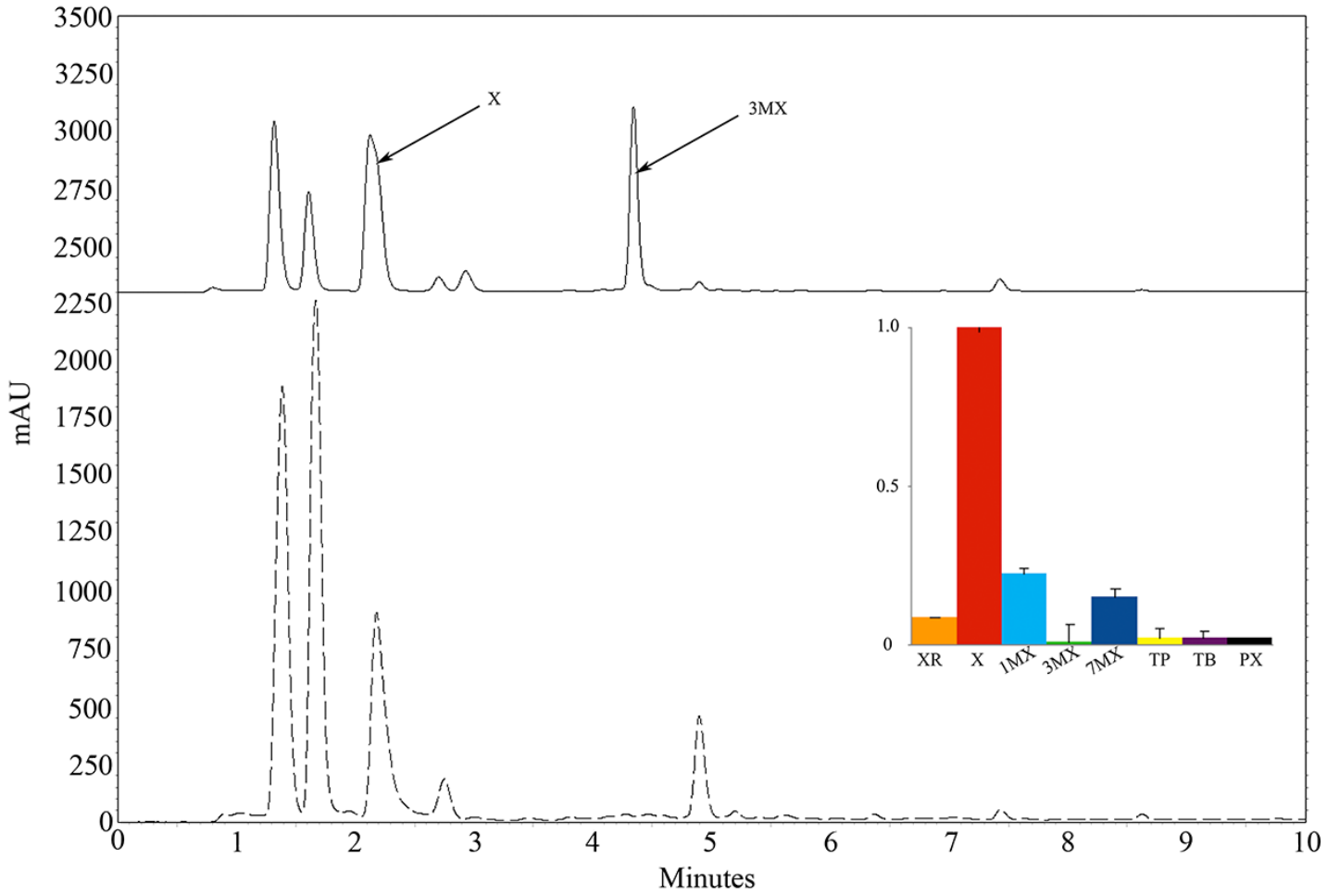


Figure 6a. HPLC trace following reaction of TcXMT1 with xanthine. The bar chart in the upper right-hand corner represents the 14C-SAM assays. Each colored bar represents a different xanthine alkaloid. The y-axis shows the amount of activity as a proportion of the highest bar.

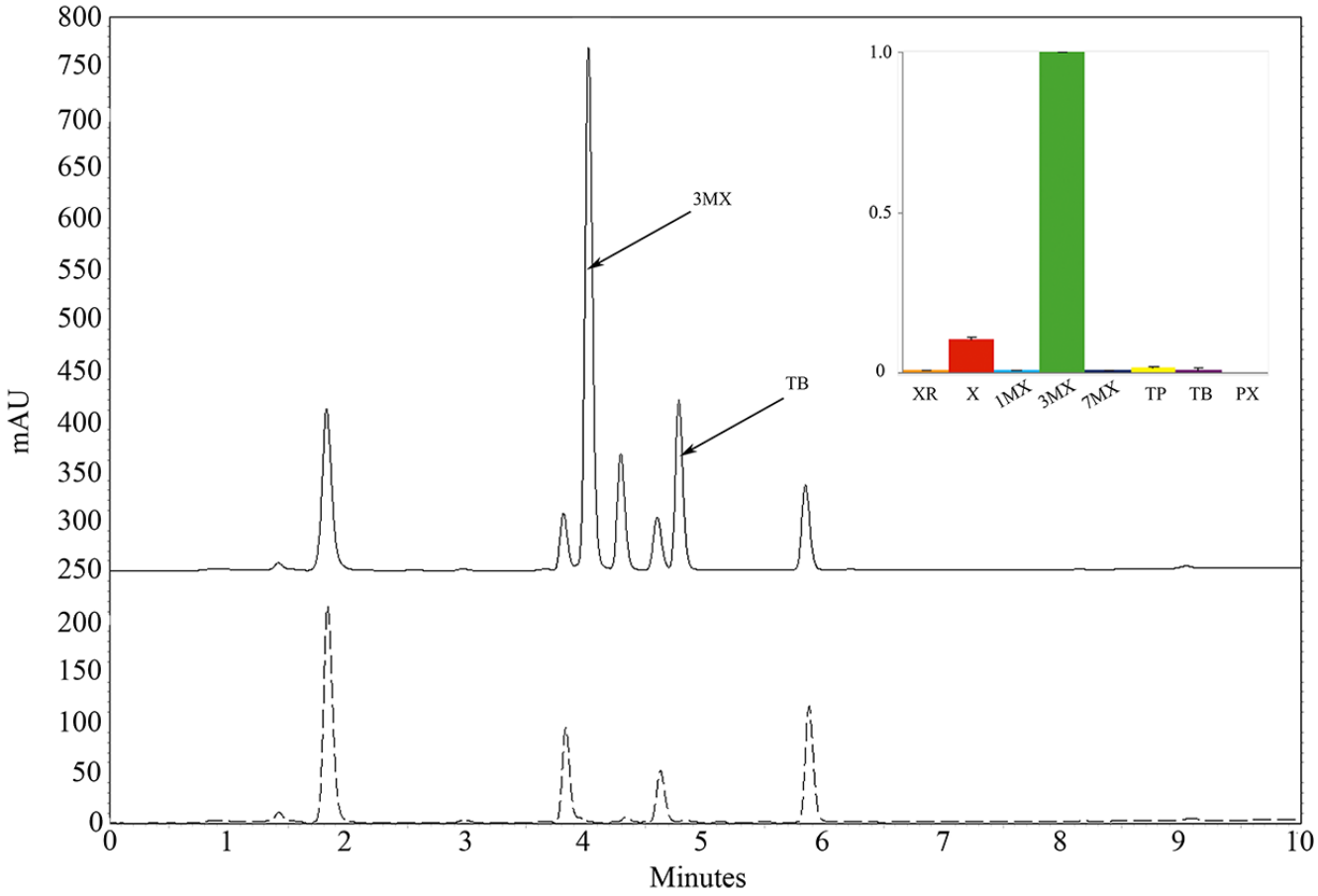


Figure 6b. HPLC trace following reaction of TcXMT2 with 3-methylxanthine. The bar chart in the upper right-hand corner represents the 14C-SAM assays. Each colored bar represents a different xanthine alkaloid. The y-axis shows the amount of activity as a proportion of the highest bar.

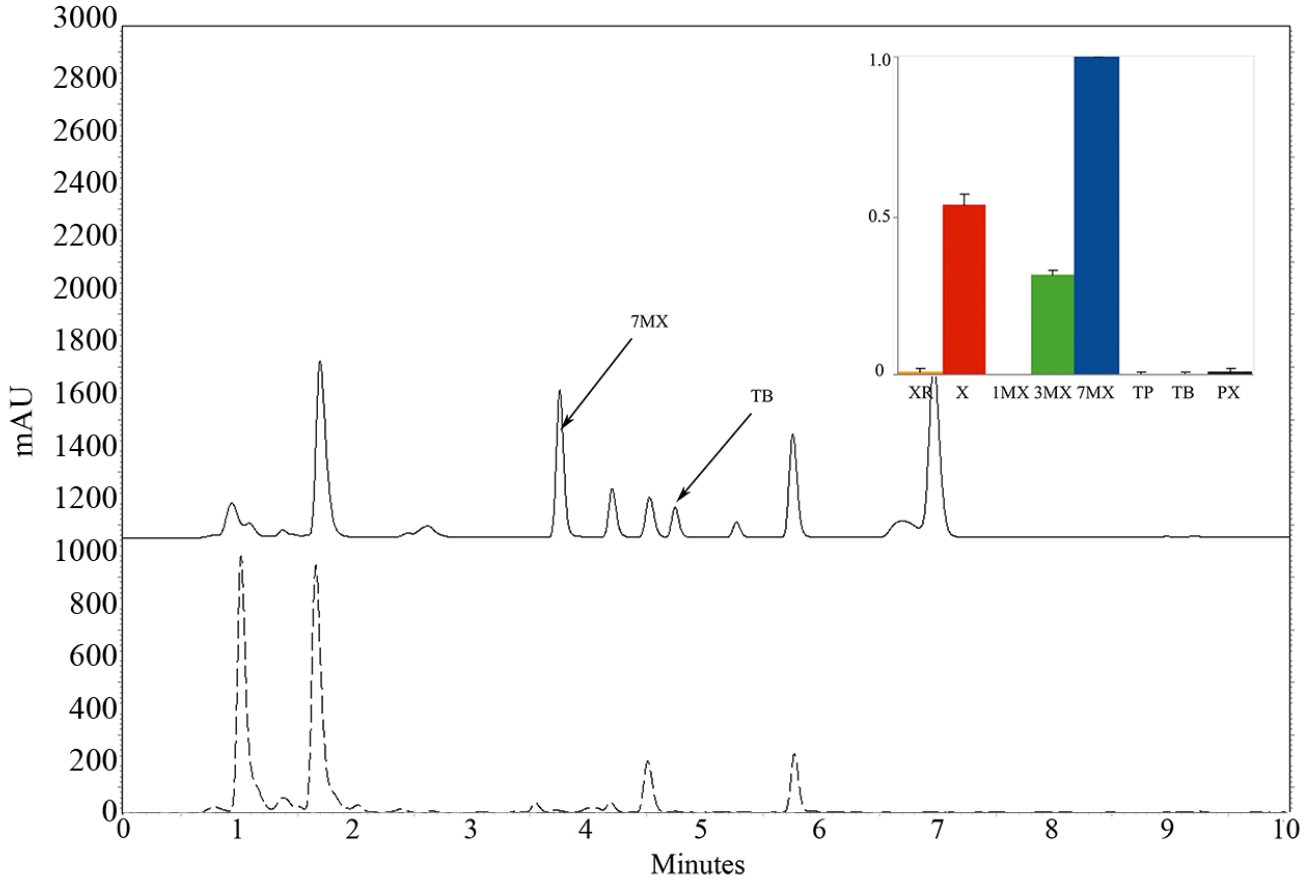


Figure 6c. HPLC trace following reaction of AncXMT1 with 7-methylxanthine. The bar chart in the upper right-hand corner represents the 14C-SAM assays. Each colored bar represents a different xanthine alkaloid. The y-axis shows the amount of activity as a proportion of the highest bar.

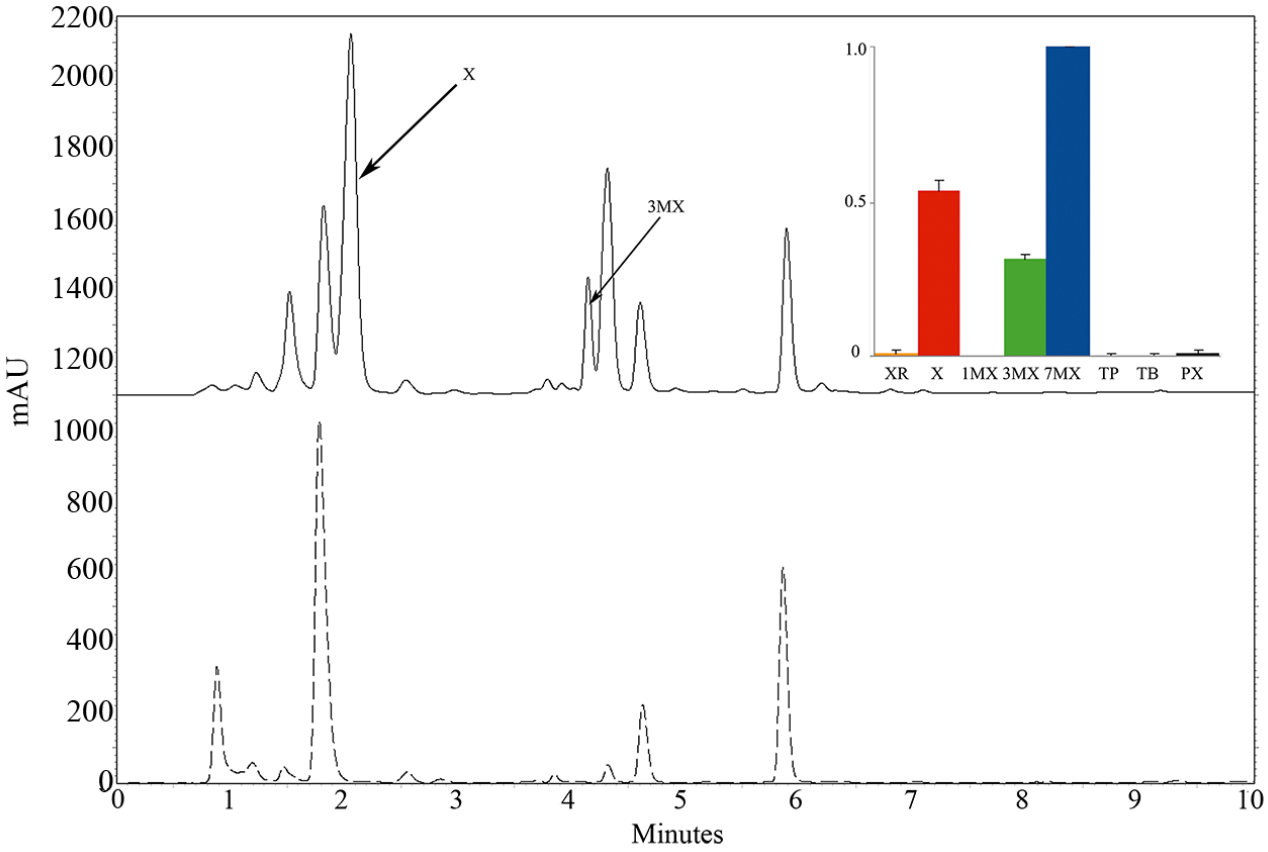


Figure 6d. HPLC trace following reaction of AncXMT1 with xanthine. The bar chart in the upper right-hand corner represents the 14C-SAM assays. Each colored bar represents a different xanthine alkaloid. The y-axis shows the amount of activity as a proportion of the highest bar.

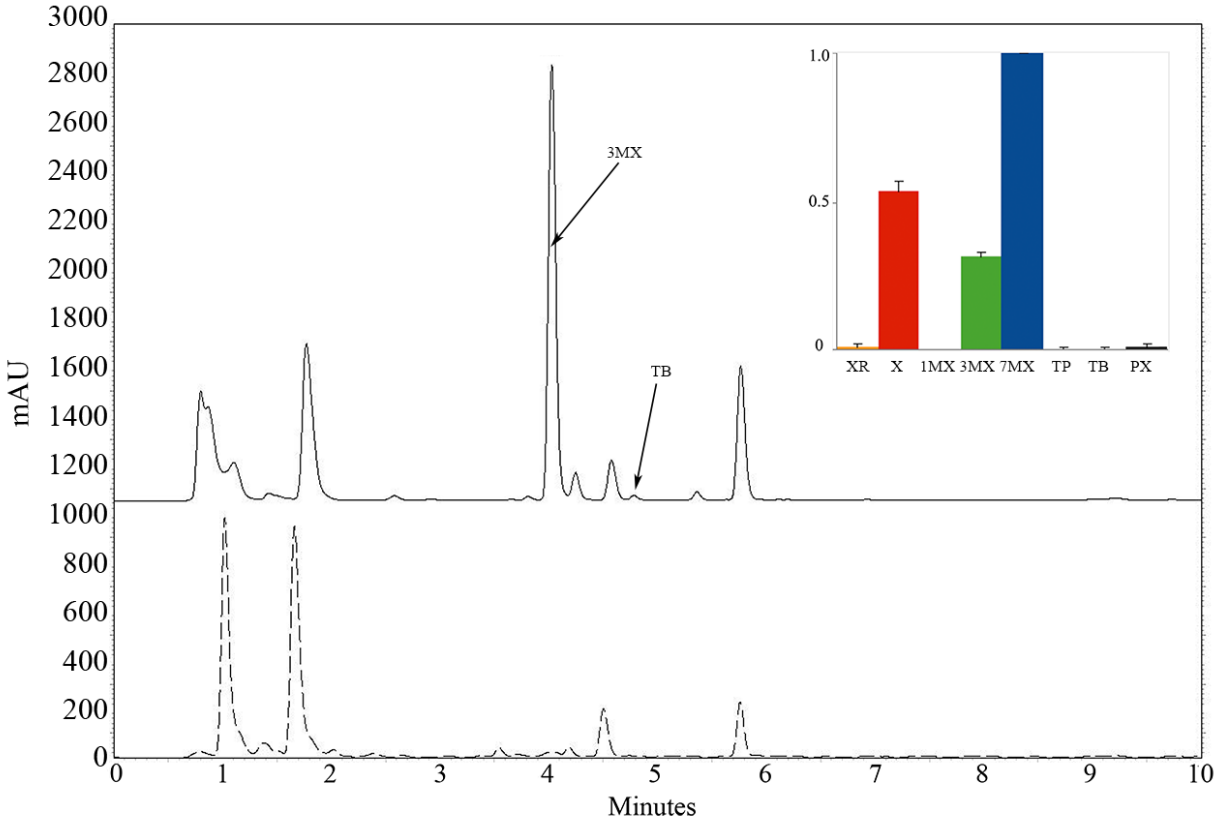


Figure 6e. HPLC trace following reaction of AncXMT1 with 3-methylxanthine. The bar chart in the upper right-hand corner represents the 14C-SAM assays. Each colored bar represents a different xanthine alkaloid. The y-axis shows the amount of activity as a proportion of the highest bar.

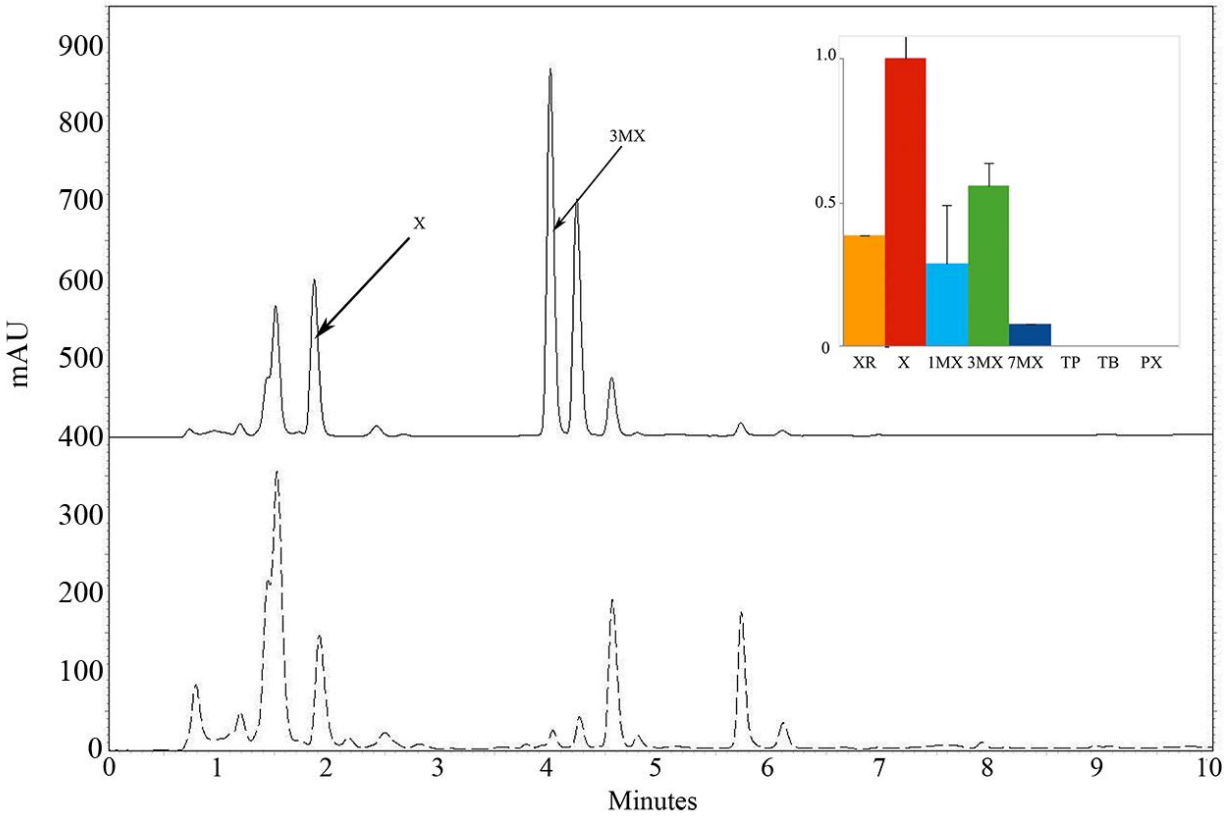


Figure 6f. HPLC trace following reaction of AncXMT2 with xanthine. The bar chart in the upper right-hand corner represents the 14C-SAM assays. Each colored bar represents a different xanthine alkaloid. The y-axis shows the amount of activity as a proportion of the highest bar.

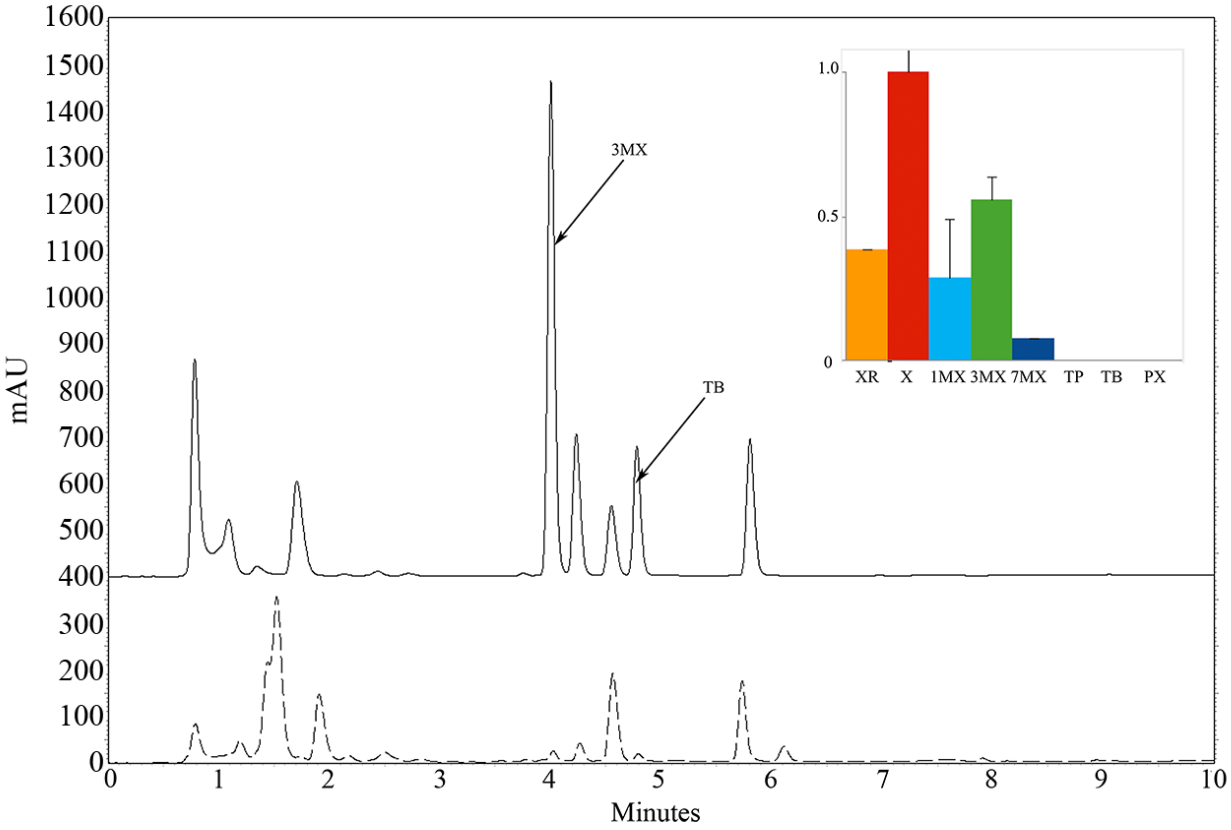


Figure 6g. HPLC trace following reaction of AncXMT2 with 3-methylxanthine. The bar chart in the upper right-hand corner represents the 14C-SAM assays. Each colored bar represents a different xanthine alkaloid. The y-axis shows the amount of activity as a proportion of the highest bar.

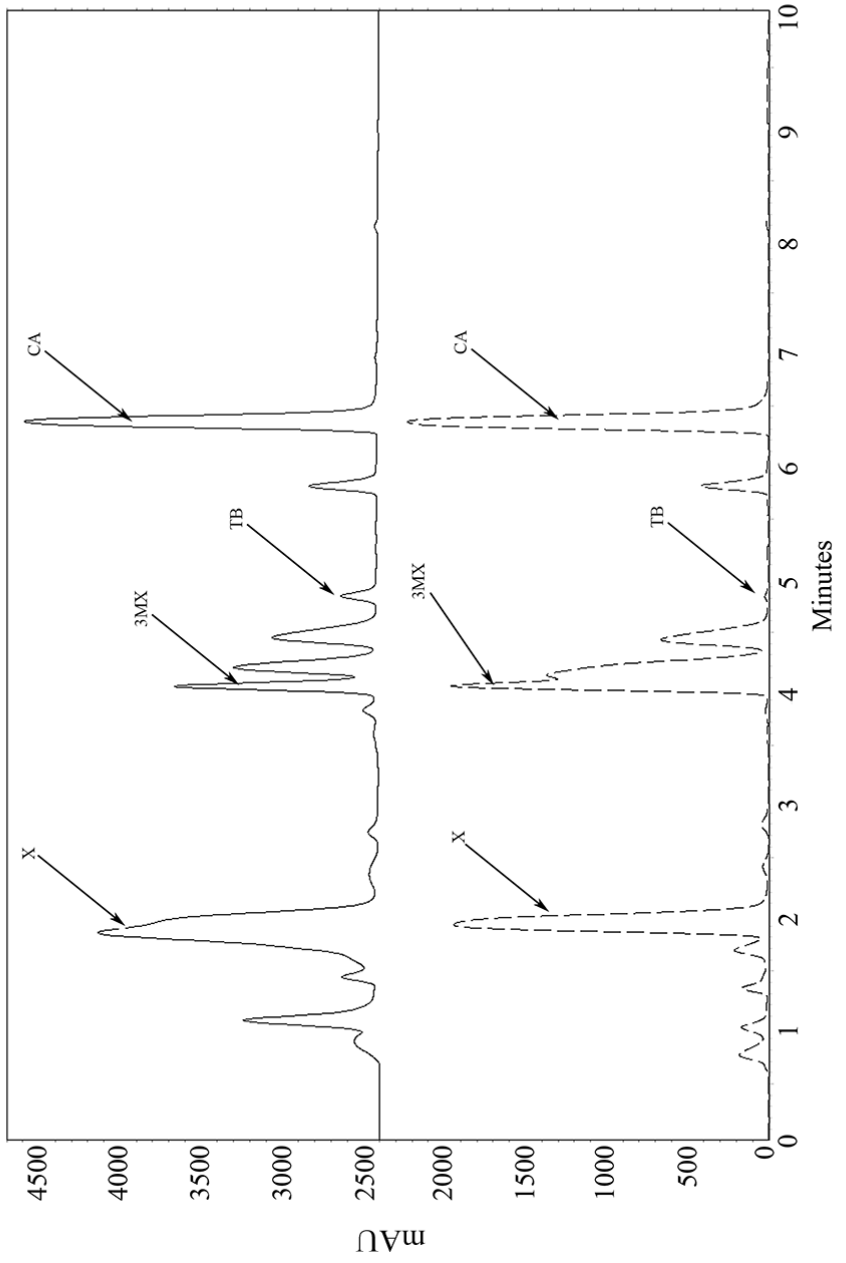


Figure 6h. HPLC trace following reaction of TcXMT1 and TcXMT2 (solid trace) with xanthine and AncXMT2 alone assayed with xanthine (dotted trace).

Figure 6. Relative activity profiles and product verification of each enzyme characterized in this study. The chromatograms are HPLC traces of reaction products after enzymes were assayed with one of the caffeine precursors

shown in Figure 1 and SAM. Figures 6b-6g, the solid trace is purified protein assayed with substrate and the dotted traces are purified protein assayed with NaOH instead of substrate. Each bar chart inset represents 8 different ^{14}C-SAM reactions using a single enzyme. Assayed substrates are shown below each bar. Abbreviations are XR, xanthosine; X, xanthine; 1MX, 1-methylxanthine; 3MX, 3-methylxanthine; 7MX, 7-methylxanthine; TB, theobromine; CA, caffeine internal standard.

Ancestral XMTs in T. cacao were promiscuous in terms of relative substrate preference and gave rise to specialized, modern-day XMTs

To examine the ancestral conditions that preceded the convergently-evolved xanthine alkaloid pathway in modern-day *T. cacao*, two ancestral enzymes that preceded the duplications of *T. cacao* XMT-B genes were resurrected and characterized *in vitro*: the ancestor of all XMT-B genes in *T. cacao* (AncXMT1), and the ancestor of only TcXMT1 and TcXMT2 (AncXMT2) (Figure 7).

AncXMT1 was found to prefer 7-methylxanthine, converting it to theobromine (Figure 6c). This ancestor also showed secondary activities with xanthine and 3-methylxanthine, converting them to 3-methylxanthine and theobromine, respectively (Figure 6d and 6e). As shown in the relative activity chart of figures 6c-e, the activity of AncXMT1 with 7-methylxanthine was about twice that of its activity with xanthine, and more than three times higher than activity with 3-methylxanthine.

AncXMT2 was also found to be promiscuous. The highest relative activity of this enzyme was with xanthine, converting it to 3-methylxanthine (Figure 6f). AncXMT2 also

had considerable secondary activity with 3-methylxanthine, converting it to theobromine (Figure 6g). Activity of AncXMT2 with 3-methylxanthine equaled about half of its activity with xanthine. AncXMT2 had lesser activity with xanthosine, 1-methylxanthine, and 7-methylxanthine. Thus, it appears that all of the activities of modern-day XMTs were present in the two ancestral XMTs.

When relative activities of ancestral enzymes were viewed in a phylogenetic context (Figure 7), a switch in highest relative activity from 7-methylxanthine in AncXMT1 to xanthine in AncXMT2 was observed. This could suggest that this lineage of enzymes has evolved participation in a novel pathway (conversion of xanthine to 3-methylxanthine) along the branch leading from AncXMT1 to AncXMT2 (Figure 7). Because XMT-B enzymes in *T. cacao* are orthologous to those in *Camellia* and *Paullinia* (Figure 5), this functional switch to xanthine methylation in AncXMT2 is inferred to have occurred after the divergence of the Malvales and Sapindales (Figure 2). Thus, it appears that an ancestor of *T. cacao*, independently of other caffeine-producing lineages, engineered the machinery necessary for theobromine production. However, because the methylation activities reported thus far are only relative to each other, it is unclear whether or not any kinetic improvements were required for this switch in relative activity.

A second exciting trend occurred after duplication of AncXMT2. AncXMT2, which evolved a switch to primary activity with xanthine, and could convert it to 3-methylxanthine, was also capable of methylating 3-methylxanthine and converting it to theobromine (Figures 6f and 6g). It could therefore be hypothesized based on the

aforementioned results that AncXMT2 alone was converting xanthine to theobromine in an ancestor of *T. cacao* by performing two subsequent methylation reactions. It is curious, then, why AncXMT2 appears to have partitioned its two major activities (methylation of xanthine and 3-methylxanthine) into its two descendants, TcXMT1 and TcXMT2 (Figure 7). It is possible that gene duplication of AncXMT2 provided an initial enzyme dosage effect, thereby increasing the amount of 3-methylxanthine methylation in an ancestor of *T. cacao*. Subsequent specialization and kinetic improvement upon 3-methylxanthine methylation along the branch leading to TcXMT2 could then provide an advantage in theobromine production. This situation would follow the predictions of the IAD model of protein evolution in the presence of continued selection for theobromine accumulation (38). In addition to characterizing the kinetic parameters of AncXMT1, those of AncXMT2 were characterized in order to determine whether kinetic improvements accompanied the evolution of 3-methylxatnhine specialization in TcXMT2.

Kinetic properties of ancestral and modern-day XMTs in T. cacao reveal that major biochemical changes accompanied the evolution of xanthine alkaloid methylation

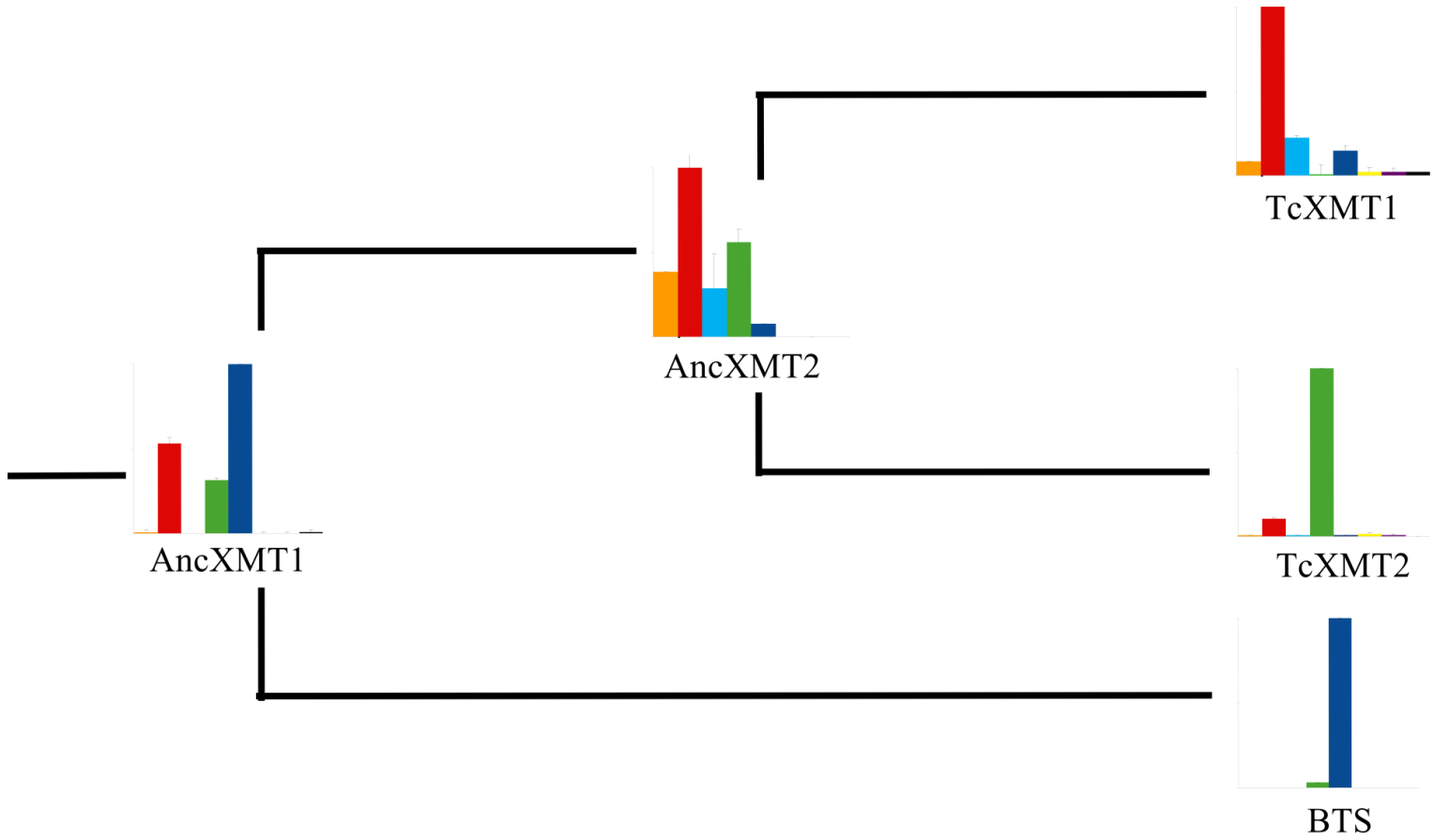


Figure 7. Evolution of relative activity among the XMTs from this study and BTS1 from Yoneyama at al. Each bar chart inset represents 8 different ^{14}C-SAM reactions using a single enzyme, where each reactions contains a different xanthine alkaloid substrate from Figure 1. The height of each bar is proportional to the number methylations ("activity") that occurred over a 20-minute period, and the maximum activity an enzyme had across all tested substrates was set to 1. XR (xanthosine), X (xanthine), 1MX (1-methylxanthine), 3MX (3-methylxanthine), 7XM (7-methylxanthine), TP (theophylline), TB (theobromine), PX (paraxanthine).

There are a few kinetic mechanisms that could explanation the switch in relative activity from 7-methylxanthine to xanthine between AncXMT1 and AncXMT2. One possibility is that AncXMT1 simply lost its ability to methylate 7-methylxanthine, while retaining its activities with xanthine and 3-methylxanthine. Alternatively, it could have improved on the ability to methylate xanthine in combination with a loss in catalytic abilities with 7-methylxanthine. Additionally, whether or not kinetic changes required for the specialization on substrates of TcXMT1 and TcXMT2 are unclear. If kinetic improvements occurred along branches leading to descendants of AncXMT2, this would help explain why *T. cacao* has maintained paralogs following gene duplication. To assess the specific changes to kinetic parameters of xanthine alkaloid methylation within this lineage of XMTs, the kinetic properties of AncXMT1, AncXMT2, TcXMT1, and TcXMT2 were characterized with their respective substrates (Table 2 and Figure 8).

Table 2. Kinetic parameters of ezymes characterized in this study

Enzyme	***K* m (μM)**	**Kcat (sec^{-1})**	**Kcat/*K* m ($sec^{-1}M^{-1}$)**
AncXMT1 (7MX)	34.6	2.47E-04	7.2
AncXMT1 (X)	53.4	8.65E-05	1.6
AncXMT1 (3MX)	138.6	1.28E-04	0.923
AncXMT2 (X)	4.14	3.39E-04	82.1
AncXMT2 (3MX)	154	3.32E-04	2.2
TcXMT1 (X)	95.8	8.37E-05	0.87
TcXMT2 (3MX)	49.1	9.81E-05	2.00

Interestingly, AncXMT2 showed increased affinity for xanthine binding through a >12-fold decrease in *K*m from 53.4 to 4.14 μM along the branch leading from AncXMT1

to AncXMT2. AncXMT2 also showed an approximate four-fold increase in product turnover rate (Kcat) with xanthine of $3.39x10^{-4}sec^{-1}$, compared to $8.65x10^{-5}$ sec^{-1} in AncXMT1. Assuming equal proportions of active enzyme, it also appears that AncXMT2 is about 50 times more efficient with xanthine than AncXMT1, as determined by Kcat/*K*m (1.6 and, 82 $s^{-1}M^{-1}$, respectively). Thus, it appears that the major changes in kinetic abilities of AncXMT1 were a loss in activity with 7-methylxanthine as a substrate, and kinetic improvements on xanthine methylation. Moreover, AncXMT2 had a Kcat/*K*m with 3-methylxanthine of 2.2 $s^{-1}M^{-1}$.

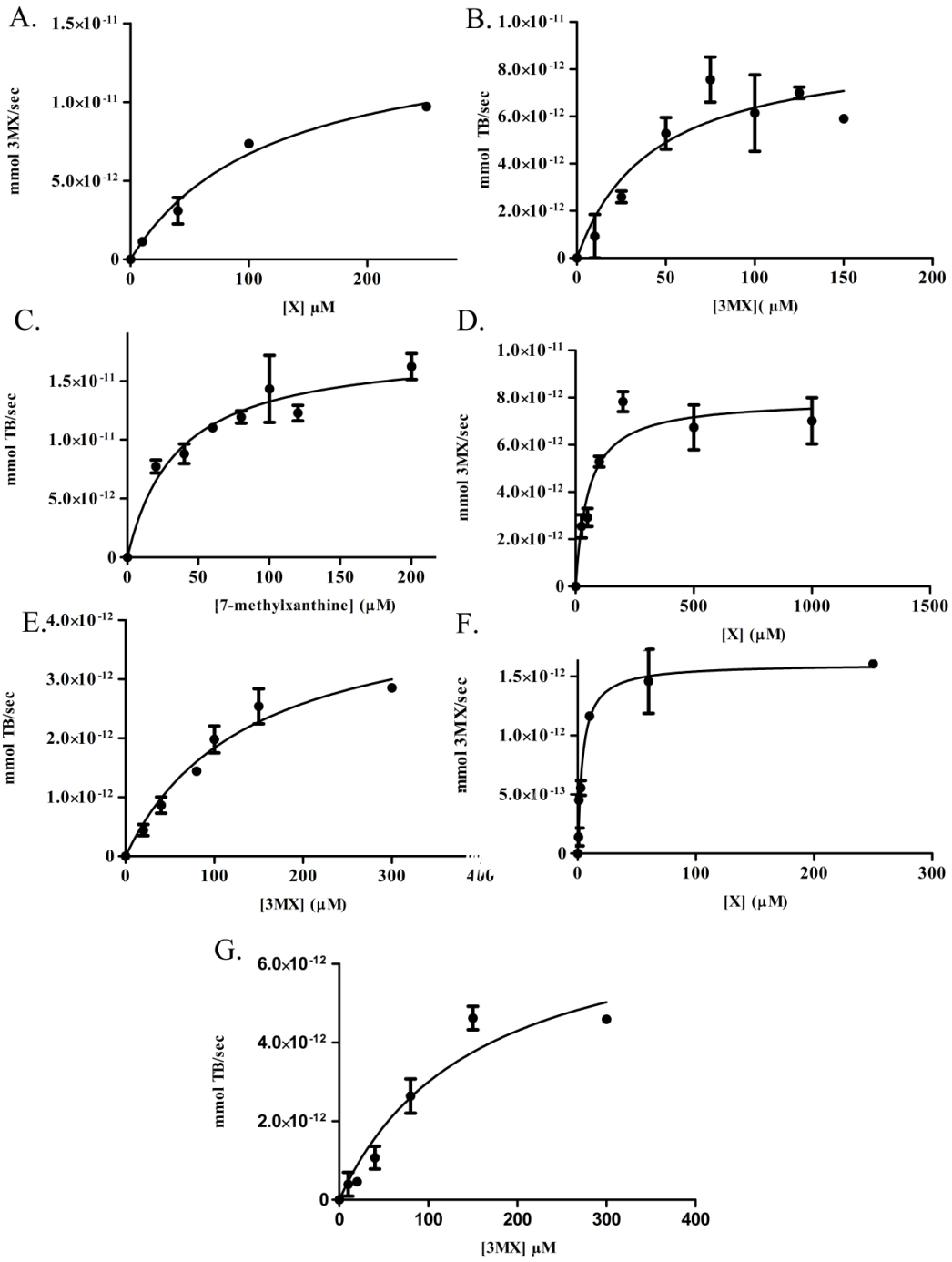


Figure 8. Michaelis-Menten graphs of enzymes characterized in this study. **A.** TcXMT1, **B.** TcXMT2, **C.-E.** AncXMT1 **F.-G.** AncXMT2. X (xanthine), 3MX (3-

methylxanthine), 7MX (7-methylxanthine), TB (theobromine).

In contrast to our observation of evolution of improved kinetic parameters between AncXMT1 and AncXMT2, we found that both TcXMT1 and TcXMT2 generally have poorer kinetic properties for each of their substrates with respect to AncXMT2. TcXMT1 had a *K*m of 95.8 μM with xanthine, a striking 23-fold decrease in binding affinity along the branch leading from AncXMT2 to TcXMT1. TcXMT1 also had a four-fold decrease in Kcat ($8.37x10^{-5}$ sec^{-1}) from AncXMT2. Assuming equal proportions of active enzyme, TcXMT1 was about 80 times less efficient with xanthine than its ancestor, with a Kcat/*K*m of 0.87 $s^{-1}M^{-1}$.

In contrast to the decrease in binding affinity seen in TcXMT1, TcXMT2 had a *K*m of 49.1 μM with 3-methylxanthine. This result indicates a three-fold increase in binding affitinity for 3-methylxanthine from 154 μM, as observed in AncXMT2. However, TcXMT2 showed loss in product turnover and specificity with 3-methylxanthine ($9.81x10^{-5}$ sec^{-1} and 2.0 $s^{-1}M^{-1}$, respectively).

The specificity constant, Kcat/*K*m, for AncXMT2 with xanthine was about 40 times higher than with 3-methylxanthine. This result implies that 3-methylxanthine would have to accumulate to 40 times the concentration of xanthine before efficient subsequent conversion to theobromine could occur if these two substrates exist together in solution. Unless equilibrium of this reaction lies in extreme favor of 3-methylxanthine, a condition for which we have no evidence, we would predict from these kinetic data that endogenous xanthine in *T. cacao* fruits would be more likely to occupy

the active site of AncXMT2 than 3-methylxanthine, thus causing *in vivo* discrimination by this enzyme in favor of xanthine. Moreover, any minor methylation of 3-methylxanthine that does occur in the presence of xanthine would result in depletion of the 3-methylxanthine pool relative to xanthine, thus constraining AncXMT2 from performing the conversion of 3-methylxanthine to theobromine.

In contrast to the situation of AncXMT2, TcXMT1 and TcXMT2 each appear to be less efficient than their ancestor with each of their substrates, but we would nevertheless expect the modern-day XMTs to provide improved pathway flux with respect to AncXMT1. Thus, in spite of the expectation that specialization was accompanied by improved kinetic parameters in TcXMT1 and TcXMT2, it appears that gene duplication resulted in improved production of theobromine because TcXMT2 simply lost activity with xanthine. To determine whether the single-enzyme system comprised of AncXMT2 in the ancestor of *T. cacao* was less efficient than the two-enzyme modern system in *T. cacao*, the ability of AncXMT2 to convert xanthine to theobromine was compared directly to that of TcXMT1 and TcXMT2 together.

When incubated with only xanthine and SAM as substrate, AncXMT2 produced about 2.6 micromoles of theobromine in a 1 ml reaction. As expected, TcXMT1 and TcXMT2 together produced more theobromine (12.1 micromoles) (Table 3 and Figure 6h). The complementary partitioning of AncXMT2's activities into its two descendants therefore appears to have resulted in a more efficient system for theobromine accumulation.

Table 3. Starting activities and results of pathway flux analyses. An amount of enzyme proportional to that used in acquiring starting activities was used for pathway flux analysis. Showing starting DPMs with 3-methylxanthine

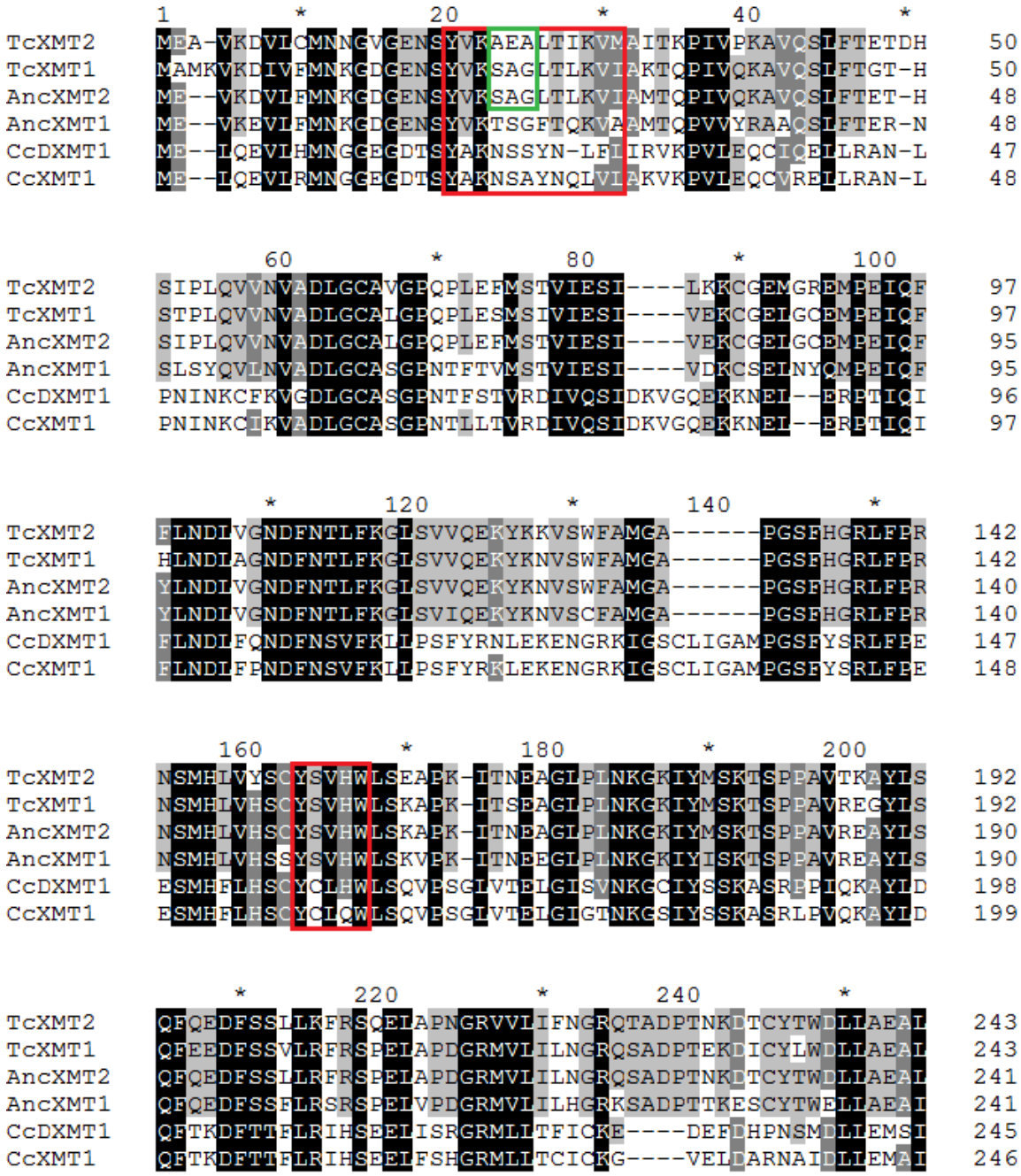
enzyme	starting activity (DPM)	micromoles of Tb produced
AncXMT2	3741	2.6
TcXMT1 + TcXMT2	2290	12.1

Ancestral XMTs were unable to improve on existing secondary activities without sacrificing activity with other xanthine alkaloids

The observation that *T. cacao* has evolved specialized paralogs following gene duplication raises another question: why did an ancestor of *T. cacao* maintain paralogs rather than simply improve activity of AncXMT2 with 3-methylxanthine? The latter fate should be the favorable scenario since an optimized, multifunctional enzyme would require maintenance of only a single coding gene rather than two. This led to the hypothesis that AncXMT2 was adaptively conflicted, as in the EAC model of protein evolution, and gene duplication was an escape for a constrained ancestral enzyme. That is, adaptive substitutions that would confer improvements to 3-methylxanthine methylation relative to xanthine would have resulted in functional loss of its primary activity of xanthine methylation, so the *T. cacao* lineage instead evolved two specialized enzymes. To test this hypothesis, and to determine whether there is a general trend in pleiotropic mutational effects, historical mutations that occurred along the branch leading from AncXMT1 to AncXMT2, and from AncXMT2 to TcXMT2, were introduced into the background of ancestral enzymes, and the evolutionary intermediate were characterized *in vitro.*

In a related and unpublished study, mutations in an ancestor of XMTs from *P. cupana* that were predicted to be near the active site based on sequence homology to the structurally-characterized XMT and DXMT enzymes from *C. canephora* were shown to have profound effects on substrate discrimination in favor of either xanthine or 3-methylxanthine (data not shown). Assuming that a similar trend of mutation and functional evolution in *P. cupana* occurred in *T. cacao*, residues predicted to be near the active sites in XMT-B enzymes in *T. cacao* were targeted for mutagenesis experiments (Figure 9). The NLRS region in AncXMT1 near site 340 acquired three amino acid substitutions, because the same region in AncXMT2 is GHRC. Based on sequence homology to XMT and DXMT from *C. canephora*, these three concentrated changes were predicted to occur near the active site. Furthermore, mutations in this region were shown to shift the substrate preference of the multifunctional ancestor in *P. cupana* in favor of xanthine. We therefore introduced the NL-S to GH-C mutation near this site.

As expected, when NLRS near site 340 was mutated to GHRC in the AncXMT1 background, there was a shift in relative activity so that xanthine methylation appeared to


1 * 20 * 40 *
TcXMT2 MEA-VKDVLCMNNGVGENSYVKAEALTIKVMAITKPIVPKAVQSLFTETDH 50
TcXMT1 MAMKVKDIVFMNKGDGENSYVKSAGLTLKVIAKTQPIVQKAVQSLFTGT-H 50
AncXMT2 ME--VKDVLFMNKGDGENSYVKSAGLTLKVIAMTQPIVQKAVQSLFTET-H 48
AncXMT1 ME--VKEVLFMNKGDGENSYVKTSGFTQKVAAMTQPVVYRAAQSLFTER-N 48
CcDXMT1 ME--LQEVLHMNGGEGDTSYAKNSSYN-LFLIRVKPVLEQCIQELLRAN-L 47
CcXMT1 ME--LQEVLRMNGGEGDTSYAKNSAYNQLVIAKVKPVLEQCVREILRAN-L 48
60 * 80 * 100
TcXMT2 SIPLQVVNVADLGCAVGPQPLEFMSTVIESI----LKKCGEMGREMEEIQF 97
TcXMT1 STPLQVVNVADLGCALGPQPLESMSIVIESI----VEKCGELGCEMEEIQF 97
AncXMT2 SIPLQVVNVADLGCALGPQPLEFMSTVIESI----VEKCGELGCEMEEIQF 95
AncXMT1 SLSYQVLNVADLGCASGPNTFTVMSTVIESI----VDKCSELNYQMEEIQF 95
CcDXMT1 PNINKCFKVGDLGCASGPNTFSTVRDIVQSIDKVGQEKKNEL--ERETIQI 96
CcXMT1 PNINKCIKVADLGCASGPNTLLTVRDIVQSIDKVGQEKKNEL--ERETIQI 97
* 120 * 140 *
TcXMT2 FLNDLVGNDFNTLFKGLSVVQEKYKKVSWFAMGA------PGSFHGRLFPR 142
TcXMT1 HLNDLAGNDFNTLFKGLSVVQEKYKNVSWFAMGA------PGSFHGRLFPR 142
AncXMT2 YLNDLVGNDFNTLFKGLSVVQEKYKNVSWFAMGA------PGSFHGRLFPR 140
AncXMT1 YLNDLVGNDFNTLFKGLSVIQEKYKNVSCFAMGA------PGSFHGRLFPR 140
CcDXMT1 FLNDLFQNDFNSVFKLLPSFYRNLEKENGRKIGSCLIGAMPGSFYSRLFPE 147
CcXMT1 FLNDLFPNDFNSVFKLLPSFYRKLEKENGRKIGSCLIGAMPGSFYSRLFPE 148
160 * 180 * 200
TcXMT2 NSMHLVYSCYSVHWLSEAPK-ITNEAGLPLNKGKIYMSKTSPPAVTKAYLS 192
TcXMT1 NSMHLVHSCYSVHWLSKAPK-ITSEAGLPLNKGKIYMSKTSPPAVREGYLS 192
AncXMT2 NSMHLVHSCYSVHWLSKAPK-ITNEAGLPLNKGKIYMSKTSPPAVREAYLS 190
AncXMT1 NSMHLVHSSYSVHWLSKVPK-ITNEEGLPLNKGKIYISKTSPPAVREAYLS 190
CcDXMT1 ESMHFLHSCYCLHWLSQVPSGLVTELGISVNKGCIYSSKASRFPIQKAYLD 198
CcXMT1 ESMHFLHSCYCLQWLSQVPSGLVTELGIGTNKGSIYSSKASRLPVQKAYLD 199
* 220 * 240 *
TcXMT2 QFQEDFSSLLKFRSQELAPNGRVVLIFNGRQTADPTNKDTCYTWDLLAEAL 243
TcXMT1 QFEEDFSSVLRFRSPELAPDGRMVLILNGRQSADPTEKDICYLWDLLAEAL 243
AncXMT2 QFQEDFSSLLRFRSPELAPDGRMVLILNGRQSADPTNKDTCYTWDLLAEAL 241
AncXMT1 QFQEDFSSFLRSRSPELVPDGRMVLILHGRKSADPTTKESCYTWELLAEAI 241
CcDXMT1 QFTKDFTTFLRIHSEELISRGRMLITFICKE----DEFLHPNSMDLLEMSI 245
CcXMT1 QFTKDFTTFLRIHSEELFSHGRMLITCICKG----VELLARNAIDLLEMAI 246

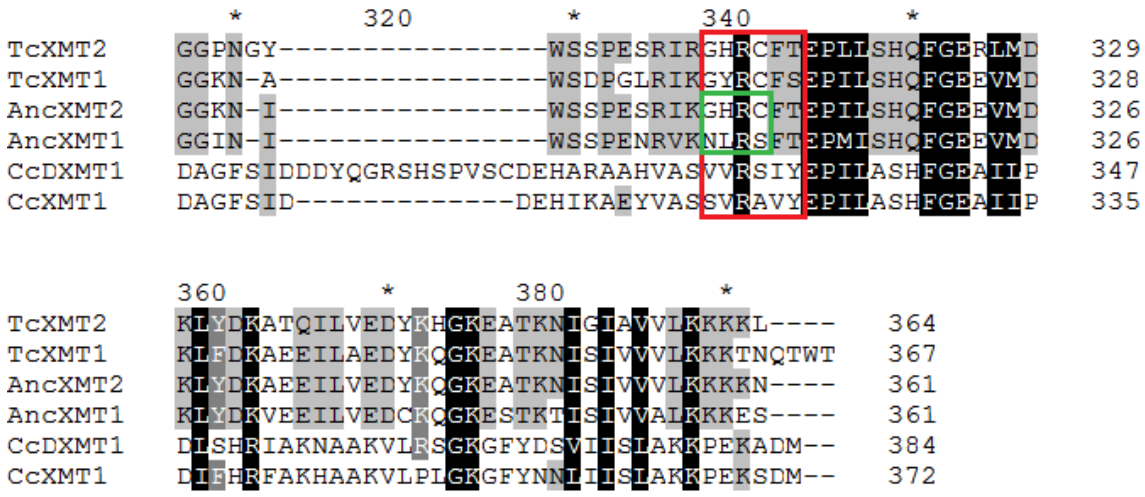


Figure 9. Multiple sequence alignment of the XMTs characterized in this study and the two XMTs structurally characterized from Coffea (McCarthy and McCarthy 2007). Black, dark gray, and light gray shading represents totally conserved, highly conserved, and moderately conserved regions, respectively. Red boxes represent regions considered to be near the xanthine alkaloid binding pocket based on structures of CcXMT1 and CcDXMT1. Green boxes show the regions that were the focus of our mutagenesis studies.

be the preferred function of this enzyme (Figure 10a). Two secondary activities of this intermediate enzyme were with 7-methylxanthine (72%) and 3-methylxanthine (31%), and secondary activities with the rest of the tested substrates of less than 11%. Thus, it appears that mutations that allowed for an apparent improvement on xanthine methylation caused a relative decrease in its putative fortuitous secondary function of 7-methylxanthine methylation.

To test the hypothesis that AncXMT2 was constrained from acquiring function-shifting mutations in favor of 3-methylxanthine methylation, mutations that

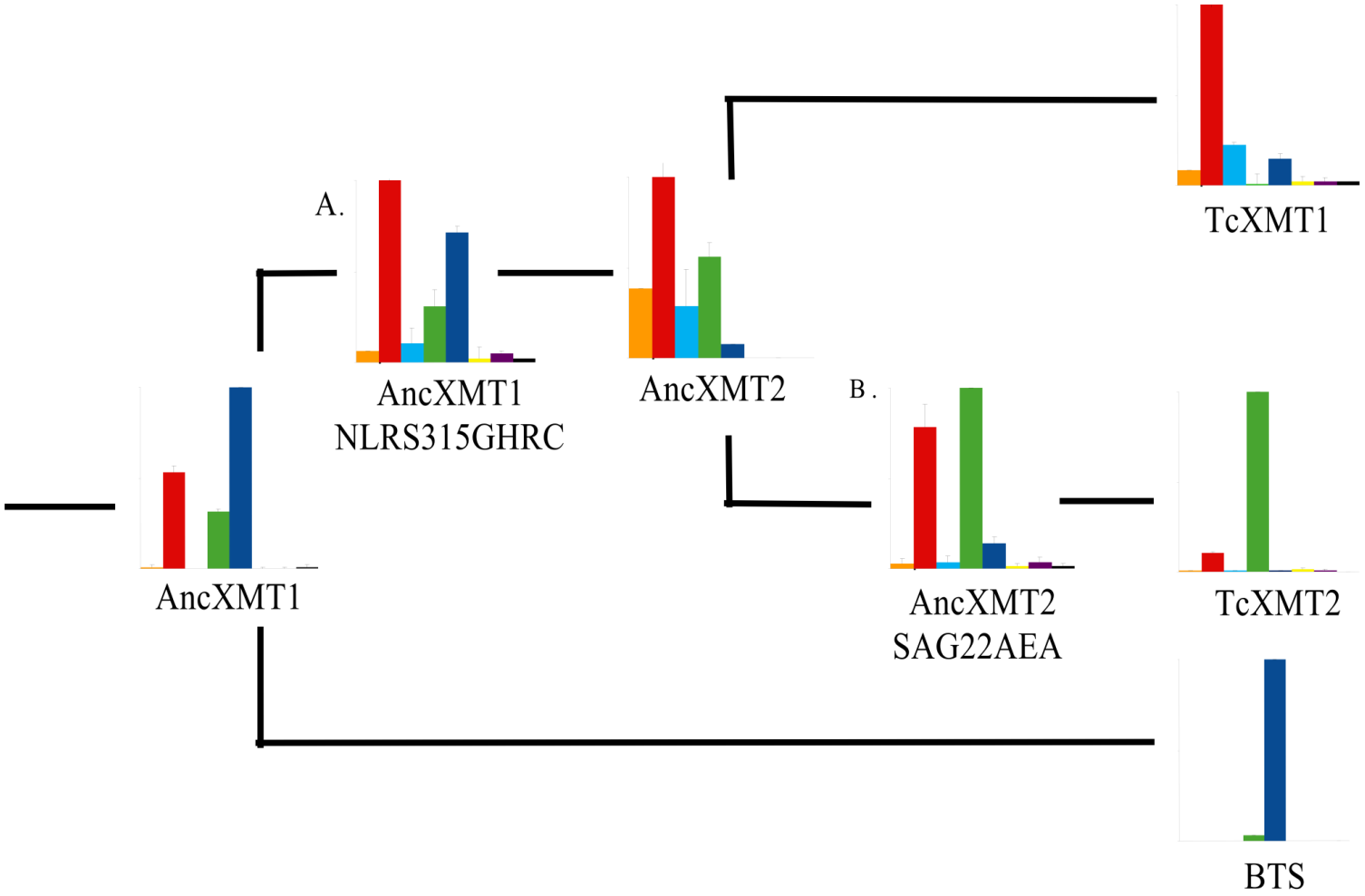


Figure 10. Effects of historical mutations on the evolution of relative activities within XMTs from this study **A.** Changing NLS near residue 315 shifted the primary activity of AncXMT1 from 7-methylxanthine to xanthine. **B.** Changing SAG to AEA near site 22 shifted the activity of AncXMT2 to TcXMT2.

occurred between this ancestor and TcXMT2 were also targeted. Along this branch, SAG near site 22 is substituted with AEA (Figure 9). Because this site is also predicted to occur near the active site based on homology with XMTs from *Coffea*, and due to the function-shifting effects mutations in a nearby region in the *P. cupana* ancestral enzyme, AEA was introduced into the background of AncXMT2. As expected, mutating SAG to AEA in AncXMT2 caused a shift in relative preference in favor of 3-methylxanthine, and a relative loss in activity with xanthine, thus nearly completely recapitulating the evolution of modern-day enzyme preferences (Figure 10b).

DISCUSSION

In summary, this study characterized two enzymes in *T. cacao* fruits that not only sufficiently explain the required enzymatic machinery for xanthine alkaloid production in fruits, but also provide compelling evidence for a previously-unexplored major route for caffeine biosynthesis in plants. Additionally, the question of to what extent convergence explains the distribution of the caffeine phenotype across the angiosperm phylogeny was addressed. Finally, this study was concluded with a set of experiments that demonstrate how gene duplication resolved an ancient adaptive conflict and permitted evolutionary improvements to flux through the caffeine pathway.

Preferred activities of TcXMT1 and TcXMT2 explain most, but not all, of xanthine alkaloid metabolism in T. cacao

As discussed above, the two best-studied tissues of *T. cacao* in terms of xanthine alkaloid metabolism are the leaves and fruits (18, 19). Due to the presence of xanthine and 3-methylxanthine in all growth stages of both tissue types, the primary metabolic route suggested by TcXMT1 and TcXMT2 for the accumulation of theobromine is supported. Conspicuously, when fruit tissues were incubated with radiolabeled xanthosine, some of the precursor was incorporated into xanthine and 3-methylxanthine, but neither 7-methylxanthosine nor 7-methylxanthine were detected as products (18). This further supports that the canonical route to caffeine in plants, which consists of a first committed step of conversion of xanthosine to 7-methylxanthine (Figure 1), is not the major pathway in operation in *T. cacao* tissues.

The above results present the puzzle of why BTS1, a putative theobromine synthase characterized by Yoneyama *et al*. (2006), is expressed in leaves and tissues and can convert 7-methylxanthine to caffeine. The *K*m of BTS1 with 7-methylxanthine is 2.4 mM, which is probably too high to be physiologically relevant since all of the XMTs characterized in this report, and XMTs from *Camellia*, fall between 4 and 200 μM (Table 2) (31). Furthermore, the importance of BTS1 in xanthine alkaloid biosynthesis appears to be greatly undermined by the number of ESTs that support TcXMT1 and TcXMT2 (Figure 4). BTS1 could be functioning in a minor pathway for caffeine biosynthesis via 7-methylxanthine. Indeed, there is some 7-methylxanthine present in all stages of *T. cacao* fruits and leaves, and 7-methylxanthine does not appear to be part of xanthine alkaloid catabolism in either tissue type (18, 19). It is possible that TcXMT2 is

providing minor amounts of 7-methylxanthine as substrate for BTS1 *in vivo* through its secondary activity with xanthine (figure 6b). This is highly feasible because TcXMT2 is capable of N7 methylation in the conversion of 3-methylxanthine to theobromine and could also be methylating xanthine in the N7 position.

One conspicuous result of our relative activity experiments is that activity of both TcXMT1 and TcXMT2 with theobromine appeared to be nearly completely absent. Thus, the identity of an enzyme that produces the relatively low amounts of caffeine in *T. cacao* fruits remains elusive. It is important to note that TcXMT1 had trace amounts of activity with theobromine in radiochemical assays (Figure 6a). However, when TcXMT1 was incubated in solution containing 100 mM theobromine overnight, we were unable to confirm the production of caffeine using HPLC analysis (data not shown). Given the fact that concentrations of theobromine exceed those of caffeine in nearly all stages of *T. cacao* fruits, as discussed previously, it is likely that any enzyme catalyzing the conversion of theobromine to caffeine *in vivo* is not very efficient in doing so. Thus, it is possible that TcXMT1 is responsible for the *in vivo* conversion of theobromine to caffeine, and that the amount of product formed over the time span of our assays is below our level of detection.

Convergence of the caffeine biosynthetic pathway is supported by both genetic and biochemical diversity

As shown in Figure 5, the genus *Coffea* appears to have recruited a separate SABATH member than did *Camellia*, *P. cupana*, and *T. cacao* for caffeine production. This result was highly unexpected because *Camellia* and *Coffea* each belong to orders

within the Asterids, while *P. cupana* and *T. cacao* are Rosids. Furthermore, it was shown phylogenetic analysis (Figure 5), along with the evolution of relative activities in this study (Figure 7), shows that duplications and functional divergence in caffeine biosynthetic machinery in *Coffea, Camellia,* and *T. cacao* occurred independently within each lineage. Thus, it is clear that historical, lineage-independent genetic changes strongly support the hypothesis that caffeine biosynthesis has evolved independently in caffeine-producing lineages.

Further support for this hypothesis is found in the biochemical activities of modern-day XMTs from *Coffea*, *Camellia*, and *T. cacao*. Despite the fact that *Coffea* and *Camellia* use paralogs to accumulate xanthine alkaloids, each lineage is known to produce caffeine via methylation of xanthosine and 7-methylxanthine (61). In contrast, *Camellia* and *T. cacao* use orthologs to accumulate xanthine alkaloids (Figure 5), yet the major route implicated by XMTs from *T. cacao* support a major pathway via xanthine and 3-methylxanthine. Thus, *Coffea*, *Camellia*, and *T. cacao* have all independently evolved the production of caffeine via convergent genetic and biochemical means.

The evolution of XMTs in T. cacao is explained by multiple models of protein evolution

As described previously, the IAD model of protein evolution posits that new environmental pressures, which render a secondary activity favorable, can lead to selection for gene duplicate maintenance due to their combined dosage effects (38). This model also predicts that divergence following duplication occurs in the specialization of

paralogous proteins. The study by Nasvall et al (2012) showed that IAD explains the fate of an ancestral, bifunctional isomerase in bacteria. The possibility should be explored that AncXMT1was under some sort of selection for its relatively minor ability to convert xanthine to 3-methylxanthine. If this is the case, then the IAD model could explain how paralogs following gene duplication of this ancestor were maintained for long enough for one to accumulate function-shifting mutations in favor of xanthine methylation. In this setting, enzymes that existed immediately after duplication of AncXMT1, possibly having primary roles outside of xanthine alkaloid methylation, would have provided a selective advantage if the resultant increase in enzyme product increased the pool of converted 3-methylxanthine *in vivo*. This is not unlikely because intermediates of the caffeine pathway have been found to antagonize adenosine receptors of mammals and could have had an ecological effect in deterring herbivores (62, 63). Following duplication, any mutations that improved catalysis 3-methylxanthine production along the branch leading to AncXMT2 would be fixed quickly in the presence of strong positive selection.

Studies also show that evolution of novel metabolic relationships can occur by taking advantage of existing ancestral activities (64-66). Surprisingly, this process may also explain part of the evolution of XMTs in *T. cacao*. If the observed switch in relative activity of AncXMT1 to xanthine following duplication conferred high enough concentrations of 3-methylxanthine, then the secondary activity of conversion of 3-methylxanthine to theobromine, a previously useless secondary activity, would have a

newly-derived functional role. The role of gene duplication in this model, whether or not it was previously selected for its increase in gene dosage, was likely to release AncXMT1 from any constraint in performing its original, unknown role.

The model of protein evolution implicated by duplication of AncXMT2 is clearer. As shown in Figure 7, AncXMT2 partitioned its two major activities into its two descendents (Figure 10). It appears that this was the only way in which *T. cacao* could improve upon its ability to convert 3-methylxanthine to theobromine because mutations that improve relative activity with 3-methylxanthine cause a relative loss in xanthine activity (Figure 10b). Thus, if theobromine accumulation were a selected trait, AncXMT2 would have been adaptively conflicted. The only release from this constraint would have been to provide duplicate gene copies so that one is able to specialize on the conversion of 3-methylxanthine to theobromine. It therefore appears that EAC explains the role of gene duplication following this ancestor.

REFERENCES

1. FAO (2003). "Medium-term prospects for agricultural commodities." www.fao.org. Website. 22 March 2015.

2. Fausto B. (1999). *A Concise History of Brazil*. New York, USA. Cambridge University Press. 324.

3. Sunarharum W., Williams D., Smyth H. (2014). Complexity of coffee flavor: A compositional and sensory perspective. *Food Research International.* 62: 315-325.

4. Luna F., Crouzillat D., Cirou L., Bucheli P. (2002). Chemical composition and flavor of ecuadorian cocoa liquor. *Journal of Agricultural and Food Chemistry*. 50(12): 3527-3532.

5. Ransom F. (1912). The effects of caffeine upon the germination and growth of seeds. *Biochemical Journal*. 6(2): 151-155.

6. Fernandes F. L., Picanco M. C., Fernandes M. E. S., Queiroz R. B., Xavier V. M., Martinez H. E. P. (2012). The effects of nutrients and secondary compounds of coffea arabica on the behavior and development of coccus viridis. *Environmental Entomology*. 41(2): 333-341.

7. Mathavan S., Premalatha Y., Christopher M. (1985). Effects of caffeine and theophylline on the fecundity of 4 lepidopteran species. *Experimental Biology*. 44(2): 133-138.

8. Aneja M., Gianfagna T. (2001). Induction and accumulation of caffeine in young, actively growing leaves of cocoa (Theobroma cacao L.) by wounding or infection with crinipellis perniciosa. *Physiological and Molecular Plant Pathology*. 59(1): 13-16.

9. Naef R. Jaquier A., Velluz A., Bachofen B. (2004). From the linden flower to linden honey--volatile constituents of linden nectar, the extract of bee-stomach and ripe honey. *Chemistry & Biodiversity*. 1(12): 1870-1879.

10. Takeda Y. (1994). Differences in caffeine and tannin contents between tea cultivars, and application to tea breeding. *Japan Agricultural Research Quarterly*. 28(2): 117-123.

11. Aashihara H., Shimizu H., Takeda Y., Suzuki T., Gillies F., Crozier A. (1995). Caffeine metabolism in high and low caffeine containing cultivars of camellia sinensis. *Zeitschrift für Naturforschung*. 50(9-10): 602-607.

12. Stewart I. (1985). Identification of caffeine in citrus flowers and leaves. *Journal of Agricultural and Food Chemistry*. 33(6): 1163-1165.

13. Mazzafera P., Crozier A., Magalhaes A. C. (1991). Caffeine metabolism in Coffea arabica and other species of coffee. *Phytochemistry*. 30(12): 3913-3916.

14. Ashihara H., Monteiro A., Gillies F., Crozier A. (1996). Biosynthesis of caffeine in leaves of coffee. *Plant Physiology*. 111(3): 747-753.

15. Zheng X., Ashihara H. (2004). Distribution, biosynthesis and function of purine and pyridine alkaloids in Coffea arabica seedlings. *Plant Science*. 166(3): 807-813.

16. Burdock G., Carabin I. (2009). Safety assessment of kola nut extract as a food ingredient. *Food and Chemical Toxicology*. 47(8): 1725-1732.

17. Murakami N., Amboni R., Prudencio E., Amante E., Freire C., Boaventura B., Munoz I., Branco C., Salvador M., Maraschin M. (2013). Concentration of biologically active compounds extracted from Ilex paraguariensis St. Hil. by nanofiltration. *Food Chemistry*. 141(1): 60-65.

18. Zheng X. Q., Koyama Y., Nagai C., Ashihara H. (2004). Biosynthesis, accumulation and degradation of theobromine in developing Theobroma cacao fruits. *Journal of Plant Physiology*. 161(4): 363-369.

19. Koyama Y., Tomoda Y., Kato M., Ashihara H. (2003). Metabolism of purine bases, nucleosides and alkaloids in theobromine-forming Theobroma cacao leaves. *Plant Physiology and Biochemistry*. 41(11-12): 977-984.

20. Scarpari L., Meinhardt L., Mazzafera P., Pomella A., Schiavinato M., Cascardo J., Pereira G. (2005). Biochemical changes during the development of witches' broom: the most important disease of cocoa in Brazil caused by Crinipellis perniciosa. *Journal of Experimental Botany*. 56(413): 865-877.

21. Weckerle C., Stutz M., Baumann T. (2003). Purine alkaloids in Paullinia. *Phytochemistry*. 64(3): 735-742.

22. Wright G. A., Baker D. D., Palmer M. J., Stabler D., Mustard J. A., Power E. F., Borland A. M., Stevenson P. C. (2013). Caffeine in floral nectar enhances a pollinator's memory of reward. *Science*. 339(6124): 1202-1204.

23. Thomson J. D., Draguleasa M. A., Tan M. G. (2015). Flowers with caffeinated nectar receive more pollination. *Arthropod-Plant Interactions*. 9(1): 1-7.

24. Ashihara H., Suzuki T. (2004). Distribution and biosynthesis of caffeine in plants. *Frontiers in Bioscience*. 9(1-3): 1864-1876.

25. Mazzafera P., Crozier A., Magalhaes A. (1991). Caffeine metabolism in Coffea arabica and other species of coffee. *Phytochemistry*. 30(12): 3913-3916.

26. Ogawa M., Herai Y., Koizumi N., Kusano T., Sano H. (2001). 7-methylxanthine methyltransferase of coffee plants - gene isolation and enzymatic properties. *Journal of Biological Chemistry*. 276(11): 8213-8218.

27. Uefuji H., Ogita S., Yamaguchi Y., Koizumi N., Sano H. (2003). Molecular cloning and functional characterization of three distinct N-methyltransferases involved in the caffeine biosynthetic pathway in coffee plants. *Plant Physiology*. 132(1): 372-380.

28. Kato M., Mizuno K., Fujimura T., Iwama M., Irie M., Crozier A., Ashihara H. (1999) Purification and characterization of caffeine synthase from tea leaves. *Plant Physiology*. 120(2): 579-586.

29. Kato M., Kanehara T., Shimizu H., Suzuki T., Gillies F. M., Crozier A., Ashihara H. (1996). Caffeine biosynthesis in young leaves of Camellia sinensis: in vitro studies on N-methyltransferase activity involved in the conversion of xanthosine to caffeine. *Physiologia Plantarum*. 98(3): 629-636.

30. Kato M., Mizuno K., Crozier A., Fujimura T., Ashihara H. (2000). Plant biotechnology: caffeine synthase gene from tea leaves. *Nature*. 406(6799): 956-957.

31. Yoneyama N., Morimoto H., Ashihara C. X., Mizuno K., Kato M. (2006). Substrate specificity of N-methyltransferase involved in purine alkaloidssynthesis is dependent upon one amino acid residue of the enzyme. *Molecular Genetics and Genomics*. 275(2): 125-135.

32. Ashihara H., Kato M., Chuang-Xing Y. (1998). Biosynthesis and metabolism of purine alkaloids in leaves of cocoa tea (Camellia ptilophylla). *Journal of Plant Research*. 111(1104): 599-604.

33. LiJia Q., Shuang L., ShuFan X. (2010). Methylation of phytohormones by the SABATH methyltransferases. *Chinese Science Bulletin*. 55(21): 2211-2218.

34. Zubieta C., Ross J., Koscheski P., Yang Y., Pichersky E., Noel J. (2003). Structural basis for substrate recognition in the salicylic acid carboxyl methyltransferase family. *Plant Cell*. 15(8): 1704-1716.

35. McCarthy A. A., McCarthy J. G. (2007). The structure of two N-methyltransferases from the caffeine biosynthetic pathway. *Plant Physiology*. 144(2): 879-889.

36. Zhao N., Ferrer J., Ross J., Guan J., Yang Y., Pichersky E., Noel J. P., Chen F. (2008). Structural, biochemical, and phylogenetic analyses suggest that indole-3-acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family. *Plant Physiology*. 146(2): 455-467.

37. Ashihara H., Kubota H. (1987). Biosynthesis of purine alkaloids in Camellia plants. *Plant and Cell Physiology*. 28(3): 535-539.

38. Bergthorsson U., Andersson D. I., Roth J. R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences USA*. 104(43): 17004-17009.

39. Force A., Lynch M., Pickett F., Amores A., Yan Y., Postlethwait J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. 151(4): 1531-1545.

40. Finnigan G. C., Hanson-Smith V., Stevens T. H., Thornton J. W. (2012). Evolution of increased complexity in a molecular machine. *Nature*. 481(7381): 360-364.

41. Nasvall J., Sun L., Roth J. R., Andersson D. I. (2012). Real-time evolution of new genes by innovation, amplification, and divergence. *Science*. 338(6105): 384-387.

42. Deng C., Cheng C.-H. C., Ye H., He X., Chen L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences USA*. 107(50): 21593-21598.

43. Des Marais D. L., Rausher M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*. 454(7205): 762-765.

44. Hittinger C. T., Carroll S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*. 449(7163): 677-681.

45. Kliebenstein D. J., Osbourn A. (2012). Making new molecules – evolution of pathways for novelmetabolites in plants. *Current Opinion in Plant Biology*. 15(4): 415-423.

46. Boada J., Roig T., Perez X., Gamez A., Bartrons R., Cascante M., Bermudez J. (2000). Cells overexpressing fructose-2,6-bisphosphatase showed enhanced pentose phosphate pathway flux and resistance to oxidative stress. *FEBS Letters*. 480(2-3): 261-264.

47. Lin J., Mazarei M., Zhao N., Zhu J. J., Zhuang X., Liu W., Pantalone V. R., Arelli P. R., Stewart C. N. Jr., Chen F. (2013). Overexpression of a soybean salicylic acid methyltransferase gene confers resistance to soybean cyst nematode. *Plant Biotechnology Journal*. 11(9): 1135-1145.

48. Parry M. A. J., Madgwick P. J., Carvalho J. F. C., Andralojc P. J. (2007). Prospects for increasing photosynthesis by overcoming the limitations of Rubisco. *Journal of Agricultural Science*. 145(1): 31-43.

49. Harpel M. R., Hartman F. C. (1992). Enhanced C02/02 specificity of a site-directed mutant of Ribulosebisphosphate Carboxylase/Oxygenase. *The Journal of Biological Chemistry*. 267(10): 6475-6478.

50. Ishikawa C., Hatanaka T., Misoo S., Miyake C., Fukayama H. (2011). Functional incorporation of sorghum small subunit increases the catalytic turnover rate of Rubisco in transgenic rice. *Plant Physiology*. 156(3): 1603-1611.

51. Weng J., Philippe R. N., Noel J. P. (2012). The rise of chemodiversity in plants. *Science*. 336(6089): 1667-1670.

52. Boycheva S., Daviet L., Wolfender J., Fitzpatrick T. (2014). The rise of operon-like gene clusters in plants. *Trends in Plant Science*. 19(7): 447-459.

53. Olson-Manning C. F., Lee C., Rausher M. D., Mitchell-Olds T. (2013). Evolution of flux control in the glucosinolate pathway in arabidopsis thaliana. *Molecular Biology and Evolution*. 30(1): 14-23.

54. Kliebenstein D., Kroymann J., Brown P., Figuth A., Pedersen D., Gershenzon J., Mitchell-Olds T. (2001). Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiology*. 126(2): 811-25.

55. Thornton J.W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews Genetics*. 5: 336-375.

56. Harms M. J., Thornton J. W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology*. 20(3): 360-366.

57. Chang B. S. W., Donoghue M. J. (2000). Recreating ancestral proteins. *Trends in Ecology and Evolution*. 15(3): 109-114.

58. Katoh K., Standley D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 30(4): 772-780.

59. Yang Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*. 24(8): 1586-1591.

60. Wikstrom N., Savolainen V., Chase M. (2003). Angiosperm divergence times: congruence and incongruence between fossils and sequence divergence estimates. *Systematics Association Special Volume Series*. 66:142-165.

61. Ashihara H., Sano H., Crozier A. (2008). Caffeine and related purine alkaloids: biosynthesis, catabolism, function and genetic engineering. *Phytochemistry*. 69(4): 841-856.

62. DeSarro A., Grasso S., Zappala M., Nava F., DeSarro G. (1997). Convulsant effects of some xanthine derivatives in genetically epilepsy-prone rats. *Naunyn-Schmiedeberg's Archives of Pharmacology*. 356(1): 48-55.

63. Shi D., Daly J. (1999). Chronic effects of xanthines on levels of central receptors in mice. *Cellular and Molecular Neurobiology*. 19(6): 719-732.

64. Thornton J. (2001). Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proceedings of the National Academy of Sciences USA*. 98(10): 5671-5676.

65. Bridgham J., Carroll S., Thornton J. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science*. 312(5770): 97-101.

66. Carroll S. M., Bridgham J. T., Thornton J. W. (2008). Evolution of hormone signaling in elasmobranchs by exploitation of promiscuous receptors. Molecular Biology and Evolution. 25(12): 2643-2652.

Appendix A

Project approval certification

WESTERN MICHIGAN UNIVERSITY

Recombinant DNA Biosafety Committee

Centennial 1903-2003 Celebration

Recombinant DNA Biosafety Committee

Project Approval Certification

For rDNA Biosafety Committee Use Only

Project Title: Ghosts of Evolution Past: Resurrecting Ancestral Enzymes to Understand the Evolution of Modern-Day Enzyme Activities

Principal Investigator: Todd Barkman

RDBC Project Number: 12-TBa

Date Received by the rDNA Biosafety Committee: October 31, 2011

☒ Reviewed by the rDNA Biosafety Committee

☒ Approved

☐ Approval not required

/ Karin Essani

Vice Chair of rDNA Biosafety Committee Signature

12/01/2011
Date

Revised 11/2011 WMU RDBC
All other copies obsolete

Walwood Hall, Kalamazoo MI 49008-5456
PHONE: (616) 387-8293 FAX: (616) 387-8276

Appendix B

Approval of radioactive material usage

WESTERN MICHIGAN UNIVERSITY

Office of the Vice President for Research

March 25, 2015

Dear Mr. Andrew J O'Donnell:

Congratulations on completing your thesis. I wish you good luck in your future endeavors.

Please keep a copy of this letter with your records as proof of your successful completion of Western Michigan University's (WMU) Basic Radiation Safety Program. The topics in which you demonstrated proficiency are radiation safety terms and definitions, radiation types and sources, radiation signs and symbols, biological effects of ionizing radiation, contamination control and work practices, ALARA (Time, Distance, & Shielding), waste management, emergency response, regulatory agencies, radiation detection instrument theory and use, quality control, and university specific procedures, policies, and practices.

In addition, I commend you on the safe handling, use, and storage/security of radioactive materials throughout your time here at WMU.

In the event you need information concerning your approval or use of radioactive material, please direct your correspondence to:

Radiation Safety Officer
Western Michigan University
Office of Vice President For Research
1903 W. Michigan Ave.
Kalamazoo, MI., 49008-5410

Respectfully,

James F. Center

Mr. James F. Center
Radiation Safety Officer
3922 Wood Hall
Western Michigan University
1903 W. Michigan Ave.
Kalamazoo, MI 49008-5410

Ph: 269.387.5933 Fax: 269.387.5888

Office of the Vice President for Research
1903 W. Michigan Ave., Kalamazoo, MI 49008-5456
PHONE: (269) 387-8298, FAX: (269) 387-8276
WEBSITE: wmich.edu/research

LOCATION: Walwood Hall