8-2017

Evolution of Caffeine Biosynthetic Enzymes and Pathways in Flowering Plants

Ruiqi Huang
Western Michigan University, ruiqi.huang@hotmail.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations
Part of the Biology Commons, and the Molecular and Cellular Neuroscience Commons

Recommended Citation
Huang, Ruiqi, "Evolution of Caffeine Biosynthetic Enzymes and Pathways in Flowering Plants" (2017). Dissertations. 3169.
https://scholarworks.wmich.edu/dissertations/3169

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
EVOLUTION OF CAFFEINE BIOSYNTHETIC ENZYMES AND PATHWAYS IN FLOWERING PLANTS

by

Ruiqi Huang

A dissertation submitted to the Graduate College in partial fulfillment of the requirements for the degree of Doctor of Philosophy
Biological Sciences
Western Michigan University
August 2017

Doctoral Committee:

Todd Barkman, Ph.D., Chair
Yan Lu, Ph.D.
Pamela Hoppe, Ph.D.
James Kiddle, Ph.D.
Convergent evolution generally refers to the independent evolution of similar biological function more than once in unrelated species. Caffeine is thought to have evolved by convergence, and is naturally produced through secondary metabolism in plants to defend against pathogen attack and insect feeding or to attract pollinators. The same caffeine biosynthetic pathway has been elucidated in *Camellia* (tea) and *Coffea* (coffee), in which xanthosine is sequentially methylated to caffeine via 7-methylxanthine and theobromine. However, although the same catalysis pathway is used, different (paralogous) enzymes in the SA\textsubscript{MT}/BA\textsubscript{MT}/theobromine synthase (SABATH) multigene family are used in the two species. In my dissertation, first, novel three-step caffeine biosynthetic pathways were characterized in *Citrus* (kumquat) flowers and *Paullinia* (guarana) seeds, and although both species belong to Sapindales, their caffeine synthases that catalyze the new pathways are paralogous, rather than orthologous genes. Interestingly, the caffeine synthases of *Paullinia* are the same as (orthologous) *Camellia* while those in *Citrus* are orthologous to *Coffea*.

Secondly, I used a paleomolecular biology technique to resurrect ancestral caffeine synthases at different evolutionary time points to see whether the ability to synthesize caffeine was an ancestral trait that was inherited only in a small number of modern-day species while lost...
in many other species, or whether the caffeine biosynthetic ability evolved more recently in a limited number of plant species that we see today. Most ancestral caffeine synthases were found to possess activities with a broad range of substrates, and may therefore have been maintained long-term for possibly alternative biochemical functions before they specialized for caffeine biosynthesis, a phenomenon termed exaptation. In my experiments, ancestral caffeine synthases in *Citrus* and *Paullinia* were also experimentally mutated towards their modern-day descendants to show how the three-step caffeine biosynthetic pathways were assembled in these two species, and to test three hypotheses regarding the evolution of metabolic pathways.

Lastly, I characterized the caffeine biosynthetic enzymes and pathway in *Tilia*, where caffeine is produced in *Tilia* flowers, possibly for pollinator attraction. The *Tilia* transcriptome was assembled using the RNA-Seq technique, and one caffeine synthase gene, which is orthologous to the caffeine synthase genes in *Theobroma* (chocolate), *Paullinia* and *Camellia*, was identified and functionally characterized. Based upon xanthine alkaloids found in *Tilia* flowers, possible caffeine biosynthetic pathways functioning in *Tilia* were proposed.
ACKNOWLEDGEMENTS

It gives me great pleasure in acknowledging the influence of my advisor, Dr. Todd Barkman, whose encouragement, supervision and support since 2008 inspired me to pursue the subject of molecular evolution, especially the evolutionary story of caffeine biosynthesis, which ultimately led to the project contained in this dissertation. Besides guiding my research project, he was also a great mentor in my career who taught me how to think like a scientist and how to be a good scientist.

Secondly, I would like to thank my committee members: Dr. Yan Lu, who taught me detailed knowledge of plant physiology and real-time RT-PCR; Dr. Pamela Hoppe, who helped me with DNA digestion & ligation and SDS-PAGE; and Dr. James Kiddle, for his professional suggestions regarding properties of different chemicals and enzymatic reactions. They provided me with invaluable suggestions for my research project during the past five years.

Thirdly, I would like to thank other graduate and undergraduate students in the laboratory, Andrew O’Donnell and Jessica Barboline, who worked together with me to perform most of the experiments in this dissertation and in our newly published paper, and Breck Davis and Troy Burtchett who helped me with the latter part of my research project.

Fourthly, special thanks to the Chemistry department for allowing me to use the GC-MS and HPLC instruments, especially Dr. Kevin Blair who trained me in the use of these instruments with great patience. I also thank Mr. James Center and Mr. John Center, who helped maintain the scintillation counter and disposal of the radioactive waste during the past six years.
Lastly, I would like to thank all those who supported me in any respect during the completion of the project, especially my parents and my wife, who support me all the time overseas.

Ruiqi Huang
Copyright by Ruiqi Huang
2017
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................... ii

LIST OF TABLES .................................................................................................................... viii

LIST OF FIGURES ................................................................................................................... ix

CHAPTER I: INTRODUCTION .................................................................................................. 1

CHAPTER II: EVOLUTION OF CAFFEINE BIOSYNTHETIC ENZYMES AND
PATHWAYS IN MODERN-DAY PLANTS ................................................................................. 7

INTRODUCTION ....................................................................................................................... 7

MATERIALS AND METHODS ................................................................................................. 13

Gene identification and assembly ......................................................................................... 13

Phylogenetic analyses ............................................................................................................. 14

Gene synthesis and cloning .................................................................................................... 14

Protein expression .................................................................................................................. 15

Protein purification ................................................................................................................ 15

In vitro functional analysis .................................................................................................... 16

Calculations of extraction efficiencies for xanthine alkaloids .............................................. 16

Extraction of metabolites from caffeine-producing plant tissues ......................................... 17

High-performance liquid chromatography (HPLC) ............................................................... 17

Liquid Chromatography–Tandem Mass Spectrometry (LC-MS) .......................................... 18

Gas chromatography-mass spectrometry (GC-MS) analysis ................................................ 19

Enzyme kinetics ................................................................................................................... 20

RESULTS ................................................................................................................................. 21
NEW CAFFEINE BIOSYNTHETIC PATHWAYS EVOLVED IN *Citrus japonica* (kumquat) ........ 21

A new caffeine biosynthetic pathway evolved in *Paullinia cupana* (guarana) .......... 28

TCS2 is responsible for the first methylation step of the caffeine biosynthetic pathway in *Camellia sinensis* (tea) ................................................................. 35

DISCUSSION................................................................................................................. 37

CHAPTER III: EVOLUTIONARY ASSEMBLY OF CAFFEINE BIOSYNTHETIC PATHWAYS IN MODERN-DAY PLANTS ................................................................. 43

INTRODUCTION .............................................................................................................. 43

Evolution of metabolic pathways .................................................................................. 43

Ancestral sequence resurrection .................................................................................... 47

MATERIALS AND METHODS ....................................................................................... 50

Resurrection of ancestral caffeine synthases ................................................................ 50

Gene synthesis and cloning ......................................................................................... 51

Generation of ancestral gene variants and forward mutagenesis ............................... 51

Protein expression ........................................................................................................ 52

Protein purification ....................................................................................................... 53

In vitro functional analysis ........................................................................................... 53

High-performance liquid chromatography (HPLC) ..................................................... 54

Liquid Chromatography–Tandem Mass Spectrometry (LC-MS) ............................... 54

Gas chromatography-mass spectrometry (GC-MS) analysis ...................................... 56

Enzyme kinetics ........................................................................................................... 56

RESULTS ...................................................................................................................... 57

Two distinct types of caffeine synthases evolved convergently in rosids and asterids .................................................................................................................. 57
Ancestral CS enzymes were maintained for more than 100 million years with fortuitous secondary activity with 7-methylxanthine .............................................. 57

Historical maintenance of ancestral XMT enzymes for O-methylation for over 100 million years facilitated the evolution of their N-methylation abilities .......................... 63

Exapted multifunctional ancestral CS enzymes facilitate the evolution of novel caffeine biosynthetic pathway in *Paullinia* ................................................................. 66

Exapted multifunctional ancestral XMT enzymes facilitate the evolution of novel caffeine biosynthetic pathways in *Citrus* ................................................................. 70

DISCUSSION ........................................................................................................ 75

CHAPTER IV: FUNCTION OF SABATH GENES IN *Tilia* FLOWERS USING RNA-SEQ AND BIOCHEMICAL STUDIES ................................................................. 79

INTRODUCTION ............................................................................................... 79

MATERIALS AND METHODS ........................................................................... 80

RNA extraction ................................................................................................. 80

RNA-Seq and transcriptome assembly .............................................................. 80

Gene synthesis and cloning .............................................................................. 81

Protein expression ............................................................................................ 81

Protein purification ........................................................................................... 82

In vitro functional analysis .............................................................................. 83

High-performance liquid chromatography (HPLC) .......................................... 83

Heterologous gene expression and gas chromatography-mass spectrometry (GC-MS) analysis ................................................................................................. 84

Quantitative reverse transcription PCR (RT-qPCR) ........................................ 85

RESULTS ......................................................................................................... 86

Phylogenetic analysis predicted possible functions of some full-length SABATH genes in *Tilia* .................................................................................................. 86
Table of Contents – Continued

Tilia9944 is expressed in *Tilia* developing flower buds ........................................... 89

Heterologously expressed Tilia2608 demonstrated higher relative activity with salicylic acid than with benzoic acid .......................................................................................... 90

Heterologously expressed Tilia7163 demonstrated specific activity with indole-3-acetic acid............................................................................................................................. 90

Tilia8820 likely methylates unknown substrates ......................................................... 92

Tilia13434 likely produces insoluble proteins when heterologously expressed........ 92

Tilia14622 codes for a new MT that has high relative activity with phenylacetic acid (PAA) ................................................................................................................................. 94

Tilia9944 methylates xanthine alkaloids ..................................................................... 94

Potential caffeine biosynthetic pathways evolved in *Tilia cordata* (linden) .......... 94

DISCUSSION ...................................................................................................................... 100

CHAPTER V: EXPECT SIGNIFICANCE OF THE STUDY .......................................................... 103

REFERENCES ..................................................................................................................... 104

APPENDICES

A. Sequence alignment of ancestral and modern-day caffeine synthases ............... 116

B. Project approval certification ...................................................................................... 120

C. Approval of radioactive material usage ................................................................. 123
LIST OF TABLES

1. LC-MS scans of products of enzymatic reactions of CitrusXMT1 and CitrusXMT2 with their optimal substrates .......................................................................................................................... 27

2. Kinetic parameters calculated for CitrusXMT1 with different substrates ................................................. 27

3. LC-MS scans of products of enzymatic reactions of PaulliniaCS1 and PaulliniaCS2 with their optimal substrates .......................................................................................................................... 35

4. Kinetic parameters calculated for PaulliniaCS1 and PaulliniaCS2 with their respective substrates .......................................................................................................................... 35

5. LC-MS scans of products of enzymatic reactions of PaulliniaAncCS1, PaulliniaAncCS2, CamelliaAncCS and RAAncCS with different substrates .................................................................................. 60

6. LC-MS scans of products of enzymatic reactions of CitrusAncXMT2 and CoffeaAncXMT with different substrates .......................................................................................................................... 67

7. Kinetic parameters calculated for ancestral CS enzymes with selected substrates ....................... 69

8. LC-MS scans of products of enzymatic reactions of PaulliniaAncCS2 forward mutants with preferred substrates .......................................................................................................................... 70

9. Kinetic parameters calculated for ancestral XMT enzymes with selected substrates ............ 72

10. LC-MS scans of products of enzymatic reactions of CitrusAncXMT2 forward mutants with preferred substrates .......................................................................................................................... 75

11. Posterior probabilities of original and mutated sites in alternative ancestral alleles ........... 119
LIST OF FIGURES

1. Methyl transfer reactions catalyzed by representative SABATH enzymes ....................... 4
2. Phylogenetic relationships of functionally characterized SABATH members .................... 5
3. An angiosperm phylogenetic tree [60] showing that caffeine production evolved at least five times in flowering plant species. .................................................................................. 8
4. Possible caffeine biosynthetic pathways ............................................................................ 9
5. Xanthosine used for caffeine biosynthesis is synthesized from four different routes. ......... 10
6. Phylogenetic relationships of full-length SABATH genes (red lineages) in Citrus with other functionally characterized SABATH members. .............................................. 22
7. Expression levels and tissues of all full-length genes in Citrus genome represented by EST counts ......................................................................................................................... 23
8. Citrus 1g017514m codes for an SAMT that methylates salicylic acid to form methyl salicylate. ....................................................................................................................... 24
9. Mean relative activities (from 0 to 100) of CitrusXMT1 and CitrusXMT2 with eight xanthine alkaloids. ....................................................................................................................... 24
10. LC-MS scans show that CitrusXMT1 methylates xanthine (X) to form 1-methylxanthine (1X), 3-methylxanthine (3X) and theophylline (Tp). ................................................................. 25
11. LC-MS scans show that CitrusXMT1 methylates 1-methylxanthine (1X) to theophylline (Tp). ............................................................................................................................... 26
12. HPLC UV absorbance spectra of metabolites extracted from Citrus japonica developing flower buds of two different developing stages ................................................................. 29
13. LC-MS scans of metabolites extracted from Citrus japonica developing flower buds ....... 30
14. Caffeine is synthesized using XMT genes via two new pathways in Citrus flowers .......... 31
15. PaulliniaCS0 to PaulliniaCS4 are highly represented by ESTs in the fruits/seeds of Paullinia where caffeine accumulation occurred. ................................................................. 32
16. Phylogenetic relationships of potential caffeine synthase genes (red lineages) in *Paullinia* with other functionally characterized SABATH members. ................................................................. 33

17. Mean relative activities (from 0 to 100) of PaulliniaCS1 to PaulliniaCS4 with eight xanthine alkaloids. ......................................................................................................................... 34

18. Caffeine is synthesized using CS genes via a new pathway in *Paullinia* seeds. .............. 34

19. Phylogenetic relationships of *Camellia* TCS1 and TCS2 in the context of other SABATH family members. .................................................................................................................. 36

20. TCS1 and TCS2 are highly represented by ESTs from leaves and buds of *Camellia sinensis*. .................................................................................................................................. 38

21. Mean relative activities (from 0 to 100) of *Camellia* TCS2 with eight xanthine alkaloids. 38

22. HPLC analysis of product from the enzymatic reaction of TCS2 with xanthosine shows that TCS2 methylated xanthosine to form 7-methylxanthine. .................................................................... 39

23. TCS2 catalyzes the first methylation step of the caffeine biosynthetic pathway in *Camellia sinensis*. .......................................................................................................................... 39

24. Schematic comparison of three hypotheses on the origin and evolution of metabolic pathways. ................................................................................................................................. 44

25. Maximum likelihood phylogenetic tree of representative SABATH members with dense samplings of CS (green colored lineages) and XMT (purple colored lineages) sequences showing nodes of estimated ancestral caffeine synthases. ........................................ 58

26. Mean relative activities (from 0 to 100) of ancestral CS enzymes with eight xanthine alkaloids. ................................................................................................................................. 59

27. Mean relative activities (from 0 to 100) of ancestral XMT enzymes with eight xanthine alkaloids. ................................................................................................................................. 64

28. GC-MS spectra show that RAAneXMT methylates benzoic acid and salicylic acid to form methyl benzoate and methyl salicylate, respectively. .......................................................... 65

29. Forward stepwise evolution of caffeine biosynthetic pathway in *Paullinia* and facile mutational basis for the functional diversification of *Paullinia* CS enzymes. ........................................ 69

30. Simultaneous evolution of caffeine biosynthetic pathways in *Citrus* and facile mutational basis for the functional diversification of *Citrus* caffeine synthases. ........................................... 72
31. HPLC UV absorbance spectra show important evolutionary changes by introducing one large-effect mutation P25S into CitrusAncXMT2. ................................................................. 73

32. Tilia9944 contig2 is an alternatively spliced version of contig1. ................................................... 87

33. Phylogenetic analysis of Tilia SABATH genes with functionally characterized SABATH members predicts their possible functions (substrates). ................................................................. 88

34. Relative expression levels of full-length SABATH genes in Tilia developing buds (a) given by the FPKM values and (b) quantified using RT-qPCR. ................................................................. 89

35. GC-MS spectra show that TiliaSAMT primarily methylates salicylic acid to methyl salicylate. ............................................................................................................................... 91

36. Mean relative activities (from 0 to 100) of TiliaIAMT with benzoic acid, salicylic acid, indole-3-acetic acid and eight xanthine alkaloids................................................................. 92

37. SDS-PAGE gels of purified 8820 (a) and Tilia13434 (b). .................................................................... 93

38. GC-MS spectra show that TiliaPAAMT methylates phenylacetic acid to methyl phenylacetate............................................................................................................................ 95

39. Mean relative activities (0 to 100) of crude and purified TiliaCS1 with eight xanthine alkaloids. ................................................................................................................................. 95

40. HPLC spectra show that TiliaCS1 methylates (a) xanthine (X) to form 3-methylxanthine (3X) and (b) 3-methylxanthine (3X) to form theobromine (Tb). ................................................. 96

41. HPLC analysis shows that crude TiliaCS1 methylates (a) 1-methylxanthine (1X) to form theophylline (Tp) and (b) theophylline (Tp) to form caffeine (Cf). ........................................ 97

42. Two potential caffeine biosynthetic pathways (orange arrows and blue arrows) evolved in Tilia based on enzymatic activities of TiliaCS1. ................................................................. 98

43. Accumulated xanthine alkaloids in Tilia (a) developing buds and (b) open flowers. ............ 99

44. Another caffeine synthase gene may exist to complete the caffeine biosynthetic pathway evolved in Tilia flowers. ........................................................................................................ 101
CHAPTER I: INTRODUCTION

How novel gene functions emerged and evolved over evolutionary time has been a popular research area for a long time; gene duplication was proposed to provide raw genetic material for the evolution of new gene functions under natural selection [1-3]. Gene duplication has been documented in genomes of all three domains of life forms (bacteria, archaea and eukarya) [4]. New gene copies may be generated by DNA recombination events such as unequal crossing over, or by RNA-mediated retroposition, or through whole genome duplication [4, 5]. Duplicated genes have different evolutionary fates: 1) genes are maintained by providing an extra amount of beneficial genetic product (increased dosage) [4], 2) genes acquire novel functions [6, 7], 3) genes inherit partial ancestral functions which are complementary to the functions of the other gene copy [8-11], or 4) genes become pseudogenes by accumulating deleterious mutations and may eventually be lost from the genome [4, 12]. In my graduate study, I investigated the molecular basis of functional evolution of duplicated genes by using a paleomolecular biology technique to investigate their functions in the past (ancestor resurrection and functional characterization) and “replaying evolutionary tapes” (mutating resurrected ancestors towards modern-day descendants) [13].

Duplicated genes form various gene families that are seen within genomes today. In my graduate study, I chose the SABATH multigene family in plants to study the functional evolution of duplicated genes. This family has been identified and named based on its three functionally characterized members (SAMT, BAMT, and theobromine synthase) [14], and there were three reasons for choosing it: First, it exists in many plant species (monocots, dicots and mosses), and the number of family members in these plant species varies due to different gene duplication and recombination events. For example, there are 24 SABATH members in Arabidopsis thaliana
41 SABATH members in *Oryza sativa* (rice) [15], 33 SABATH members in *Populus trichocarpa* (poplar) [16] and only 4 SABATH members in *Physcomitrella patens* (moss) [17].

Second, many SABATH family members have been functionally characterized and participate in various secondary metabolism pathways in plants, producing different compounds (see below). Third, three SABATH members have protein crystal structures available [15, 18, 19]. These crystal structures are useful for homology modeling to identify important amino acid residues and for investigating enzyme interactions with different substrates.

In general, the SABATH family encodes a group of methyltransferases (MTs) that are able to transfer a methyl group (-CH\(_3\)) from the methyl donor, S-adenosyl-L-methionine (SAM), to either the carboxyl group, the sulfur group, or ring nitrogen of a substrate, forming S-adenosyl-L-homocysteine (SAH) and O-methylated ester, like methyl benzoate and methyl salicylate, S-methylated ester, like methyl thiobenzoate, or N-methylated compounds such as caffeine [14, 17]. Members in this family that are capable of methylating the carboxyl group of a given substrate include loganic acid methyltransferase (LAMT) [20], salicylic acid methyltransferase (SAMT) [21], benzoic acid methyltransferase (BAMT) [22], jasmonic acid methyltransferase (JMT) [23], indole-3-acetic acid methyltransferase (IAMT) [18], farnesonic acid methyltransferase (FAMT) [24], gibberellic acid methyltransferase (GAMT) [25], cinnamate/p-coumarate methyltransferase (CCMT) [26], anthranilic acid methyltransferase (AAMT) [27], nicotinic acid methyltransferase (NAMT) [28], and p-methoxybenzoic acid methyltransferase (MBMT) [29]. The SABATH member that methylates the sulfur group of a given substrate is the recently characterized moss thiol MT [17]. SABATH members that methylate the ring nitrogen of a given substrate include enzymes that are involved in caffeine biosynthesis, e.g. caffeine synthase (TCS1) characterized in *Camellia* [30], theobromine synthase.
(BTS1) characterized in *Theobroma* [31], 7-methylxanthine methyltransferase (MXMT) [32], xanthosine methyltransferase (XMT) and 3, 7-dimethylxanthine methyltransferase (DXMT) [33] found in *Coffea*, as well as the enzyme that is responsible for the biosynthesis of trigonelline characterized in *Coffea* [34]. Typical chemical reactions catalyzed by SABATH members are shown in Figure 1 and phylogenetic relationships of these members are shown in Figure 2.

The compounds produced by SABATH family members play important roles in plant secondary metabolism. For example, methylated gibberellins and indole-3-acetate regulate plant growth [25, 35]. Methyl jasmonate, methyl salicylate and caffeine are induced plant defense responses against pathogen attack, insect feeding and wounding [36-39]. Methyl benzoate, methyl salicylate and caffeine also attract pollinator or seed-dispersing animals [40] [41] [42].

In my master’s thesis, I illustrated the functional evolution of duplicated SAMT genes in Apocynaceae and Solanaceae lineages in flowering plants [43]. In this dissertation, I focus on the functional evolution of XMT and CS genes in the SABATH family (Figure 2), because genes in the two clades are involved in the biosynthesis of caffeine, which is an important commercial compound and is widely consumed in tea, coffee, chocolate, maté and guarana by humans as a central nervous system stimulant to fight drowsiness [44, 45]. Caffeine is also proposed to be a new type of pest control repellent for slugs and snails because it is more environmentally acceptable [46]. Another focus of this dissertation is how the ability to synthesize caffeine evolved in flowering plants.

This dissertation consists of four chapters. Chapter II describes the characterization of the caffeine biosynthetic pathways and genes present in modern-day *Citrus* (kumquat) and *Paullinia* (guarana). In this chapter, caffeine biosynthetic pathways in the two species were established
Figure 1. Methyl transfer reactions catalyzed by representative SABATH enzymes. All reactions use SAM as the methyl group donor, and the methyl group receptors are either the carboxyl group (a, c, e) or the sulfur group (b) of the substrate, forming methyl esters, or the ring nitrogen of the substrate, forming either trigonelline (e) or caffeine and its precursors (f). Nicotinic acid can be methylated either at the carboxyl group by NAMT or at the ring nitrogen by TgS.
Figure 2. Phylogenetic relationships of functionally characterized SABATH members. The XMT and CS clades include enzymes that participate in caffeine biosynthesis. The name of each member in the two clades has the genus name as the prefix. The prefixes in the names of members in other clades come from the first letter of each word in their Latin names. For example, CoffeaXMT refers to an XMT enzyme characterized from Coffea arabica, while CbSAMT refers to an SAMT enzyme characterized from Clarkia breweri. This naming convention is consistent throughout the dissertation for easy identification of species that possess CS and XMT genes.
from three aspects: First, identified caffeine synthases are highly represented by expressed sequence tags (ESTs) from the tissues where caffeine is accumulated; Second, potential caffeine precursors are the preferred substrates of characterized caffeine synthases; Third, these caffeine precursors suggested by the relative enzymatic activities are also found as metabolites from the caffeine producing tissues. Chapter III discusses how the caffeine biosynthetic pathways were assembled in *Citrus* and *Paullinia*. In this chapter, ancestral caffeine synthase genes were resurrected and experimentally mutated to test three different hypotheses regarding the evolution of multistep metabolic pathways. Chapter IV focuses on the use of RNA-Seq and quantitative reverse transcription PCR (RT-qPCR) techniques as well as metabolomics, gene cloning and enzymatic assays to explore possible caffeine biosynthetic pathways evolved in *Tilia* (linden).
CHAPTER II: EVOLUTION OF CAFFEINE BIOSYNTHETIC ENZYMES AND PATHWAYS IN MODERN-DAY PLANTS

INTRODUCTION

Caffeine has been found in large amounts in coffee seeds, tea leaves, guarana seeds and maté leaves, and there are also some plants which produce trace or large amounts of caffeine or other purine alkaloids (precursors of caffeine) in specific tissues, e.g. flowers of orange [47] and linden [48], and seeds of cocoa [49] and kola [50]. Despite its existence in at least five plant orders in both rosids and asterids (Sapindales, Paullinia and Citrus; Malvales, Theobroma, Cola and Tilia; Ericales, Camellia; Aquifoliales, Ilex; and Gentianales, Coffea) (Figure 3), how plants produce caffeine was only investigated in Camellia and Coffea.

The caffeine biosynthetic pathway and caffeine synthases were elucidated in Camellia sinensis [51, 52] and Coffea arabica [53, 54] using radioactive tracer feeding experiments and crude enzyme extracts from Camellia and Coffea leaves. The major four-step pathway starts with xanthosine, which is converted to 7-methylxanthosine by 7-methylxanthosine synthase [33, 55, 56]; in the second step, 7-methylxanthosine is converted to 7-methylxanthine by a putative N-methyl nucleosidase (uncharacterized); in the third step, 7-methyxanthine is converted to 3,7-dimethylxanthine (theobromine) by theobromine synthase [30, 32]; and in the last step, 3,7-dimethylxanthine is converted to 1,3,7-trimethylxanthine (caffeine) by caffeine synthase [30, 33, 57] (Figure 4). Minor biosynthetic steps, which consist of formation of 3-methylxanthine from xanthine, formation of theophylline and theobromine from 3-methylxanthine, formation of paraxanthine from 7-methylxanthine, or formation of caffeine from theophylline and paraxanthine, were also proposed [45, 58, 59] (Figure 4). These reactions may occur in plants, but have not been identified as major pathways, and xanthine, 3-methylxanthine and theophylline
Figure 3. An angiosperm phylogenetic tree [60] showing that caffeine production evolved at least five times in flowering plant species. Plant orders in which certain species were known to produce caffeine are colored with red branches. It should be noted that not all species in these orders are able to synthesize caffeine [61].
Figure 4. Possible caffeine biosynthetic pathways. The same canonical pathway has been characterized in *Camellia* and *Coffea*, as shown in solid arrows, and proposed minor pathways are shown in dashed arrows. Three distinct enzymes (XMT, MXMT and DXMT) have been characterized in *Coffea* to catalyze each methylation step in the pathway, and a dual-function enzyme (TCS1) has been characterized in *Camellia* to catalyze the last two steps in the pathway.

are generally considered as degradation products of caffeine [52, 62]. In addition, the caffeine biosynthetic route via 1-methylxanthine has never been reported (Figure 4).

Xanthosine is generally considered as an entry point of caffeine biosynthesis in *Camellia* and *Coffea*, and can be supplied in plants through four different routes (Figure 5). The first route (de novo route) is from inosine monophosphate (IMP) generated via de novo purine biosynthesis. IMP is converted to xanthosine monophosphate (XMP) by IMP dehydrogenase, and XMP is then
Figure 5. Xanthosine used for caffeine biosynthesis is synthesized from four different routes. The de novo route is shown with red arrows, the AMP route is shown with green arrows, the SAM route is shown with blue arrows, and the GMP route is shown with purple arrows. The first reaction in each route is labeled with colored numbers.

converted to xanthosine by 5’-nucleotidase [63]. The second route (AMP route) is from adenosine monophosphate (AMP) in the adenine nucleotide pool, AMP is converted to IMP by AMP deaminase, and IMP is then sequentially converted to xanthosine in the same as in the first route [64]. The third route (SAM route) starts from SAM, a universal methyl donor. SAM is firstly converted to SAH after losing the methyl group, and SAH is further hydrolyzed to homocysteine and adenosine by SAH hydrolase. Adenosine is converted to AMP through two routes. In the major route, adenosine is converted to adenine by adenosine nucleosidase, and next, adenine is converted to AMP by adenine phosphoribosyltransferase. In the minor route,
adenosine is converted directly to AMP by adenosine kinase. AMP from both adenine and adenosine is continually converted to xanthosine via the above mentioned AMP route [64]. The last route (GMP route) is from guanosine monophosphate (GMP) in the guanine nucleotide pool. GMP is converted to guanosine by 5’-nucleotidase, and guanosine is converted to xanthosine by guanosine deaminase [65]. Of these four routes, the SAM route was proposed to be able to provide enough xanthosine needed for the caffeine biosynthetic pathway by itself because three molecules of SAH were provided during the three methylation steps in the pathway [62]. In addition, besides its role as the xanthosine precursor, XMP was also hypothesized to be another starting point for caffeine biosynthesis in coffee leaves, and it was methylated to form 7-methyl XMP, which was then converted to 7-methylxanthine quickly [66]. However, the enzyme that is responsible for the methylation of XMP has not been characterized.

To date, genes responsible for the three-methylation steps in the major pathway in Coffea were identified and characterized, and they were named as the XMT-type genes [32, 33, 56, 57] (Figure 4). However, only one gene was characterized in the caffeine biosynthetic pathway in Camellia and was named as the CS-type gene, which codes for a dual functional enzyme (TCS1) that catalyzes the latter two methylation steps of the pathway [30] (Figure 4). The gene that is responsible for the first methylation step (from xanthosine to 7-methylxanthosine) is still unknown, even though a crude extract of enzymes from Camellia leaves has been shown to methylate xanthosine to form 7-methylxanthosine [67]. A candidate gene named TCS2 [accession number: AB031281], which is paralogous to TCS1, was cloned from Camellia sinensis and reported to have no activity with xanthosine [31]. Therefore, the gene responsible for the first step of the caffeine biosynthetic pathway in Camellia sinensis was characterized.
There are no reported studies regarding the caffeine biosynthetic pathways in *Citrus*, *Paullinia* and *Tilia*, although one gene that encodes a dual function enzyme, which catalyzes 7-methylxanthine to form theobromine and then caffeine, was reported recently in *Paullinia* [68]. Even though the caffeine biosynthetic pathways in *Camellia* and *Coffea* are the same, genes responsible for the pathways evolved in parallel and form separate clades in phylogenetic analysis [31] (Figure 2).

To expand understanding of caffeine biosynthesis in other plants, I intended to address two questions in this chapter: Is the caffeine biosynthetic pathway that evolved in *Camellia* and *Coffea* universal in each caffeine producing plant? Are there more different types of caffeine synthases that have evolved in the SABATH family? To answer the two questions, I functionally characterized the enzymes of the caffeine biosynthetic pathways in the flower of *Citrus japonica* (kumquat) and in the seed of *Paullinia cupana* (guarana). One hypothesis is that the caffeine biosynthetic pathways that evolved in *Citrus* and *Paullinia* are the same as the canonical pathway in *Camellia* and *Coffea*, and that genes involved in the pathways are orthologous to the caffeine synthase genes in *Camellia* or in *Coffea*. A second hypothesis is that caffeine biosynthesis in *Citrus* and *Paullinia* utilizes alternative routes, but the genes responsible for the new pathways are orthologous to the caffeine synthase genes in *Camellia* or in *Coffea*. A third hypothesis is that caffeine biosynthesis in *Citrus* and *Paullinia* utilizes alternative routes, and that enzymes that catalyze these pathways are new SABATH members, which are paralogous to the caffeine synthases genes in *Camellia* and *Coffea*. 

MATERIALS AND METHODS

Gene identification and assembly

The *Camellia* caffeine synthase gene TCS1 [accession number: AB031280], was used to search the *Citrus* genome on Phytozome and Genbank using BLAST. The obtained *Citrus* SABATH genes with complete open reading frames (ORFs) from Phytozome and Genbank were then used to search the *Citrus* EST database on Genbank. Sequences of these *Citrus* SABATH genes and their matching ESTs were assembled into contigs in Sequencher (Gene Codes Corp.), and primers were then designed to amplify the target genes from *Citrus* RNA.

Because *Paullinia* genomic sequences are not available, full-length SABATH genomic sequences from *Citrus*, a close relative to *Paullinia*, were used to search the *Paullinia* EST database on Genbank. The obtained *Paullinia* ESTs were assembled into contigs using Sequencher, and sequences of full-length *Paullinia* SABATH genes were determined if these contigs contained sufficient number of ESTs to cover the entire open reading frame (ORF) and possessed 98% identity over a 100-bp window of the coding sequence. Phylogenetic analyses were performed to verify the orthology of each assembled sequence to the query sequence. EST number and tissue type were recorded after orthology was established.

Because there is also no published genome for *Camellia sinensis*, full-length SABATH genomic sequences from *Mimulus* as well as published *Camellia* TCS1 and TCS2 sequences were used to search the *Camellia* transcriptome and EST database on Genbank. The obtained *Camellia* ESTs were assembled into contigs using Sequencher. EST number and tissue type were recorded for each contig.
Phylogenetic analyses

The amino acid sequences of functionally characterized representative SABATH family members, the SABATH genes from *Paullinia* and *Camellia* transcriptomes, as well as *Theobroma* and *Citrus* genomes were obtained from GenBank and the Phytozome database. In addition, several XMT-type and CS-type sequences were obtained from the oneKP database (www.onekp.com) to provide detailed branching relationships of these caffeine synthase genes in *Camellia, Coffea, Paullinia* and *Citrus*. Sequences were aligned using MAFFT version 7 [69] using auto strategy and default parameters, and a maximum-likelihood phylogenetic tree of these sequences was then estimated using PhyML [70] assuming the Jones, Taylor, Thornton (JTT) amino acid substitution model [71] with gamma distribution as chosen by Modeltest [72]. 100 replicates of bootstrap analyses were conducted.

Gene synthesis and cloning

The nucleotide sequences of *Paullinia* caffeine synthase-like genes and MangiferaXMT were synthesized by GenScript Corp. with codons optimized for expression in *E. coli* and cloned into the pUC57 vector. They were either subcloned into pET15b vector (Novagen) or expresso SUMO vector (Lucigen) with a N-terminal His-tag for subsequent protein expression and purification. RNA was extracted from tea leaves and *Citrus* flowers following a published protocol [73]; gene sequences of TCS2, CitrusXMT1 – 4, CitrusCS and CitrusSAMT genes were then amplified through Reverse Transcription (RT)-PCR using specific primers for each gene and cloned into either pET15b vector with a N-terminal His-tag (Novagen) or pTricHis2 TOPO vector with a C-terminal His-tag (Invitrogen) for protein expression and purification. All recombinant plasmids were checked on an agarose gel for their size and sequenced by GENEWIZ Corp to ensure that all cloned sequences were in the correct open reading frame.
Protein expression

Recombinant pET15b plasmids were transformed into *E. coli* strain BL21(DE3) Star™ cells (Invitrogen), recombinant pTricHis2 plasmids were transformed into *E. coli* strain Top10 cells (Invitrogen), and recombinant expresso SUMO plasmids were transformed into HI-Control™ BL21(DE3) cells (Lucigen) for overexpression of His-tagged proteins. Protein expression started from a 5 ml starter culture of transformed *E. coli* cells, which were grown at 37°C in LB broth containing 100 µg/ml ampicillin (or 50 µg/ml Kanamycin if the gene was cloned into the SUMO vector) overnight. On the next day, the 5 ml starter culture was used to inoculate a fresh 100 ml LB broth containing the same concentration of antibiotics and allowed to grow until the OD_{600} reached 0.6 to 0.8, at which point they were induced with 1.0 mM isopropyl 1-thio-galactopyranoside (IPTG) for 6 hours at 23°C. The cells were harvested by centrifugation at 4°C, 4000 rpm for 15 minutes. To isolate proteins from the *E. coli* cells, pellets were resuspended in chilled 1X equilibration buffer (50 mM Na$_3$PO$_4$, 300 mM NaCl, 10 mM imidazole, 10 mM β-mercaptoethanol, and 12% (v/v) glycerol, pH 8), treated with 0.75 mg/ml lysozyme on ice for 30 minutes, and then sonicated on ice in 20-second bursts, 3 times, with a 20-second pause on ice between each burst. The suspension after sonication was centrifuged at 10,000 g for 20 minutes at 4°C to get rid of cell debris. The supernatant containing the soluble crude proteins were either used directly for enzyme catalytic assays or subsequently purified.

Protein purification

Protein purification was achieved using cobalt spin columns (Clontech) according to the manufacturer’s instructions. Purified proteins were eluted in 1X elution buffer (50 mM Na$_3$PO$_4$, 300 mM NaCl, 200 mM imidazole, 10 mM β-mercaptoethanol, and 12% (v/v) glycerol, pH 8), aliquoted and flash frozen in a dry ice-ethanol bath and stored at -80°C. The purity and
molecular mass of purified proteins was evaluated by running on pre-cast NuPAGE 4-12% Bis-Tris gels (Invitrogen), and protein concentration was calculated by the Bradford assay [74] using bovine serum albumin (BSA) as the standard.

**In vitro functional analysis**

Relative enzymatic activities of heterologously expressed SABATH proteins with eight different xanthine alkaloid substrates as well as benzoic acid and salicylic acid were tested using SAM with a $^{14}$C-labeled methyl group. For each substrate, a 50 µl reaction containing 10 or 20 µl purified proteins (or 30 µl crude extracted total cellular proteins), 100 µM methyl acceptor substrate and 0.01 µCi $^{14}$C-labeled SAM was carried out in 50 mM Tris-HCl buffer at room temperature for 20 minutes (or 30 minutes if using crude cellular proteins). The methylated product was extracted using 200 µl ethyl acetate, and 120 µl of the ethyl acetate (top) phase was removed and mixed with 4 ml scintillation counting cocktail. Radioactive decay was measured using a liquid scintillation counter (PerkinElmer) and disintegrations per minute (DPM) values were reported. Assays were done in duplicate and standard deviations were calculated.

**Calculations of extraction efficiencies for xanthine alkaloids**

As different xanthine alkaloids have different solubility parameters in ethyl acetate, (the solvent for extractions of methylated products from our enzymatic assays,) extraction efficiencies were calculated for each mono- and di-methylxanthine as well as caffeine. Extraction efficiency is found irrespective of the xanthine alkaloid concentration, and was calculated as follows: Two 200 µl identical mixtures (A and B) containing 10 µM xanthine alkaloid in 50 mM Tris-HCl buffer (pH 6.5) were prepared. A was left untreated, while B was mixed thoroughly with 800 µl ethyl acetate. After centrifugation of B, 640 µl of the upper ethyl acetate phase was transferred into a new tube (C). The ethyl acetate in B and C was completely
evaporated using vacuum centrifugation. The volume of B was brought back to 200 µl using H₂O and this was the leftover sample. 200 µl of 50 mM Tris-HCl (pH 6.5) solution was added to C to dissolve the precipitates, and this was the extracted sample. 1 µl 10 mM internal standard was added to A, B and C, respectively, and all three of them were then filtered through a 0.2 µm nylon membrane and analyzed using HPLC. Areas of xanthine alkaloid peaks were quantified from the UV spectrum and adjusted to a comparable level using the internal standard. Adjusted area of A (total amount of xanthine alkaloid) should equal to the sum of adjusted areas of B (leftover xanthine alkaloid) and C (extracted xanthine alkaloid), extraction efficiency = area(C) / (area(A) + area(B)).

**Extraction of metabolites from caffeine-producing plant tissues**

Up to 100 mg of *Citrus* developing flower buds were ground up in a 1.5 ml tube, and 1 ml methanol was then added into the tube. The tube containing the ground tissues was vortexed for 1 minute and then incubated at 50°C for 30 minutes. After the incubation, the tube was centrifuged at 13,200 rpm for 5 minutes to pellet the tissue waste, and the methanol phase containing the metabolites was transferred to a new 1.5 ml tube. Methanol in the new tube was evaporated using vacuum centrifugation, and 500 µl 0.1% TFA was added to dissolve the pellets in the new tube. The solution containing the extracted metabolites were centrifuged through a 0.45 µM nylon filter before being analyzed by HPLC.

**High-performance liquid chromatography (HPLC)**

Product identity was determined using HPLC on reactions that were scaled up 10-fold using non-radioactive SAM for two hours at room temperature. The whole reaction was acidified to pH 5~6, then filtered through Vivaspin columns (Sartorius Stedim Biotech) to remove proteins and impurities, and 10 µl of the processed reaction (or metabolite extract) was injected into
HPLC. Mixtures were separated by HPLC using a two-solvent system with a 250 mm × 4.6 mm Kinetex 5µm EVO C18 column (Phenomenex). Solvent A was 99.9% (V/V) water with 0.1% TFA and solvent B was 80% (V/V) acetonitrile, 19.9% (V/V) water, and 0.1% TFA, and a 0–16% solvent B gradient was generated over 16 minutes with a flow rate of 1.0 ml/min. Subsequently, buffer B was increased to 100% and then held at that percentage for 20 minutes. Equilibration back to 0% buffer B was achieved over a 20-minute period. Two UV wavelengths (254 and 272 nm) were used for absorption measurements. The methylated products of enzyme reactions or metabolites from caffeine-producing tissues of plants were verified by comparing its absorbance peak/retention time with xanthine alkaloid standards. Negative control reactions without the substrate added were always run in parallel.

**Liquid Chromatography–Tandem Mass Spectrometry (LC-MS)**

Identities of methylated products from enzymatic assays based on UV absorbance were also confirmed using LC-MS. LC-MS/MS analyses were performed using an Agilent 1100 HPLC inline with a QuattroMicro mass spectrometer. The same C18 column and a different two-solvent system were used to separate and analyze the reaction products: Solvent A was 0.1% formic acid/0.01% trifluoroacetic acid/water, solvent B was 0.1% formic acid/0.01% trifluoroacetic acid/acetonitrile, and a 0–16% solvent B gradient was generated over 16 minutes with a flow rate of 0.5 ml/min followed by 2-minute of 95% B for a run time of 20 minutes. A postcolumn addition of 0.1% formic acid in acetonitrile was added via a PEEK tee at a flow rate of 100 µL/min. Under the above conditions, detection was initially optimized using pure standards of different xanthine alkaloids diluted to 1 µM in 0.1% formic acid/50% (V/V) acetonitrile and infused directly onto a Waters Quattro Micro mass spectrometer via an electrospray ion source, and mass spectrometry scans successfully separated and identified
xanthine, 1-, 3-, 7-methylxanthine, theophylline, theobromine, paraxanthine and caffeine based on their retention time and parent ion - fragment ion fragmentation pattern differences. First, xanthine is clearly separated from all other alkaloids and shows a unique 153.4 > 110 fragmentation pattern and a retention time of 4.67 minutes. Second, 1-methylxanthine has a different retention time (9.36 minutes) from all other alkaloids and a unique 167.4 > 110 fragmentation pattern. Third, 3-methylxanthine and 7-methylxanthine may be distinguished from their different retention times (8.82 and 8.23 minutes, respectively) even though they both show a 167.4 > 124 fragmentation pattern. Fourth, the 181.5 > 138 fragmentation pattern is specific for theobromine, which also has a retention time (12.09 minutes) different from other alkaloids. Fifth, because paraxanthine and theophylline are difficult to separate under these LC conditions (the retention time is 12.70 minutes for paraxanthine and 12.92 minutes for theophylline) and both have a common 181.5 > 124 fragment, a specific 181.5 > 96 fragment was scanned for theophylline and a 181.5 > 55 fragment was scanned for paraxanthine. Lastly, caffeine separated far from all other alkaloids with a retention time of 15.69 minutes, and had a specific 195 > 138 fragmentation pattern that allowed its unique identification. For each identified xanthine alkaloid peak, only when raw peak area was larger than 25, (an arbitrary cutoff value, which was 1/10 the area of the lowest intensity standard,) was it recorded as the real xanthine alkaloid peak.

Gas chromatography-mass spectrometry (GC-MS) analysis

Reaction products of SAMT with salicylic acid and benzoic acid were verified using GC-MS. A 50 ml transformed E. coli Top10 cell culture containing the recombinant pTricHis2 plasmid was grown to optical density (OD) 0.6 - 0.8 prior to addition of IPTG. After induction with 1 mM IPTG for 4 hours at 23°C, 200 µl of 50 mM salicylic acid or benzoic acid was added to the culture, which continued to grow for another 2 hours at 23°C. Following the 2-hour
incubation with the substrate, cells were pelleted at 4°C, 4000 rpm for 15 minutes, and the supernatant (spent growth medium) was collected. Volatiles were extracted from the supernatant with 4 ml hexane and 2 ml recovered hexane phase was transferred to a vial and ready for GC-MS analysis. The negative control was set up the same way except that no substrate was added after IPTG induction. One µl concentrated product was analyzed by GC-MS on an HP6890GCSystem coupled to an HP5973 Mass Selective Detector using a DB-5 capillary column with the following oven conditions: 40°C for 2 minutes, ramping 20°C / minute to 300°C with a 2-minute hold. Product was identified by comparing its mass spectra and retention time with available standards and reference spectra from the library of National Institute of Standards and Technology.

**Enzyme kinetics**

Kinetic parameters ($k_{\text{cat}}$ and $K_M$) of these methyltransferases with a given substrate were determined using the 50 µl radioactive assay described above. Appropriate enzyme concentration and incubation time were determined in time-course assays with low unsaturated substrate concentrations, such that the reaction velocity was linear during the assay period. When varying substrate concentration, the SAM concentration was held constant and saturated at 320 µM. Initial velocities versus substrate concentration were plotted using GraphPad Prism (GraphPad Software, La Jolla, CA) to fit the hyperbolic Michaelis-Menten equation to calculate $V_{\text{max}}$ and $K_M$. $V_{\text{max}}$ was corrected using extraction efficiency and converted to $k_{\text{cat}}$ based on estimated protein concentrations and expressed in units of s$^{-1}$. Assays were done in duplicate and standard errors were calculated using GraphPad Prism.
RESULTS

New caffeine biosynthetic pathways evolved in *Citrus japonica* (kumquat)

Nineteen full-length SABATH genomic sequences, which were named using their genome accession numbers, were found in the *Citrus* genome, and six of them (1g018139m, 1g036911m, 1g044727m, 1g047625m, XM_006469386/7, XM_006469416) were identified as caffeine synthase-like genes as they were either phylogenetically close to the CS genes in *Camellia* or XMT genes in *Coffea* (Figure 6). The number of ESTs that were assembled into the same contig with each full-length gene was counted and their tissue origins were recorded (Figure 7). This EST chart showed that only 1g047625m, 1g044727m and 1g017514m have large numbers of ESTs from flowers where caffeine is reported to accumulate (Figure 7). Six genes (1g017514m, 1g018139m, 1g044727m, 1g047625m, XM_006469386/7, XM_006469416) were successfully amplified from the *Citrus japonica* flower mRNA (no amplification occurred for 1g036911m). I demonstrated that 1g017514m codes for an SAMT without any activities with xanthine alkaloids (Figure 8). In contrast, heterologously expressed 1g047625m and 1g044727m have demonstrated noticeable activities with some caffeine precursors tested, thus they are renamed as CitrusXMT1 and CitrusXMT2 (Figure 9).

My enzymatic assays showed that CitrusXMT1 methylates xanthine to form both 1- and 3-methylxanthine (Figure 10), and it also methylates 1- methylxanthine (Figure 11) and 3-methylxanthine (Table 1) to form theophylline. CitrusXMT2 methylates theophylline to form caffeine (Table 1). Although CitrusXMT1 can convert both 1- and 3-methylxanthine to theophylline, 1-methylxanthine is suggested to be a preferred substrate over 3-methylxanthine due to the kinetic properties of CitrusXMT1 with 1- and 3-methylxanthine (Table 2).
Figure 6. Phylogenetic relationships of full-length SABATH genes (red lineages) in *Citrus* with other functionally characterized SABATH members. One CS gene and five XMT genes are potentially involved in caffeine biosynthesis in *Citrus*. Lineages with functionally characterized sequences are labeled by enzyme names, whereas those without known functions are arbitrarily numbered from MT1 to MT3.
Figure 7. Expression levels and tissues of all full-length genes in *Citrus* genome represented by EST counts. Only lg047625m, lg044727m and lg017514m have more than five ESTs expressed in flowers.

The $K_M$ of 3-methylxanthine is 1.7 times higher than the $K_M$ of 1-methylxanthine, and the $k_{cat}/K_M$ of 1-methylxanthine is two times higher than that of 3-methylxanthine, further suggesting that if both 1- and 3- methylxanthine are present in the flower, 1-methylxanthine will preferably be converted into theophylline first.
Figure 8. *Citrus* 1g017514m codes for an SAMT that methylates salicylic acid to form methyl salicylate. a. Mean relative activities (from 0 to 100) of CitrusSAMT with benzoic acid, salicylic acid and eight xanthine alkaloid substrates. b. GC-MS analysis of reaction products of CitrusSAMT and salicylic acid.

Figure 9. Mean relative activities (from 0 to 100) of CitrusXMT1 and CitrusXMT2 with eight xanthine alkaloids. a. CitrusXMT1 displays high relative activities with 1- and 3-methylxanthine, and low activity with xanthine. b. CitrusXMT2 is highly specific for theophylline.
Figure 10. LC-MS scans show that CitrusXMT1 methylates xanthine (X) to form 1-methylxanthine (1X), 3-methylxanthine (3X) and theophylline (Tp). Tp is possibly methylated from both 1X and 3X. Scans for authentic standards are shown above to confirm product identities. Xanthine alkaloid abbreviations: 7X, 7-methylxanthine; Tb, theobromine; Px, paraxanthine; Cf, caffeine.
Figure 11. LC-MS scans show that CitrusXMT1 methylates 1-methylxanthine (1X) to theophylline (Tp). Scans for authentic standards are shown above to confirm product identities. Xanthine alkaloid abbreviations: X, xanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tb, theobromine; Px, paraxanthine; Cf, caffeine.
Table 1. LC-MS scans of products of enzymatic reactions of CitrusXMT1 and CitrusXMT2 with their optimal substrates

<table>
<thead>
<tr>
<th>Enzyme + Substrate</th>
<th>X Area</th>
<th>1X Area</th>
<th>3X Area</th>
<th>7X Area</th>
<th>Tp Area</th>
<th>Tb Area</th>
<th>Px Area</th>
<th>Cf Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard eight xanthine alkaloids mix</td>
<td>329</td>
<td>1474</td>
<td>292</td>
<td>625</td>
<td>550</td>
<td>281</td>
<td>721</td>
<td>2088</td>
</tr>
<tr>
<td>CitrusXMT1 + 3X</td>
<td>155</td>
<td>334</td>
<td></td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CitrusXMT1 + Solvent</td>
<td>127</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CitrusXMT2 + Tp</td>
<td>54</td>
<td></td>
<td></td>
<td>1010</td>
<td></td>
<td></td>
<td></td>
<td>74</td>
</tr>
<tr>
<td>CitrusXMT2 + Solvent</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Numbers represent raw peak areas. Substrate areas are colored in blue and product areas are colored in red. The presence of xanthine in all reactions is probably due to other chemical reactions in bacteria. Xanthine alkaloid abbreviations: X, xanthine; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tp, theophylline; Tb, theobromine; Px, paraxanthine; Cf, caffeine.

Table 2. Kinetic parameters calculated for CitrusXMT1 with different substrates

<table>
<thead>
<tr>
<th>Enzyme + substrate</th>
<th>$K_M$ (µM)</th>
<th>$k_{cat}$ (s$^{-1}$)</th>
<th>$k_{cat}/K_M$ (sec$^{-1}$·M$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CitrusXMT1 + xanthine</td>
<td>758.5</td>
<td>8.74E-05</td>
<td>0.12</td>
</tr>
<tr>
<td>CitrusXMT1 + 1-methylxanthine</td>
<td>657.6</td>
<td>6.32E-04</td>
<td>0.97</td>
</tr>
<tr>
<td>CitrusXMT1 + 3-methylxanthine</td>
<td>1144</td>
<td>2.99E-04</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Because XM_006469386/7, XM_006469416 are closely related to CitrusXMT1 and CitrusXMT2, they are renamed as CitrusXMT3 and CitrusXMT4, respectively, and 1g018139m is in the same clade with the caffeine synthase genes in *Camellia*, thus is renamed as CitrusCS (Figure 6). However, no activities were detected from heterologously expressed CitrusXMT3, CitrusXMT4 and CitrusCS with eight xanthine alkaloids (data not shown), and none of them are represented by large numbers of flower ESTs (Figure 7), so they may not participate in the caffeine biosynthetic pathway in *Citrus*.  

27
HPLC spectra and LC-MS scans of metabolites from *Citrus japonica* developing flower buds show the accumulation of xanthine, 1- and 3- methylxanthine, theophylline and caffeine (Figures 12 and 13). These metabolites detected in *Citrus* flower buds match the substrates in the enzymatic assays that are catalyzed by these two *Citrus* caffeine synthases (CitrusXMT1 and CitrusXMT2), suggesting two alternative pathways for caffeine biosynthesis in *Citrus*: Xanthine is firstly methylated to form both 1- and 3-methylxanthine, which are then methylated to form theophylline, and these two steps are catalyzed by a single enzyme - CitrusXMT1, and theophylline is methylated to form caffeine by a second enzyme - CitrusXMT2 (Figure 14).

**A new caffeine biosynthetic pathway evolved in *Paullinia cupana* (guarana)**

A previously characterized dual-functional enzyme PcCS [accession number: DAA64605] is reported to methylate 7-methylxanthine to form theobromine and theobromine to form caffeine in *Paullinia* [75]. However, the enzyme responsible for the biosynthesis of 7-methylxanthine is still unknown, and the caffeine biosynthetic pathway in *Paullinia* is still unclear. By using *Citrus* SABATH sequences to BLAST against the *Paullinia* EST database on Genbank, five full-length caffeine synthase-like sequences were assembled in contigs, and named PaulliniaCS0 to PaulliniaCS4 (Figure 15), and these ESTs are sampled from either fruits or seeds of *Paullinia*, where caffeine is reported to accumulate. No XMT-type gene is found in *Paullinia*’s transcriptome (Figures 15 and 16). Phylogenetic analysis of these *Paullinia* caffeine synthase-like sequences shows that all of them are closely related to the caffeine synthase genes in *Camellia, Theobroma* and *Ilex* (Figure 16). PaulliniaCS1 putatively represented the previously published GRN06, which was cloned from *Paullinia* without any functional characterization [76]. PaulliniaCS3 is a minor sequence variant of PaulliniaCS1, and PaulliniaCS4 is a minor
Figure 12. HPLC UV absorbance spectra of metabolites extracted from *Citrus japonica* developing flower buds of two different developing stages. Potential caffeine (Cf) precursors detected include xanthine (X), 1-methylxanthine (1X), 3-methylxanthine (3X) and theophylline (Tp). These precursors are consistent with the substrates of the characterized CitrusXMT1 and CitrusXMT2. Xanthine alkaloid abbreviations: 7X, 7-methylxanthine; Tb, theobromine; Px, paraxanthine; Cf, caffeine.
Figure 13. LC-MS scans of metabolites extracted from Citrus japonica developing flower buds. Detected xanthine alkaloid peaks are enclosed in blue columns. Potential caffeine (Cf) precursors accumulated include xanthine (X), 1-methylxanthine (1X), 3-methylxanthine (3X) and
theophylline (Tp). These precursors are consistent with the substrates of the characterized CitrusXMT1 and CitrusXMT2. No 7-methylxanthine (7X), theobromine (Tb) and paraxanthine (Px) were found.

Figure 14. Caffeine is synthesized using XMT genes via two new pathways in *Citrus* flowers. Pathways start from xanthine, instead of xanthosine, and the intermediates are 1-methylxanthine, 3-methylxanthine and theophylline, completely different from the canonical pathway in *Camellia* and *Coffea*. Two enzymes (CitrusXMT1 and CitrusXMT2) are responsible for the three methylation steps and CitrusXMT1 is a multifunctional enzyme that catalyzes the first two methylation steps in each pathway.

sequence variant of PaulliniaCS2. Heterologously expressed PaulliniaCS1 and PaulliniaCS3 showed similar relative activities, and the same for PaulliniaCS2 and PaulliniaCS4 (Figure 17), so PaulliniaCS3 and PaulliniaCS4 were not further investigated. I couldn’t detect any in vitro activity of PaulliniaCS0 with the eight xanthine alkaloids (data not shown). LC-MS scans of reaction products show that PaulliniaCS1 methylates xanthine to 3-methylxanthine, and PaulliniaCS2 methylates 3-methylxanthine to form theobromine (Table 3). Both PaulliniaCS1 and PaulliniaCS2 have substrate affinities (as reflected by $K_M$) comparable to characterized caffeine synthases that are involved in the caffeine biosynthetic pathway in *Coffea* and *Camellia*,

1-methylxanthine

3-methylxanthine

Xanthine

CitrusXMT1

CitrusXMT1

CitrusXMT1

CitrusXMT2

Theophylline

Caffeine

H$_3$C

H$_3$C

H$_3$C
e.g. XMTs in Coffea [33] and TCS1 in Camellia [58] (Table 4). Although no xanthine, xanthosine, or mono-methylxanthine (e.g. 3-methylxanthine) were ever reported to accumulate in any Paullinia tissues, theobromine and theophylline are reported in Paullinia seeds, where most caffeine accumulates [68, 77, 78]. Relative activities of PaulliniaCS1, PaulliniaCS2 and PcCS, together with the metabolites found in Paullinia seeds suggest that the caffeine production in Paullinia may have evolved a novel biosynthetic pathway, which is different from the pathways in Camellia, Coffea and Citrus: First, PaulliniaCS1 methylates xanthine to form 3-methylxanthine; Second, PaulliniaCS2 methylates 3-methylxanthine to form theobromine; Third, theobromine is methylated to form caffeine by PcCS (Figure 18).

Figure 15. PaulliniaCS0 to PaulliniaCS4 are highly represented by ESTs in the fruits/seeds of Paullinia where caffeine accumulation occurred. No EST matches to other SABATH members are found in fruits/seeds.
Figure 16. Phylogenetic relationships of potential caffeine synthase genes (red lineages) in *Paullinia* with other functionally characterized SABATH members. PaulliniaCS0 to PaulliniaCS4 are closely related to CS genes in *Camellia, Theobroma* and *Ilex*.
Figure 17. Mean relative activities (from 0 to 100) of PaulliniaCS1 to PaulliniaCS4 with eight xanthine alkaloids. Both PaulliniaCS1 and PaulliniaCS3 display highest relative activity with xanthine, and both PaulliniaCS2 and PaulliniaCS4 have highest relative activity with 3-methylxanthine.

Figure 18. Caffeine is synthesized using CS genes via a new pathway in Paullinia seeds. The pathway starts from xanthine, instead of xanthosine, and the intermediates are 3-methylxanthine and theobromine, different from the pathways in Camellia, Coffea and Citrus. Three distinct enzymes (PaulliniaCS1, PaulliniaCS2 and PcCS) are responsible for each methylation step in the pathway, respectively.
Table 3. LC-MS scans of products of enzymatic reactions of PaulliniaCS1 and PaulliniaCS2 with their optimal substrates

<table>
<thead>
<tr>
<th>Enzyme + Substrate</th>
<th>X Area</th>
<th>1X Area</th>
<th>3X Area</th>
<th>7X Area</th>
<th>Tp Area</th>
<th>Tb Area</th>
<th>Px Area</th>
<th>Cf Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard eight xanthine alkaloids mix</td>
<td>329</td>
<td>1474</td>
<td>292</td>
<td>625</td>
<td>550</td>
<td>281</td>
<td>721</td>
<td>2088</td>
</tr>
<tr>
<td>PaulliniaCS1 + X</td>
<td>554</td>
<td>129</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaCS1 + Solvent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaCS2 + 3X</td>
<td>311</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>242</td>
<td></td>
</tr>
<tr>
<td>PaulliniaCS2 + Solvent</td>
<td>286</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Numbers represent raw peak areas. Substrate areas are colored in blue and product areas are colored in red. No 3-methylxanthine area were shown in PaulliniaCS2 + 3X reaction probably because most 3-methylxanthine were converted to theobromine. Xanthine alkaloid abbreviations: X, xanthine; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tp, theophylline; Tb, theobromine; Px, paraxanthine; Cf, caffeine.

Table 4. Kinetic parameters calculated for PaulliniaCS1 and PaulliniaCS2 with their respective substrates

<table>
<thead>
<tr>
<th>Enzyme + substrate</th>
<th>$K_M$ (µM)</th>
<th>$k_{cat}$ (s$^{-1}$)</th>
<th>$k_{cat}/K_M$ (sec$^{-1}$·M$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaulliniaCS1 + xanthine</td>
<td>95.38</td>
<td>1.52E-03</td>
<td>15.94</td>
</tr>
<tr>
<td>PaulliniaCS2 + 3-methylxanthine</td>
<td>677</td>
<td>9.33E-04</td>
<td>1.38</td>
</tr>
</tbody>
</table>

**TCS2 is responsible for the first methylation step of the caffeine biosynthetic pathway in Camellia sinensis (tea)**

As a close relative of the caffeine biosynthetic enzyme TCS1, TCS2 was cloned from *Camellia* leaves. The phylogenetic analysis of both genes with other SABATH members clearly indicates that TCS2 is a reasonable candidate responsible for the first methylation step of the caffeine biosynthetic pathway in *Camellia*, because it is closely related to all other CS enzymes [31] (Figure 19). Furthermore, caffeine is found mainly in leaves and buds of *Camellia sinensis*, and both TCS1 and TCS2 have many represented EST sequences from leaves exclusively,
Figure 19. Phylogenetic relationships of *Camellia* TCS1 and TCS2 in the context of other SABATH family members. Branches leading to *Camellia* CS enzymes are colored red.
whereas other SABATH sequences do not (Figure 20). Heterologously expressed TCS2 has the highest relative activity with xanthosine, suggesting that it is a 7-methylxanthosine synthase (Figure 21). The methylated product 7-methylxanthosine has not yet been verified by HPLC because there is no commercially available 7-methylxanthosine to use as the standard. However, a 7-methylxanthine peak shows up in the HPLC trace of crude extracted TCS2 incubated with a xanthosine reaction mixture, and is missing in the negative control assay without xanthosine added (Figure 22). It is possible that 7-methylxanthosine is not a stable compound, so it breaks down to 7-methylxanthine when the reaction mixture is acidified for HPLC analysis or through action of other bacterial enzymes in the crude extract which cleave the ribose group off from 7-methylxanthosine. A third possibility is that TCS2 may be a dual-functional enzyme that is also responsible for cleaving the ribose group off the 7-methylxanthosine, because the same result is also found in the HPLC analysis of the reaction product of purified *Coffea canephora* XMT with xanthosine [19]. Nevertheless, it can be concluded that TCS2 is able to methylate at the N-7 position of xanthosine as expected, and it is the missing caffeine synthase that catalyzes the first step in the caffeine biosynthetic pathway in *Camellia* (Figure 23).

**DISCUSSION**

New caffeine biosynthetic pathways have been found in *Citrus* and *Paullinia*. Compared to the canonical pathway that starts from xanthosine in *Coffea* and *Camellia*, both new pathways start from xanthine, yet end in caffeine through different xanthine alkaloid intermediates. Based on the caffeine synthase genes characterized from the four caffeine producing species (*Camellia, Coffea, Citrus* and *Paullinia*), caffeine biosynthesis in plants makes use of either the CS or XMT genes in the SABATH family, but I did not explore whether a third type of SABATH gene or even some non-SABATH genes have evolved to produce caffeine.
Figure 20. TCS1 and TCS2 are highly represented by ESTs from leaves and buds of *Camellia sinensis*. No orthologous XMT gene is known from *Mimulus*, which was used to blast the *Camellia* transcriptome and EST database.

Figure 21. Mean relative activities (from 0 to 100) of *Camellia* TCS2 with eight xanthine alkaloids. TCS2 has the highest relative activity with xanthosine.
Figure 22. HPLC analysis of product from the enzymatic reaction of TCS2 with xanthosine shows that TCS2 methylated xanthosine to form 7-methylxanthine. The reaction without xanthosine (solvent only) was used as the negative control, in which no 7-methylxanthine peak was detected.

Figure 23. TCS2 catalyzes the first methylation step of the caffeine biosynthetic pathway in *Camellia sinensis*. 
The main purine alkaloid in *Citrus* flower (primarily in anthers and pollen) is reported to be caffeine, following by a significant amount of theophylline as well as negligible amounts of theobromine and paraxanthine [47]. Kretschmar and Baumann (1999) also sampled the purine alkaloid content in *Citrus* flowers of different development stages and found that only theophylline is detectable in small buds. When flowers grow from buds to full anthesis, a decrease in theophylline level is accompanied by an increase in caffeine level, while no change occurs in levels of theobromine and paraxanthine. These findings indicate that theophylline is a pathway precursor for caffeine, instead of theobromine, and are congruent with enzymatic activities of CitrusXMT2 and the pathway characterized in *Citrus*. The caffeine biosynthetic pathway via theophylline also changes the long-term view that theophylline is a degradation product of caffeine, as is found in *Camellia, Coffea*, and *Theobroma* using radiolabeled tracers [49, 79, 80]. In addition, of the two caffeine biosynthetic pathways in *Citrus*, the pathway via 1-methylxanthine seems to be a major route to caffeine, as suggested by the kinetic parameters of CitrusXMT1 with 1- and 3-methylxanthine (Table 2) and the higher level of 1-methylxanthine than 3-methylxanthine in the *Citrus* flower buds (Figure 12). Although the caffeine biosynthetic pathway via xanthine → 3-methylxanthine → theophylline was also found as a minor pathway in leaves of young tea seedlings [81], the pathway through xanthine → 1-methylxanthine → theophylline has never been reported in any caffeine producing plants. Yet, *Citrus* appears to have evolved the ability to synthesize caffeine through this major novel pathway.

Besides CitrusXMT1 and CitrusXMT2, which are phylogenetically close to the XMT genes in *Coffea, Citrus* also possesses a CS gene (1g018139m in Figure 6) that is closely related to the CS genes in *Camellia, Paullinia* and *Ilex*. Although it is not represented by ESTs in flowers (Figure 7) and it does not have any detectable activity with eight xanthine alkaloids
tested, it can be amplified from RNA from *Citrus* flowers. Its role in caffeine biosynthesis in *Citrus* flowers is mysterious, because the complete caffeine biosynthetic pathway is already catalyzed by XMTs. In addition, CS genes are also found in some plants which are not known to produce caffeine, e.g. *Kirkia*, *Mangifera*, *Eucalyptus* (Figure 6). Functional characterization of these CS genes in the future may help to understand the role of CitrusCS in *Citrus* flowers.

Although there is no available metabolic information regarding xanthine and monomethylxanthines in any *Paullinia* tissues, the first two steps of the caffeine biosynthetic pathway characterized in *Paullinia* (xanthine → 3-methylxanthine → theobromine) have also been reported as one of the routes found in *Theobroma* for the biosynthesis of theobromine [59]. It is entirely possible that other caffeine biosynthetic pathways also exist in *Paullinia* seeds. First, in addition to theobromine and caffeine, theophylline accumulates at a considerable level in vegetative and floral tissues of *Paullinia*, in which the level of theophylline is higher than the level of caffeine, and when caffeine accumulated in the seeds, both levels of theobromine and theophylline are almost undetectable [68]. Thus, besides theobromine, theophylline may also be a precursor to caffeine. An undiscovered CS enzyme may exist to convert theophylline to caffeine. Second, the previously characterized PcCS also converts 7-methylxanthine into theobromine. If there is an undiscovered enzyme that methylates xanthosine or xanthine to 7-methylxanthine, another caffeine biosynthetic pathway may be formed in *Paullinia*. One candidate enzyme is the PaulliniaCS0, which is represented by 15 ESTs in fruit/seed but showed no activity with xanthine alkaloids when heterologously expressed. It is possible that PaulliniaCS0 didn’t fold correctly when expressed in *E. coli*, so that it lost activity to methylate xanthine alkaloids, or that PaulliniaCS0 needs special assay conditions to be active with xanthine
alkaloids. The radiolabeled tracer feeding experiment may be a better way to explore all possible caffeine biosynthetic pathways in *Paullinia*.

Since both *Citrus* and *Paullinia* belong to Sapindales, it was predicted that their caffeine biosynthetic pathways would be the same, and genes responsible for their pathways should be more closely related to each other than to the caffeine synthase genes in *Camellia* or *Coffea*, both of which are asterids. However, *Citrus* and *Paullinia* evolved to use different sets of genes (XMT versus CS) and different pathways to synthesize caffeine in different tissues (flowers versus seeds). These four different caffeine biosynthetic pathways and two different sets of genes (XMT and CS) involved in these pathways further suggest the convergent evolution of caffeine biosynthetic abilities in different plants. As there are many other plants that were reported to produce caffeine as well, e.g. *Euonymus* in Celastrales and *Erodium* in Geraniales [82], it would be interesting to know which types of SABATH genes and what pathways they evolved to synthesize caffeine. Phylogenetic analysis of caffeine synthase genes may easily identify which type of genes a plant may use to produce caffeine, but it would still be difficult to predict which pathway a plant might evolve as there are many possible combinations of precursors to form caffeine (Figure 4).
CHAPTER III: EVOLUTIONARY ASSEMBLY OF CAFFEINE BIOSYNTHETIC PATHWAYS IN MODERN-DAY PLANTS

INTRODUCTION

Evolution of metabolic pathways

Many chemicals are produced via metabolic pathways that are composed of a series of enzyme catalyzed reactions instead of simple single-step reactions. For example, pyruvate is produced from glucose through glycolysis which is a 9-step metabolic pathway, and glyceraldehyde-3-phosphate (G3P) is synthesized via 12 steps from CO$_2$ through the Calvin Cycle. However, in spite of the hundreds of studies of these pathways, there are only a few studies investigating how multistep metabolic pathways were evolutionarily assembled; in other words, in what order individual enzymes for each step evolved. For my dissertation, I am interested in understanding how the caffeine biosynthetic pathway evolved. Three hypotheses were proposed to explain the assembly of metabolic pathways.

The first hypothesis regarding the origin of metabolic pathways was the retrograde hypothesis [83, 84], which states that the assembly of a metabolic pathway occurred in reverse order of the direction of biosynthesis in a stepwise manner. Specifically, the last step and its associate enzyme in a pathway evolved first, then gene duplication and natural selection facilitated the evolution of a second enzyme that catalyzed the penultimate step, and the process kept going until finally the first step was assembled (Figure 24a). Horowitz mentioned a special environmental background is needed for this hypothesis: a primitive environment with non-enzymatically produced pathway intermediates, since the cells would not have had the ability to synthesize them directly [83]. This hypothesis has its limitation in that such an environment was
Figure 24. Schematic comparison of three hypotheses on the origin and evolution of metabolic pathways. The retrograde hypothesis (a) is opposite to the cumulative hypothesis (b) in the order (in time) in which steps were assembled. However, they both state that the pathway was assembled in a stepwise manner. The patchwork hypothesis (c) focuses on the recruitment of enzymes that catalyzed analogous reactions into a new pathway by making use of their promiscuous activities. Gene duplication and functional divergence of these enzymes provided the new pathway with newly specialized enzymes.
unlikely to exist, and it fails to consider whether intermediates in a pathway were stable or not and how these chemicals were transported through a cellular network even if they were available [85].

The second hypothesis was termed the cumulative hypothesis (or Granick hypothesis), which was contrary to the retrograde hypothesis [86]. With evidence from the evolution of the photosynthetic pathway from a primitive energy-conversion unit, Granick proposed that the assembly of a metabolic pathway was in the same direction as product synthesis in a stepwise manner [86]. Specifically, earlier reactions and associated enzymes in a pathway evolved first and started from compounds of simple structure. Subsequently, the second and third steps sequentially evolved with increasing complexity in the structures of intermediates and end-product as well as recruitment of more efficient enzymes through gene duplications (Figure 24b). Although the cumulative hypothesis was successfully applied to explain the evolution of cholesterol from terpenoids [87], it has the assumption (or limitation) that intermediates in a pathway must be useful to the organism, because multistep enzyme recruitment into a pathway would rarely happen simultaneously.

Unlike the first two hypotheses, the patchwork hypothesis [88, 89] didn’t focus on the order of steps developed (or enzymes recruited) in a pathway. Instead, it focuses on the substrate promiscuity of ancestral enzymes and assembly of new metabolic pathways by recruiting such promiscuous enzymes that fortuitously catalyze existing reactions (Figure 24c). The key importance of the patchwork hypothesis is that promiscuous activities of ancestral enzymes allowed the instantaneous development of a new pathway by making use of these minor activities. Once the product of the new pathway was favored by natural selection, gene duplications and functional mutations of ancestral enzymes then provided newly specified
enzymes to complete the new pathway [89] (Figure 24c). The patchwork hypothesis took into account the pervasive nature of gene duplication events [4] and commonly observed promiscuous activities of enzymes [90], thus it has much broader applicability than either of the first two hypotheses [85, 91, 92], especially for the evolution of plant secondary metabolic pathways [93].

Whereas enzymes catalyzing consecutive steps in a metabolic pathway are usually not paralogous [89, 94], thereby favoring the patchwork hypothesis, there are several examples showing that enzymes coded by duplicated genes were recruited to catalyze consecutive steps in the same metabolic pathway, as is required of the retrograde hypothesis. For example, duplicated metB and metC genes that code for γ-synthase and β-cystathionase in E. coli, respectively, evolved to catalyze consecutive steps in methionine biosynthesis [95]. Similarly, four paralogous carbon-nitrogen ligases were assembled to catalyze a four-step peptidoglycan biosynthetic pathway in E. coli [96, 97]. The three-step caffeine biosynthetic pathways characterized in Camellia, Citrus, Paullinia and Coffea are also catalyzed by duplicated paralogous CSs or XMTs (see Chapter II) [33].

To test the three hypotheses and find out how these three methylation steps were evolutionarily assembled in each species, it is necessary to look for answers in the past. Specifically, I asked the following questions: 1) what reactions were catalyzed by ancestral CS and XMT enzymes, and 2) what evolutionary changes prompted the assembly of each methylation step? The development of paleomolecular biology techniques (ancestral sequence resurrection coupled with experimental mutagenesis) allows me to answer these questions and unravel the evolutionary history of caffeine biosynthesis in plants.
Ancestral sequence resurrection

The idea of ancestral sequence resurrection (ASR) was first brought up in 1963 [98]. Typical ASR makes use of statistical methods (e.g. parsimony, maximum likelihood, or Bayesian inference) to estimate the ancestral gene/protein sequence in a phylogenetic context, then chemically synthesize the gene and directly test its function through heterologous gene expression [99]. ASR was used as early as in 1990 to study the property of an ancient ribonuclease in the extinct bovid ruminant [100], and later it was gradually and widely applied in studying absorbance spectra of ancient visual pigments [101-103], the palaeoenvironment for ancient bacteria [104, 105], the functional evolution of corticoid receptors [106], and the color development of the fluorescent proteins in corals [107]. These studies (and several others) demonstrated that ASR is an invaluable tool to understand ancient life, especially the behavior and physiology of ancestral organisms as well as their interactions with the ancient environment, and provides a useful supplement to the fossil record.

Results from ASR studies show that some resurrected ancestral proteins were functionally specific. For example, earliest ancestral esterases that gave rise to both modern-day esterases and acetone cyanohydrin lyases, were found to be highly specific like modern-day esterases, which splits esters into an alcohol and a carboxylic acid through a two-step mechanism; yet they lack the acetone cyanohydrin cleavage ability [108]. In contrast, most studies of ancestral proteins show that they were functionally promiscuous, because they exhibited more than one function or had activities with more than one substrate. For example, the ancestral corticoid receptor of vertebrate could be activated by 11-deoxycorticosterone, aldosterone and cortisol [106], the pre-duplicated ancestral maltase in yeast had activities with both maltose and isomaltose-like substrates [109], the Precambrian bacterial β-lactamases could
degrade more than one type of antibiotic [110], and the ancestral polar amino acid-binding protein in bacteria was able to bind both L-arginine and L-glutamine [111]. In my previous study, resurrected ancestors of SAMTs in the Apocynaceae and Solanaceae lineage were also multifunctional and could methylate benzoic acid, salicylic acid and nicotinic acid [112]. Importantly, these promiscuous activities were hypothesized to provide the starting point for the evolution of new gene functions or pathways. Therefore, these studies support the patchwork hypothesis and predict that the ancient organism would acquire selective advantages in changing environments because of the fortuitous promiscuous activities, and subsequent gene duplication or amplification would allow the specialization of new gene functions through mutation [88, 89]. Considerable evidence has been found to support this hypothesis, including the evolution of the above mentioned ancestral maltase, β-lactamases, and polar amino acid-binding protein, as well as studies from directed laboratory evolution using E. coli [113, 114].

In general, these studies indicate that ancestral proteins had one primary function that was essential to the ancient organism, while other activities might not have any definite roles. However, one of these minor activities might later be selected to become the primary activity of a newly evolved protein, a phenomenon termed exaptation [115]. For example, feathers in birds might be initially used for regulating body temperature, but have evolved for flight nowadays. Therefore, feathers were exapted for the ability to fly in ancestral birds. Likewise, the multifunctional ancestral enzyme was exapted for minor activities which later became selectively advantageous.

ASR also helps to better understand the functional divergence of genes through evolutionary time by introducing historical mutations to the resurrected ancestors (by “forward mutagenesis”) to re-evolve the descendants’ activities. Such experimental mutagenesis studies
were performed to study the evolution of vertebrate glucocorticoid receptors’ hormone specificity from a mineralocorticoid receptor-like ancestor [116], the evolution of a red fluorescent protein from an ancestral green fluorescent protein in coral [117], and the evolution of substrate specificity of the dihydroflavonol-4-reductase from an anthocyanin pathway in modern-day red flowers in *Iochroma* from ancestral blue state [118]. One important finding from these studies is that protein functional evolution usually shows sign epistasis, in which the fitness effect of one mutation on a protein is contingent on the effect of other existing mutations, thus it constrains the possible evolutionary paths of the protein [119]. For instance, some function-shifting mutations were deleterious to the ancestor by themselves, thus “permissive” mutations (mutations of no immediate functional consequence) needed to be introduced to the ancestor first (or at the same time) to tolerate the negative effect of the functional change mutations [116]. In addition, sign epistasis also decreases the probabilities of mutational convergence and parallelism of orthologous proteins in different populations or species because of their different genetic backgrounds [120]. However, the study on the functional evolution of SAMT genes in Solanaceae and Apocynaceae showed that epistasis didn’t affect the functional evolution of these two SABATH enzymes, and a single site was solely responsible for the substrate preference switch between benzoic acid and salicylic acid regardless of genetic background [112]. Another finding from ASR and experimental mutagenesis studies is that although major evolutionary shifts in protein function are caused by a few large-effect mutations, which drastically alter the function of an ancestral protein towards the descendant, in a punctuated manner [121, 122], there are also small “fine-tuning” additive mutations, which may have relatively small effect on the function of the ancestral protein, to complete the functional change from the ancestor to the descendent after large-effect mutations took place [117].
The existence of caffeine in different plant species indicates that the ability to synthesize caffeine has evolved more than eight times during the flowering plant history (Figure 3). To answer the previously asked questions: First, I resurrected and functionally characterized the ancestral CS and XMT in each caffeine-producing species (*Camellia, Coffea, Citrus*, and *Paullinia*) as well as the ancient CS and XMT shared by rosids and asterids to see when the ability to synthesize caffeine arose. Next, I mutated ancestral CS in *Paullinia* and ancestral XMT in *Citrus* towards their modern-day descendants, respectively, to see how the caffeine biosynthetic pathway in each species was assembled. The first hypothesis is caffeine was produced by ancient plants, in which all ancestral CSs and XMTs could methylate all caffeine precursors in a caffeine biosynthetic pathway, and these ancestral functions were inherited and partitioned in their descendants after gene duplications in each caffeine producing species. The second hypothesis is caffeine was not produced by ancient plants, in which all ancestral CSs and XMTs were used for functions other than methylating xanthine alkaloids, and the ability to produce caffeine evolved independently in a small portion of plant lineages recently.

**MATERIALS AND METHODS**

**Resurrection of ancestral caffeine synthases**

To accurately and efficiently estimate ancestral CS and XMT, two phylogenetic trees were constructed separately. Both trees included amino acid sequences of all functionally characterized SABATH family members, with one tree containing a dense sampling of CS-like sequences and the other tree containing a dense sampling of XMT-like sequences. These CS-like and XMT-like sequences were sampled from all available angiosperm sequences in the oneKP database (www.onekp.com). Sequences were aligned using MAFFT version 7 [69] using auto strategy and default parameters, followed by manual adjustment to remove uninformative gap
regions. The two maximum-likelihood phylogenetic trees were then estimated using PhyML [70] assuming the Jones, Taylor, Thornton (JTT) amino acid substitution model [71] with gamma distribution as chosen by Modeltest [72].

Ancestral sequences were estimated using the CODEML module in PAML [123] assuming the JTT + gamma model of amino acid substitution based on the constructed phylogenetic trees using the maximum likelihood algorithm. Regions with alignment gaps were analyzed with a parsimony algorithm to determine the number of ancestral residues. To test the robustness of these estimated ancestral sequences, at least two ancestral variant alleles were created and functionally tested. Each ancestral variant differs from the original one at amino acid positions with low posterior probabilities. The alternative amino acid residues at these sites were estimated using either different subsets of sequences or different parameters.

**Gene synthesis and cloning**

The nucleotide sequences of ancestral CSs and XMTs were synthesized by GenScript Corp. with codons optimized for expression in *E. coli* and cloned into pUC57 vector. They were subcloned into pET15b vector (Novagen) with a N-terminal His-tag for subsequent protein expression and purification. All recombinant plasmids were checked on an agarose gel for their size and sequenced by GENEWIZ Corp to ensure that all cloned sequences were in the correct open reading frame.

**Generation of ancestral gene variants and forward mutagenesis**

Ancestral gene variants and forward mutants were obtained using the QuikChange Site-Directed Mutagenesis Kit (Stratagene). Plasmids containing the target gene were used as PCR templates and high fidelity *PfuTurbo* DNA polymerase was used to run the PCR reactions. Mutagenic primers were designed to be between 25 and 45 bases in length, and the melting
temperature ($T_m$) of the primers was calculated to be greater than or equal to 78°C by using the following equation: $T_m = 81.5 + 0.41(\%GC) - 675/N - \%mismatch$ ($N$ is the primer length in bases) for point mutations, or $T_m = 81.5 + 0.41(\%GC) - 675/N$ ($N$ does not include the bases to be inserted or deleted) for insertions or deletions. Cycling parameters were as follows: template denaturation at 95°C for 30 seconds, cycling then proceeded to denaturation at 95°C for 30 seconds, annealing at 55°C for 1 minute, and extension for 1 minute / kb of plasmid length at 68°C, for 16 cycles. PCR synthesized plasmids containing the mutant gene after the DpnI digestion of the templates were transformed into XL10 - gold supercompetent E. coli cells. Mutant plasmids were miniprepped from transformed bacterial colonies and all mutated sites were verified by DNA sequencing (GENEWIZ INC.).

**Protein expression**

Recombinant pET15b plasmids were transformed into E. coli strain BL21(DE3) Star™ cells (Invitrogen) for overexpression of His-tagged proteins. Protein expression started from a 5 ml starter culture of transformed E. coli cells, which were grown at 37°C in LB broth containing 100 µg/ml ampicillin overnight. On the next day, the 5 ml starter culture was used to inoculate a fresh 100 ml LB broth containing the same concentration of antibiotics and continued to grow until OD$_{600}$ reached 0.6 to 0.8, then induced with 1.0 mM IPTG for 6 hours at 23°C. The cells were harvested by centrifugation at 4°C, 4000 rpm for 15 minutes. To isolate proteins from E. coli cells, pellets were resuspended in chilled 1X equilibration buffer (50 mM Na$_3$PO$_4$, 300 mM NaCl, 10 mM imidazole, 10 mM β-mercaptoethanol, and 12% (v/v) glycerol, pH 8), treated with 0.75 mg/ml Lysozyme on ice for 30 minutes, and then sonicated on ice for 20 seconds, 3 times with a 20-second pause on ice between each burst. The suspension after sonication was centrifuged at 10,000 g for 20 minutes at 4°C to get rid of cell debris. The supernatant contained
the soluble crude proteins which were either used directly for enzyme catalytic assays or subsequently purified.

**Protein purification**

Protein purification was achieved using cobalt spin columns (Clontech) according to the manufacturer’s instructions. Purified proteins were eluted in 1X elution buffer (50 mM Na₃PO₄, 300 mM NaCl, 200 mM imidazole, 10 mM β-mercaptoethanol, and 12% (v/v) glycerol, pH 8), aliquoted and flash frozen in a dry ice-ethanol bath and stored at -80°C. The purity and molecular mass of purified proteins was evaluated by running on pre-cast NuPAGE 4-12% Bis-Tris gels (Invitrogen), and protein concentration was calculated by the Bradford assay [74] using bovine serum albumin (BSA) as the standard.

**In vitro functional analysis**

Relative enzymatic activities of heterologously expressed ancestral proteins with eight xanthine alkaloids as well as benzoic acid and salicylic acid were tested using SAM with a ¹⁴C-labeled methyl group. For each substrate, a 50 µl reaction containing 10 or 20 µl purified proteins (or 30 µl crude extracted total cellular proteins), 100 µM methyl acceptor substrate and 0.01 µCi ¹⁴C-labeled SAM was carried out in 50 mM Tris-HCl buffer at room temperature for 20 minutes (or 30 minutes if using crude cellular proteins). The methylated product was extracted using 200 µl ethyl acetate, and 120 µl of the ethyl acetate (top) phase was removed and mixed with 4 ml scintillation counting cocktail. Radioactive decay was measured using a liquid scintillation counter (PerkinElmer) and disintegrations per minute (DPM) values were reported. Assays were done in duplicate or triplicate and standard deviations were calculated.
High-performance liquid chromatography (HPLC)

Product identity was determined using HPLC on reactions that were scaled up 10-fold using non-radioactive SAM for two hours at room temperature. The whole reaction was acidified to pH 5–6, then filtered through Vivaspin columns (Sartorius Stedim Biotech) to remove proteins and impurities, and 10 µl of the processed reaction (or metabolite extract) was injected into HPLC. Mixtures were separated by HPLC using a two-solvent system with a 250 mm × 4.6 mm Kinetex 5µm EVO C18 column (Phenomenex). Solvent A was 99.9% (V/V) water with 0.1% TFA and solvent B was 80% (V/V) acetonitrile, 19.9% (V/V) water, and 0.1% TFA, and a 0–16% solvent B gradient was generated over 16 minutes with a flow rate of 1.0 ml/min. Subsequently, buffer B was increased to 100% and then held at that percentage for 20 minutes. Equilibration back to 0% buffer B was achieved over a 20-minute period. Two UV wavelengths (254 and 272 nm) were used for absorption measurements. The products of reactions catalyzed by ancestral enzymes were verified by comparing its absorbance peak/retention time with xanthine alkaloid standards. Negative control reactions without the substrate added were always run in parallel.

Liquid Chromatography–Tandem Mass Spectrometry (LC-MS)

Identities of methylated products from non-radioactive enzymatic assays were also confirmed using LC-MS. LC-MS/MS analyses were performed using an Agilent 1100 HPLC inline with a QuattroMicro mass spectrometer. The same C18 column and a different two-solvent system were used to separate and analyze the reaction products: Solvent A was 0.1% formic acid/0.01% trifluoroacetic acid/water, solvent B was 0.1% formic acid/0.01% trifluoroacetic acid/acetonitrile, and a 0–16% solvent B gradient was generated over 16 minutes with a flow rate of 0.5 ml/min followed by 2-minute of 95% B for a run time of 20 minutes. A postcolumn
addition of 0.1% formic acid in acetonitrile was added via a PEEK tee at a flow rate of 100 µL/min. Under the above conditions, detection was initially optimized using pure standards of different xanthine alkaloids diluted to 1 µM in 0.1% formic acid/50% (V/V) acetonitrile and infused directly onto a Waters Quattro Micro mass spectrometer via an electrospray ion source, and mass spectrometry scans successfully separated and identified xanthine, 1-, 3-, 7-methylxanthine, theophylline, theobromine, paraxanthine and caffeine based on their retention time and parent ion - fragment ion fragmentation pattern differences. First, xanthine is clearly separated from all other alkaloids and shows a unique 153.4 > 110 fragmentation pattern and a retention time of 4.67 minutes. Second, 1-methylxanthine has a different retention time (9.36 minutes) from all other alkaloids and a unique 167.4 > 110 fragmentation pattern. Third, 3-methylxanthine and 7-methylxanthine may be distinguished from their different retention times (8.82 and 8.23 minutes, respectively) even though they both show a 167.4 > 124 fragmentation pattern. Fourth, the 181.5 > 138 fragmentation pattern is specific for theobromine, which also has a retention time (12.09 minutes) different from other alkaloids. Fifth, because paraxanthine and theophylline are difficult to separate under these LC conditions (the retention time is 12.70 minutes for paraxanthine and 12.92 minutes for theophylline) and both have a common 181.5 > 124 fragment, a specific 181.5 > 96 fragment was scanned for theophylline and a 181.5 > 55 fragment was scanned for paraxanthine. Lastly, caffeine separated far from all other alkaloids with a retention time of 15.69 minutes, and had a specific 195 > 138 fragmentation pattern that allowed its unique identification. For each identified xanthine alkaloid peak, only when raw peak area was larger than 25, (an arbitrary cutoff value, which was 1/10 the area of the lowest intensity standard,) was it recorded as the real xanthine alkaloid peak.
Gas chromatography-mass spectrometry (GC-MS) analysis

Reaction products of RAAnCXM with salicylic acid and benzoic acid as the substrates were verified using GC-MS. Enzymatic reactions were scaled up 10-fold using non-radioactive SAM for two hours at room temperature. Products were extracted from each reaction using 800 µl ethyl acetate and 500 µl of the ethyl acetate (top) phase was recovered and transferred into a vial and was ready for GC-MS analysis. One µl reaction product was analyzed by GC-MS on an HP6890GC System coupled to an HP5973 Mass Selective Detector using a DB-5 capillary column with the following oven conditions: 40°C for 2 minutes, ramping 20°C / minute to 300°C with a 2-minute hold. Product was identified by comparing its mass spectra and retention time with available standards and reference spectra from the library of National Institute of Standards and Technology. The negative control reaction without the substrate added was run in parallel.

Enzyme kinetics

Kinetic parameters (k_{cat} and K_{M}) of ancestral Cs and XMTs with a given substrate were determined using the 50 µl radioactive assay described above. Appropriate enzyme concentration and incubation time were determined in time-course assays with low unsaturated substrate concentrations, such that the reaction velocity was linear during the assay period. When varying substrate concentration, the SAM concentration was held constant and saturated at 320 µM. Initial velocities versus substrate concentration were plotted using GraphPad Prism (GraphPad Software, La Jolla, CA) to fit the hyperbolic Michaelis-Menten equation to calculate V_{max} and K_{M}. V_{max} was corrected using extraction efficiency and converted to k_{cat} based on estimated protein concentrations and expressed in units of s^{-1}. Assays were done in duplicate and standard errors were calculated using GraphPad Prism.
RESULTS

Two distinct types of caffeine synthases evolved convergently in rosids and asterids

As shown in Figure 25, phylogenetic relationships of the caffeine synthases in *Camellia*, *Paullinia*, *Coffea* and *Citrus* show that caffeine synthases within each species are more closely related to each other than to those in other species. This indicates that gene duplications have happened within each species after speciation events and such duplicates are expected to be retained only if they confer selective advantage. When compared among different species, caffeine synthases in *Camellia* and *Paullinia* are most closely related (coded by orthologous genes in the CS clade), while caffeine synthases in *Citrus* and *Coffea* are most closely related (coded by orthologous genes in the XMT clade). This is surprising since at the species level, *Paullinia* and *Citrus* are more closely related to each other than to either *Camellia* or *Coffea*. The above evidence suggests that CSs and XMTs may have evolved convergently to produce caffeine in these four species.

Ancestral CS enzymes were maintained for more than 100 million years with fortuitous secondary activity with 7-methylxanthine

Ancestral CS enzymes in *Paullinia* (PaulliniaAncCS1 and PaulliniaAncCS2, ~37-million-year old) and *Camellia* (CamelliaAncCS, > 20-million-year old) as well as an ancient CS enzyme shared by rosids and asterids (RAAncCS, >100-million-year old) were resurrected (see green dots in the phylogenetic tree in Figure 25). Their alignment with modern-day caffeine synthases is shown in Appendix A.

Heterologously expressed PaulliniaAncCS1, PaulliniaAncCS2, CamelliaAncCS and RA AncCS all have highest relative activity with 7-methylxanthine as well as low relative
Figure 25. Maximum likelihood phylogenetic tree of representative SABATH members with dense samplings of CS (green colored lineages) and XMT (purple colored lineages) sequences showing nodes of estimated ancestral caffeine synthases. Bootstrap values (100 replicates) are shown for nodes with higher than 50% support.
whether theobromine or paraxanthine is the major product of CamelliaAncCS2 and
them to have highest relative activity with 7 sequence identity
activities with xan
among them is only 0.0002 \[P = (1/8)^4\]. LC-MS scans of reaction products showed that PaulliniaAncCS1 and RAAncCS methylate 7-methylxanthine to form theobromine, while CamelliaAncCS and PaulliniaAncCS2 methylate 7-methylxanthine to form both theobromine and paraxanthine (Table 5). However, whether theobromine or paraxanthine is the major product of CamelliaAncCS2 and

Figure 26. Mean relative activities (from 0 to 100) of ancestral CS enzymes with eight xanthine alkaloids. All ancestral CS enzymes display highest relative activity with 7-methylxanthine, as well as minor activities with xanthine and 3-methylxanthine. Average posterior probability (PP) for each resurrected ancestral CS is shown in parenthesis. Ancestral CS enzymes have no activities with benzoic acid and salicylic acid (data not shown).
PaulliniaAncCS2 with 7-methylxanthine was difficult to distinguish from Table 5, because these peak areas are not directly comparable to each other, and further absolute quantifications of peak areas using serial diluted standards of each xanthine alkaloid is needed because of ionization efficiency differences.

Table 5. LC-MS scans of products of enzymatic reactions of PaulliniaAncCS1, PaulliniaAncCS2, CamelliaAncCS and RAAncCS with different substrates

<table>
<thead>
<tr>
<th>Enzyme + Substrate</th>
<th>X Area</th>
<th>1X Area</th>
<th>3X Area</th>
<th>7X Area</th>
<th>Tp Area</th>
<th>Tb Area</th>
<th>Px Area</th>
<th>Cf Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard eight xanthine alkaloids mix</td>
<td>329</td>
<td>1474</td>
<td>292</td>
<td>625</td>
<td>550</td>
<td>281</td>
<td>721</td>
<td>2088</td>
</tr>
<tr>
<td>PaulliniaAncCS1 + X</td>
<td>761</td>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS1 + 3X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS1 + 7X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS1 + Solvent</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS2 + X</td>
<td>610</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS2 + 3X</td>
<td>112</td>
<td>483</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS2 + 7X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS2 + Solvent</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CamelliaAncCS + X</td>
<td>1000</td>
<td>44</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CamelliaAncCS + 7X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CamelliaAncCS + Solvent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA AncCS + 7X</td>
<td>4740</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>RA AncCS + Solvent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Numbers represent raw peak areas. Substrate areas are colored in blue and product areas are colored in red. The presence of xanthine in negative control reactions is probably due to other chemical reactions in bacteria. Xanthine alkaloid abbreviations: X, xanthine; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tp, theophylline; Tb, theobromine; Px, paraxanthine; Cf, caffeine.

In addition to activity with 7-methylxanthine, LC-MS scans of reaction products showed that PaulliniaAncCS1 and PaulliniaAncCS2 methylate xanthine to form 3-methylxanthine, and
CamelliaAncCS methylates xanthine to form both 1-methylxanthine and 3-methylxanthine (Table 5). However, whether 1-methylxanthine or 3-methylxanthine is the major product of CamelliaAncCS with xanthine was also difficult to distinguish due to lack of absolute quantification data. It should be noted that, although RAAncCS also has minor activity with xanthine (Figure 26), the product level is too low to be detectable by HPLC (data not shown).

In addition, I couldn’t detect any reaction products of RAAncCS, CamelliaAncCS and PaulliniaAncCS1 with 3-methylxanthine from HPLC spectra (data not shown), probably due to their low activities with 3-methylxanthine (Figure 26). In contrast, relative activity of PaulliniaAncCS2 with 3-methylxanthine increased at least five-fold compared to the other three ancestral CS enzymes, and the product was verified to be theobromine by LC-MS analysis (Table 5).

Based on the enzymatic activities of these ancestors, it is unlikely that the ancient plants could produce caffeine, because none of these ancestral CSs can catalyze all the required precursors to form a complete pathway. It is unknown why all of them have the highest relative activity with 7-methylxanthine as the ancient plants should be able to synthesize 7-methylxanthine in the first place; otherwise, the ability to methylate 7-methylxanthine would presumably be useless. Yet all four ancestral CSs showed no activity with xanthosine (Figure 26), and xanthine is converted to 3-methylxanthine by CamelliaAncCS, PaulliniaAncCS1 and PaulliniaAncCS2, as well as 1-methylxanthine by CamelliaAncCS (Table 5), but not 7-methylxanthine. Nonetheless, RAAncCS has been maintained for more than 100 million years instead of becoming a pseudogene and getting lost. One possibility is that another unknown ancestral enzyme may exist to methylate xanthine or xanthosine to form 7-methylxanthine, such that the ancestral CS enzymes could produce useful products. However, the only enzymes that
are known to produce 7-methylxanthine are the modern-day TCS2 from *Camellia* (Figure 22) and XMT from *Coffea* [56], and their ability to convert xanthosine to 7-methylxanthine were likely evolved only recently. Thus, it was unlikely that such an ancestral enzyme existed. A second possibility is that the ancestral CS enzymes might have an unknown primary function, which was essential for the survival of the ancient plants, while activities with xanthine, 3-methylxanthine and 7-methylxanthine were their fortuitous secondary activities. One evidence to support the second hypothesis comes from descendant CS enzymes expressed in many purine alkaloid-free *Camellia* species. Heterologously expressed CS genes in *Camellia japonica*, *Camellia grandhamiana*, *Camellia kissi*, *Camellia lutchuensis* methylate 7-methylxanthine to form theobromine, but these *Camellia* species are not able to produce 7-methyxanthine in the first place [61]. Therefore, ancestral CS enzymes were more likely exapted for reactions currently used for various steps of caffeine biosynthesis in *Camellia and Paullinia*.

After gene duplications happened in *Camellia* and *Paullinia*, duplicated descendants of the ancestral CSs evolved the abilities to catalyze consecutive reactions in the caffeine biosynthetic pathways in both species. In *Camellia*, CamelliaAncCS’s activity with 7-methylxanthine likely became physiologically significant when one of its descendants, TCS2 evolved the ability to methylate xanthosine to form 7-methylxanthine (Figure 22), while the other descendant, TCS1 inherited the ancestral activity with 7-methylxanthine and evolved new abilities to methylate theobromine and paraxanthine (Figure 26). In *Paullinia*, the relative enzymatic activity with 3-methylxanthine increased from PaulliniaAncCS1 to PaulliniaAncCS2, and the ancestral activities with xanthine and 3-methylxanthine subsequently specialized in two descendant enzymes, PaulliniaCS1 and PaulliniaCS2, respectively, while the ancestral activity with 7-methylxanthine was lost in both (Figure 26).
Historical maintenance of ancestral XMT enzymes for O-methylation for over 100 million years facilitated the evolution of their N-methylation abilities

Ancestral XMT enzymes in *Citrus* (CitrusXMT1 and CitrusXMT2, ~3.6-million-year old) and *Coffea* (CoffeaAncXMT, ~30-million-year old) as well as an ancient XMT enzyme shared by rosids and asterids (RAAncXMT, >100-million-year old) were resurrected (purple dots in the phylogenetic tree in Figure 25), and their alignment with modern-day caffeine synthases is shown in Appendix A.

Heterologously expressed RAANCXMT and CitrusAncXMT1 have major activities with benzoic acid and salicylic acid (Figure 27), and the reaction products of RAANCXMT with benzoic acid and salicylic acid were verified to be methyl benzoate and methyl salicylate, respectively (Figure 28). In contrast, they only have minor (about 9-fold less) activities with some xanthine alkaloids (Figure 27), suggesting that the ancient plants were not likely to produce caffeine using XMT enzymes. Thus, RAANCXMT and CitrusAncXMT1 were likely used for O-methylation. The activity of RAANCXMT with benzoic acid was conserved in its modern-day descendant, MangiferaXMT, which lost most activity with salicylic acid and all minor activities with xanthine alkaloids (Figure 27). In contrast, its descendant enzymes in *Citrus* and *Coffea* have lost all O-methylation abilities and evolved to specialize in caffeine biosynthesis (Figure 27). Therefore, RAANCXMT was likely exapted for caffeine biosynthesis, which was later evolved in *Citrus* and *Coffea*. 
Figure 27. Mean relative activities (from 0 to 100) of ancestral XMT enzymes with eight xanthine alkaloids. Average posterior probability (PP) for each resurrected ancestral XMT is shown in parenthesis. N-methylation abilities in modern-day *Citrus* and *Coffea* were evolved from ancestral O-methylation abilities of RAAncXMT. Relative activity data for *Coffea* XMTs is obtained from Uefuji et al. (2003).
Figure 28. GC-MS spectra show that RAAnXMT methylates benzoic acid and salicylic acid to form methyl benzoate and methyl salicylate, respectively. The mass spectra for methyl salicylate and methyl benzoate are shown in insets in the first two spectra, respectively.
The exaptation of ancestral XMT enzymes for caffeine biosynthesis was further evidenced by CitrusAncXMT2, which seems to be a transitional enzyme which had activities with xanthine alkaloids comparable to benzoic acid and salicylic acid (Figure 27). Although the relative activities with 1-methylxanthine and 3-methylxanthine are higher than benzoic acid and salicylic acid, the primary function of CitrusAncXMT2 might still have been O-methylation because its specific activities with benzoic acid and salicylic acid are 0.6 and 2.3 pkat/mg, respectively, which are comparable to functionally characterized modern-day SAMTs and BSMTs [28]. LC-MS scans of reaction products of CitrusAncXMT2 with 1-methylxanthine, 3-methylxanthine, and theophylline show that CitrusAncXMT2 converts 1-methylxanthine to paraxanthine, converts 3-methylxanthine to theophylline and converts theophylline to caffeine (Table 6). These ancestral activities with xanthine, 1- & 3-methylxanthine and theophylline were later inherited and partitioned into two modern-day descendant enzymes, CitrusXMT1 and CitrusXMT2 after gene duplication in Citrus (Figure 27).

The resurrected CoffeaAncXMT had already lost the activities with benzoic acid and salicylic acid, and specialized to methylate 1- methylxanthine to form theophylline (Figure 27) (Table 6). Its activity with 1-methylxanthine may also be an exapted fortuitous activity because none of its descendant enzymes have high relative activity with 1-methylxanthine, and both 1-methylxanthine and the methylated product - theophylline, are not the pathway intermediates known in Coffea [32, 124], but may have been in ancestral Coffea plant.

**Exapted multifunctional ancestral CS enzymes facilitate the evolution of novel caffeine biosynthetic pathway in Paullinia**

To understand how co-opted multifunctional ancestral CS enzymes evolved to catalyze
Table 6. LC-MS scans of products of enzymatic reactions of CitrusAncXMT2 and CoffeaAncXMT with different substrates

<table>
<thead>
<tr>
<th>Enzyme + Substrate</th>
<th>X Area</th>
<th>1X Area</th>
<th>3X Area</th>
<th>7X Area</th>
<th>Tp Area</th>
<th>Tb Area</th>
<th>Px Area</th>
<th>Cf Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard eight xanthine alkaloids mix</td>
<td>329</td>
<td>1474</td>
<td>292</td>
<td>625</td>
<td>550</td>
<td>281</td>
<td>721</td>
<td>2088</td>
</tr>
<tr>
<td>CitrusAncXMT2 + 1X</td>
<td></td>
<td>2720</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CitrusAncXMT2 + 3X</td>
<td></td>
<td></td>
<td>883</td>
<td></td>
<td>140</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CitrusAncXMT2 + Tp</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1780</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>CitrusAncXMT2 + Solvent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoffeaAncXMT + 1X</td>
<td>2530</td>
<td></td>
<td>880</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoffeaAncXMT + Solvent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Numbers represent raw peak areas. Substrate areas are colored in blue and product areas are colored in red. The presence of xanthine in some reactions is probably due to other chemical reactions in bacteria. Xanthine alkaloid abbreviations: X, xanthine; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tp, theophylline; Tb, theobromine; Px, paraxanthine; Cf, caffeine.

As mentioned earlier, 7-methylxanthine appears to be a fortuitous secondary activity of RA AncCS, and the methylated product is theobromine (Table 5). Although RA AncCS also exhibits low activity with xanthine, the $K_M$ with xanthine is larger than 2 mM, which is much higher than the reported in vivo concentration range (10 to 1000 µM) of most caffeine precursors in tissues of *Theobroma*, *Paullinia* and *Citrus* [47, 59, 77], suggesting that its activity with xanthine is probably physiologically irrelevant. Thus, the caffeine biosynthetic pathway probably has not evolved at stage A (Figure 29).

When PaulliniaAncCS1 evolved from RA AncCS, its relative activity with xanthine increased from 11% to 25%, and more importantly, the enzymatic activity with xanthine became
physiologically relevant as the $K_M$ with xanthine was reduced to 417.5 µM, which is within the in vivo concentration range [47, 59, 77], and its $k_{cat}/K_M$ with xanthine increased almost eight-fold compared to RAAnCS with xanthine (Table 7). The reaction product of PaulliniaAncCS1 with xanthine is verified to be 3-methylxanthine (Table 5). Thus, although the caffeine biosynthetic pathway was still incomplete at stage B, the initial reaction in the pathway has successfully evolved in PaulliniaAncCS1 from the long-term maintained RAAnCS (Figure 29). However, due to the very low activity with 3-methylxanthine (Figure 29), the product of PaulliniaAncCS1 with 3-methylxanthine is unable to verify using LC-MS scan (Table 5), making it unlikely that the second reaction in the pathway was formed at stage B. The relative activity with 3-methylxanthine evolved to be the second highest of PaulliniaAncCS2 and is methylated to form theobromine (Table 5). Although 7-methylxanthine was still the preferred substrate of PaulliniaAncCS2, the first two methylation steps of the caffeine biosynthetic pathway were successfully connected by a single PaulliniaAncCS2 at stage C (Figure 29). The functional evolution of the ancestral CS enzymes suggested that the caffeine biosynthetic pathway evolved in a forward stepwise manner in Paullinia, starting from the first reaction in the pathway. The third step of the pathway (from theobromine to caffeine) is not shown Figure 29, but probably evolved after additional gene duplication events in Paullinia [68].

As modern-day PaulliniaCS1 and PaulliniaCS2 are highly specific for their own substrate and catalyze the first and second steps of the caffeine biosynthetic pathway, respectively, I further investigated how they partitioned and inherited the ancestral functions from PaulliniaAncCS2 by mutating PaulliniaAncCS2 towards its modern-day descendants. Two active site residues (① and ③ in Appendix A) were mutated. These sites are suggested by the
Out of twenty substitutions from PaulliniaAncCS2 to PaulliniaCS2, one single substitution T25S abolished the ancestral activities with 3- and 7-methylxanthine, similar to the activity of modern-day PaulliniaCS1 (Figure 29).

Table 7. Kinetic parameters calculated for ancestral CS enzymes with selected substrates

<table>
<thead>
<tr>
<th>Enzyme + substrate</th>
<th>$K_M$ (μM)</th>
<th>$k_{cat}$ (s$^{-1}$)</th>
<th>$k_{cat}/K_M$ (sec$^{-1}$·M$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAAncCS + xanthine</td>
<td>2014</td>
<td>2.20E-04</td>
<td>0.11</td>
</tr>
<tr>
<td>PaulliniaAncCS1 + xanthine</td>
<td>417.5</td>
<td>3.76E-04</td>
<td>0.90</td>
</tr>
<tr>
<td>CamelliaAncCS + 7-methylxanthine</td>
<td>26.7</td>
<td>1.05E-03</td>
<td>39.33</td>
</tr>
</tbody>
</table>

Crystal structures of *Clarkia* SAMT [18] and *Coffeea* DXMT [19] to be important for substrate discrimination. Out of nineteen substitutions from PaulliniaAncCS2 to PaulliniaCS1, one single mutation N314Y (site 3) in PaulliniaAncCS2 resulted in an enzyme that specifically methylated xanthine to form 3-methylxanthine (Table 8), and abolished the ancestral activities with 3- and 7-methylxanthine, similar to the activity of modern-day PaulliniaCS1 (Figure 29).
(site ①) in PaulliniaAncCS2 resulted in a huge increase in the enzyme’s relative activity with 3-methylxanthine to form theobromine (Table 8), largely like the activity of modern-day PaulliniaCS2 (Figure 29). The connected pathway catalyzed by these two mutants at stage D was similar to the one catalyzed by the modern-day enzymes at stage E (Figure 29).

Table 8. LC-MS scans of products of enzymatic reactions of PaulliniaAncCS2 forward mutants with preferred substrates

<table>
<thead>
<tr>
<th>Enzyme + Substrate</th>
<th>X Area</th>
<th>1X Area</th>
<th>3X Area</th>
<th>7X Area</th>
<th>Tp Area</th>
<th>Tb Area</th>
<th>Px Area</th>
<th>Cf Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard eight xanthine alkaloids mix</td>
<td>329</td>
<td>1474</td>
<td>292</td>
<td>625</td>
<td>550</td>
<td>281</td>
<td>721</td>
<td>2088</td>
</tr>
<tr>
<td>PaulliniaAncCS2T25S + 3X</td>
<td>111</td>
<td></td>
<td></td>
<td></td>
<td>526</td>
<td></td>
<td></td>
<td>107</td>
</tr>
<tr>
<td>PaulliniaAncCS2T25S + Solvent</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS2N314Y + X</td>
<td>309</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>PaulliniaAncCS2N314Y + Solvent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Numbers represent raw peak areas. Substrate areas are colored in blue and product areas are colored in red. The presence of xanthine in some reactions is probably due to other chemical reactions in bacteria. Xanthine alkaloid abbreviations: X, xanthine; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tp, theophylline; Tb, theobromine; Px, paraxanthine; Cf, caffeine.

**Exapted multifunctional ancestral XMT enzymes facilitate the evolution of novel caffeine biosynthetic pathways in *Citrus***

To understand how multifunctional ancestral XMT enzymes evolved to catalyze the novel three-step caffeine biosynthetic pathways in *Citrus*, pathway connections catalyzed by these ancestral XMT enzymes are mapped to the possible caffeine biosynthetic routes shown in Figure 4.
As described earlier, historically maintained RAAncXMT, CitrusAncXMT1 and CitrusAncXMT2 are hypothesized to be used for O-methylation in plants (stage A, B and C in Figure 30), and CitrusAncXMT2 was likely exapted to catalyze the last two reactions of the caffeine biosynthetic pathway (3-methylxanthine $\rightarrow$ theophylline $\rightarrow$ caffeine) in Citrus (Table 6). It should be noted that although CitrusAncXMT2 also exhibits a trace level of activity with xanthine, its $K_M$ of xanthine is quite high, outside the reported in vivo concentration range (10 to 1000 µM) of most caffeine precursors in tissues of Theobroma, Paullinia and Citrus (Table 9) [47, 59, 77], so the reaction was unlikely to happen in ancient Citrus plant (stage C in Figure 30). Even if the ancient Citrus plant could provide enough amount of xanthine for the methylation to happen, the methylated product would be 1-methylxanthine (Figure 31a), which was then converted to paraxanthine, not theophylline (Table 6). Thus, a complete caffeine biosynthetic pathway couldn’t be catalyzed by CitrusAncXMT2 alone, unless ancient Citrus plants had a now-extinct enzyme that could have converted xanthine to 3-methylxanthine.

A functional caffeine biosynthetic pathway likely evolved after gene duplication in Citrus, when two modern-day descendants CitrusXMT1 and CitrusXMT2 inherited and partitioned ancestral functions from CitrusAncXMT2 (Figure 30). To trace the molecular basis for the functional divergence between CitrusXMT1 and CitrusXMT2, CitrusAncXMT2 was mutated towards the two descendants, respectively. Like the functional divergence between PaulliniaCS1 and PaulliniaCS2, single active site residue substitutions (① and ② in Appendix A) were also sufficient for the functional divergence between CitrusXMT1 and CitrusXMT2. Out of eighteen amino acid changes from CitrusAncXMT2 to CitrusXMT1 (Figure 30) (Appendix A), one active site replacement P25S (site ①) resulted in three important
Figure 30. Simultaneous evolution of caffeine biosynthetic pathways in *Citrus* and facile mutational basis for the functional diversification of *Citrus* caffeine synthases. Pathway connections are enclosed in boxes, and each evolutionary stage is shown on the top left corner of the box. Ancestral XMT enzymes are hypothesized to be used for O-methylation. Mean relative activities (0 to 100) of ancestral XMT enzymes and forward mutants demonstrated how duplicated genes inherited and partitioned ancestral functions in *Citrus*.

Table 9. Kinetic parameters calculated for ancestral XMT enzymes with selected substrates

<table>
<thead>
<tr>
<th>Enzyme + substrate</th>
<th>$K_M$ (µM)</th>
<th>$k_{cat}$ (s⁻¹)</th>
<th>$k_{cat}/K_M$ (sec·M⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ancCitrusXMT2 + xanthine</td>
<td>2740</td>
<td>2.48E-04</td>
<td>0.09</td>
</tr>
<tr>
<td>ancCitrusXMT2P25S + xanthine</td>
<td>1690</td>
<td>3.97E-04</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Figure 31. HPLC UV absorbance spectra show important evolutionary changes by introducing one large-effect mutation P25S into CitrusAncXMT2. a. CitrusAncXMT2 converts xanthine (X) to 1-methylxanthine (1X), while the P25S mutant converts xanthine (X) to both 1-
methylxanthine (1X) and 3-methylxanthine (3X) as well as trace amount of theophylline (Tp). b. The P25S mutant converts 1-methylxanthine (1X) to theophylline (Tp) instead of paraxanthine (Px). Substrates are enclosed in blue rectangle and methylated products are enclosed in red rectangle. Enzyme assays were conducted using 2mM xanthine and 200µM 1-methylxanthine, and products were measured by absorbance at 272 nm. Xanthine alkaloid abbreviations: 7X, 7-methylxanthine; Tb, theobromine; Cf, caffeine.

Evolutionary changes. First, a great loss of ancestral activities with benzoic acid, salicylic acid and theophylline occurred, so that the mutant was specialized to methylate xanthine, 1- and 3-methylxanthine, like the relative activities of modern-day CitrusXMT1 (Figure 30). Second, compared with AncCitrusXMT2, both the affinity and catalytic efficiency of the P25S mutant with xanthine increased a lot (Table 9), so that xanthine became potentially physiologically relevant and was methylated to form both 1-methylxanthine and 3-methylxanthine, the same as CitrusXMT1 (Figure 31a). Third, the P25S mutant now methylated 1-methylxanthine to form theophylline instead of paraxanthine (Figure 31b), also the same as CitrusXMT1. Therefore, the first and second steps of the caffeine biosynthetic pathway in Citrus evolved simultaneously by introducing this large-effect mutation into the duplicated CitrusAncXMT2 (stage D in Figure 30). Next, there are seventeen amino acid substitutions on the lineage from CitrusAncXMT2 to CitrusXMT2 (Figure 30) (Appendix A). One mutation, H157N (site ②), which is also at the active site and is important for SABATH enzyme’s activity change as suggested by other studies [112, 125], resulted in a great loss of activities with xanthine, 1- and 3-methylxanthine, while maintaining high relative activity with theophylline to convert it to caffeine (Table 10) (stage D in Figure 30), such that it resembled the modern-day citrusXMT2. Even though additional fine-tune mutations would be needed to fully evolve the descendants’ activities, the two intermediate
enzymes constituted a complete caffeine biosynthetic pathway that was much like the pathway in modern-day *Citrus* (stage E in Figure 30).

Table 10. LC-MS scans of products of enzymatic reactions of CitrusAncXMT2 forward mutants with preferred substrates

<table>
<thead>
<tr>
<th>Enzyme + Substrate</th>
<th>X Area</th>
<th>1X Area</th>
<th>3X Area</th>
<th>7X Area</th>
<th>Tp Area</th>
<th>Tb Area</th>
<th>Px Area</th>
<th>Cf Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard eight xanthine alkaloids mix</td>
<td>329</td>
<td>1474</td>
<td>292</td>
<td>625</td>
<td>550</td>
<td>281</td>
<td>721</td>
<td>2088</td>
</tr>
<tr>
<td>CitrusAncXMT2P25S + 3X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>835</td>
<td></td>
<td></td>
<td>217</td>
</tr>
<tr>
<td>CitrusAncXMT2P25S + Solvent</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CitrusAncXMT2H157N + 1X</td>
<td></td>
<td>1980</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2400</td>
</tr>
<tr>
<td>CitrusAncXMT2H157N + Tp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2230</td>
<td>127</td>
</tr>
<tr>
<td>CitrusAncXMT2H157N + Solvent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Numbers represent raw peak areas. Substrate areas are colored in blue and product areas are colored in red. The presence of xanthine in some reactions is probably due to other chemical reactions in bacteria. Xanthine alkaloid abbreviations: X, xanthine; 1X, 1-methylxanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tp, theophylline; Tb, theobromine; Px, paraxanthine; Cf, caffeine.

**DISCUSSION**

Many modern-day characterized SABATH members have activities with several structurally similar substrates [30, 57, 112], therefore, it is not surprising that ancestral caffeine synthases, like most other resurrected ancestors, were also “generalists”, which could methylate more than one xanthine alkaloid, or even had activities with benzoic acid and salicylic acid (such as RAAncXMT, CitrusAncXMT1 and CitrusAncXMT2). Both ancestral CS and XMT enzymes were maintained for more than 100 million years for alternative biochemical functions, and their exapted activities with xanthine alkaloids finally facilitated the convergent evolution of caffeine biosynthetic pathways in *Camellia, Paullinia, Coffea* and *Citrus*. After gene duplication within
each species, modern-day descendant enzymes either inherited or evolved new functions from these ancestral enzymes.

The caffeine biosynthetic pathway in *Paullinia* evolved in a stepwise cumulative manner from *PaulliniaAncCS1*. Assuming each xanthine alkaloid intermediate in the caffeine biosynthetic pathway conferred a selective advantage for *Paullinia*, the evolution of caffeine biosynthetic pathway in *Paullinia* fits the cumulative (Granick) hypothesis. The first step, in which xanthine was converted to 3-methylxanthine by *PaulliniaAncCS1* evolved earliest, then came up with the second step, in which 3-methylxanthine was converted to theobromine by *PaulliniaAncCS2*, and finally the third step, in which theobromine was converted to caffeine by *PcCS* evolved. Although theobromine is reported to be toxic to insect larvae, the same as caffeine, thus protecting plants from insect feeding [126], there are no published reports regarding the physiological role of 3-methylxanthine other than caffeine synthesis in *Paullinia* or in other plant species that produce caffeine. If 3-methylxanthine had no selective advantage, the evolution of caffeine biosynthetic pathway in *Paullinia* may be explained using a combination of the patchwork and cumulative hypotheses, with the assumption that xanthine, 3-methylxanthine and 7-methylxanthine were fortuitous minor substrates for the functionally promiscuous *PaulliniaAncCS1*. It was maintained for presumably analogous methylation reactions, until higher level of activity with 3-methylxanthine evolved in *PaulliniaAncCS2*, and the biosynthesis of theobromine became selectively advantageous for *Paullinia*. Natural selection that favored the production of theobromine would then prompt maintenance of duplicated genes, and subsequent functional specializations of enzymes that catalyze the first two methylation steps in the caffeine biosynthetic pathway in *Paullinia* could happen. The third step in the caffeine biosynthetic pathway latter evolved in *Paullinia* also in a cumulative manner.
The evolution of the caffeine biosynthetic pathways in *Citrus* can be explained by all three hypotheses. First, CitrusAncXMT1 and CitrusAncXMT2 were multifunctional enzymes whose primary role appeared to be methylating benzoic acid and/or salicylic acid, and they were also exapted to methylate several xanthine alkaloids, which were intermediates in the caffeine biosynthetic pathway in modern-day *Citrus*. Later, the descendant enzymes evolved to only methylate xanthine alkaloids and lost the ability for O-methylation. Therefore, it fits the patchwork hypothesis. Second, CitrusAncXMT2 was likely exapted to catalyze the last two steps of the caffeine biosynthetic pathway (3-methylxanthine → theophylline → caffeine), and the initial step(s) (xanthine → 3-methylxanthine or xanthine → 1-methylxanthine → theophylline) evolved after the gene duplication and functional divergence of CitrusAncXMT2. Therefore, it also fits the retrograde hypothesis. Third, the assembly of the first two steps (xanthine → 1-methylxanthine → theophylline) in the pathway likely evolved in a cumulative manner. Because one dilemma in the cumulative hypothesis was that intermediates in a pathway should have selective advantages or should exist in the primitive environment before multiple recruitments of enzymes for each step happened. However, that would not be possible until the intact pathway was formed. The existence of exapted CitrusAncXMT2 solved this dilemma because the two steps in the caffeine biosynthetic pathway evolved simultaneously after gene duplication and evolutionary divergence of CitrusAncXMT2, without the need of assuming the selectively advantageous 1-methylxanthine.

Although sign epistasis was found in many studies that used ASR and forward mutagenesis [116, 117], it might happen rarely to members in the SABATH family. Like the single active site replacement (histidine to methionine or *vice versa*) in enzymes in the SAMT/BSMT clade controlled the substrate preference switch between benzoic acid and
salicylic acid without any epistatic interactions [112], single large-effect mutations were also responsible for the functional divergence of PaulliniaAncCS2 and CitrusAncXMT2 without the need of introducing any permissive mutations first as illustrated in previous studies [116, 117].

The ancient ancestral enzyme, RAAncXMT in the XMT clade was found to prefer benzoic acid and salicylic acid to xanthine alkaloids, and its activity with benzoic acid was inherited in the modern-day *Mangifera*. Perhaps not only in *Mangifera*, because XMT genes are found in many species in flowering plants, like *Kirkia wilmsii*, *Eucalyptus leucoxylon*, and *Acer negundo* in Figure 25, and they are probably able to methylate benzoic acid as well (need to functionally characterize these genes in the future). The ability to produce methyl benzoate is important for pollinator attraction in plants [127], it is not surprising that *Mangifera* uses the XMT enzyme to methylate benzoic acid, because there is no BAMT gene identified in *Mangifera* when BLAST the *Mangifera* transcriptome using either an SAMT or BAMT sequence (data not shown). In contrast, although *Citrus* has no BAMT gene in its genome, it expresses an SAMT gene that can methylate benzoic acid (Figure 8), so that its ancestral XMT gene might be free to evolve the ability to synthesize caffeine.
CHAPTER IV: FUNCTION OF SABATH GENES IN TILIA FLOWERS USING RNA-SEQ AND BIOCHEMICAL STUDIES

INTRODUCTION

*Tilia*, generally called linden (*Tilia cordata*, European species) or basswood (*Tilia americana*, North American species), is a genus in Malvales, and closely related to *Theobroma* and *Cola* [60, 128]. Species in *Tilia* are mostly large, deciduous trees, which are commonly used as ornamentals to provide shade because of their mass foliage. *Tilia* trees also produce fragrant flowers with various volatile compounds in the nectar to attract honeybees [48], thus it is also cultivated by beekeepers to produce monofloral honey. Caffeine is produced by *Tilia* flowers [48], and it is probably used to attract honeybees by enhancing the memory of reward, the same role found in *Citrus* flowers [42]. In fact, the caffeine from *Tilia* is also detected in the honey stomach and linden honey from the honeybees [48]. Therefore, I have characterized the caffeine biosynthetic pathway and expressed genes in *Tilia* flowers as a supplement to Chapter II regarding the convergent evolution of caffeine biosynthetic pathways in various caffeine-producing species.

Given my studies in Chapter II and Chapter III, I had three hypotheses about the evolution of caffeine biosynthetic pathway in *Tilia* flowers. On one hand, as caffeine is produced in flowers of *Tilia* and *Citrus*, it is possible that their caffeine biosynthetic pathways are the same, and *Tilia* also evolved to use XMT enzymes for caffeine biosynthesis. On the other hand, as *Tilia* is a close relative of *Theobroma*, which produces theobromine and caffeine in seeds, the second possibility is that their caffeine biosynthetic pathways are the same and both use CS enzymes to produce caffeine [129]. The third possibility is that a new caffeine biosynthetic pathway has evolved in *Tilia*, probably using a third type of SABATH enzyme. As there is rarely any genetic information available for *Tilia*, to test these hypotheses, the transcriptome of *Tilia*...
was constructed using the RNA-Seq technique. Assembled full-length SABATH genes were then functionally characterized, and metabolites in different development stages of *Tilia* flowers were sampled.

**MATERIALS AND METHODS**

**RNA extraction**

Total RNA was extracted from *Tilia cordata* developing flower buds following a published protocol [73]. Total RNA was then digested using RNase-free DNase Kit (Qiagen) to remove residual genomic DNA followed by a column cleanup using RNeasy Plant Mini Kit (Qiagen). RNA quantity and quality were measured using a NanoDrop (Thermo Scientific) before being sent out for reverse transcription and sequencing.

**RNA-Seq and transcriptome assembly**

Total RNA was sent to BGI AMERICAS CORPORATION for sequencing. Subsequent de novo transcriptome assembly, Unigene functional annotation and expression level calculation were all performed by this company using the following platform and methods: RNA sequencing was performed on Hiseq 4000 platform with 150 bp paired-end read length and 30 million reads. Raw sequencing reads were filtered to remove low-quality and adaptor-polluted reads. Filtered clean reads were de novo assembled into contigs using Trinity (v2.0.6) [130] with the following parameters: \(--\text{min\_contig\_length}\ 150\ --\text{CPU}\ 8\ --\text{min\_kmer\_cov}\ 3\ --\text{min\_glue}\ 3\ --\text{bfly\_opts}\ \text{'}-\text{V} 5\ --\text{edge\_thr}=0.1\ --\text{stderr}\'. Contigs were then clustered to Unigenes using Tgicl (v2.0.6) [131] with the following parameters: \(-\text{I}\ 40\ -\text{c}\ 10\ -\text{v}\ 25\ -\text{O}\ \text{'}-\text{repeat\_stringency}\ 0.95\ -\text{minmatch}\ 35\ -\text{mins}\text{core}\ 35\'. Unigenes were used to BLAST [132] the NT (NCBI: nucleotide sequence database), NR (NCBI: non-redundant protein sequence database), COG (NCBI: Cluster of Orthologous Groups of proteins, phylogenetic classification of proteins encoded in complete genomes), KEGG
(Kyoto Encyclopedia of Genes and Genomes) and SwissProt databases to get the functional annotations. Further annotations were done by using Blast2GO with NR annotation to get the GO (Gene Ontology project) database annotation [133] and using InterProScan5 to get the InterPro database annotation [134]. Finally, clean reads were mapped to assembled Unigenes using Bowtie2 (v2.2.5) [135] with the following parameters: q --phred64 --sensitive --dpad 0 --gbar 99999999 --mp 1,1 --np 1 --score-min L,0,-0.1 -I 1 -X 1000 --no-mixed --no-discardant -p 1 -k 200, and relative gene expression levels were calculated with RSEM (v1.2.12) [136] with default parameters.

**Gene synthesis and cloning**

Full-length SABATH genes (except Tilia9944 contig2) in *Tilia* were amplified using SuperScript III One-Step Reverse Transcription (RT)-PCR Kit (Invitrogen) using gene specific primers and cloned into either pET15b vector (Novagen) (Tilia9944 contig1) with a N-terminal His-tag or pTricHis2 TOPO vector (other full-length genes) with a C-terminal His-tag (Invitrogen) for protein expression and purification. Tilia8820 was also cloned into the expresso SUMO vector (Lucigen) for protein expression and purification. The nucleotide sequence of Tilia9944 contig2 was synthesized by GenScript Corp. with codons optimized for expression in *E. coli* and cloned into pUC57 vector. It was subcloned into pET15b vector (Novagen) with a N-terminal His-tag for subsequent protein expression and purification. All recombinant plasmids were checked on an agarose gel for their size and sequenced by GENEWIZ Corp to ensure that all cloned sequences were in the correct open reading frame.

**Protein expression**

Recombinant pET15b plasmids were transformed into *E. coli* strain BL21(DE3) Star™ cells (Invitrogen), recombinant pTricHis2 plasmids were transformed into *E. coli* strain Top10
cells (Invitrogen), and recombinant expresso SUMO plasmid was transformed into HI-Control™ BL21(DE3) cells (Lucigen) for overexpression of His-tagged proteins. Protein expression started from a 5 ml starter culture of transformed E. coli cells, which were grown at 37°C in LB broth containing 100 µg/ml ampicillin (or 50 µg/ml Kanamycin if the gene was cloned into the SUMO vector) overnight. On the next day, the 5 ml starter culture was used to inoculate a fresh 100 ml LB broth containing the same concentration of antibiotics and continued to grow until OD$_{600}$ reached 0.6 to 0.8, then induced with 1.0 mM IPTG for 6 hours at 23°C. The cells were harvested by centrifugation at 4°C, 4000 rpm for 15 minutes. To isolate proteins from E. coli cells, pellets were resuspended in chilled 1X equilibration buffer (50 mM Na$_3$PO$_4$, 300 mM NaCl, 10 mM imidazole, 10 mM β-mercaptoethanol, and 12% (v/v) glycerol, pH 8), treated with 0.75 mg/ml Lysozyme on ice for 30 minutes, and then sonicated on ice for 20 seconds, 3 times with a 20-second pause on ice between each burst. The suspension after sonication was centrifuged at 10,000 g for 20 minutes at 4°C to get rid of cell debris. The supernatant contained the soluble crude proteins which were either used directly for enzyme catalytic assays or subsequently purified.

**Protein purification**

Protein purification was achieved using cobalt spin columns (Clontech) according to the manufacturer’s instructions. Purified proteins were eluted in 1X elution buffer (50 mM Na$_3$PO$_4$, 300 mM NaCl, 200 mM imidazole, 10 mM β-mercaptoethanol, and 12% (v/v) glycerol, pH 8), aliquoted and flash frozen in a dry ice-ethanol bath and stored at -80°C. The purity and molecular mass of purified proteins was evaluated by running on pre-cast NuPAGE 4-12% Bis-Tris gels (Invitrogen), and protein concentration was calculated by the Bradford assay [74] using bovine serum albumin (BSA) as the standard.
In vitro functional analysis

Relative enzymatic activities of heterologously expressed *Tilia* proteins with eight xanthine alkaloids as well as benzoic acid, salicylic acid, indole-3-acetic acid and phenyl acetic acid were tested using SAM with a $^{14}$C-labeled methyl group. For each substrate, a 50 µl reaction containing 10 or 20 µl purified proteins (or 30 µl crude extracted total cellular proteins), 100 µM methyl acceptor substrate and 0.01 µCi $^{14}$C-labeled SAM was carried out in 50 mM Tris-HCl buffer at room temperature for 20 minutes (or 1 hour if using crude cellular proteins). The methylated product was extracted using 200 µl ethyl acetate, and 120 µl of the ethyl acetate (top) phase was removed and mixed with 4 ml scintillation counting cocktail. Radioactive decay was measured using a liquid scintillation counter (PerkinElmer) and disintegrations per minute (DPM) values were reported. Assays were done in duplicate or triplicate and standard deviations were calculated.

High-performance liquid chromatography (HPLC)

Identification of products from each enzymatic reaction with xanthine alkaloids as the substrates was determined using HPLC on reactions that were scaled up 10-fold using non-radioactive SAM and 2 mM methyl acceptor substrates for two hours at room temperature. The whole reaction was acidified to pH 5–6, then filtered through Vivaspin columns (Sartorius Stedim Biotech) to remove proteins and impurities, and 10 µl of the processed reaction (or metabolite extract) was injected into HPLC. Mixtures were separated by HPLC using a two-solvent system with a 250 mm × 4.6 mm Kinetex 5µm EVO C18 column (Phenomenex). Solvent A was 99.9% (V/V) water with 0.1% TFA and solvent B was 80% (V/V) acetonitrile, 19.9% (V/V) water, and 0.1% TFA, and a 0–16% solvent B gradient was generated over 16 minutes with a flow rate of 1.0 ml/min. Subsequently, buffer B was increased to 100% and then held at
that percentage for 20 minutes. Equilibration back to 0% buffer B was achieved over a 20-minute period. Two UV wavelengths (254 and 272 nm) were used for absorption measurements. The products were verified by comparing its absorbance peak/retention time with xanthine alkaloid standards. Negative control reactions without the substrate added were always run in parallel.

**Heterologous gene expression and gas chromatography-mass spectrometry (GC-MS) analysis**

As some heterologously expressed *Tilia* proteins lost activities or were insoluble after extraction from bacterial cells, a different approach was used to characterize the enzyme’s function. A 50 ml transformed *E. coli* Top10 cell culture containing the recombinant pTricHis2 plasmid was grown to optical density (OD) 0.6 - 0.8 prior to addition of IPTG. After induction with 1 mM IPTG for 4 hours at 23°C, 200 µl of 50 mM substrate (benzoic acid or salicylic acid) was added to the culture, which continued to grow for another 2 hours at 23°C. Following the 2-hour incubation with the substrate, cells were pelleted at 4°C, 4000 rpm for 15 minutes, and the supernatant (spent growth medium) was collected. Volatiles were extracted from the supernatant with 4 ml hexane and 2 ml recovered hexane phase was transferred to a vial and ready for GC-MS analysis. Negative controls were set up the same way except that no substrate was added after IPTG induction. One µl concentrated product was analyzed by GC-MS on an HP6890GCSystem coupled to an HP5973 Mass Selective Detector using a DB-5 capillary column with the following oven conditions: 40 °C for 2 minutes, ramping 20°C / minute to 300°C with a 2-minute hold. Products were identified by comparing their mass spectra and retention times with available standards and reference spectra from the library of National Institute of Standards and Technology. The same approach was also used to identify the products of enzymatic reactions with indole-3-acetic acid or phenylacetic acid as the substrate.
Quantitative reverse transcription PCR (RT-qPCR)

Two-step RT-qPCR using the Power SYBR Green PCR Master Mix Kit (Applied Biosystems) was used to quantify the relative expression levels of assembled full-length SABATH genes in developing *Tilia* flower buds. Prior to quantification, primers spanning the estimated intron regions of Tilia9944 contig1 were designed and used to test whether there was any residual genomic DNA left in total RNA after DNase digestion. Primer pairs for each gene were designed using PerlPrimer (v1.1.21) [137] with amplicon lengths selected between 103 bp and 110 bp, melting temperature (Tm) between 58°C and 60°C, and low chance of primer-dimer formation. mRNA was firstly reverse transcribed into complementary DNA (cDNA) by SuperScript III reverse transcriptase using oligo d(T)$_{20}$ at 37°C for 60 minutes, followed by inactivation of reverse transcriptase by heating at 70°C for 15 minutes. DNA of each full-length gene was amplified from its recombinant plasmid mentioned above, gel purified and quantified using a NanoDrop (Thermo Fisher Scientific). The gel purified DNA was then used as the stock solution of the quantification standard for that gene. An equal amount of hybrid cDNA / mRNA from the reverse transcription reaction was added as the template for the second step PCR amplification for each gene, which was quantified based on its own standard curve, using the StepOne Plus Real-Time PCR System (Applied Biosystems). A 10-fold serial dilution of the stock standard of each gene (with the concentration range from $10^{-4}$ µM to $10^{-10}$ µM) was amplified together to create the standard curve to quantify the amount of each amplicon in the hybrid cDNA / mRNA pool. Parameters for the PCR amplification were as follows: samples were held at 95°C for 10 minutes first, then cycled at 95°C for 15 seconds, 55°C for 30 seconds, and 72°C for 30 seconds for a total of 45 cycles. Fluorescent signals were collected at step 3 of each cycling stage. The melt curve analysis, which is an assessment of the dissociation-
characteristics of double-stranded DNA during heating, was performed immediately following the PCR amplification to check for the existence of primer dimers and possible contamination in the no template control sample. The parameters used for the melt curve analysis were: 95°C for 15 seconds, 60°C for 1 minute, ramp temperature from 60°C to 95°C with 0.3°C increments. Fluorescent signals were collected at ramp temperature from 60°C to 95°C as well as at 95°C.

RESULTS

Phylogenetic analysis predicted possible functions of some full-length SABATH genes in Tilia

Six full-length SABATH genes were assembled from the sequenced Tilia transcriptome and were temporarily named according to their assembled identification numbers: Tilia2608, Tilia7163, Tilia8820, Tilia9944, Tilia13434, and Tilia14622. Tilia9944 has two contigs, in which contig2 is an 8-amino-acid splice variant of contig1 (Figure 32), thus only contig1 was phylogenetically analyzed. A phylogenetic analysis of proteins encoded by these six genes together with characterized representative SABATH members in other species predicted possible functions/substrates of four genes: Tilia2608 falls in the SAMT clade, Tilia7163 falls in the IAMT clade, Tilia9944 falls in the CS clade, and Tilia13434 is in the same clade with BAMT/BSMT (Figure 33). Substrates of Tilia8820 and Tilia14622 couldn’t be predicted as they are not closely related to other functionally characterized enzyme branches in the phylogenetic tree, but are in clades MT3 and MT1, respectively (Figure 33). Therefore, Tilia9944 is likely a candidate gene for caffeine biosynthesis.
Figure 32. Tilia9944 contig2 is an alternatively spliced version of contig1. 

a. Nucleotide sequence alignment of Tilia9944 contig1 and contig2 with closely related Theobroma genomic and exon sequences showing the splice position. 

b. The alternatively spliced site in Tilia9944 contig2 is very close to the active site (positions 190 – 199) of SABATH enzymes.
Figure 33. Phylogenetic analysis of Tilia SABATH genes with functionally characterized SABATH members predicts their possible functions (substrates). Branches leading to Tilia SABATH genes are colored in red. CitrusMT1 to CitrusMT3 were Citrus genomic sequences without known functions that were arbitrarily numbered in Chapter II.
Tilia9944 is expressed in *Tilia* developing flower buds

Because caffeine synthase genes in *Citrus*, *Paullinia* and *Camellia* were found to be highly represented by EST counts in tissues where caffeine is accumulated (Figures 7, 15 and 20), relative expression levels of Tilia9944 together with five other assembled full-length SABATH genes were initially quantified using the FPKM (fragments per kilobase of transcript per million mapped reads) values given by the RNA-Seq technique (Figure 34a). The FPKM value for Tilia9944 was 0, making the role of Tilia9944 for caffeine biosynthesis unclear. To accurately quantify the relative expression level of Tilia9944 and screen other possible candidate genes for caffeine biosynthesis, relative expression levels of six full-length SABATH genes in *Tilia* developing buds were quantified using Quantitative Reverse Transcription PCR (RT-qPCR) (Figure 34b). Relative expression levels of Tilia2608, Tilia8820, Tilia7163, and Tilia13434 are largely consistent between the two quantification methods, while the relative expression levels of Tilia9944 and Tilia14622 differ substantially. Results from RT-qPCR showed that Tilia9944 is expressed in the developing buds.

Figure 34. Relative expression levels of full-length SABATH genes in *Tilia* developing buds (a) given by the FPKM values and (b) quantified using RT-qPCR. Relative expression levels of Tilia9944 and Tilia14622 differ between the two quantification methods. b. Tilia8820 has the
highest relative expression level within a given amount of RNA, which is set arbitrarily to 100. Expression levels of the other five genes are normalized to the value of Tilia8820. Tilia9944 is colored in green.

**Heterologously expressed Tilia2608 demonstrated higher relative activity with salicylic acid than with benzoic acid**

As crude cellular proteins of Tilia2608 lost enzymatic activity after being extracted from bacterial cells, potential substrates (salicylic acid and benzoic acid) were added directly into the bacterial culture that was expressing the target gene. Methylated products were extracted using hexane and verified using GC-MS. Tilia2608 primarily methylates salicylic acid to form methyl salicylate, and it also has a weak activity to methylate benzoic acid to form methyl benzoate, thus it is renamed as TiliaSAMT (Figure 35). To further verify its substrate preference between benzoic acid and salicylic acid, equimolar benzoic acid and salicylic acid solutions were added to the bacterial culture expressing TiliaSAMT, and only methyl salicylate could be detected (Figure 35).

**Heterologously expressed Tilia7163 demonstrated specific activity with indole-3-acetic acid**

Crude total cellular proteins extracted from heterologously expressed Tilia7163 showed activity only with indole-3-acetic acid (Figure 36), and the product was methyl indole-3-acetate. Its activity is consistent with its the lineage position on the phylogenetic tree (Figure 33), thus is renamed as TiliaIAMT.
Figure 35. GC-MS spectra show that TiliaSAMT primarily methylates salicylic acid to methyl salicylate. TiliaSAMT also has weak activity with benzoic acid, and the product is methyl benzoate. The mass spectra for methyl salicylate and methyl benzoate are shown in insets in the first two spectra, respectively.
Figure 36. Mean relative activities (from 0 to 100) of TiliaIAMT with benzoic acid, salicylic acid, indole-3-acetic acid and eight xanthine alkaloids. TiliaIAMT is highly specific for the substrate indole-3-acetic acid.

**Tilia8820 likely methylates unknown substrates**

Tilia8820 falls into the same clade with CitrusMT3, which is not functionally characterized (Figure 33). Tilia8820 has been cloned into both the pTricHis2 vector and the SUMO vector, heterologously expressed and functionally tested, yet no activity was detected with benzoic acid, salicylic acid or xanthine alkaloids. Since the SDS-PAGE gel showed that heterologously expressed SUMO Tilia8820 yielded soluble proteins (Figure 37a), the substrate for Tilia8820 remains unknown.

**Tilia13434 likely produces insoluble proteins when heterologously expressed**

Since Tilia13434 is closely related to BSMT in *Arabidopsis*, it was predicted to be capable of methylating benzoic acid and salicylic acid. However, neither the radioactive assay nor the GC-MS analysis of heterologously expressed Tilia13434 detected any activity with benzoic acid, salicylic acid or xanthine alkaloids (data not shown). When I tried to purify heterologously expressed Tilia13434 crude cellular proteins, there was no visible target band on
the SDS-PAGE gel (Figure 37b), indicating that Tilia13434 likely produces insoluble proteins when heterologously expressed in *E. coli*.

**a. SDS-PAGE gel of purified Tilia8820**

- Lane M: AMRESCO Protein Molecular Weight Markers
- Lane 1: Soluble crude cellular proteins
- Lane 2: Flow through
- Lane 3: First wash
- Lane 4: Second wash
- Lane 5: First eluate
- Lane 6: Second eluate

**b. SDS-PAGE gel of purified Tilia13434**

Figure 37. SDS-PAGE gels of purified 8820 (a) and Tilia13434 (b). a. Tilia8820 was expressed in the expresso SUMO plasmid with a N-terminal His-tag and a SUMO tag, and the molecular weight of Tilia8820 plus the 6xHis-tag and SUMO tag is about 53 kDa. Bands of target protein are indicated in red circles. b. Tilia13434 was expressed in the pTricHis2 plasmid with a C-terminal His-tag, and the molecular weight of Tilia13434 plus 6xHis-tag is about 45 kDa. There is no visible target protein band in either of the two elution lanes or the soluble crude cellular protein lane.
Tilia14622 codes for a new MT that has high relative activity with phenylacetic acid (PAA)

Although Tilia14622 falls into the same clade with CitrusMT1 (Figure 33), which is not functionally characterized, heterologously expressed Tilia14622 showed activity with PAA, and the product was verified to be methyl phenylacetate, thus it is renamed as PAAMT (Figure 38).

Tilia9944 methylates xanthine alkaloids

Heterologously expressed Tilia9944 contig1 had noticeable activities with xanthine, 1-methylxanthine, 3-methylxanthine and theophylline (Figure 39), thus it is renamed as TiliaCS1. HPLC spectra of reactions of TiliaCS1 with the four xanthine alkaloids showed that TiliaCS1 methylates xanthine to form 3-methylxanthine (Figure 40a), 3-methylxanthine to form theobromine (Figure 40b), 1-methylxanthine to form theophylline (Figure 41a) and theophylline to form caffeine (Figure 41b). In contrast, Tilia9944 contig2, which is an alternatively spliced variant of TiliaCS1 (Figure 32), had no activities with any xanthine alkaloids (data not shown).

Potential caffeine biosynthetic pathways evolved in Tilia cordata (linden)

After mapping the activities of TiliaCS1 with the four xanthine alkaloids on the possible caffeine biosynthetic routes, it appears that TiliaCS1 can catalyze two consecutive steps of two different potential caffeine biosynthetic pathways (Figure 42).

To further investigate the caffeine biosynthetic pathway evolved in Tilia and identify possible pathway intermediates, metabolites from Tilia developing buds and open flowers were extracted and examined using HPLC. In addition, I also sampled the whole stamens (including anther, pollen and filament) of Tilia developing buds and open flowers, because caffeine and theophylline were reported to primarily exist in male reproductive organs in Citrus flowers [47]. Xanthine alkaloids identified in the HPLC spectra of stamens were found to be very similar to
Figure 38. GC-MS spectra show that TiliaPAAMT methylates phenylacetic acid to methyl phenylacetate. The mass spectrum for methyl phenylacetate is shown in the inset of the first spectrum.

Figure 39. Mean relative activities (0 to 100) of crude and purified TiliaCS1 with eight xanthine alkaloids. TiliaCS1 has noticeable activities with xanthine, 1-methylxanthine, 3-methylxanthine and theophylline. Because enzymatic activities with xanthosine and theobromine were too low for product identification through HPLC, relative activities of xanthosine and theobromine were corrected for extraction efficiency assuming the products were 7-methylxanthine and caffeine, respectively.

<table>
<thead>
<tr>
<th>Substrate Abbreviations:</th>
</tr>
</thead>
<tbody>
<tr>
<td>XR, xanthosine</td>
</tr>
<tr>
<td>X, xanthine</td>
</tr>
<tr>
<td>1X, 1-methylxanthine</td>
</tr>
<tr>
<td>3X, 3-methylxanthine</td>
</tr>
<tr>
<td>7X, 7-methylxanthine</td>
</tr>
<tr>
<td>Tp, theophylline</td>
</tr>
<tr>
<td>Tb, theobromine</td>
</tr>
<tr>
<td>Px, paraxanthine</td>
</tr>
</tbody>
</table>
Figure 40. HPLC spectra show that TiliaCS1 methylates (a) xanthine (X) to form 3-methylxanthine (3X) and (b) 3-methylxanthine (3X) to form theobromine (Tb). Reactions with solvent (0.5M NaOH), in which substrates were dissolved, were used as negative controls. Substrates are enclosed in blue rectangles and products are enclosed in red rectangles. Xanthine alkaloid abbreviations: 1X, 1-methylxanthine; 7X, 7-methylxanthine; Tp, theophylline; Px, paraxanthine; Cf, caffeine.
Figure 41. HPLC analysis shows that crude TiliaCS1 methylates (a) 1-methylxanthine (1X) to form theophylline (Tp) and (b) theophylline (Tp) to form caffeine (Cf). Reactions with solvent (0.5M NaOH), in which substrates were dissolved, were used as negative controls. Substrates are enclosed in blue rectangles and products are enclosed in red rectangles. Xanthine alkaloid abbreviations: X, xanthine; 3X, 3-methylxanthine; 7X, 7-methylxanthine; Tb, theobromine; Px, paraxanthine.
Figure 42. Two potential caffeine biosynthetic pathways (orange arrows and blue arrows) evolved in *Tilia* based on enzymatic activities of TiliaCS1.

Those identified in the HPLC spectra of entire buds (Figure 43a) or flowers (Figure 43b). The primary xanthine alkaloids were found to be 3-methylxanthine, theobromine, theophylline and caffeine, and minor xanthine alkaloids were 7-methylxanthine and paraxanthine (Figure 43). The existence of xanthine was not clear from these spectra (Figure 43). However, 1-methylxanthine was undetectable in these spectra (Figure 43), indicating that it was either unlikely to be a pathway intermediate or was completely converted to theophylline.

It should be noted that different relative quantities of 3-methylxanthine, theophylline and caffeine were found as metabolites in *Tilia* developing buds compared to open flowers (Figure 43). Caffeine is the most abundant xanthine alkaloid in the developing buds (Figure 43a), while theophylline is the most abundant xanthine alkaloid in the open flowers (Figure 43b). Therefore, the roles of 3-methylxanthine and theophylline were unclear in open flowers, because they may be degradation products of caffeine, rather than precursors [49, 79, 80]. Metabolites found in *Tilia* developing buds and open flowers failed to provide definitive support for either of the two potential caffeine biosynthetic pathways catalyzed by TiliaCS1 (Figure 42).
Figure 43. Accumulated xanthine alkaloids in *Tilia* (a) developing buds and (b) open flowers. Existing xanthine alkaloids are enclosed in red columns. Xanthine alkaloids identified in the HPLC spectra of stamens were found to be very similar to those identified in the spectra of entire buds or flowers. 3-methylxanthine (3X), theophylline (Tp) and caffeine (Cf) are more abundant than other xanthine alkaloids. No 1-methylxanthine (1X) was detected in developing buds and flowers, and the existence of xanthine (X) is unclear. Xanthine alkaloid abbreviations: 7X, 7-methylxanthine; Px, paraxanthine.
DISCUSSION

Based on gene expression level and enzymatic activities of TiliaCS1, three possible caffeine biosynthetic pathways may exist in *Tilia* flowers. Relative activities of TiliaCS1 with xanthine, 1-methylxanthine, 3-methylxanthine and theophylline show that it is capable of methylating at the N-3 and N-7 positions of the purine ring. However, there is little evidence that this enzyme can methylate the N-1 position (Figures 44a). A second caffeine synthase (either an CS enzyme or an XMT enzyme), which is potentially able to methylate at the N-1 position of the purine ring, may complete the caffeine biosynthetic pathway by either methylating xanthine to form 1-methylxanthine (Figure 44b, I) and/or methylating theobromine to caffeine (Figure 44b, II), and/or methylating 3-methylxanthine to form theophylline (Figure 44b, III). Pathway II is the most plausible route if theophylline is the precursor of caffeine instead of its degradation product, because 1-methylxanthine was not detected in *Tilia* developing buds and flowers, and TiliaCS1 has low relative activity of methylating 3-methylxanthine to form theobromine. Pathway III is the second possible route because although theobromine is converted inefficiently from 3-methylxanthine, theobromine exists at low levels in the developing buds and flowers. Pathway III may also be possible if the failure to detect 1-methylxanthine is due to its rapid conversion to theophylline. In addition, all three pathways may exist at the same time with one pathway being the primary one and the other two being the minor ones, given the possible substrate promiscuity of the potential N-1 methyltransferase.

The alternatively spliced TiliaCS1 (Tilia9944 contig2) was initially proposed to be a candidate for another CS enzyme. However, it was later found incapable of methylating any of these eight xanthine alkaloids. One possible explanation for its existence relates to the reduced
Figure 44. Another caffeine synthase gene may exist to complete the caffeine biosynthetic pathway evolved in *Tilia* flowers. a. Enzymatic activities of TiliaCS1 show that it can methylate at the N-3 and N-7 positions of the purine ring, but not N-1 position. b. Three possible caffeine biosynthetic pathways might have evolved in *Tilia* flower with another potential caffeine synthase. The hypothesized caffeine synthase should be able to methylate at N-1 position of the purine ring; therefore, it may complete the caffeine biosynthetic pathway in three different ways. The steps catalyzed by the characterized TiliaCS1 are labeled with orange and blue arrows, while the ones catalyzed by the potential caffeine synthase are labeled with gray arrows. Xanthine alkaloid abbreviations: 7X, 7-methylxanthine; Px, paraxanthine.

level of caffeine production in *Tilia* flowers comparing to the developing buds, *Tilia* may simply stop continuously producing caffeine from theophylline for its pollination needs by alternatively splicing Tilia9944 contig1 to contig2, as alternatively spliced mRNA sometimes results in inactive transcribed proteins [138].

Caffeine and theophylline were previously reported in the nectar extract of *Tilia* flowers, and caffeine is more abundant than theophylline in the nectar [48]. In addition, only caffeine existed in the honey stomach of bees that fed on the nectar as well as in the linden honey [48]. If
theophylline is a pathway precursor to caffeine, these findings may help to explain the lower level of caffeine than theophylline in *Tilia* open flowers: It is possible that caffeine produced from stamens and other organs of *Tilia* flowers was secreted into the nectar, and a good amount of caffeine had already been consumed by honeybees or other insects that fed on the nectar before I sampled the *Tilia* open flowers. Further sampling of “about to open” flower buds before their contact with honeybees and insects may help to confirm the role of theophylline. In addition, radioactive tracer feeding experiments [51, 53, 139] may be another way to investigate intermediates in the caffeine biosynthetic pathway in *Tilia*. 
CHAPTER V: EXPECT SIGNIFICANCE OF THE STUDY

As caffeine is an important economic compound worldwide, study of caffeine production in various plants and how these plants evolved the abilities to produce caffeine would benefit the caffeine production industry in terms of designing transgenic coffee/tea plants with high caffeine yields or creating “naturally decaffeinated” coffee/tea plants by silencing or mutating the caffeine biosynthesis genes to interrupt the normal caffeine biosynthetic pathways, or inventing new types of caffeinated drinks from *Citrus* or *Tilia* flowers or even non-caffeinated plants.

This study applied the paleomolecular biology approach (functional characterizations of ancestral caffeine synthases and re-evolution of modern-day enzymatic activities through forward mutagenesis) to investigate the molecular basis of protein functional diversification and evolution of metabolic pathway. This further proved the importance and usefulness of ancestral sequence reconstruction and forward mutagenesis in studying protein functional evolution and pathway assembly scenarios, and methods and results of this research will also shed light on studies of functional divergence and pathway evolution in other gene families.
REFERENCES

1. Bridges CB. The bar "gene" - a duplication. Science. 1936;83:210-211.
16. Zhao N. Comparative functional genomics of the SABATH family of methyltransferases in plants [PhD diss.]: University of Tennessee - Knoxville; 2008.


129. O'Donnell AJ. Evolutionary convergence of the caffeine biosynthetic pathway in chocolate followed duplication of a constrained ancestral enzyme [Master's Thesis]. Kalamazoo, MI: Western Michigan University; 2015.


APPENDIX A

Sequence alignment of ancestral and modern-day caffeine synthases
<table>
<thead>
<tr>
<th>160</th>
<th>200</th>
<th>220</th>
<th>240</th>
<th>260</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAAnCtallele1</td>
<td>FPR6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
<tr>
<td>RAAnCtallele2</td>
<td>FPP6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
<tr>
<td>TheobromaCtallele1</td>
<td>FPGA6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
<tr>
<td>TheobromaCtallele2</td>
<td>FPGA6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
<tr>
<td>CamelliaCtallele1</td>
<td>FPGA6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
<tr>
<td>CamelliaCtallele2</td>
<td>FPGA6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
<tr>
<td>PaulliniaCtallele1</td>
<td>FPGA6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
<tr>
<td>PaulliniaCtallele2</td>
<td>FPGA6LHIVDFQVFLGLQFVGIQFETEESL</td>
<td>$L_{NAGK1}$</td>
<td>$N_{GFP}$</td>
<td>$V_{M}<em>{WL}</em>{II}$</td>
</tr>
</tbody>
</table>

**References:**

1. PGA6LHIVDFQVFLGLQFVGIQFETEESL
2. $L_{NAGK1}$
3. $N_{GFP}$
4. $V_{M}_{WL}_{II}$
To ensure the robustness of ancestral sequence estimation, two alleles of RAAnCS and CoffeaAncXMT, which differ from each other by over 18 amino acids were functionally tested. Variants of other ancestors were mutated from the original version through site-directed mutagenesis, and the mutated sites are colored in green. Posterior probabilities of original and mutated sites are shown in Table 11. Amino acid sites in PaulliniaAncCS2 and CitrusAncXMT2 that were experimentally replaced to re-evolve modern-day caffeine synthases’ activities in the Paullinia and Citrus lineages, respectively, are highlighted in light blue.

Table 11. Posterior probabilities of original and mutated sites in alternative ancestral alleles

<table>
<thead>
<tr>
<th>Mutation made</th>
<th>Original amino acid</th>
<th>Posterior probability</th>
<th>Mutant amino acid</th>
<th>Posterior probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAAnCSXMT241N</td>
<td>H</td>
<td>0.584</td>
<td>N</td>
<td>0.844</td>
</tr>
<tr>
<td>CitrusAncXMTF66L</td>
<td>F</td>
<td>0.542</td>
<td>L</td>
<td>0.535</td>
</tr>
<tr>
<td>CitrusAncXMT1M221I</td>
<td>M</td>
<td>0.405</td>
<td>I</td>
<td>0.355</td>
</tr>
<tr>
<td>CitrusAncXMT2I1263M</td>
<td>I</td>
<td>0.52</td>
<td>M</td>
<td>0.405</td>
</tr>
<tr>
<td>CamelliaAncCSN231S</td>
<td>N</td>
<td>0.81</td>
<td>S</td>
<td>0.218</td>
</tr>
<tr>
<td>CamelliaAncCSN302K</td>
<td>N</td>
<td>0.941</td>
<td>K</td>
<td>0.357</td>
</tr>
<tr>
<td>PaulliniaAncCSF123Y</td>
<td>F</td>
<td>0.517</td>
<td>Y</td>
<td>0.532</td>
</tr>
<tr>
<td>PaulliniaAncCSV116M</td>
<td>V</td>
<td>0.622</td>
<td>M</td>
<td>0.971</td>
</tr>
</tbody>
</table>
APPENDIX B

Project approval certification
Western Michigan University
Institutional Biosafety Committee

Registration for Recombinant or Synthetic Nucleic Acid Molecules
Research – 2016

This form must be submitted for all research involving recombinant or synthetic DNA molecules. Renewal of approval is required annually.

In this form, Guidelines means: NIH Guidelines for Research Involving Recombinant or Synthetic Nucleic Acid Molecules (NIH Guidelines)

General Registration Information

Principal Investigator: Dr. Todd Barkman
Office Phone: 269-387-5610 Laboratory Phone: 269-387-2776 Department: BIOS
Electronic Mail Address: todd.barkman@wmich.edu
Office Address: Room 3437, Building: Wood Hall

Laboratory Where Research is to be Conducted:
Building: Haenicke Hall Room: 3456

Project Title: Ghosts of evolution past: resurrecting ancestral enzymes to understand the evolution of modern-day enzyme activities

Current status (check one):

- Initiated Date:
- Will be initiated Date:
- Continuing (no changes) 2015 IBC Project Number: 15
- Continuing (modifications) 2015 IBC Project Number: 15
  Please provide a cover letter stating modifications.
- Will not be initiated or will be discontinued
- Completed Date:

If Part of a Grant Proposal, List Agency/Agencies: NSF
WMU Proposal Tracking Number (if applicable):

Revised 10/15 WMU IBC
All other copies obsolete
III-E-3 was selected because the host system is E. coli B. Biosafety Level 1 was selected, because it is suitable for work involving agents of unknown or minimal potential hazard to laboratory personnel and the environment. The laboratory is separated from the general traffic patterns in the building. Work is generally conducted on open bench tops. Special containment equipment is not required or generally used. Laboratory personnel have specific training in the procedures conducted in the laboratory and are supervised by a scientist with general training in microbiology or a related science." (Appendix G - III - M). Standard microbiological practices are being used for the bacterial work and include disinfection of laboratory benches and materials used. The same methods will be used for accidental spills. Transgenic bacteria are autoclaved before disposal.
APPENDIX C

Approval of radioactive material usage
To Whom It May Concern,

This letter is provided to verify that Ruiqi Huang has completed the Western Michigan University's Radiation Safety Training course. Ruiqi has shown the required knowledge to maintain compliance with Western Michigan University's Radiation Safety Program, state regulations, and Federal regulations. He has demonstrated the ability to work safely around radioactive materials and/or radiation producing machines as it pertains to the ALARA (as low as reasonably achievable) radiation exposure practices and follow general laboratory safety practices.

Ruiqi had his initial training in July, 2011 with John G. Center. He scored 100 percent on his written exam. His next annual refresher training will be held in September, 2012.

With respect,

[Signature]

John G. Center, Jr.

Research Safety Officer
3922 Wood Hall
Western Michigan University
1903 West Michigan Ave
Kalamazoo, MI 49008-5410
(269) 387-5933