Molecular Plant
Research Article

CellPress
PARTNER JOURNAL

The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis

En-Hua Xia[1,2,3,12], Hai-Bin Zhang[1,2,12], Jun Sheng[4,12], Kui Li[1,2,12], Qun-Jie Zhang[1,5,12], Changhoon Kim[6], Yun Zhang[1], Yuan Liu[1,2], Ting Zhu[1,7], Wei Li[1,2], Hui Huang[1,2], Yan Tong[1], Hong Nan[1,3], Cong Shi[1,3], Chao Shi[1,2], Jian-Jun Jiang[1,2], Shu-Yan Mao[1], Jun-Ying Jiao[1], Dan Zhang[1,2], Yuan Zhao[4], You-Jie Zhao[1], Li-Ping Zhang[1], Yun-Long Liu[1], Ben-Ying Liu[8], Yue Yu[6], Sheng-Fu Shao[9], De-Jiang Ni[10], Evan E. Eichler[11] and Li-Zhi Gao[1,2,*]

[1]Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwestern China, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China

[2]Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou 510642, China

[3]University of the Chinese Academy of Sciences, Beijing 100039, China

[4]Yunnan Agricultural University, Kunming 650204, China

[5]Agrobiological Gene Research Center, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China

[6]Macrogene Inc., Seoul 08511, South Korea

[7]College of Life Science, Liaoning Normal University, Dalian 116081, China

[8]National Tea Tree Germplasm Bank, Tea Research Institute, Yunnan Academy of Agricultural Sciences, Menghai 666201, China

[9]Jinhua International Camellia Germplasm Bank, Jinhua 321000, China

[10]Department of Tea Science, Key Lab for Horticultural Plant Biology, Huazhong Agricultural University, Wuhan 430070 China

[11]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

[12]These authors contributed equally to this article.

***Correspondence: Li-Zhi Gao (lgao@mail.kib.ac.cn)**

http://dx.doi.org/10.1016/j.molp.2017.04.002

ABSTRACT

Tea is the world's oldest and most popular caffeine-containing beverage with immense economic, medicinal, and cultural importance. Here, we present the first high-quality nucleotide sequence of the repeat-rich (80.9%), 3.02-Gb genome of the cultivated tea tree *Camellia sinensis*. We show that an extraordinarily large genome size of tea tree is resulted from the slow, steady, and long-term amplification of a few LTR retrotransposon families. In addition to a recent whole-genome duplication event, lineage-specific expansions of genes associated with flavonoid metabolic biosynthesis were discovered, which enhance catechin production, terpene enzyme activation, and stress tolerance, important features for tea flavor and adaptation. We demonstrate an independent and rapid evolution of the tea caffeine synthesis pathway relative to cacao and coffee. A comparative study among 25 *Camellia* species revealed that higher expression levels of most flavonoid- and caffeine- but not theanine-related genes contribute to the increased production of catechins and caffeine and thus enhance tea-processing suitability and tea quality. These novel findings pave the way for further metabolomic and functional genomic refinement of characteristic biosynthesis pathways and will help develop a more diversified set of tea flavors that would eventually satisfy and attract more tea drinkers worldwide.

Key words: Tea tree genome, Comparative genomics, Tea flavor, Tea-proccessing suitability, Global adaptation, Caffeine biosynthesis

Xia E.-H., Zhang H.-B., Sheng J., Li K., Zhang Q.-J., Kim C., Zhang Y., Liu Y., Zhu T., Li W., Huang H., Tong Y., Nan H., Shi C., Shi C., Jiang J.-J., Mao S.-Y., Jiao J.-Y., Zhang D., Zhao Y., Zhao Y.-J., Zhang L.-P., Liu Y.-L., Liu B.-Y., Yu Y., Shao S.-F., Ni D.-J., Eichler E.E., and Gao L.-Z. (2017). The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. Mol. Plant. ■ ■, 1–12.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

INTRODUCTION

Socially and habitually consumed by more than 3 billion people across 160 countries, tea is the oldest (since 3000 BC) and most popular nonalcoholic caffeine-containing beverage in the world (Banerjee, 1992; Mondal et al., 2004). Besides its attractive aroma and pleasant taste, the tea beverage has numerous healthful and medicinal benefits for humans due to many of the characteristic secondary metabolites in tea leaves, such as polyphenols, caffeine, theanine, vitamins, polysaccharides, volatile oils, and minerals (Yamamoto et al., 1997; Cabrera et al., 2006; Rogers et al., 2008; Chacko et al., 2010). The tea plant *Camellia sinensis* is the source of commercially grown tea and a member of the genus *Camellia* in the tea family Theaceae, which also contains several other economically important species, including well-known camellias with their attractive flowers (e.g., *C. japonica*, *C. reticulata*, and *C. sasanqua*) and the traditional oil tree *C. oleifera* that produces high-quality edible seed oil (Ming and Bartholomew, 2007). The first credible record of tea as a medicinal drink occurred during the Shang dynasty of China and dates back to the third century AD (Weinberg and Bealer, 2001; Heiss and Heiss, 2007). The global expansion of tea is long and complex, spreading across multiple cultures over the span of thousands of years and expanding worldwide to more than 100 countries (Heiss and Heiss, 2007; Liu et al., 2015). Today, tea is commercially cultivated on more than 3.80 million hectares of land on a continent-wide scale, and 5.56 million metric tons of tea worldwide were produced annually in 2014.

As one of the most popular beverages worldwide, tea has well-established nutritional and medicinal properties derived from the three major characteristic secondary metabolites: catechins, theanine, and caffeine. These phytochemical compounds, especially catechins, are beneficial for human health (Khan and Mukhtar, 2007), the contents and component proportions of which in large part determine the flavor of tea. The genus *Camellia*, consisting of ~119 species (Ming and Bartholomew, 2007) with differential metabolite profiles, provides a uniquely powerful system for dissecting the variation and evolution of flavonoid, theanine, and caffeine biosynthesis pathways that define tea-processing suitability. Thousands of years of continental introduction and conventional selective breeding efforts have resulted in a large number of land race and elite cultivars that adapt to globally diverse habitats, thus ensuring different tea productivity and quality worldwide. The rich metabolite constituents within the tea tree may play an important role in adaptations to diverse ecological niches on Earth. Unraveling the genomic basis of these global adaptations remains an unsolved mystery. Although it is well recognized that the differential accumulation of the three major characteristic constituents in tea tree leaves largely determines the quality of tea, little genomic information is currently available regarding the complex transcriptional regulation of catechins, theanine, and caffeine metabolic pathways. Sequencing of the tea tree genome would facilitate to uncover the molecular mechanisms underlying secondary metabolic biosynthesis with the promise to improve breeding efficiency and thus develop better tea cultivars with even higher quality.

Here, we report a high-quality genome assembly of Yunkang 10 (2n = 2x = 30 chromosomes), a diploid elite cultivar of *C. sinensis* var. *assamica* widely grown in Southwestern China, based on sequence data from whole-genome shotgun sequencing. Together with comparative transcriptomic and phytochemical analyses for the representative *Camellia* species, we aim to obtain new insights into the molecular basis of the biosynthesis of the three characteristic secondary metabolites with an emphasis on the suitability of tea-processing and the formation of tea flavor.

RESULTS

Genome Sequencing, Assembly, and Annotation

We sequenced the tea tree genome (cultivar Yunkang 10) from Yunnan Province, China. We performed a whole-genome shotgun sequencing analysis with the Illumina next-generation sequencing platform (HiSeq 2000). This generated raw sequence data sets of ~707.88 Gb, thus yielding approximately 159.43-fold high-quality sequence coverage (Supplemental Table 1). Using two orthogonal methods, we estimated that the genome size of Yunkang 10 is between 2.9 and 3.1 Gb (Supplemental Figures 1 and 2; Supplemental Table 2). The tea tree genome was assembled using Platanus (Kajitani et al., 2014), followed by scaffolding preassembled contig sequences and paired-read next-generation sequencing data using SSPACE (Boetzer et al., 2011). This finally yielded a ~3.02-Gb genome assembly that spans ~98% of the estimated genome size and contains 37 618 scaffolds (N50 = 449 kb) and 258 790 contigs (N50 = 20.0 kb) (Table 1 and Supplemental Table 3). To validate the genome assembly quality, we first aligned all available DNA and expressed sequence tags of the tea tree from public databases and obtained mapping rates of 75.56% and 88.30%, respectively (Supplemental Table 5); secondly, we mapped all high-quality reads (~339.49 Gb) to the assembled genome sequences, which show good alignments with a mapping rate of 93.96% (Supplemental Table 5); and thirdly, the transcripts we assembled also showed excellent alignments/sequence identities to the assembled genome: out of 198 175 transcripts, 76.23% were mapped (transcript coverage ≥90% and identity ≥90%; Supplemental Table 5 and Supplemental Section 1.6).

To further aid in genome annotation, we generated ~29.7 Gb of RNA sequencing (RNA-seq) data obtained from a total of eight libraries representing major tissue types and developmental stages, including young leaf, tender shoot, flower bud, flower, stem, root, seed, and seedling (Supplemental Tables 6 and 7). In combination with *ab initio* and EvidenceModeler prediction, protein and public expressed sequence tag alignments, and further filtering, we defined 36 951 protein-coding genes (Supplemental Table 8 and Supplemental Figure 3). Of these, 33 415 (~90.43%) and 26 861 (~72.69%) could be functionally classified and supported by transcripts, protein and/or expressed sequence tags (Supplemental Tables 9–11), respectively. In addition, we performed homology searches and annotated noncoding RNA (ncRNA) genes (Supplemental Table 12), yielding 700 transfer RNA (tRNA) genes, 2860 ribosomal RNA (rRNA) genes, 454 small nucleolar RNA (snoRNA) genes, 223 small nuclear RNA (snRNA) genes, and 233 microRNA (miRNA) genes. The annotation of repeat sequences showed that transposable elements accounted for

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

Assembly	
Genome-sequencing depth (×)	159.43
Estimated genome size (Gb)	3.0
No. of scaffolds	37 618
Total length of scaffolds (bp)	3 021 230 785
N50 of scaffolds (bp)	449 457
Longest scaffolds (bp)	3 505 831
No. of contigs	258 790
Total length of contigs (bp)	2 575 242 646
N50 of contigs (bp)	19 958
Longest contigs (bp)	257 648
Predicted coverage of the assembled sequences (%)	100
GC content of the genome (%)	42.31
Annotation	
No. of predicted protein-coding genes	36 951
Average gene length (bp)	3549
Percentage of gene length in the genome (%)	4.49
Mean exon length (bp)	237
Average exon per gene	4.8
Mean intron length (bp)	640
tRNAs	700
rRNAs	2860
snoRNAs	454
snRNAs	223
miRNAs	233
Masked repeat sequence length (Mb)	2083
Percentage of repeat sequences (%)	80.89

Table 1. Summary of Genome Assembly and Annotation for the Tea Tree *Camellia sinensis*.

approximately 80.89% of the assembled genome (Supplemental Tables 13 and 14). The GC content was ~42.31% across the genome and ~44.55% in coding sequences (Supplemental Tables 3 and 8). We annotated ~867 339 simple sequence repeats, which will provide valuable genetic markers to assist tea tree breeding programs (Supplemental Table 15).

Repetitive Nature of Tea Tree Genome and Repeat-Driven Genome Expansion

Among the five sequenced species of asterids, including potato (Xu et al., 2011), tomato (Sato et al., 2012), coffee (Denoeud et al., 2014), and pepper (Qin et al., 2014), the tea tree not only has the largest genome but also harbors the highest content of repetitive DNA (Table 1; Supplemental Tables 13 and 16; Supplemental Figure 6). Long terminal repeat (LTR) retrotransposons represent the majority (~67.21%) of the tea tree genome (Supplemental Table 13), of which about 55.09% of the assembly was identified as belonging to two types of LTR retrotransposons: Ty1/*copia* and Ty3/*gypsy*. Dating transposable elements shows that both retrotransposons and DNA transposons were almost simultaneously amplified, with a peak substitution rate around 16% compared with the consensus (Supplemental Figure 7). Comparative analyses of the Ty1/*copia* and Ty3/*gypsy* LTR retrotransposon families suggest that both repeat classes have experienced multiple retrotransposition bursts over the last 50 million years (Supplemental Figure 7), perhaps the result of a relatively longer half-life when compared with smaller genomes such as *Arabidopsis thaliana* (Devos et al., 2002) and rice (Gao et al., 2004). Ty3/*gypsy* LTR retrotransposon families predominate and contribute most to the tea tree genome expansion (~47.08%) (Supplemental Table 13). Unlike small plant genomes, long-standing and incessant LTR retrotransposon bursts may have persisted due to the lack of efficient DNA removal mechanisms leading to genome size growth, as previously reported for *Picea abies* (Nystedt et al., 2013). *TL001* is the largest of the Ty3/*gypsy* retrotransposon families, and accounts for ~66.70% of LTR retrotransposons and ~36.79% of the whole tea tree genome (Figure 1; Supplemental Figures 9 and 11; Supplemental Table 14). We report a single LTR retrotransposon family propagating and persisting for more than 50 million years of evolution, leading to such a dramatic expansion in genome size (Supplemental Figure 11).

A comprehensive survey based on RNA-seq data from seven tissues generally showed that expression levels of LTR retrotransposon families are positively correlated to the copy number of elements (Figure 1; Supplemental Figures 13 and 14). Ty3/*gypsy* LTR retrotransposons (~0.94% on average) exhibit two-fold higher expression levels than Ty1/*copia* LTR retrotransposons (~0.48% on average) across all seven tissues (Supplemental Table 18). Among reads that mapped to intact retroelements, the most abundant Ty3/*gypsy* family, *TL001* (Supplemental Figure 9), was also the most highly expressed across the seven tissues, accounting for ~30.18% of all expressed LTR retrotransposons and ~51.36% of LTR retrotransposon-related DNA sequencing reads (Supplemental Figures 12 and 13). We observe differential expression of LTR retrotransposons among tissues, evidenced by ~1.78% (on average) of RNA-seq reads from the seven tissues originated from LTR retrotransposons, with the proportions ranging from ~1.46% in young leaf to ~2.30% in the flower (Supplemental Figures 12 and 13; Supplemental Table 18). Tissue-specific expression differences were particularly remarkable for some families (Supplemental Figures 12 and 13; Supplemental Table 18). For example, nonautonomous retrotransposon family *TL026* was highly expressed in seeds corresponding to ~13.38% of the expressed LTR retrotransposon reads. This is ~32-fold larger than an average rate of ~0.42% across the other six tissues. Tissue-specific expression profiling of LTR retrotransposons holds true after testing using RNA-seq datasets from twice-repeated sampling of tea tree leaf tissues (Supplemental Figure 14; Supplemental Table 19). These results suggest that retrotransposition expression levels may be indicative of retrotransposon activity in the tea tree genome.

Gene Family Evolution and Whole-Genome Duplication

Defining gene families evolving rapidly among flowering plants has been useful in identifying the genomic bases underlying species adaptation and physiological changes of metabolite

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

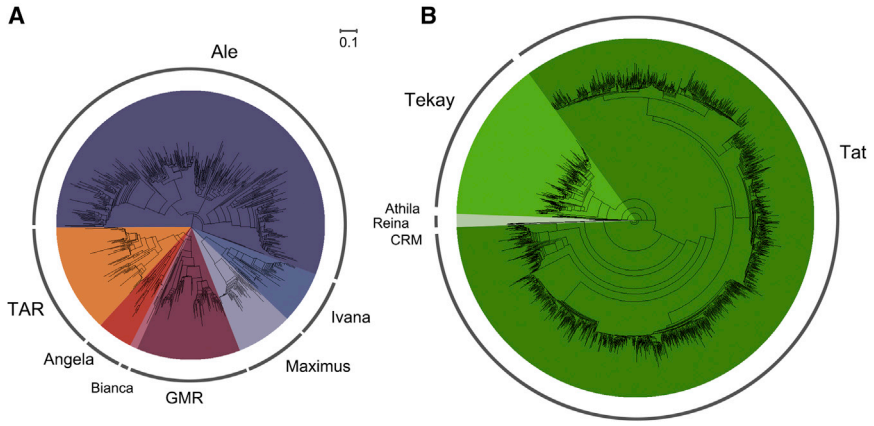

Figure 1. Phylogenetic Analysis of the LTR Retrotransposon Sequences in the Tea Tree Genome.

The neighbor-joining and unrooted phylogenetic trees were constructed on the basis of 678 Ty1/*copia* **(A)** and 1795 Ty3/*gypsy* **(B)** aligned sequences corresponding to the RT domains without premature termination codon. LTR family names and proportion of each are indicated.

constituents during evolution (Mitreva et al., 2011; Denoeud et al., 2014). We compared the predicted proteomes of the tea tree, kiwifruit, potato, tomato, coffee, *A. thaliana*, cacao, poplar, grape, and lotus, yielding a total of 26 024 orthologous gene families that comprised 246 457 genes (Supplemental Table 20; Supplemental Figure 15). This revealed a core set of 113 439 genes belonging to 6730 clusters that were shared among all 10 plant species, representing ancestral gene families (Figure 2A). We found a total of 714 gene clusters containing 2170 genes unique to the tea tree, potentially related to environmental adaptation and phytochemical properties within the tea lineage (Figure 2A). Functional enrichment analyses of tea tree-specific genes by both gene ontology (GO) terms and PFAM domains together revealed functional categories related to biosynthetic processes associated with major tea characteristic secondary metabolites (e.g., catechins). The latter included flavonoid biosynthetic process (GO: 0009813, $P < 0.001$) and secondary metabolite catabolic process (GO: 0090487, $P < 0.001$) (Supplemental Tables 21 and 22). PFAM analysis further revealed that gene functions involved in flavonoid biosynthesis are enriched in 2OG-Fe(II) oxygenase superfamily (PF03171, $P < 0.001$), which encodes enzymes associated with the production of anthocyanidin and flavonol (flavanone 3-hydroxylase, anthocyanidin synthase, and flavonol synthase) (Supplemental Tables 21 and 23). Terpenoids constitute a large family of natural compounds and are major components of resins, essential oils, and aromas (Zwenger and Basu, 2008). Remarkably, we found that the tea tree-specific gene families were also significantly enriched in functions related to terpene synthase activity (GO: 0010333, $P < 0.001$) that may be associated with the tea aroma, further evidenced by PFAM annotation with enriched functional domain of terpene synthase (PF01397; $P < 0.001$) (Supplemental Tables 21, 22, and 23).

In flowering plants, the expansion or contraction of gene families is an important driver of lineage splitting and phenotypic diversification (Ohno, 1970; Chen et al., 2013). We characterized gene families that underwent discernible changes and divergently evolved along different branches, with particular emphasis on those involved in tea tree traits and tea flavor (Figure 2B). Our results showed that, of the 13 476 gene families inferred to be present in the most recent common ancestor of the ten studied plant species, 1857 comprising 2048 genes exhibited significant expansions ($P < 0.001$) in the tea tree lineage (Figure 2B). Functional annotation of these genes demonstrates that they were mainly enriched in functional categories involved in flavonoid metabolic processes, including flavonoid metabolic process (GO: 0009812, $P < 0.001$) and flavonoid biosynthetic process (GO: 0009813, $P < 0.001$) (Supplemental Tables 24, 25, and 26). Notably, gene families were significantly enriched in a number of functions related to the modification of flavonoid metabolic compounds, such as quercetin 3-*O*-glucosyltransferase activity (GO: 0080043, $P < 0.001$), UDP-glucosyltransferase activity (GO: 0035251, $P < 0.001$; PF00201, $P < 0.001$), UDP-glycosyltransferase activity (GO: 0008194, $P < 0.001$), and flavonoid glucuronidation (GO: 0052696, $P < 0.001$) (Supplemental Tables 24, 25, and 26). The glucosyltransferase activities are well known to affect tea flavor and quality by controlling the content and formation of important secondary metabolites, for example, galloylated catechins and flavonol 3-*O*-glycosides, which largely determine the astringency of tea flavor (Lim and Bowles, 2004; Liu et al., 2012; Cui et al., 2016).

Among the tea tree-specific and expanded gene families, we found that defense genes were among one of the most highly enriched functional categories including plant disease defense response, e.g., NB-ARC domain (PF00931; $P < 0.001$) and leucine-rich repeat (LRR) (PF13516, PF07725, PF12799, PF00560, PF13855; $P < 0.001$) (Supplemental Tables 21–, 22, 23, 24, 25, and 26). These findings suggest that strong natural selection for enhanced disease resistance in the tea tree potentiated global adaptations to the diverse habitats of Asia, Africa, Europe, North America, South America, and Oceania. To further assess this, we thoroughly explored the disease resistance genes, including the nucleotide-binding site with leucine-rich repeat (NBS-LRR) and pattern-recognition receptor (RLK-LRR) genes in the tea tree together with four other eudicots (kiwifruit, tomato, cacao, and *A. thaliana*). Results showed that tea tree harbored a total of 313 NBS-LRR encoding genes, which is larger than those in kiwifruit (104), *A. thaliana* (207), tomato (263), and cacao (297) (Figure 2C and Supplemental Table 27). NBS-LRR genes in plants are mainly responsible for recognizing specific pathogen effectors (Jones and Dangl, 2006); thus, the observation of a large expansion of this type of genes implies selection pressures in response to pathogenic challenge. We also characterized a total of 272 putative RLK-LRR genes that encode receptor-like kinases with an LRR domain (RLK-LRR) in the tea tree genome (Supplemental Table 27). This number is slightly larger than that found in kiwifruit (254), tomato (231), potato (261), cacao (238), and *A. thaliana* (224), suggesting that pattern-triggered immunity, another type of ancient innate immunity in plants, is

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

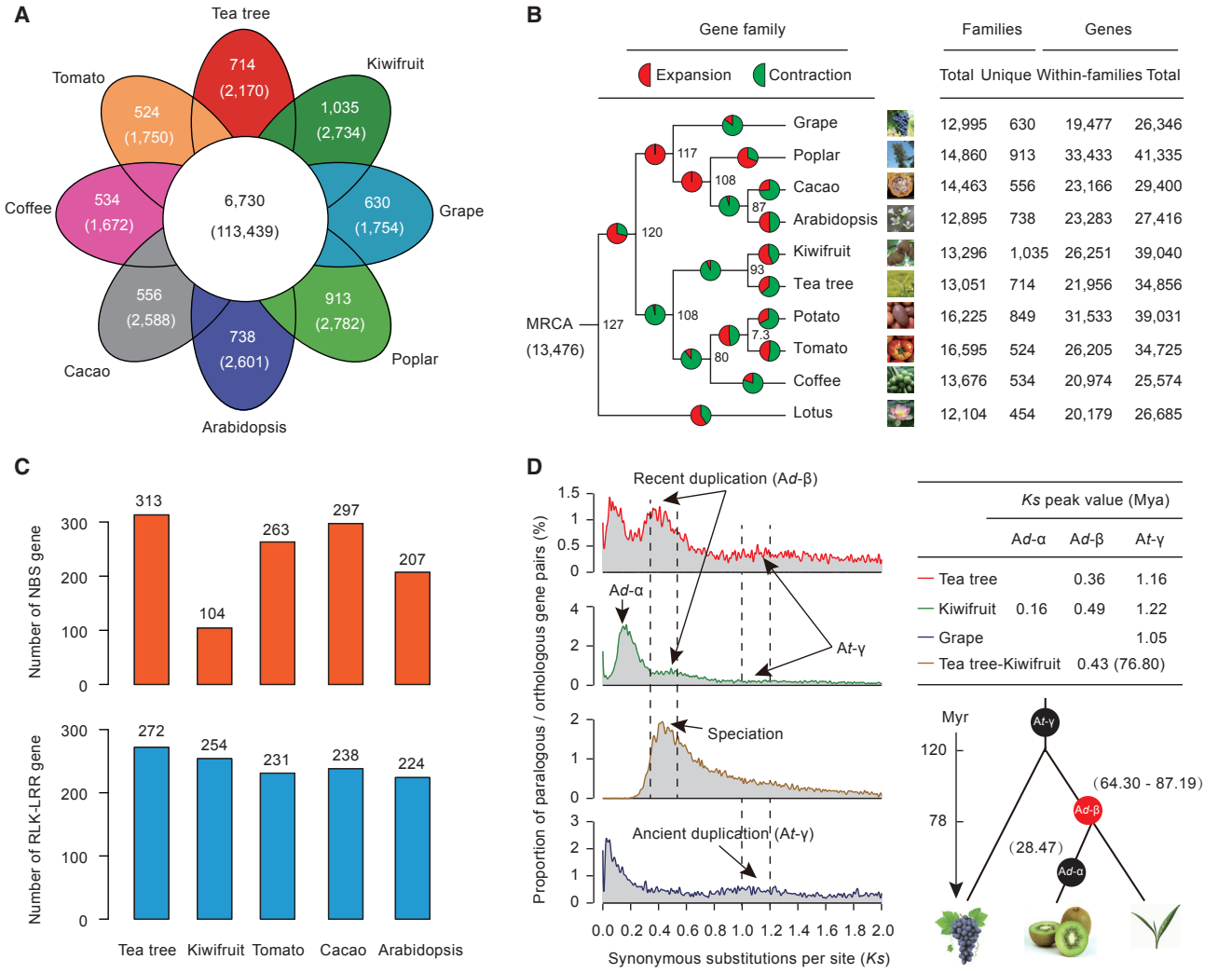

Figure 2. Evolution of the Tea Tree Genome and Gene Families.

(A) Venn diagram shows the shared and unique gene families among the tea tree and seven other plant species. Each number in parentheses represents the number of genes within corresponding families (without parentheses).

(B) Expansion and contraction of gene families among the 10 plant species. Phylogenetic tree was constructed based on 597 high-quality 1:1 single-copy orthologous genes using sacred lotus (*Nelumbo nucifera*) as outgroup. Pie diagram on each branch of the tree represents the proportion of genes undergoing gain (red) or loss (green) events. Number at root (13 476) denotes the total number of gene families predicted in the most recent common ancestor (MRCA) (see Supplemental Information). The numerical value beside each node shows the estimated divergent time of each node (myr).

(C) Comparisons of disease-resistant genes among the five plant species.

(D) Whole-genome duplication events detected in the tea tree.

more conserved in the tea tree and may play an important role in pathogen defense.

Previous studies on the sequenced plant genomes have shown that polyploidy has been a prominent feature in the evolutionary history of angiosperms and that whole-genome duplication (WGD) events, in particular, have had major impacts on crop gene and genome evolution (Bennett, 2004; Jaillon et al., 2007; Huang et al., 2013; Salman-Minkov et al., 2016). We identified 16 520 paralogous gene pairs that spanned 47.6% of the protein-coding genes in the tea tree genome (Supplemental Table 28). On the basis of these duplicated gene pairs, we calculated an age distribution of synonymous substitution rates (*Ks*) that peaked around 0.36 and 1.16 (Figure 2D; Supplemental Figure 16; Supplemental Table 29), suggesting that two rounds of WGD events occurred in the tea tree genome. We compared the tea tree genome with two other eudicot genome sequences (kiwifruit and grape), respectively, based on the distribution of *Ks* values of paralogous gene pairs (Figure 2D). Our results confirm that the ancient WGD (Ad-γ), referenced as γ in the literature for eudicots, was shared among tea tree, grape (Jaillon et al., 2007), and kiwifruit (Huang et al., 2013). The recent WGD event (referenced as Ad-β) that occurred in tea tree was also observed in four other *Camellia* species (*C. sinensis* var. *sinensis*, *C. taliensis*, *C. reticulata*, and *C. impressinervis*) based on *Ks* values of paralogous genes derived from their high-quality transcriptome data (Supplemental Figure 17). This WGD event, thus, occurred

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

in the common ancestor of these investigated *Camellia* species. To determine whether A*d*-β was a genus-specific event in the tea tree or shared with the WGD reported in kiwifruit (Figure 2D), we computed *Ks* values to date the speciation time based on orthologous gene pairs from syntenic blocks between tea tree and kiwifruit. However, our results still failed to clearly conclude whether this recent WGD event occurred before or after the tea tree-kiwifruit divergence (Figure 2D and Supplemental Figure 18) because the estimated dates for tea tree and kiwifruit WGD events are quite close to their speciation time. We further adopted the PUG (Phylogenetic Placement of Polyploidy Using Genomes) pipeline (McKain et al., 2016) and obtained a small proportion (~16%) of gene trees derived from the A*d*-β event of tea tree and kiwifruit, supporting that this A*d*-β event occurred prior to the divergence between tea tree and kiwifruit (Supplemental Figure 19). Further efforts are needed to sequence and compare eudicot genomes between tea tree and kiwifruit lineages to exactly determine whether they are the same or distinct WGD events.

Genomic Basis of Tea-Processing Suitability and Quality

Tea tree leaves are often used to produce tea with nutritious and health-giving properties. The quality of tea is primarily determined by three major characteristic constituents: polyphenols (mainly catechins), theanine, and caffeine. To uncover the accumulation of the characteristic secondary metabolites in tea tree leaves, which are key to defining tea-processing suitability and tea quality, we performed comparative phytochemical analysis of most species from section *Thea* as well as 10 species from the representative non-*Thea* sections of the genus *Camellia* (Supplemental Table 30A). A high-performance liquid chromatography (HPLC) analysis revealed significantly higher contents of total catechins (~7.40-fold on average, P = 7.78E-07) and caffeine (~9.50-fold on average, P = 2.56E-03), in the tea tree and other species from section *Thea* when compared with species from non-*Thea* sections (Figure 3A; Supplemental Figure 20; Supplemental Tables 30 and 31). Noticeably, characteristic catechin constituents, including EGCG (~8.2-fold on average, P < 5.56E-05), EGC (~7.7-fold on average, P < 6.79E-04), C (~3.3-fold on average, P < 1.07E-02), EC (~2.6-fold on average, P < 1.92E-02), and ECG (~19.5-fold on average, P < 2.08E-04), which largely varied among section *Thea* species, were significantly higher than species from non-*Thea* sections (Figure 3A; Supplemental Tables 30 and 31). In contrast, we observed no significant difference in theanine content between these species (~1.3-fold on average, P < 2.44E-01) (Figure 3A; Supplemental Tables 30 and 31). Our results suggest that the tea tree and other related species from section *Thea* are remarkably abundant in catechins and caffeine.

To gain novel insights into the molecular mechanisms underlying the phytochemical characteristics of major secondary metabolites in the tea tree and other *Camellia* species, we performed an integrated analysis based on comparative transcriptomic and phytochemical data for the same panel of *Camellia* species grown under the same conditions (Supplemental Tables 30 and 33). On the basis of the annotation of genes encoding enzymes potentially involved in catalyzing these reactions of flavonoid, theanine, and caffeine pathways in our assembled tea tree genome, we first obtained homologous genes from the respective transcriptomes of the other 23 *Camellia* species (Supplemental Table 32), including the 14 catechin biosynthesis-related genes (*PAL*, *C4H*, *4CL*, *CHS*, *CHI*, *F3′H*, *F3′5′H*, *FNS II*, *FLS*, *DFR*, *LCR*, *ANS*, *ANR*, and *F4′ST*), six theanine biosynthesis-related genes (*TS*, *GS*, *GDH*, *ADC*, *SAMDC*, and *Fe-GOGAT*), and four caffeine biosynthesis-related genes (*IMPDH*, *SAMS*, *AMPDA*, and *TCS*), respectively. Our analysis revealed that species from section *Thea* as well as other relatives from non-*Thea* sections possessed all of these important genes encoding enzymes involved in the biosynthesis of catechins, theanine, and caffeine in cultivated tea tree. This suggests that the three characteristic metabolic pathways were already present in the common ancestor of *Camellia* and have remained well conserved for ~6.3 million years (Supplemental Section 7.5).

Nevertheless, the 24 characteristic metabolite-related genes exhibited considerably distinct expression patterns in mature leaves across the 24 examined *Camellia* species (Figure 3B and Supplemental Tables 33, 34, and 35). The detected genes responsible for catechin biosynthesis, in particular, and caffeine biosynthesis rather than theanine biosynthesis were differentially expressed between *Camellia* species from section *Thea* and non-*Thea* sections (Figure 3B and Supplemental Table 35). For example, four genes encoding enzymes involved in the last few steps of catechin biosynthesis, *ANR* (anthocyanidin reductase) (~1.38-fold on average, P = 1.88E-01), *F3′5′H* (flavonoid 3′,5′-hydroxylase) (~16.86-fold on average, P = 5.17E-02), *CHI* (chalcone isomerase) (~3.79-fold on average, 2.19E-02), and *FNS II* (flavone synthase II) (~6.64-fold on average, P = 1.27E-01), were more highly expressed in section *Thea* species where higher content of catechins was observed when compared with non-*Thea* sections. Similar expression patterns were also observed in three of the four examined genes encoding key enzymes involved in caffeine biosynthesis, including *TCS* (tea caffeine synthase) (~12.17-fold on average, P = 2.81E-03), *IMPDH* (inosine-5′-monophosphate dehydrogenase) (~3.30-fold on average, P = 4.71E-05), and *AMPDA* (AMP deaminase) (~2.42-fold on average, P = 1.08E-04); they were significantly more highly expressed in section *Thea* species, which also contains significantly higher content of caffeine when compared with non-*Thea* sections (Figure 3B and Supplemental Table 35). Notably, gene expression levels of *TCS* encoding the enzyme that catalyzes the final step in caffeine biosynthesis largely differed among species from either section *Thea* or non-*Thea* sections (P < 0.001), corresponding to their variable amounts of caffeine (Figure 3B and Supplemental Table 35). Sequence variation of these 24 characteristic metabolite-related genes correlates well with phytochemical differentiation of the three major secondary metabolic pathways among these representative *Camellia* species (Figure 3C).

The 24 characteristic metabolite-related genes exhibit distinct expression patterns in the eight tissues of the tea tree (Supplemental Figures 23–25). Our results showed that the majority of genes encoding enzymes involved in flavonoid biosynthesis pathways were highly expressed in tender shoots (Supplemental Figure 23), indicating that flavonoid biosynthesis actively occurs early during shoot differentiation. Genes

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

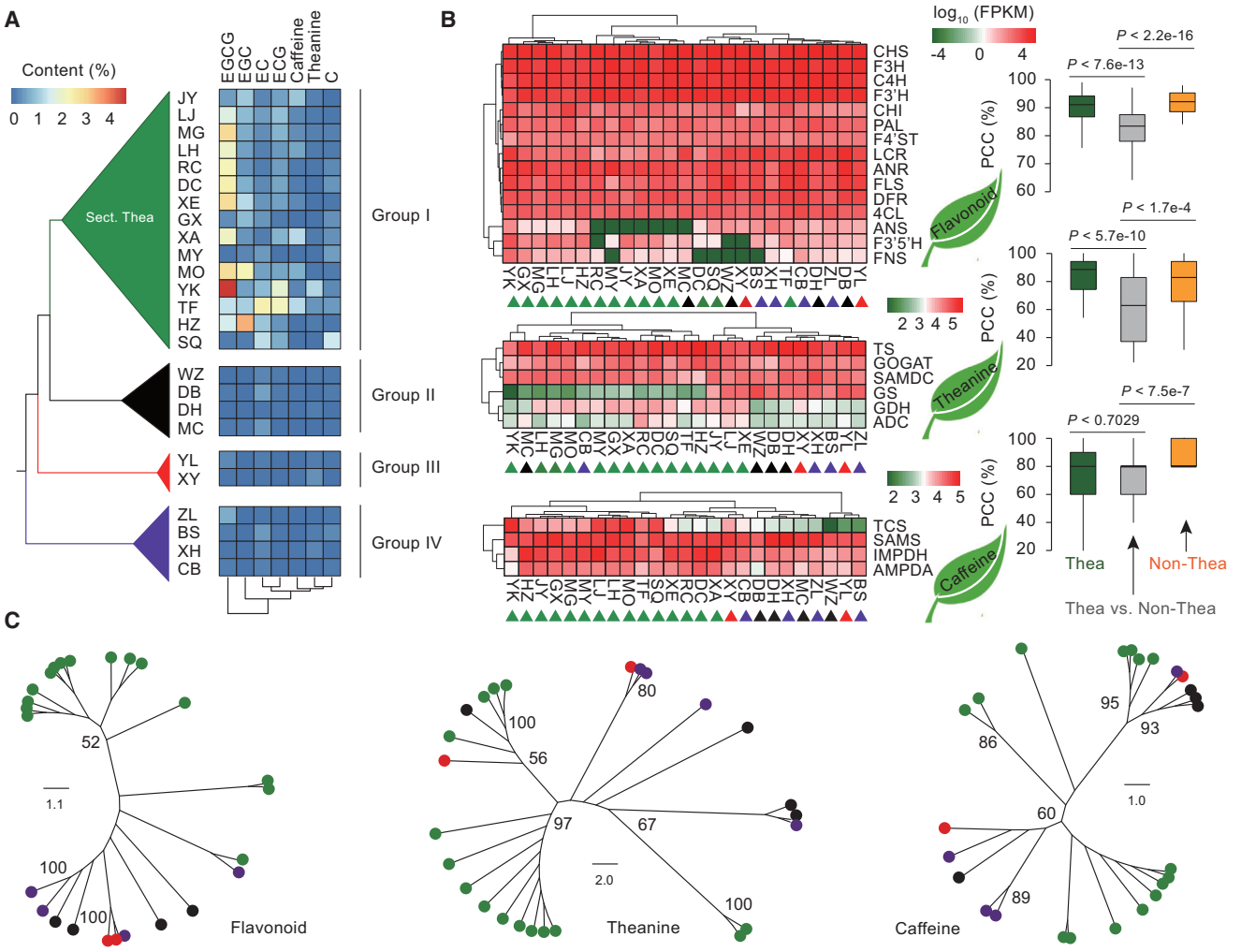

Figure 3. Evolutionary Differences of Three Important Metabolic Pathways Associated with Tea-Processing Suitability and Quality among the 25 *Camellia* Species.

(A) Left panel represents the phylogenetic relationship of the 25 *Camellia* species constructed using whole-transcriptome sequencing data. Right panel shows the percent content of seven characteristic metabolites detected in the leaves of each *Camellia* species using HPLC (see Supplemental Information for abbreviation details).

(B) Expression profiles in FPKM (fragments per kilobase per million reads mapped) of key functional genes (rows) for each species (columns) related to three metabolic pathways in the tea tree. Data are plotted as $\log_{10}$ values. Right box plot indicates the expression correlations within section *Thea* (Thea; green), non-*Thea* sections (Non-Thea; orange), or between Thea and Non-Thea (gray).

(C) Sequence variations of genes involved in flavonoid, theanine, and caffeine metabolic pathways. Orthologous genes in each *Camellia* species were identified and multi-aligned to construct their phylogenetic topology.

involved in the theanine metabolic pathway were expressed in all tissues, but were more highly expressed in the seedling, agreeing with previous findings (Deng et al., 2008) (Supplemental Figure 24). We also observed that genes encoding enzymes responsible for caffeine biosynthesis were highly expressed in seeds except for *TCS*, which is much more highly expressed in tender shoots and flowers than in other tissues, suggesting that caffeine may also be synthesized in seeds besides leaves (Supplemental Figure 25).

Independent Evolution of Tea Caffeine Biosynthesis

Caffeine (1,3,7-trimethylxanthine) is one of the most well-known purine alkaloids in plants (Ashihara and Crozier, 2001). It is synthesized by some eudicot plants, such as tea, coffee, cacao (*Theobroma cacao*), and maté (*Ilex paraguariensis*) from the holly family (Suzuki and Waller, 1988). The caffeine of the tea tree is synthesized from xanthosine via a key pathway that has three methylation steps catalyzed by SAM-dependent *N*-methyltransferases (NMTs) (Figure 4A) (Kato et al., 1996; Kato and Mizuno, 2004). With the aid of the completely sequenced tea tree genome, we identified a total of 13 NMT genes. We found that tea tree has fewer NMT genes than cacao (21) and coffee (23) (Figure 4B; Supplemental Tables 36 and 37). The gene expression profiles of NMTs at different tea tree developmental stages showed that most NMT genes (~77%) were prone to be expressed in the leaves and flowers—two primary tissues for caffeine accumulation—while the tender shoots exhibited a slightly higher gene expression level in comparison with young leaves (Supplemental Table 38 and Figure 4D). With the three

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

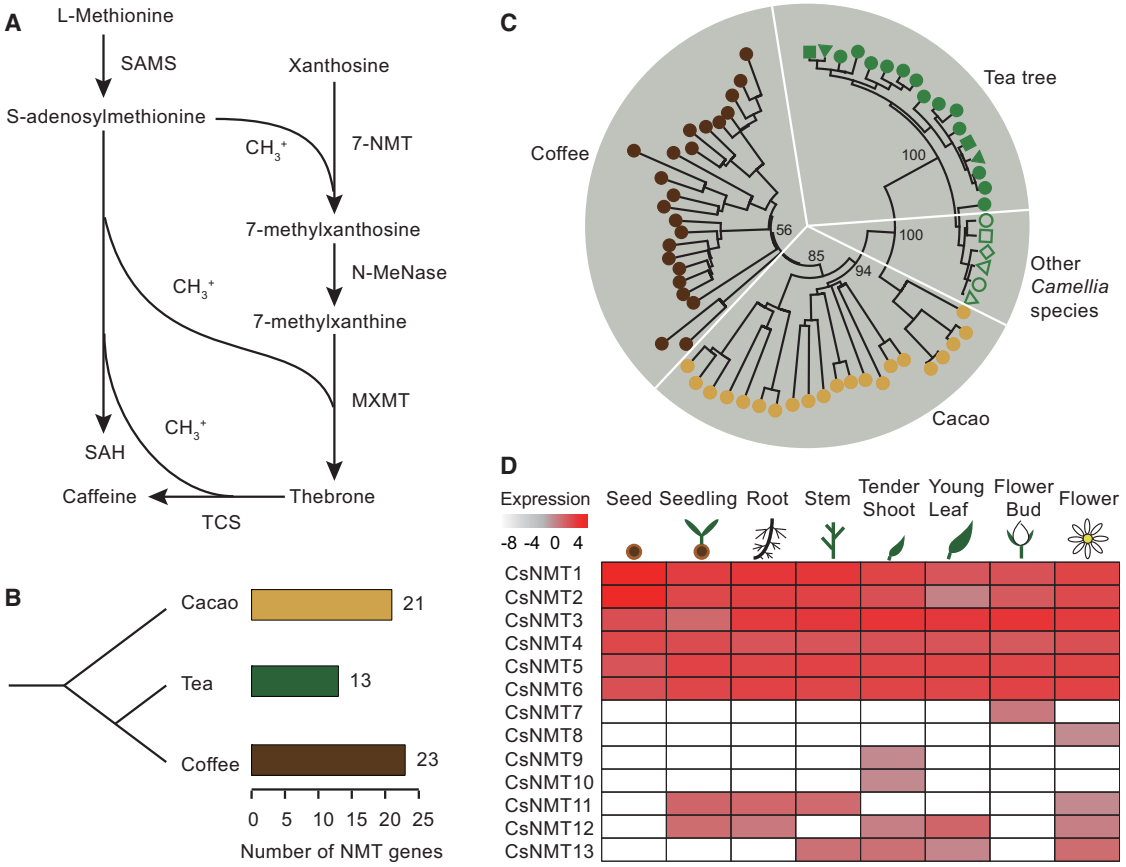

Figure 4. Evolution of Caffeine Biosynthesis.

(A) The most essential and last three methylation steps for caffeine biosynthesis in plants. These methylation steps are catalyzed by a series of *N*-methyltransferases (NMTs), including xanthosine methyltransferase (7-NMT), theobromine synthase (7-methylxanthine methyltransferase; MXMT), and caffeine synthase (3, 7-dimethylxanthine methyltransferase; TCS). SAMS represents *S*-adenosylmethionine synthetase, while SAH indicates *S*-adenosylhomocysteine.

(B) Distribution of genes encoding NMTs among tea tree, coffee, and cacao genomes.

(C) Neighbor-joining (NJ) phylogenetic tree of NMT genes from tea tree (green solid dots), coffee (brown solid dots), and cacao (orange solid dots) NMTs. The 10 NMT genes cloned in seven wild relatives of tea tree, including *C. irrawadiensis* (green solid squares), *C. ptilophylla* (green solid lower triangles), *C. granthamiana* (green circles), *C. lutchuensis* (green squares), *C. chrysantha* (green diamonds), *Camellia kissi* (green lower triangles) and *C. japonica* (green upper triangles), are also shown and listed in Supplemental Table 36. The phylogeny shows high bootstrap support for independent evolution of the 13 caffeine biosynthesis genes.

(D) Expression profiles of 13 tea tree NMT genes (rows) based on RNA-seq data from eight different tissues (columns) (see Supplemental Information for names of abbreviated NMT genes).

completely sequenced tea, coffee (Denoeud et al., 2014), and cacao (Argout et al., 2011) genomes now in hand, we are able to comprehensively investigate the evolutionary landscape of caffeine biosynthesis by comparing genome-wide sampling of NMT genes from coffee, cacao, and tea tree and its wild relatives. Phylogenetic analyses show that the NMTs from tea tree and cacao apparently separate from coffee with strong bootstrap support (Figure 4C and Supplemental Figure 26), indicating an independent evolution of the caffeine synthetic pathway in tea tree and cacao relative to coffee. Notably, all NMT genes from tea tree and its relatives form a single gene clade with a strong bootstrap support, and are monophyletic with the five NMT genes from cacao (Figure 4C and Supplemental Figure 26). This suggests that the caffeine synthetic pathway of the tea tree and its related *Camellia* species may have originated from a common tea tree–cacao ancestor but diverged later and evolved independently. We demonstrate an independent, recent, and rapid evolution of caffeine biosynthesis in the tea tree, supporting multiple origins of caffeine biosynthetic NMT activity as proposed previously (Kato and Mizuno, 2004; Pichersky and Lewinsohn, 2011; Denoeud et al., 2014; Huang et al., 2016).

DISCUSSION

We present a high-quality genome sequence for the cultivated tea tree. The tea tree offers advantages as an ideal system for functional genomics to understand the formation of a large number of secondary metabolites in many medicinal plants. This draft genome sequence thus provides the foundation for revealing the genetic basis of agronomically important traits and the characteristic physiological, medicinal, and nutritional properties of the tea tree. The availability of the first genome in the genus *Camellia* will facilitate in-depth fundamental comparative studies on tea tree

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

biology, addressing a wealth of questions about the *Camellia* gene and genome evolution. This is particularly important for enhancing the breeding programs of the most productive oil-bearing crop *C. oleifera* and the horticulturally distinguished camellias comprising *C. japonica*, *C. reticulata*, and *C. sasanqua*.

The genome sequence obtained reveals some of the unique biology of the tea tree. For instance, the tea tree possesses an extraordinarily large genome when compared with most sequenced plant species. We show that this results from the slow, steady, and long-term amplification of a few LTR retrotransposon families. It is possible that efficient DNA removal mechanisms (i.e., unequal homologous recombination and illegitimate recombination) are less prevalent in the tea tree genome, as previously described in other flowering plants (e.g., *P. abies* [Nystedt et al., 2013], *Amborella trichopoda* [Albert et al., 2013]), when compared with *A. thaliana* (Devos et al., 2002) and rice (Devos et al., 2002; Ma et al., 2004). We observe a positive relationship between expression levels of LTR retrotransposons and copy number of elements, in sharp contrast to an inverse correlation previously reported for maize (Meyers et al., 2001) and pineapple (Ming et al., 2015). DNA methylation differences may, instead, play a role in suppressing retrotransposition activation, leading to the increase of LTR retrotransposons in the tea tree genome (Mirouze et al., 2009). The detection of a relatively recent WGD that occurred in the tea tree and other relatives indicates the contribution of genome duplication to the evolution of the genus *Camellia*. Such a WGD event together with massive segmental duplications have potentially facilitated the expansion of gene families relevant to the activation of major secondary metabolic biosynthesis (e.g., flavonoids and terpenoids) as well as disease resistance and abiotic stress tolerance. The accumulation of abundant metabolic constituents, such as flavonoids and terpenoids, has apparently played a significant role in supporting environmental adaptations of the tea tree. The rapid expansion of disease resistance-related and abiotic stress tolerance-related genes suggests a strong selection for enhanced disease resistance in the tea tree that may be attributable to the potential adaptation to globally diverse habitats, providing a large number of candidate stress tolerance and disease resistance loci for further study to generate even more environmentally resilient tea tree varieties. We thus hypothesize that these genomic features enabled the tea tree to widely adapt to varied climates and become a ubiquitous worldwide beverage plant.

We have identified lineage-specific genes that likely control the quality of tea, in particular genes encoding enzymes involved in the flavonoids, theanine, and caffeine biosynthesis pathways. The tea tree-expanded genes related to flavonoid metabolic processes and terpene synthase activity that regulate tea flavor and quality are significantly enriched GO terms. Our comparative analyses indicate that the three major characteristic metabolic pathways are extremely conserved among the tea tree and other *Camellia* plants. Large amounts of catechins and caffeine in the tea tree and other members from section *Thea* is a feature that has distinguished these species from those from non-*Thea* sections. Although catechins, theanine, and caffeine are typically thought to be key characteristic metabolic compounds to determine tea-processing suitability and tea quality, tea flavor is also affected by many other known (e.g., terpenoids) and unknown secondary metabolic compounds. Extremely low contents of catechins and caffeine in the *Camellia* species from non-*Thea* sections, likely with some other particular secondary metabolites, degrade tea quality.

The high content of catechins and caffeine detected in the tea tree and other species belonging to section *Thea* provide the fundamental basis of tea flavor. This provides possible answers to the basic biological question of tea-processing suitability and addresses, among hundreds of *Camellia* species, why favorable flavor generated by section *Thea* plants have long been appreciated, from which the cultivated tea tree *C. sinensis* was domesticated in human history, instead of those from non-*Thea* sections. Our findings imply that wild relatives of cultivated tea tree present a huge reservoir for novel gene discovery toward the improvement of tea quality-related traits, whereas further studies are needed to dissect how differential expression of flavonoid and caffeine-related genes are related to regulating the accumulation of secondary metabolic compounds (Yan et al., 2015). The tea-processing industries in tea-drinking countries have developed numerous tea products with diverse tea flavor. Besides advances in tea-processing technologies, these depend on the development of a number of tea tree varieties containing diverse combinations of such characteristic secondary metabolites. We present an integrative data framework based on large-scale phytochemical, transcriptomic, and functional data, which will enable further metabolomic and functional genomic refinement of characteristic biosynthetic pathways including secondary metabolites, to form a more diversified set of tea flavors that will eventually satisfy and attract more tea drinkers worldwide. Considering the rapid extinction and severely endangered status of natural wild tea tree populations due to leaf over-harvesting creating tea products at high market prices (Liu et al., 2012), the genome assembly of the tea tree and transcriptomic variation data presented here will offer valuable information to aid the global conservation of these precious wild tea tree species.

METHODS

Plant Materials, DNA Extraction, and Library Construction

An individual plant of cultivar Yunkang 10 of *Camellia sinensis* var. *assamica* was collected in April 2009 from Menghai County, Yunnan Province, China. Fresh and healthy leaves were harvested and immediately frozen in liquid nitrogen after collection, followed by preservation at −80°C in the laboratory prior to DNA extraction. High-quality genomic DNA was extracted from the leaves using a modified CTAB method (Porebski et al., 1997). The quality and quantity of the isolated DNA were checked by electrophoresis on a 0.8% agarose gel and a NanoDrop D-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE), respectively. A total of 11 paired-end libraries, including four types of small-insert libraries (180 bp, 260 bp, 300 bp, 500 bp) and seven large-insert libraries (2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 8 kb, 20 kb), were prepared by following the instructions provided by Illumina.

Genome Sequencing, Assembly, and Quality Assessment

The constructed libraries were sequenced using the Illumina HiSeq2000 platform by following the standard Illumina protocols (Illumina, San Diego, CA). Raw sequencing data were preprocessed using Trimmomatic (version 0.33) (Bolger et al., 2014) to remove adapter sequences, potential contamination, and low-quality bases. High-quality sequencing data from short-insert size libraries were first assembled into contigs using Platanus (version 1.2.2) (Kajitani et al., 2014). Based on the mate-pair

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

reads from large-insert size libraries, the preassembled contigs were further elongated and eventually combined into scaffolds using SSPACE (Boetzer et al., 2011). The gaps within scaffolds were closed using GapCloser (version 1.12) (Luo et al., 2012). Haplomerger (version 20120810) (Huang et al., 2012) was finally used to remove potential heterozygous scaffolds, generating the final tea tree genome assembly. To evaluate the quality of genome assembly we used the three approaches including reads mapping, DNA alignment, and EST alignment.

Genome Annotation

Putative protein-coding genes were predicted using a combined strategy that integrates several *ab initio* gene predictors and sequence evidence, such as protein sequences from closely related plant species and the assembled ESTs in this study. Quality validation of gene models was evaluated by aligning transcriptome, EST, and homologous peptides to our predicted gene models. TEs were annotated by integrating RepeatMasker (www.repeatmasker.org), LTR_STRUCT (McCarthy and McDonald, 2003), RECON (Bao and Eddy, 2002), and LTR_Finder (Xu and Wang, 2007). Five different types of noncoding RNA genes, including miRNA, tRNA, rRNA, snoRNA, and snRNA genes, were predicted using *de novo* and/or homology search methods.

Comparative Genomic, Transcriptomic, and Phytochemical Analyses

We used the OrthoMCL package (version 2.0.9) (Li et al., 2003) to identify gene families/clusters between the tea tree and nine other plant species, including kiwifruit, potato, tomato, coffee, *Arabidopsis*, cacao, poplar, grape, and lotus. The species-specific gene families were determined according to the presence or absence of genes for a given species. We investigated the dynamic evolution of gene families using an updated version of Café (version 3.1) (Han et al., 2013) with a probabilistic graphical model. Phylogenetic relationship among these 10 plant species was resolved using the RAxML package (version 8.1.13) (Stamatakis, 2014) based on the 597 high-quality 1:1 single-copy orthologous genes. Divergence times among them were directly retrieved from TimeTree database (Hedges et al., 2015).

Fresh and healthy leaves from a total of 25 *Camellia* species, including 15 species from section *Thea* and 10 species from 10 representative non-*Thea* sections, were harvested from Jinhua International *Camellia* Germplasm Bank, Zhejiang Province, China on December 13, 2014. We measured the catechin, theanine, and caffeine content in the leaves of these 25 *Camellia* species using an Agilent 1100 HPLC system equipped with Agilent ZORBAX SB-C18 (4.6 × 250 mm, 5 μm).

The high-quality RNA was separately extracted from the fresh leaves of these 25 *Camellia* species using similar methods as described in Supplemental Section 2.1.2. Libraries were constructed by the following manufacturer's instructions and sequenced using the Illumina HiSeq2000 platform. We separately assembled the transcriptome for each species using Trinity (version r20140717) (Grabherr et al., 2011) with default parameters. To investigate the variation and evolution of the genes associated with the three (flavonoids, theanine, and caffeine) metabolic biosynthesis pathways in the 25 *Camellia* species, we annotated all genes encoding enzymes potentially involved in catalyzing the reactions of the flavonoid, theanine, and caffeine pathways in our assembled tea tree genome. Among them, we only focused on the 23 important genes that are further validated by the cloned gene sequences in diverse *Camellia* species that were submitted to the NCBI database. To calculate the expression levels of genes associated with three metabolic biosynthesis pathways in the 25 *Camellia* species, we mapped all clean RNA-seq reads generated from each species to the tea tree genome using TopHat (version 2.1.0) (Trapnell et al., 2009) with default parameters. The generated BAM format alignments together with gene GTF annotation file were then fed to Cuffdiff (version 2.2.1) (Trapnell et al., 2010) to compute the FPKM values for each gene. We visualized the gene expression levels using the "pheatmap" package (https://cran.r-project.org/web/packages/pheatmap/index.html) implemented in R.

Further detailed methods for data analyses are fully provided in Supplemental Information.

ACCESSION NUMBERS

Raw Illumina sequencing reads of tea tree have been deposited in the NCBI Sequence Read Archive Database under accession PRJNA381277. Genome assembly, gene prediction, gene functional annotations, and transcriptomic data may be accessed via the web site at: www.plantkingdomgdb.com/tea_tree/.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by the project of Yunnan Innovation Team Project, the Hundreds Oversea Talents Program of Yunnan Province, the Top Talents Program of Yunnan Province (Grant 20080A009), the Key Project of the Natural Science Foundation of Yunnan Province (201401PC00397), National Science Foundation of China (U0936603), Key Project of Natural Science Foundation of Yunnan Province (2008CC016), Frontier Grant of Kunming Institute of Botany, CAS (672705232515), Top Talents Program of Yunnan Province (20080A009), and Hundreds Talents Program of Chinese Academy of Sciences (CAS) (to L.G.).

AUTHOR CONTRIBUTIONS

L.-Z.G. designed and managed the project; H.-B.Z., T.Z., Y.T., S.-Y.M., J.-Y.J., D.Z., J.-J.J., L.-P.Z., B.-Y.L., H.N., S.-F.S., Y.Z. (Yuan Zhao), and C.S. (Cong Shi) collected materials; H. Z., Y.Y., and Ch.S. (Chao Shi) prepared and purified DNA samples; K.L. and C.K. performed the genome assembly; E.-H.X., Q.-J.Z., Y.L., Y.Z., W.L., H. Z., Yun Z. (Yun Zhang), and Y.-L.L. performed genome annotation and subsequent data analyses; H.H. performed flow cytometry analysis; L.-Z.G. and E.-H.X. wrote the paper. L.-Z.G., E.-H.X., J.S., D.-J.N., and E.E.E. revised the paper.

ACKNOWLEDGMENTS

E.E.E. is an investigator of the Howard Hughes Medical Institute. E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. and is a consultant for Kunming University of Science and Technology (KUST) as part of the 1000 China Talent Program. No conflict of interest declared.

Received: March 29, 2017
Revised: April 5, 2017
Accepted: April 7, 2017
Published: May 1, 2017

REFERENCES

Albert, V.A., Barbazuk, W.B., dePamphilis, C.W., Der, J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C., et al. (2013). The *Amborella* genome and the evolution of flowering plants. Science **342**:1–10.

Argout, X., Salse, J., Aury, J.M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., et al. (2011). The genome of *Theobroma cacao*. Nat. Genet. **43**:101–108.

Ashihara, H., and Crozier, A. (2001). Caffeine: a well known but little mentioned compound in plant science. Trends Plant Sci. **6**:407–413.

Banerjee, B. (1992). Botanical Classification of Tea (Dordrecht: Springer Netherlands).

Bao, Z., and Eddy, S.R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. **12**:1269–1276.

Please cite this article in press as: Xia et al., The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis, Molecular Plant (2017), http://dx.doi.org/10.1016/j.molp.2017.04.002

Bennett, M.D. (2004). Perspectives on polyploidy in plants—ancient and neo. Biol. J. Linn. Soc. **82**:411–423.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**:578–579.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**:2114–2120.

Cabrera, C., Artacho, R., and Gimenez, R. (2006). Beneficial effects of green tea—a review. J. Am. Coll. Nutr. **25**:79–99.

Chacko, S.M., Thambi, P.T., Kuttan, R., and Nishigaki, I. (2010). Beneficial effects of green tea: a literature review. Chin. Med. **5**:13.

Chen, S.D., Krinsky, B.H., and Long, M.Y. (2013). New genes as drivers of phenotypic evolution. Nat. Rev. Genet. **14**:645–660.

Cui, L.L., Yao, S.B., Dai, X.L., Yin, Q.G., Liu, Y.J., Jiang, X.L., Wu, Y.H., Qian, Y.M., Pang, Y.Z., Gao, L.P., et al. (2016). Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (*Camellia sinensis*). J. Exp. Bot. **67**:2285–2297.

Deng, W., Ogita, S., and Ashihara, H. (2008). Biosynthesis of theanine (γ-ethylamino-l-glutamic acid) in seedlings of *Camellia sinensis*. Phytochemistry Lett. **1**:115–119.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C.F., Alberti, A., Anthony, F., Aprea, G., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science **345**:1181–1184.

Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. **12**:1075–1079.

Gao, L., McCarthy, E.M., Ganko, E.W., and McDonald, J.F. (2004). Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequences. BMC Genomics **5**:1–18.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. **29**:644–652.

Han, M.V., Thomas, G.W., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. **30**:1987–1997.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. **32**:835–845.

Heiss, M.L., and Heiss, R.J. (2007). The Story of Tea: A Cultural History and Drinking Guide (New York: Random House).

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., and Xu, A. (2012). HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. **22**:1581–1588.

Huang, S.X., Ding, J., Deng, D.J., Tang, W., Sun, H.H., Liu, D.Y., Zhang, L., Niu, X.L., Zhang, X., Meng, M., et al. (2013). Draft genome of the kiwifruit *Actinidia chinensis*. Nat. Commun. **4**:2640.

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

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**:463–467.

Jones, J.D.G., and Dangl, J.L. (2006). The plant immune system. Nature **444**:323–329.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. **24**:1384–1395.

Kato, M., and Mizuno, K. (2004). Caffeine synthase and related methyltransferases in plants. Front. Biosci. **9**:1833–1842.

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

Khan, N., and Mukhtar, H. (2007). Tea polyphenols for health promotion. Life Sci. **81**:519–533.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**:2178–2189.

Lim, E.K., and Bowles, D.J. (2004). A class of plant glycosyltransferases involved in cellular homeostasis. EMBO J. **23**:2915–2922.

Liu, Y.J., Gao, L.P., Liu, L., Yang, Q., Lu, Z.W., Nie, Z.Y., Wang, Y.S., and Xia, T. (2012). Purification and characterization of a novel galloyltransferase involved in catechin galloylation in the tea plant (*Camellia sinensis*). J. Biol. Chem. **287**:44406–44417.

Liu, Y., Wang, D., Zhang, S., and Zhao, H. (2015). Global expansion strategy of Chinese herbal tea beverage. Adv. J. Food Sci. Technol. **7**:739–745.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., and Liu, Y. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **1**:18.

Ma, J.X., Devos, K.M., and Bennetzen, J.L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. **14**:860–869.

McCarthy, E.M., and McDonald, J.F. (2003). LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics **19**:362–367.

McKain, M.R., Tang, H., McNeal, J.R., Ayyampalayam, S., Davis, J.I., depamphilis, C.W., Givnish, T.J., Pires, J.C., Stevenson, D.W., and Leebens-Mack, J.H. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. Genome Biol. Evol. **8**:1150–1164.

Meyers, B.C., Tingley, S.V., and Morgante, M. (2001). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. **11**:1660–1676.

Ming, T., and Bartholomew, B. (2007). Theaceae. In Flora of China, Z. Wu, P. Raven, and D. Hong, eds. (Beijing and St. Louis: Science Press and Missouri Botanical Garden), pp. 367–412.

Ming, R., VanBuren, R., Wai, C.M., Tang, H.B., Schatz, M.C., Bowers, J.E., Lyons, E., Wang, M.L., Chen, J., Biggers, E., et al. (2015). The pineapple genome and the evolution of CAM photosynthesis. Nat. Genet. **47**:1435–1442.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J., and Mathieu, O. (2009). Selective epigenetic control of retrotransposition in *Arabidopsis*. Nature **461**:427–430.

Mitreva, M., Jasmer, D.P., Zarlenga, D.S., Wang, Z.Y., Abubucker, S., Martin, J., Taylor, C.M., Yin, Y., Fulton, L., Minx, P., et al. (2011). The draft genome of the parasitic nematode *Trichinella spiralis*. Nat. Genet. **43**:228–235.

Mondal, T.K., Bhattacharya, A., Laxmikumaran, M., and Ahuja, P.S. (2004). Recent advances of tea (*Camellia sinensis*) biotechnology. Plant Cell Tissue Organ. Cult. **76**:195–254.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The *Norway spruce* genome sequence and conifer genome evolution. Nature **497**:579–584.

Ohno, S. (1970). Introduction. In Evolution by Gene Duplication (Berlin, Heidelberg: Springer), pp. 1–2.

Pichersky, E., and Lewinsohn, E. (2011). Convergent evolution in plant specialized metabolism. Annu. Rev. Plant Biol. **62**:549–566.

Porebski, S., Bailey, L.G., and Baum, B.R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol. Biol. Rep. **15**:8–15.

Qin, C., Yu, C.S., Shen, Y.O., Fang, X.D., Chen, L., Min, J.M., Cheng, J.W., Zhao, S.C., Xu, M., Luo, Y., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proc. Natl. Acad. Sci. USA **111**:5135–5140.

Rogers, P.J., Smith, J.E., Heatherley, S.V., and Pleydell-Pearce, C.W. (2008). Time for tea: mood, blood pressure and cognitive performance effects of caffeine and theanine administered alone and together. Psychopharmacology **195**:569–577.

Salman-Minkov, A., Sabath, N., and Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. Nat. Plants **2**:16115.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**:635–641.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313.

Suzuki, T., and Waller, G.R. (1988). Metabolism and analysis of caffeine and other methylxanthines in coffee, tea, cola, guarana and cacao. In Analysis of Nonalcoholic Beverages, H.-F. Linskens and J.F. Jackson, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 184–220.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-seq. Bioinformatics **25**:1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**:511–515.

Weinberg, B.A., and Bealer, B.K. (2001). The World of Caffeine: The Science and Culture of the World's Most Popular Drug (Abingdon, UK: Psychology Press).

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**:W265–W268.

Xu, X., Pan, S.K., Cheng, S.F., Zhang, B., Mu, D.S., Ni, P.X., Zhang, G.Y., Yang, S., Li, R.Q., Wang, J., et al. (2011). Genome sequence and analysis of the tuber crop potato. Nature **475**:189–195.

Yamamoto, T., Juneja, L.R., and Kim, M. (1997). Chemistry and Applications of Green Tea (New York: CRC Press).

Yan, W., Wang, J., Zhou, Y., Zhao, M., Gong, Y., Ding, H., Peng, L., and Hu, D. (2015). Genome-wide identification of genes probably relevant to the uniqueness of tea plant (*Camellia sinensis*) and its cultivars. Int. J. Genomics http://dx.doi.org/10.1155/2015/527054.

Zwenger, S., and Basu, C. (2008). Plant terpenoids: applications and future potentials. Biotechnol. Mol. Biol. Rev. **3**:1–7.