RECyT
Year 25 / Nº 40 / 2023 / 14–24
DOI: https://doi.org/10.36995/j.recyt.2023.40.002

First report of pseudogenes of SABATH enzymes from xanthines metabolic pathway in *Ilex paraguariensis*

Primer reporte de pseudogenes de enzimas SABATH de la vía metabólica de las xantinas en *Ilex paraguariensis*

Sebastián Maximiliano Rossi[1, *]; Cristian Alberto, Ferri[1]; Pedro Darío Zapata[1]

1- National University of Misiones. Faculty of Exact, Chemical and Natural Sciences. Institute of Biotechnology of Misiones "María Ebe Reca". Molecular Biotechnology Laboratory.
* E-mail: maximiliano001@gmail.com

Received: 30/08/2022; Accepted: 30/11/2022

Abstract

Ilex paraguariensis St. Hil. is a dioecious tree native to the subtropical forests of South America, whose leaves and small branches are processed to prepare the stimulating and popular infusion known as "mate". On the basis of the exploration, from the transcriptomic, genomic and phylogenetic perspective of the xanthine biosynthetic pathway, it was possible to discover pseudogenes through comparative analysis with other plants of agronomic importance. The transcripts of *I. paraguariensis* were analyzed using the elite material developed by INTA EAA-Cerro Azul de Misiones as reference genome and the data provided from the transcriptome sequenced by Debat et. al in 2014. A phylogenetic examination of enzymes from a large family of SABATH genes that catalyze the methylation of oxygen atoms of a wide variety of carboxylic acids was performed. The most recently evolved genes of the SABATH family are those corresponding to the Xanthine Methyltransferases (XMT) and Caffeine Synthase (CS) pathways. Investigating the different types of methyltransferases that yerba mate presents in the metabolic process of caffeine conversion, several SABATH enzymes were categorized, of which three corresponded to pseudogenes within the caffeine synthase group.

Keywords: SABATH; Pseudogenes; Methylxanthine, Gene Duplication; *Ilex paraguariensis*.

Resumen

Ilex paraguariensis St. Hil. es un árbol dioico originario de las selvas subtropicales de América del Sur, cuyas hojas y pequeñas ramas son procesadas para preparar la infusión estimulante y popular conocida como "mate". A partir de la exploración, desde el punto de vista transcriptómico, genómico y filogenético de la vía biosintética de las xantinas, pudo realizarse la detección de pseudogenes mediante el análisis comparativo con otras plantas de importancia agronómica. Se analizaron los transcriptos de *I. paraguariensis* utilizando como genoma de referencia el material elite desarrollado por el INTA EAA-Cerro Azul de Misiones y los datos aportados por el transcriptoma secuenciado por Debat et. al en 2014. Se realizó un examen filogenético de las enzimas de una gran familia de genes SABATH que catalizan la metilación de átomos de oxígeno de una amplia diversidad de ácidos carboxílicos. Los genes más recientemente evolucionados de la familia SABATH son los que corresponden a la vía Metiltransferasas de Xantina (XMT) y Cafeína Sintasa (CS). Investigando los diferentes tipos de metiltransferasas que presenta la yerba mate en el proceso metabólico de la conversión a cafeína, se categorizaron varias enzimas SABATH de los cuales tres correspondían a pseudogenes dentro del grupo de las cafeínas sintasa.

Palabras claves: SABATH; Pseudogenes; Metilxantina; Duplicación Génica; *Ilex paraguariensis*.

Introduction

Ilex paraguariensis St. Hil. is a dioecious tree of the Aquifoliales order, Aquifoliaceae family of the Euasterids II group, native to the subtropical forests of South America. Its leaves and small dry processed branches are used to prepare the popular infusion, stimulating and with a particular flavor, known as "mate", hence its colloquial name in Latin America is "yerba mate" or "erva mate", an expression derived from the conjunction "ka`a

mati", from the Guarani "ka`a" (herb) and the Quechua "mati" (pumpkin), the latter alluding to the container of consumption. Yerba mate is also widely used to prepare other hot infusions such as "mate cocido", cold drinks such as "tereré" and more recently, as an additive in ice creams, sweets and energy drinks, as well as in dyes, cosmetics and spa ingredients. (Bracesco *et al*. 2011).

Yerba mate crops are an important agribusiness in Misiones, Argentina. The characteristics and agroecological conditions of yerba mate crops are found only in Brazil, Paraguay and Argentina. In Argentina, the world's leading producer, plantations are limited to the provinces of Misiones and Corrientes. The aerial photogrammetric survey of plantations commissioned by the INYM (National Institute of Yerba Mate), a national entity that regulates all aspects related to the cultivation and marketing of the product in 2016, determined that the cultivation of yerba mate in Argentina covers an area of 165,200 hectares, of which 144,014 hectares are located in the province of Misiones (87%), and the rest are in the province of Corrientes. In Misiones, cultivation occupies 50% of its agricultural area and yerba mate activity, in this province, occupies 22% of the workforce in the primary sector and 15% in the manufacturing sector, and constitutes its main industrial activity with the 25% of the total (Canitrot *et al.* 2011; Dolce 2012). In addition, Misiones concentrates most of the industrial stage with the largest number of dryers (94%), collectors (93%) and mills (85%). Particularly, since the creation of the National Institute of Yerba Mate (INYM) in 2002, a significant increase in the relative participation of producers and dryers in the final price of yerba has been observed (Gortari 2007; Rau *et al.* 2009).

The knowledge in genomics of *I. paraguariensis* has been increasing during the last years, as well as the sequencing of the transcriptome that represents the first evidence of the genome of this species (Debat *et al*., 2014), from which the presence of at least 32,000 genes and 12,000 variants is deduced. This work explored the vast collection of *I. paraguariensis* transcripts using the Pg538 line as reference genomic material, an elite material developed by INTA EAA-Cerro Azul de Misiones. In this work, transcripts belonging to more than 100 metabolic pathways were identified, related to osmotic stress, drought, salinity, cold stress, senescence, early flowering, and sexual determination. The large amount of information obtained in this study served as a reference framework and led to the start of the Yerba Mate Genome Sequencing Project (Pro.Mate.A.R.) in 2015 by the National University of Misiones, the University of Buenos Aires and other Argentine institutions such as the Secretariat of University Policies, National Institute of Yerba Mate, CONICET, Ministry of Science, Technology and Productive Innovation; that aim to generate a complete physical map at the sequence level of the entire genome of *I. paraguariensis*. This constitutes the starting point to discover, characterize and locate genes associated with traits of biological, agronomic and economic importance; develop efficient molecular markers, and delve into their biochemistry, evaluating the impact of the allelic variants that occur on the metabolomics of the species.

Xanthines are substances that belong to a chemical group of purine bases. From a medical and pharmacological point of view, there are three important xanthines: caffeine, theobromine and theophylline, which are the three methylated xanthines; that is the reason why they are also known as methylxanthines. They are considered alkaloids, since they are physiologically active substances, contain nitrogen and are found in higher plants. Methylxanthines and methyluric acids are plant secondary metabolites derived from purine nucleotides (Ashihara and Crozier 1999). The best characterized methylxanthines are caffeine (1,3,7-trimethylxanthine) and theobromine (3,7-dimethylxanthine) which appear in tea, coffee, cocoa, guarana, citrus, yerba mate and in a number of non-alcoholic beverages. origin in plants.

Caffeine was first isolated from tea and coffee in the early 1820s, but the main biosynthetic and catabolic pathway for caffeine was established only recently, when highly purified caffeine could be obtained from tea leaves and the genes encoding the enzyme could be cloned. (Kato *et al*. 1999; Kato *et al*. 2000). Methylxanthines have been found in about 100 species distributed in 13 orders of the plant kingdom (Ashihara and Suzuki 2004; Ashihara and Crozier 1999). Compared to other plant alkaloids such as nicotine, morphine and strychnine, purine alkaloids are widely distributed in the plant kingdom, although accumulation in high concentrations is restricted to a limited number of species including *Coffea arabica* L. (coffee), *Camellia sinensis* (L.) Kuntze (tea), *Theobroma cacao* L. (cocoa), *Paullinia cupana* Kunth (guarana) and *I. paraguariensis*. All caffeine-containing plants, except *Scilla maritima* L., belong to the Dicotyledonae. In some species, the main methylxanthine is theobromine or methyluric acids, including theacrine (1,3,7,9-tetramethyluric acid) instead of caffeine (Ashihara and Crozier 1999).

Caffeine biosynthesis evolved independently in *Coffea* and *Camellia*, involving three methylation reactions. In *Coffea*, three XMT-type enzymes (xanthine methyltransferases) of the SABATH family are used to catalyze the methylation steps of the pathway. In *Camellia*, a convergently evolved paralogous pathway, CS-type caffeine synthase enzymes, is used. Numerous studies conducted over 30 years indicate that there are several biosynthetic pathways. This caffeine biosynthesis pathway has evolved independently in *Coffea* and *Camellia*, involving three methylation reactions to sequentially convert xanthosine to 7-methylxanthosine (1) to theobromine (2) to caffeine (3) (Huang *et. al*, 2016). Most

SABATH enzymes catalyze the oxygen atom methylation of a wide variety of carboxylic acids such as anthranilic, benzoic, gibberillic, jasmonic, loganic, salicylic, and indole-3-acetic acids for flowering, defense, and hormonal modulation. XMT- and CS-mediated methylation of nitrogen atoms of xanthine alkaloids is probably a recently evolved activity (Huang *et al*. 2016). This multigene family 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, *Arabidopsis thaliana* (L.) Heynh. has 24 SABATH members, in rice (*Oryza sativa* L.) 41 members, 33 in poplar (*Populus trichocarpa* Torr. & A.Gray ex Hook.) and only 4 SABATH members in moss (*Physcomitrella patens* Bruch & W.P.Schimper). Many members of this family have been functionally characterized and participate in various secondary metabolism pathways in plants, producing different types of compounds. Three SABATH members having available protein crystal structures were characterized. These crystal structures are useful for homology modeling to identify important amino acid residues and to investigate enzyme interactions with different substrates. The SABATH family encodes a group of methyltransferases (MTs) that can transfer a methyl group (-CH3) from the methyl donor, S-adenosyl-L-methionine (SAM), to either the carboxyl group, the sulfur group, or the ring of nitrogen from a given substrate, forming *S*-adenosyl-L homocysteine (SAH) and *O*-methylated ester, such as methyl benzoate and methyl salicylate, *S*-methylated ester, such as methyl thiolbenozate, or *N*-methylated compounds such as caffeine. Members of this family that are capable of methylating the carboxyl group of a given substrate include loganic acid methyltransferase (LAMT), salicylic acid methyltransferase (SAMT), benzoic acid methyltransferase (BAMT), jasmonic acid methyltransferase (JMT), indol- 3-acetic methyltransferase (IAMT), farnesoic acid methyltransferase (FAMT), gibberellic acid methyltransferase (GAMT), cinnamate and p-coumarate methyltransferase (CCMT), anthranilic acid methyltransferase (AAMT), nicotinic acid methyltransferase (NAMT), and p-methoxybenzoic acid methyltransferase (MBMT). The SABATH member that methylates the sulfur group of a given substrate is the recently characterized moss thiol methyltransferase. Members of SABATH that methylate the nitrogen ring of a given substrate include enzymes involved in caffeine biosynthesis, such as caffeine synthase (TCS1) characterized in *Camellia*, theobromine synthase (BTS1) characterized in *Theobroma*, 7-methylxanthine methyltransferase (MXMT), xanthosine methyltransferase (XMT), and 3,7-dimethylxanthine methyltransferase (DXMT) found in *Coffea*, as well as the enzyme responsible for the biosynthesis of trigonelline. The paralogous XMT and CS lineage genes in the SABATH family have independently evolved convergently, because genes in the two clades are involved in caffeine biosynthesis.

The importance of gene duplication for the development of metabolic pathways was first discussed by Lewis (1951) and then by Ohno (1970) and has recently been confirmed by comparative analysis of complete genome sequences of archaea, bacteria and eukarya (Brilli and Fani 2004; Fani *et al*. 1994, 1995, 2007; Alifano *et al*. 1996; Fondi *et al*. 2009). Genes that descend from a common ancestor through a duplication event are called paralogs, and they can undergo successive duplications that lead to a family of paralogous genes, catalyzing different reactions, although with certain similarities. The metabolic pathways for the production of xanthine alkaloids, such as caffeine, in angiosperms have had a convergent evolution, through two biochemical pathways previously described in cocoa, citrus and guarana plants that use caffeine synthase (CS) or xanthine methyltransferase-type enzymes (XMT), and is associated with defense and pollination mechanisms of these plants. The work of Huang *et al*. 2016, through the resurrection of ancestral sequences, reveals that the convergence of these metabolic pathways required very few mutations to acquire modern enzymatic characteristics, favoring the evolution of a complete pathway. The three-step caffeine biosynthetic pathways characterized in *Camellia*, *Citrus*, *Paullinia*, and *Coffea* are also catalyzed by duplicated CS-type or XMT paralogous enzymes. To discover how these three methylation steps were evolutionarily assembled in each species, it is necessary to search for answers in the evolutionary history of these plants. Phylogenetic relationships of 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 of other species. This indicates that gene duplications have occurred within each species after speciation events and such duplicated sequences are expected to be conserved only if they confer a selective advantage. When compared between different species, the caffeine synthases in *Camellia* and *Paullinia* are more closely related (encoded by orthologous genes in the CS clade), while the caffeine synthases in *Citrus* and *Coffea* are more closely related (encoded 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 *Camellia* or *Coffea*. Evidence suggests that both the CS and the XMT enzymes may have convergently evolved to produce caffeine in these species and the metabolic pathway mediated by SABATH enzymes (Salicylic Acid, Benzoic Acid, Theobromine Methyltransferase) appears to have evolved at least 5 times in different orders during the evolutionary history of angiosperms (Huang *et al*, 2016).

The present work collects the genomic data reported to date and analyzes them bioinformatically to verify the presence of pseudogenes of the SABATH family in *I. paraguariensis*.

Materials and Methods

Bioinformatics

Mesquite software (https://www.mesquiteproject.org/) was used for the construction of phylogenetic trees and evolutionary analyses, and for genomic studies Sequencher DNA Sequence Analysis Software (http://www.genecodes.com/sequencher-features) according to the protocols and manuals downloaded from each manufacturer.

To obtain the data used in the construction of the phylogenetic tree, gene sequences of the metabolic pathway of caffeine in species of agronomic interest were used with the data provided by the Barkman Laboratory of Western Michigan University in the United States. Among the best known species used in this comparative study, we can highlight *C. arabica*, *T. cacao*, *C. sinensis*, *Citrus sinensis* (L.) Osbeck, *P. cupana* among other species of the rosids and asterids groups of the eudicots.

For the detection of SABATH sequences, the caffeine synthase sequences of *C. sinensis* TCS1 and TCS2 were used, consulting the NCBI database using the BLAST tool with transcriptome sequences against genome sequences. Genome sequences from *I. paraguariensis* were obtained by Turjanski Lab from University of Buenos Aires (UBA).

Phylogenetic analysis

Amino acid sequences of all members of the enzymatically characterized SABATH gene family and several complete land plant genomes were obtained from GenBank and the PlantTribes database. In addition, a limited sample of XMT and CS sequences were obtained from the oneKP database (www.onekp.com) to provide more detailed branching relationships of recently evolved caffeine biosynthetic pathway enzymes from *Theobroma* (Malvales), *Camellia* (Ericales), *Coffea* (Gentianales), as well as *Paullinia* and *Citrus* (Sapindales). Sequences were aligned using MAFFT version 7, using the automatic search strategy to maximize accuracy and speed. PhyML was used to generate a maximum similarity phylogenetic estimate for members of the SABATH family. We assume the JTT model for amino acid substitution with one parameter invariant and gamma for the rate of heterogeneity between sites. Bootstrap support values were obtained from 100 replicates. The graphics were made by Mesquite Software (V.3.61), as mentioned in the bioinformatic study.

Construction of the phylogenetic tree of sabath enzymes of species of productive interest

For *Citrus* and *Theobroma*, all the gene sequences of the SABATH family were obtained from the data of their respective genomic sequences obtained from the NCBI (National Center for Biotechnology Institute) and were used to perform BLAST analyzes from the EST database of the GenBank. Only complete sequences of predicted SABATH genes encoding more than 300 amino acids were included in the study. All ESTs were assembled according to reference sequences using Sequencher software (Gene Codes), and the numbers of positive ESTs were counted. In *Theobroma* only two genes were represented by >5 ESTs related to all other 23 SABATH family members: TcCs1 (Thecc1EG042578) (30 ESTs), TcCS2 (Thecc1EG042587 and Thecc1EG04290) (>120 ESTs). Thecc1EG042587 and Thecc1EG04290 were counted together because they encode identical ORFs.

TcBTS used from the Barkman laboratory data, had been isolated from leaves and reported to have activity with the substrate 7-methylxanthine (7X). This gene is represented by <5 ESTs in fruits and its *Km* for this substrate is high. Although it may participate in the production of xanthine alkaloids in leaves, any role in fruits should be relatively minor, given the kinetics of the enzyme, particularly those related to TcCS1 and TcCS2. In *Citrus*, 3 SABATH genes were used, which were represented by more than 5 ESTs in flowers, with accumulation of caffeine and theophylline. One of the most abundant ESTs is represented by SAMT, which has a preference for methylation of salicylic, benzoic and anthranilic acids, which correlates with the presence of methylanthranilate in *Citrus* flowers.

CisXMT1 (1g044727m) was represented by >30 ESTs, while CisXMT2 (1g047625m) by 7 ESTs. Since no genomic sequences were found in *Camellia* and *Paullinia*, a different strategy was used for the EST sets. Complete sequences of the SABATH family were selected from an Asterids, *Mimulus* to query for *Camellia* ESTs and as Rosids to *Citrus* to query for *Paullinia* ESTs. BLAST from the GenBank EST database was then used and matches were assembled with Sequencher. Phylogenetic analyzes were performed to verify the orthology of each putative EST, relative to the queried sequence. In *Camellia*, two genes represented by EST have been reported in leaves and shoots, where caffeine accumulates. These sequences correspond to the previously studied CS: TCS1 (AB031280) (9 ESTs) and TCS2 (AB031281) (12 ESTs). In *Paullinia*, ESTs were assembled and hypothesized to represent the same gene if they possessed at least 98% identity with a window of more than 100 bp of coding sequence. Of the more than 105 ESTs, five contigs were assembled. These were the only SABATH sequences

represented by the ESTs. One of the contigs contained 18 ESTs and putatively represents the complete sequence "CS grn006" (which was referred to by the Barkman Lab as PcCS1). The contig with the highest coverage is referred to as PCCS2 (30 ESTs). Another contig, PCCS3, was represented with 15 ESTs. Two other contigs, PcCS4 (27 ESTs) and PcCS5 (15 ESTs), are sequence variants of PcCS1 and PcCS2, respectively with the same enzymatic activity *in vitro*.

Results and discussion

PHYLOGENETIC TREE OF SABATH ENZYMES

In *I. paraguariensis*, highly conserved sequences of the SABATH family were found and analyzed, some sequences found by BLAST of the yerba mate transcriptome with the complete sequence of the *C. sinensis* TCS1 gene. Subsequently, using the data provided by the "Pro.Ma.Te.Ar Project" consortium, more sequences of the SABATH family were found, of which three corresponded to incomplete amino acid sequences. The representatives of the SABATH family found in yerba mate were the following: CS (Caffeine Synthase), MT (Methyltransferase), JMT (Jasmonic acid carboxyl Methyltransferase), SAMT (Salicylic Acid carboxyl Methyltransferase, FAMT (Farnesoic Acid carboxyl Methyltransferase), BAMT (Benzoic Acid carboxyl Methyltransferase) and IAMT (indol- 3-acetic methyltransferase) (Fig. 1 A-B).

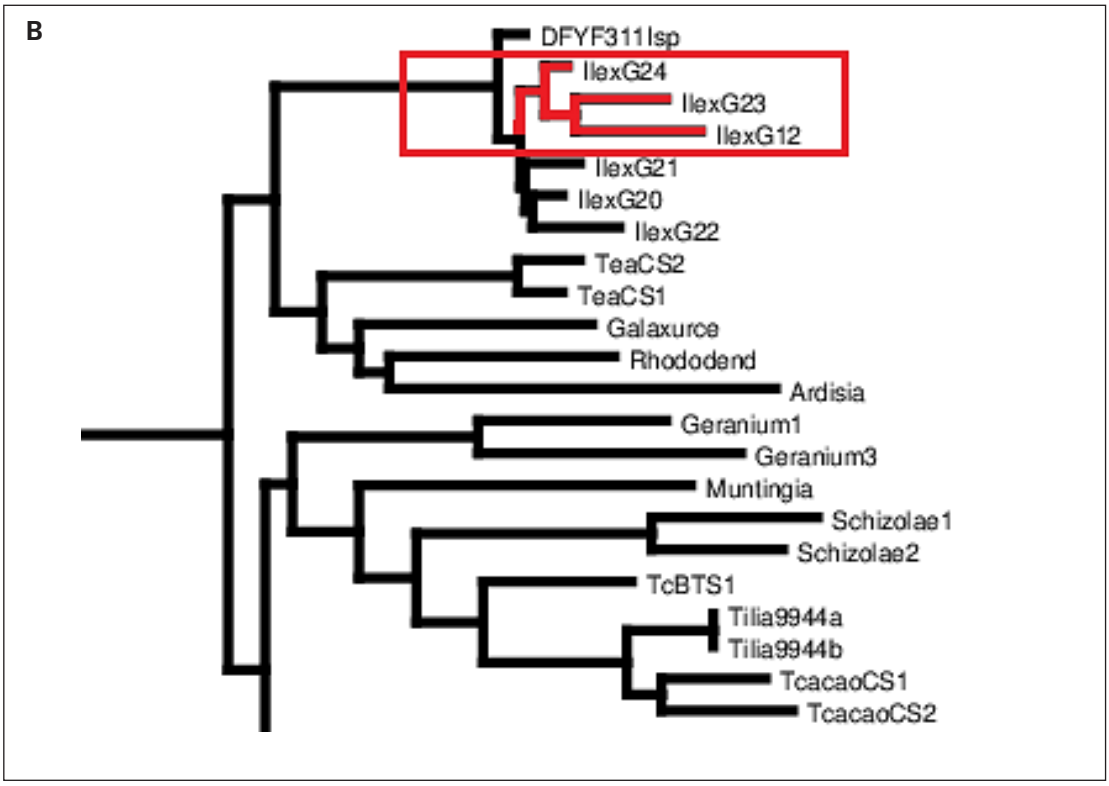

Fig. 1: A) Phylogenetic tree showing the distribution of SABATH genes in yerba mate compared with plants of productive interest that produce different xanthine alkaloids. The blue box shows CS SABATH enzymes species. **B)** Characterized CS (Caffeine Synthase) genes of different species for comparison. The red box shows the sequences of YM IlexG12, IlexG23 and IlexG24 that presented an incomplete amino acid sequence.

Gene duplication in plants is a very common phenomenon for a lot of metabolic pathways throughout angiosperm evolution. Many essential genes that participate in metabolic pathways that confer great evolutionary advantages in plants usually evolve from tandem duplications of genomic regions as operons or from duplications of the entire genome. Several duplicate genes may have acquired metabolic innovations by becoming paralogs with similar functions.

Pseudogenes characterized in the sabath family

The 3 sequences of the same clade, IlexG12, IlexG23 and IlexG24 were subjected to bioinformatic analysis. BlastP of IlexG12, IlexG23 and IlexG24 was performed against the proteome of Fay *et al*. (2018) and Aguilera *et al*. (2018). Three hits of complete proteins were obtained, paralogs of those obtained from the genome. They were annotated in PFAM and verified to be correct methyltransferases. After the alignment, it was possible to observe that the C-terminal ends were missing as well as the internal sequences. To re-obtain each protein, the genomic contigs were taken, translated into the 6 reading frames, and BlastP was performed. Proteins slightly different from the original ones (V2) were obtained: ILEXPARA_10393V2 and ILEXPARA_33366V2 are non-functional proteins according to the missing segments. ILEXPARA_10668V2 is also non-functional because it lacks a start codon and has an internal stop codon. If this area is not translated and the start codon of IlexG23 is accepted, the protein would be functional according to its structure. To finish corroborating whether the proteins derived from the three genomic contigs are expressed, a back mapping was performed with the leaf reads (where caffeine matters) by Fay *et al*. (2018) and Debat *et al*. (2014). It was shown that these contigs do not express the proteins predicted from the genome V1 (original) and V2 (predicted), since only match similar reads corresponding to the functional paralogous genes, and it was not observed that they match in all the putative exons, being able to affirm that these sequences correspond to pseudogenes (Fig. 2).

Conclusions

The bioinformatic analysis concluded that three of the total SABATH genes identified, presented loss of internal and C-terminal sequences, for which they were not active genes, being classified as pseudogenes or truncated genes. They may belong to tandemly duplicated genes for an evolutionarily conserved metabolic pathway but have lost catalytic activity.

These genes of the large SABATH family evolved in flowering plants by gene duplications for the production of xanthine alkaloids in several lineages, giving rise to

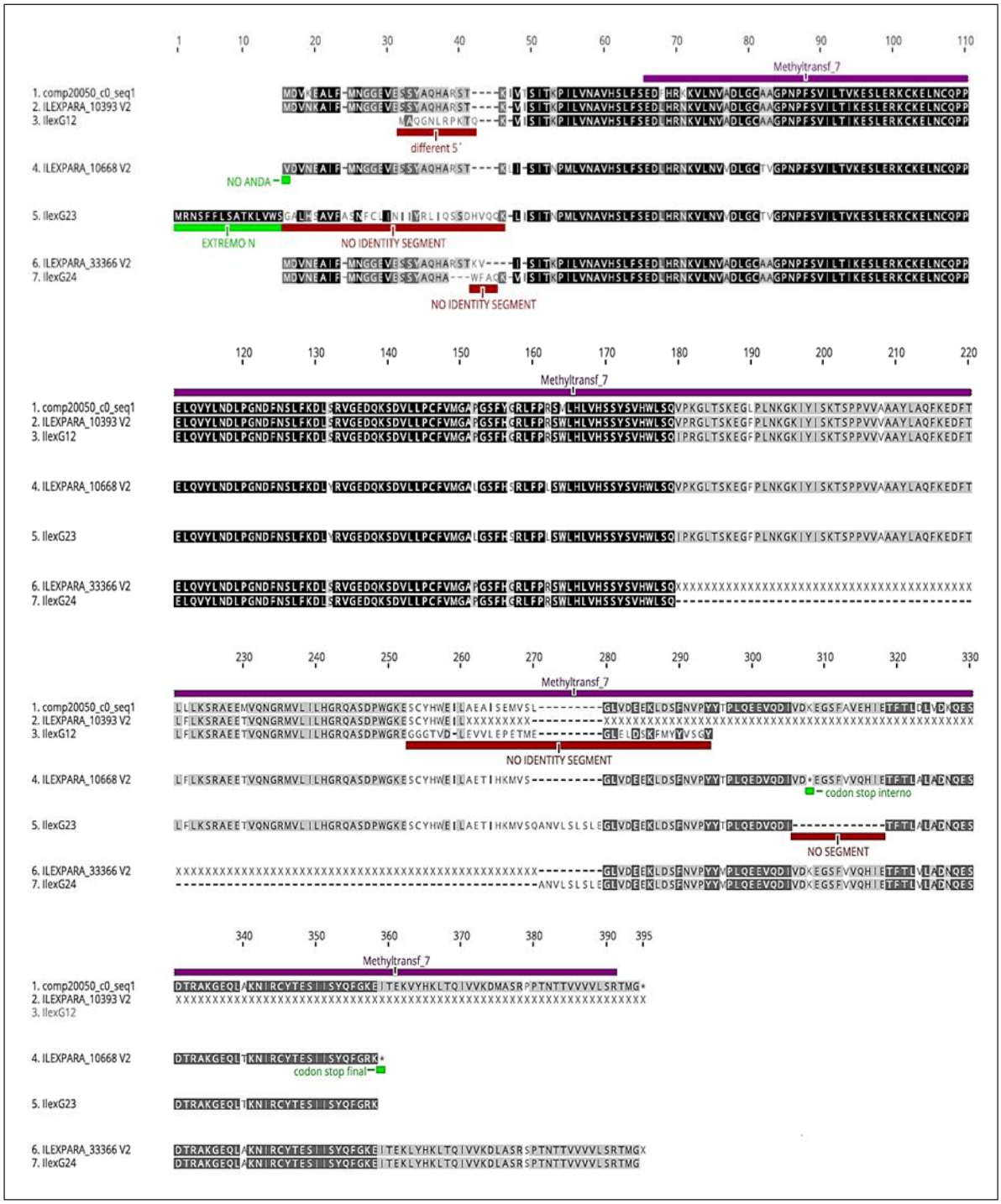

FIG. 2: Bioinformatic analysis of the genomic sequences IlexG12, IlexG23 and IlexG24 candidates for caffeine synthases according to the phylogenetic tree made. Comparisons of the original sequences (V1) with the predicted expressed ones (V2) showing missing or non-homologous sequence segments defined as pseudogenes are shown.

the most modern production pathways, the XMT and CS pathways as it has been shown in this work. The genes belonging to the CS and XMT families are genes conserved in higher eudicots and a vast collection of tandemly duplicated genes forming genomic clusters or dispersed chromosome as in *C. sinensis*, *Citrus* and *C. canephora* has been explored. This is the first report of the presence of highly conserved SABATH family genes in the evolution of the xanthine metabolic pathway in *I. paraguariensis.*

Acknowledgments

To Dr.Todd Barkman, from the Biological Sciences Laboratory of the Western Michigan University for all the SABATH enzyme sequences used in this study to compare with yerba mate genes.

To Dr. Adrián Turjanski and Federico Vignale, from the Structural Bioinformatics Laboratory, Department of Biological Chemistry of the University of Buenos Aires, for providing the provisional data of the yerba mate

genome with which all their SABATH genes of this study were performed.

Supplementary annex

Table 1. Sequence names and accession numbers used to phylogenetically classify *I. paraguariensis* genome SABATH sequences.

Name	Accession number[1]
Acer negundo CS-like	VFFP_scaffold_2043613
Antirrhinum majus BAMT	AF198492.1
Antirrhinum majus SAMT	AF515284
Arabidopsis lyrata BSMT	AY224596.1
Arabidopsis thaliana BSMT	NM_111981.5
Arabidopsis thaliana FAMT	NM_114355.3
Arabidopsis thaliana GAMT1	NM_118775.3
Arabidopsis thaliana GAMT2	NM_125013.3
Arabidopsis thaliana IAMT	NM_124907.5
Arabidopsis thaliana JMT	AY008434.1
Ardisia humilis CS-like	ODDO_scaffold_2102454
Ardisia revoluta FAMT-like	DAAD_scaffold_2041891
Atropa beladonna SAMT	AB049752
Boswellia sacra CS-like	FCCA_scaffold_2008319
Camellia sinensis CS1	AB031280
Camellia sinensis CS2	AB031281
Catharanthus roseus LAMT	KF415116.1
Citrus sinensis CS-like	XM_024186548.1
Citrus sinensis IAMT-like	XM_006474928.3
Citrus sinensis SAMT	XM_006466773.3
Citrus sinensis XMT-like_A	XM_006469416.3
Citrus sinensis XMT-like_B	XM_006469386.3
Citrus sinensis XMT1	XM_015527394.2
Citrus sinensis XMT2	XM_006469385.2
Clarkia breweri SAMT	AF133053
Coffea arabica DXMT1	AB084125.1
Coffea arabica MXMT1	XM_027230304.1
Coffea arabica MXMT2	AB084126.1
Coffea arabica XMT1	XM_027230970.1
Datura wrightii SAMT	EF472972.1
Fragaria vesca JMT1	XM_004291804.2
Fragaria vesca JMT2	XM_004291805.2
Galax urceolata CS-like	ADHK_scaffold_2052625
Geranium carolinianum CS-like	VKGP_scaffold_2120585
Geranium maculatum CS-like	YGCX_scaffold_2142952
Hoya carnosa SAMT	AJ863118.1
Ilex paraguariensis G1-G29	From Turjanski Lab
Kirkia wilmsii CS-like	BCAA_scaffold_2069913
Litchi chinensis CS-like	GCAD01028884.1
Mangifera indica CS-like	GBCV01000539.1
Melia azedarachta CS-like	VCCF_scaffold_2049817
Muntingia calabura CS-like	ATFX_scaffold_2034890
Nicotiana alata BSMT2	GU014483.1
Nicotiana alata SAMT	GU014482.1
Nicotiana gossei NAMT	GU169286.1
Nicotiana suaveolens BSMT1	AJ628349.1
Nicotiana suaveolens BSMT2	GU014480.1
Nicotiana suaveolens SAMT	GU014479.1
Nicotiana sylvestris BSMT2	GU014486.1
Nicotiana sylvestris SAMT	NM_001302610.1
Ocimum basilicum CCMT	EU033968.1
Oyrza sativa BSMT	XM_015769493.2
Oyrza sativa IAMT	EU375746.1
Paullinia cupana CS	BK008796.1
Paullinia cupana CS-like	EC774687.1
Paullinia cupana CS1	EC766748.1
Paullinia cupana CS2	EC772993.1
Populus trichocarpa IAMT	XM_002298807.3
Populus trichocarpa JMT	XM_002307635.3
Rhododendron delavayi CS-like	GFCU01041357.1
Schizolaena sp. CS-like_1	WMUK_scaffold_2093724
Schizolaena sp. CS-like_2	WMUK_scaffold_2020889
Theobroma cacao BTS	XM_007009000.2
Theobroma cacao CS1	XM_018129798.1
Theobroma cacao CS2	XM_007009008.2
Toxicodendron radicans CS-like	YUOM_scaffold_2035237
Zea mays AAMT1	HM242244.1

1- All accession numbers beginning with four letter codes are OneKP sequences from China National GeneBank while all others are from GenBank

References

1. **Ashihara, H.; Crozier, A. (1999).** *Biosynthesis and metabolism of caffeine and related purine alkaloids in plants.* Adv Bot Res 30:117–205.
2. **Ashihara, H.; Suzuki, T. (2004).** *Distribution and biosynthesis of caffeine in plants.* Front Biosci 9:1864–1876.
3. **Acevedo, R.M.; Maiale, S.J., Pessino, S.C.; Bottini, R.; Ruiz, O.A.; Sansberro, P.A. (2013).** *A succinate dehydrogenase flavoprotein subunit-like transcript is upregulated in Ilex paraguariensis leaves in response to water deficit and abscisic acid.* Plant Physiology and Biochemistry 65: 48-54.
4. **Acevedo, R. M.; Avico, E. H.; González, S.; Salvador, A. R.; Rivarola, M.; Paniego, N.; Nunes-Nesi, A.; Ruiz, O. A., & Sansberro, P. A. (2019).** *Transcript and metabolic adjustments triggered by drought in Ilex paraguariensis leaves.* Planta, 250(2), 445–462. https://doi.org/10.1007/s00425-019-03178-3
5. **Aguilera, P.M.; Grabiele, M.; DebatH, J.; Bubillo, R.E.; Martí, D.A. (2015).** *The 18S–25S ribosomal RNA unit of yerba mate (Ilex paraguariensis A. St.-Hil.).* Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology 1-9.
6. **Aguilera, Patricia; Debat, Humberto & Grabiele, Mauro (2018).** *Dataset of the first transcriptome assembly of the tree crop "yerba mate" (Ilex paraguariensis) and systematic characterization of protein coding genes.* Data in Brief. 17. 10.1016/j.dib.2018.02.015.

7. **Alifano, P., Fani, R., Liò, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S., & Bruni, C. B. (1996).** *Histidine biosynthetic pathway and genes: structure, regulation, and evolution.* Microbiological reviews, 60(1), 44–69. https://doi.org/10.1128/mr.60.1.44-69.1996
8. **Badouin, H. et al.** *The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution.* Nature 546, 148–152 (2017).
9. **Bao, W., Kojima, K. K. & Kohany, O.** *Repbase Update, a database of repetitive elements in eukaryotic genomes.* Mobile DNA vol. 6 (2015).
10. **Benson, G.** *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Research vol. 27 573–580 (1999).
11. **Bergottini, V.M.; Filippidou, S.; Junier, T.; Johnson, S.; Chain, P.S.; Otegui, M.B.; Zapata, P.D.; Junier, P.** *Genome Sequence of Kosakoniaradicincitans Strain YD4, a Plant Growth-Promoting Rhizobacterium Isolated from Yerba Mate (Ilex paraguariensis St. Hill)* Genome Announcements 2015, 3(2), e00239-15. DOI:10.1128/genomeA.00239-15
12. **Blanco, E. & Abril, J. F.** *Computational gene annotation in new genome assemblies using GeneID.* Methods Mol. Biol. 537, 243–261 (2009).
13. **Bracesco, N.; Sanchez, A. G.; Contreras, V.; Menini, T.; &Gugliucci, A. (2011).** *Recent advances on Ilex paraguariensis research: minireview.* Journal of ethnopharmacology, 136(3), 378-384
14. **Brilli, M. & Fani, R. (2004).** *Molecular evolution of hisB genes.* Journal of molecular evolution, 58(2), 225–237. https://doi.org/10.1007/s00239-003-2547-x
15. **Brilli, M. & Fani, R. (2004).** *The origin and evolution of eucaryal HIS7 genes: from metabolon to bifunctional proteins? Gene, 339, 149–160.* https://doi.org/10.1016/j.gene.2004.06.033
16. **Canitrot, L.; Grosso, M.J.; Méndez, A. (2011).** *Complejo yerbatero. Serie "Producción regional porcomplejos productivos".* Ministerio de Economía y Finanzas Públicas, Argentina. http://www.mecon.gov.ar/peconomica/docs/Complejo_yerbatero.pdf.
17. **Cardozo Junior, E. L. & Morand, C. (2016).** *Interest of mate (Ilex paraguariensis A. St.-Hil.) as a new natural functional food to preserve human cardiovascular health – A review. J. Funct*. Foods 21, 440–454
18. **Day, S.; Montagnini, F.; Eibl, B. (2011).** *Effects of native trees in agroforestry systems on the soils and yerba mate in Misiones, Argentina.* In: Montagnini F., Francesconi W. and Rossi E. (eds.). Agroforestry as a tool for landscape restoration: challenges and opportunities for success. Nova Science Publishers, New York. 201pp.
19. **Debat, H.J.; Grabiele, M.; Aguilera, P.M.; Bubillo, R.E.; Otegui, M.B.; Ducasse, D.A.; Zapata, P.D.; Marti, D.A. (2014).** *Exploring the Genes of Yerba Mate (Ilex paraguariensis A. St.-Hil.) by NGS and De Novo Transcriptome Assembly Zhang J-S (ed).* PLoS One 9:e109835.
20. **Eibl, B.; Fernández, R.; Kozarik, J.C.; Lupi, A.; Montagnini, F.; Nozzi, D. (2000).** *Agroforestry systems with Ilex paraguariensis (American holly or yerba maté) and native timber trees on small farms in Misiones, Argentina.* Agroforestry Systems, 48:1-8.
21. **El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A.; Sonnhammer, E.; Hirsh, L.; Paladin, L.; Piovesan, D.; Tosatto, S. & Finn, R. D. (2019).** *The Pfam protein families database in 2019.* Nucleic acids research, 47(D1), D427–D432. https://doi.org/10.1093/nar/gky995
22. **Emanuelsson, O.; Nielsen, H.; Brunak, S. & von Heijne, G.** Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300, 1005–1016 (2000).
23. **Errecaborde, N. (1973).** *Abonos en Yerba Mate. Oficina de publicaciones de INTA, Estación experimental agropecuaria Cerro Azul, Misiones.* Informe técnico n19.
24. **Fani, R.; Liò, P.; Chiarelli, I. & Bazzicalupo, M. (1994).** *The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and hisF genes.* Journal of molecular evolution, 38(5), 489–495. https://doi.org/10.1007/BF00178849
25. **Fani, R.; Liò, P. & Lazcano, A. (1995).** *Molecular evolution of the histidine biosynthetic pathway.* Journal of molecular evolution, 41(6), 760–774. https://doi.org/10.1007/BF00173156
26. **Fani, R. (2004).** Gene duplication and gene loading. In: Microbial evolution: gene establishment, survival, and exchange. Washington, DC: ASM.
27. **Fani, R.; Brilli, M.; Fondi, M. & Lió, P. (2007).** *The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case.* BMC evolutionary biology, 7 Suppl 2(Suppl 2), S4. https://doi.org/10.1186/1471-2148-7-S2-S4
28. **Fani, R. & Fondi, M. (2009).** *Origin and evolution of metabolic pathways.* Physics of Life Reviews, 6(1), 23-52.
29. **Fondi, M.; Emiliani, G. & Fani, R. (2009).** *Origin and evolution of operons and metabolic pathways.* Research in microbiology, 160(7), 502-512.
30. **Fay, J. V.; Watkins, C. J.; Shrestha, R. K.; Litwiñiuk, S. L.; Talavera Stefani, L. N.; Rojas, C. A.; Argüelles, C. F.; Ferreras, J. A.; Caccamo, M. & Miretti, M. M. (2018).** *Yerba mate (Ilex paraguariensis, A. St.-Hil.) de novo transcriptome assembly based on tissue specific genomic expression profiles.* BMC genomics, 19(1), 891. https://doi.org/10.1186/s12864-018-5240-6
31. **Gauer, L. & Cavalli-Molina, S. (2000).** *Genetic variation in natural populations of maté (Ilex paraguariensis A. St.-Hil., Aquifoliaceae) using RAPD markers.*Heredity, 84(6), 647-656.
32. **Glick, B.R. (2012)** *Plant Growth-Promoting Bacteria: Mechanisms and Applications.* Scientifica (Cairo) 2012:1-15.
33. **Gortari, J. (2007).** *El Instituto Nacional de la yerba mate*

(INYM) como dispositivo político de economía social: mediación intrasectorial en la distribución del ingreso, empoderamiento del sector productivo y desarrollo local en la región yerbatera. En: Realidad Económica, 232, IADE

34. **Gottlieb, A. M.; Giberti, G. C. & Poggio, L. (2011).** *Evaluación del germoplasma de Ilex paraguariensis e Ilex dumosa (Aquifoliaceae).* Boletín de la Sociedad Argentina de Botánica, 46(1-2), 113-123.

35. **Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; Chen, Z.; Mauceli, E.; Hacohen, N.; Gnirke, A; Rhind, N.; di Palma, F.; Birren, B.W.; Nusbaum, C.; Lindblad-Toh, K.; Friedman, N.; Regev, A. (2011).** *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol 29:644-652.

36. **Haft, D. H.; Loftus, B. J.; Richardson, D. L.; Yang, F.; Eisen, J. A.; Paulsen, I. T. & White, O. (2001).** *TIGRFAMs: a protein family resource for the functional identification of proteins.* Nucleic acids research, 29(1), 41–43. https://doi.org/10.1093/nar/29.1.41

37. **Heck, C.I.; De Mejia, E.G. (2007).** *Yerba Mate Tea (Ilex paraguariensis): a comprehensive review on chemistry, health implications, and technological considerations.* J Food Sci 72:R138–R151. doi:10.1111/j. 1750-3841.2007.00535

38. **Horowitz, N. H. (1945).** *On the Evolution of Biochemical Syntheses.* Proceedings of the National Academy of Sciences of the United States of America, 31(6), 153–157. https://doi.org/10.1073/pnas.31.6.153

39. **Huang, R.; O'Donnell, A. J.; Barboline, J. J. & Barkman, T. J. (2016).** *Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes.* Proc. Natl. Acad. Sci. U. S. A. 113, 10613–10618

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

41. **Isolabella, S.; Cogoi, L.; López, P.; Anesini, C.; Ferraro, G. &Filip, R. (2010).** *Study of the bioactive compounds variation during yerba mate (Ilex paraguariensis) processing.* Food Chemistry, 122(3), 695-699.

42. **Jensen, R. A. (1976).** *Enzyme recruitment in evolution of new function.* Annual review of microbiology, 30, 409–425. https://doi.org/10.1146/annurev.mi.30.100176.002205

43. **Kato, M.; Mizuno, K.; Fujimura, T.; Iwama, M.; Irie, M.; Crozier, A. & Ashihara, H. (1999).** *Purification and characterization of caffeine synthase from tea leaves.* Plant physiology, 120(2), 579–586. https://doi.org/10.1104/pp.120.2.579

44. **Kato, M.; Mizuno, K.; Crozier, A.; Fujimura, T.; Ashihara, H. (2000)** *Caffeine synthase gene from tea leaves.* Nature 406(6799):956–957.

45. **Landis, J. B.; Soltis, D. E.; Li, Z.; Marx, H. E.; Barker, M. S.; Tank, D. C. & Soltis, P. S. (2018).** *Impact of whole-genome duplication events on diversification rates in angiosperms.* American journal of botany, 105(3), 348–363. https://doi.org/10.1002/ajb2.1060

46. **Lewis E. B. (1951).** *Pseudoallelism and gene evolution.* Cold Spring Harbor symposia on quantitative biology, 16, 159–174. https://doi.org/10.1101/sqb.1951.016.01.014

47. **Maddison, W. P. and D.R. Maddison. (2021).** *Mesquite: a modular system for evolutionary analysis.* Version 3.70 http://www.mesquiteproject.org

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

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

50. **Negrin, A.; Long, C.; Motley, T. J., & Kennelly, E. J. (2019).** *LC-MS Metabolomics and Chemotaxonomy of Caffeine-Containing Holly (Ilex) Species and Related Taxa in the Aquifoliaceae.* Journal of agricultural and food chemistry, 67(19), 5687–5699. https://doi.org/10.1021/acs.jafc.8b07168

51. **Ohno, S. (2013).** *Evolution by gene duplication.* Springer Science & Business Media.

52. **One Thousand Plant Transcriptomes Initiative (2019).** *One thousand plant transcriptomes and the phylogenomics of green plants.* Nature, 574(7780), 679–685. https://doi.org/10.1038/s41586-019-1693-2

53. **Onetto, A.; Laczeski, M.; Bergottini, V.M.; Lopez, A.; Sosa, A.D.; Wiss, F.; Villalba, L.L.; Zapata, P.D.; Otegui, M.B. (2015).** *Characterization of endophytic sporulating bacteria with plant growth promoting properties isolated from ilex paraguariensis (yerba mate).* ipmb 2015 11th int plant mol biol

54. **Pagliosa, C. M.; Vieira, M. A.; Podestá, R.; Maraschin, M.; Zeni, A. L. B.; Amante, E. R. &Amboni, R. D. D. M. C. (2010).** *Methylxanthines, phenolic composition, and antioxidant activity of bark from residues from mate tree harvesting (Ilex paraguariensis A. St. Hil.).* Food Chemistry, 122(1), 173-178.

55. **Rau, V. (2009).** *La yerba mate en Misiones (Argentina).* Estructura y significados de una producción localizada. Agroalimentaria28: 49-58.

56. **Sankoff, D. & Zheng, C. (2018).** *Whole Genome Duplication in Plants: Implications for Evolutionary Analysis.* Methods Mol. Biol. 1704, 291–315.

57. **Sosa, D.A.; Munareto, N.** *Manejo nutricional del cultivo de yerba mate.* 5to. Congreso Sudamericano de la Yerba mate. Posadas, Misiones, 2011.

58. **Spannagl, M.; Nussbaumer, T.; Bader, K.; Gundlach, H., & Mayer, K. F. (2017).** *PGSB/MIPS PlantsDB Database Framework for the Integration and Analysis of Plant Genome Data.* Methods in molecular biology (Clifton, N.J.), 1533, 33–44. https://doi.org/10.1007/978-1-4939-6658-5_2

59. **Stein, J.; Luna, C.; Espasandin, F.; Sartor, M.; Espinoza, F.; Ortiz,**

J. P. & Pessino, S. C. (2014). *Construcción de un mapa genético preliminar de yerba mate (Ilex paraguariensis).* Revista de Investigaciones de la Facultad de Ciencias Agrarias-UNR, (23), 007-013.

60. **UniProt Consortium (2019).** *UniProt: a worldwide hub of protein knowledge.* Nucleic acids research, 47(D1), D506–D515. https://doi.org/10.1093/nar/gky1049
61. **Ycas, M. (1974).** *On earlier states of the biochemical system.* Journal of theoretical biology, 44(1), 145–160. https://doi.org/10.1016/s0022-5193(74)80035-4
62. **Yin, Y.; Katahira, R. & Ashihara, H. (2015).** *Metabolism of purine alkaloids and xanthine in leaves of maté (Ilex paraguariensis).* Natural product communications, 10(5), 707–712.
63. **Zhang, C.; Zhang, T.; Luebert, F.; Xiang, Y.; Huang, C. H.; Hu, Y.; Rees, M.; Frohlich, M. W.; Qi, J.; Weigend, M. & Ma, H. (2020).** *Asterid Phylogenomics/Phylotranscriptomics Uncover Morphological Evolutionary Histories and Support Phylogenetic Placement for Numerous Whole-Genome Duplications.* Molecular biology and evolution, 37(11), 3188–3210. https://doi.org/10.1093/molbev/msaa160
64. **Zubieta, C.; Ross, JR., Koscheski, P.; Yang, Y.; Pichersky, E.; Noel JP. (2003).** *Structural basis for substrate recognition in the salicylic acid carboxyl methyltransferase family.* Plant Cell 15:1704-1716.