Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality

Chaoling Wei[a,1], Hua Yang[a,1], Songbo Wang[b,1], Jian Zhao[a,1], Chun Liu[b,1], Liping Gao[a,1], Enhua Xia[a], Ying Lu[c], Yuling Tai[a], Guangbiao She[a], Jun Sun[a], Haisheng Cao[a], Wei Tong[a], Qiang Gao[b], Yeyun Li[a], Weiwei Deng[a], Xiaolan Jiang[a], Wenzhao Wang[a], Qi Chen[a], Shihua Zhang[a], Haijing Li[a], Junlan Wu[a], Ping Wang[a], Penghui Li[a], Chengying Shi[a], Fengya Zheng[b], Jianbo Jian[b], Bei Huang[a], Dai Shan[b], Mingming Shi[b], Congbing Fang[a], Yi Yue[a], Fangdong Li[a], Daxiang Li[a], Shu Wei[a], Bin Han[d], Changjun Jiang[a], Ye Yin[b], Tao Xia[a], Zhengzhu Zhang[a], Jeffrey L. Bennetzen[a,e,2], Shancen Zhao[b,2], and Xiaochun Wan[a,2]

[a]State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, 230036 Hefei, China; [b]BGI Genomics, BGI–Shenzhen, 518083 Shenzhen, China; [c]College of Fisheries and Life Science, Shanghai Ocean University, 201306 Shanghai, China; [d]National Center for Gene Research, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 200032 Shanghai, China; and [e]Department of Genetics, University of Georgia, Athens, GA 30602

Contributed by Jeffrey L. Bennetzen, March 16, 2018 (sent for review November 17, 2017; reviewed by Xiuxin Deng and Harry J. Klee)

Tea, one of the world's most important beverage crops, provides numerous secondary metabolites that account for its rich taste and health benefits. Here we present a high-quality sequence of the genome of tea, *Camellia sinensis* var. *sinensis* (CSS), using both Illumina and PacBio sequencing technologies. At least 64% of the 3.1-Gb genome assembly consists of repetitive sequences, and the rest yields 33,932 high-confidence predictions of encoded proteins. Divergence between two major lineages, CSS and *Camellia sinensis* var. *assamica* (CSA), is calculated to ∼0.38 to 1.54 million years ago (Mya). Analysis of genic collinearity reveals that the tea genome is the product of two rounds of whole-genome duplications (WGDs) that occurred ∼30 to 40 and ∼90 to 100 Mya. We provide evidence that these WGD events, and subsequent paralogous duplications, had major impacts on the copy numbers of secondary metabolite genes, particularly genes critical to producing three key quality compounds: catechins, theanine, and caffeine. Analyses of transcriptome and phytochemistry data show that amplification and transcriptional divergence of genes encoding a large acyltransferase family and leucoanthocyanidin reductases are associated with the characteristic young leaf accumulation of monomeric galloylated catechins in tea, while functional divergence of a single member of the glutamine synthetase gene family yielded theanine synthetase. This genome sequence will facilitate understanding of tea genome evolution and tea metabolite pathways, and will promote germplasm utilization for breeding improved tea varieties.

comparative genomics | genome evolution | catechins biosynthesis | theanine biosynthesis | tea quality

Tea is, after water, the world's most popular beverage, and offers a wealth of health benefits. Tea consumption possesses a history of nearly 5,000 y (1, 2). *Camellia sinensis* (L.) O. Kuntze ($2n = 2x = 30$), whose leaves are used to produce numerous kinds of tea, is a member of the Theaceae family of angiosperms. The tea plant was an endemic species in southwest China, and now is cultivated all over the world (3, 4). In the last decade, worldwide tea production has increased by ∼66% in acreage, and has reached 5.3 million tons on 3.5 million hectares across 50 tea-growing countries (Food and Agriculture Organization of the United Nations statistics; www.fao.org/faostat/). The combination of genetic variation, environmental factors, and various modes of tea processing has created a huge variety of tea products with diverse palatability, such as bitter, astringent, and sweet flavors, to meet consumer demand across the world.

Cultivated tea plant varieties mainly belong to two major groups: *Camellia sinensis* var. *sinensis* (CSS; Chinese type) and *Camellia sinensis* var. *assamica* (CSA; Assam type), with the former as the most widely distributed cultivar in China and in the world. The two tea plant types have distinct characteristics. CSS is a slower-growing shrub with a small leaf and is able to withstand colder climates, while CSA is quick-growing with large leaves and high sensitivity to cold weather. CSA thus is mainly cultivated in very warm tropical areas, distinct from the broader geography of CSS cultivation (5, 6). In agricultural practice, CSS

Significance

A high-quality genome assembly of *Camellia sinensis* var. *sinensis* facilitates genomic, transcriptomic, and metabolomic analyses of the quality traits that make tea one of the world's most-consumed beverages. The specific gene family members critical for biosynthesis of key tea metabolites, monomeric galloylated catechins and theanine, are indicated and found to have evolved specifically for these functions in the tea plant lineage. Two whole-genome duplications, critical to gene family evolution for these two metabolites, are identified and dated, but are shown to account for less amplification than subsequent paralogous duplications. These studies lay the foundation for future research to understand and utilize the genes that determine tea quality and its diversity within tea germplasm.

Author contributions: C.W., J.Z., B. Han, C.J., T.X., Z.Z., J.L.B., S. Zhao, and X.W. designed research; H.Y., S. Wang, C.L., Y.T., G.S., J.S., H.C., Y. Li, X.J., H.L., P.L., C.S., B. Huang, and Y. Yin performed research; E.X., Y. Lu, Q.G., J.J., D.S., and S. Wei contributed new reagents/analytic tools; W.T., W.D., W.W., Q.C., S. Zhang, J.W., P.W., F.Z., M.S., C.F., Y. Yue, F.L., and D.L. analyzed data; and C.W., H.Y., S. Wang, J.Z., C.L., L.G., E.X., J.L.B., S. Zhao, and X.W. wrote the paper.

Reviewers: X.D., Huazhong Agricultural University; and H.J.K., University of Florida.

The authors declare no conflict of interest.

Data deposition: Raw Illumina sequencing reads of tea tree reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive database (accession no. SRA536878). Genome assembly, gene prediction, gene functional annotations, and transcriptomic data can be accessed at and downloaded from pcsb.ahau.edu.cn:8080/CSS/. The sequence data of five serine carboxypeptidase-like genes have been deposited at pcsb.ahau.edu.cn:8080/CSS/.

[1]C.W., H.Y., S. Wang, J.Z., C.L., and L.G. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: xcwan@ahau.edu.cn, zhaoshancen@genomics.cn, or maize@uga.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1719622115/-/DCSupplemental.

Published online April 20, 2018.

can be grown in high-latitude areas for quality green tea production, while CSA is usually processed into black tea (7). Most current elite tea plant cultivars in China (~67%) belong to CSS, and CSS provides the germplasm for most of the recent increase in tea production (8).

The rich flavors and various health-promoting functions of tea are mainly attributable to ~700 bioactive compounds (9). Of these, the most characteristic are catechins (a subgroup of flavan-3-ols), theanine, caffeine, and volatiles. The catechins contribute 12 to 24% of dry weight in young leaves, and are the principal flavonoids. Catechins in tea consist of a mixture of (+)-catechins (C), (−)-epicatechin (EC), (+)-gallocatechin (GC), (−)-epigallocatechin (EGC), (−)-epicatechin-3-gallate (ECG), and (−)-epigallocatechin-3-gallate (EGCG) (10). The most active and abundant catechin is EGCG in green tea, while in black tea the catechins are polymerized to theaflavins and thearubigins by a "fermentation" that leads to catechins oxidation (11). Catechins mainly confer astringent taste to tea, while caffeine offers a bitter taste. The nonprotein amino acid theanine (γ-glutamylethylamide) contributes to the umami and sweet tastes of tea infusions, and also has been associated with relaxation and neuroprotection (12). Theanine accounts for more than 50% of total free amino acids and 1 to 2% of the dry weight of tea leaves. Tea plants also synthesize volatile terpenoids and phenolics in the form of glycosides. Their hydrolytic products, together with lipid and carotenoid oxidation products released during tea processing, provide different types of teas with a variety of pleasant scents and characteristic flavors.

In the last two decades, efforts have been made to study the biosynthesis of these natural compounds in tea plants from biochemical, physiological, or molecular genetic perspectives (13, 14). However, the genetic bases for the rich production of all types of these bioactive compounds in tea, and for the responses to disease and insects that threaten tea production, are not yet understood. The lack of a reference genome sequence is a major obstacle for basic and applied biology on the tea plant. While recently a preliminary draft genome of a CSA cultivar was reported (15), no CSS genome has yet been reported. We herein present a high-quality draft genome sequence of CSS, and provide information on how tea plants produce the abundant and diverse flavonoids and theanine that synergistically contribute to tea palatability and health benefits.

Results and Discussion

Assembly and Annotation of the CSS Genome and Comparison with the CSA Genome. The tea plant is self-incompatible and thus highly heterozygous, which elevates the complexity of genome sequencing and assembly. A panel of tea plants was screened by restriction site associated DNA sequencing (RAD-seq) (16) (Dataset S1), and the commercial variety "Shuchazao" (accession no. GS2002008; Agricultural Plant Variety Name Retrieval System, Ministry of Agriculture and Rural Affairs, China) was found to exhibit a relatively low level of heterozygosity (2.7%). Hence, Shuchazao was selected for genome sequencing. Its genome size was estimated to be ~2.98 Gb by flow cytometry and *K*-mer analyses (*SI Appendix*, Fig. S1), consistent with that of an earlier report for CSA (15).

We generated a total of 1,325 Gb (~436-fold genome coverage) of clean Illumina reads from 20 representative sequencing libraries (Dataset S2). The reads were iteratively assembled into contigs and subsequently into scaffolds using a modified assembly pipeline (*SI Appendix*, section 1.4). For gap closure, we further generated ~125.4 Gb of long reads using the single-molecule real-time (SMRT) sequencing platform (Dataset S3). The final assemblies are ~3.1 Gb that consists of 2.89 Gb of contigs (coverage 93%). Contig and scaffold N50 lengths are 67.07 kb and 1.39 Mb, respectively (Table 1). The accuracy and completeness of the assembly were assessed by comparison with bacterial artificial chromosome (BAC) sequences derived from Shuchazao, other DNA sequences, and expressed sequence tags (ESTs) generated by Sanger sequencing (Datasets S4–S6). The 18 fully assembled BACs were separately represented by 1 to 5 scaffolds each (median 3) in our assembly and 5 to 50 scaffolds each (median 9) in the previous CSA assembly. Overall coverage of these BACs was 98.3% for our assembly and 84.6% for the earlier CSA assembly (*SI Appendix*, Fig. S2). PCR was used to investigate the discordance between CSS and CSA assemblies. Ten primer pairs were designed for regions that were identical in both the CSS and CSA genotypes, and all of these were found to yield the appropriate size PCR fragment with DNA from either cultivar. A further 14 primer pairs were designed for regions that were adjacent in CSS but not predicted to be amplifiable by PCR in CSA. Of these, 12 yielded the appropriate size PCR fragment in CSS and 7 yielded approximately the same size PCR fragment in CSA. A final 10 primer pairs were designed for regions that were predicted to be adjacent in CSA but not predicted to be amplifiable by PCR in CSS. Of these, 7 yielded the appropriate size PCR fragment in CSA and 1 yielded the predicted size PCR fragment in CSS (Dataset S7). Hence, these results indicate that regions with collinearity between the CSS and CSA assemblies are most likely to be correct, while about half of the differences in collinearity would be errors in at least one of the assemblies, with errors in the CSA assembly about three times more likely than errors in the CSS assembly. In addition, analysis of the conserved core eukaryotic genes derived from the benchmarking universal single-copy ortholog (BUSCO) plant lineage indicated that 2% of these genes were missing from our assembly and slightly over 5% were missing from the earlier CSA assembly (Dataset S7).

Table 1. Global statistics for assembly and annotation of two tea genomes

Characteristics	CSS	CSA
Scaffold length, Gb	3.14	3.02
Contig length, Gb	2.89	2.58
Scaffold number, >2 kb	14,051	22,255
Longest scaffold, Mb	7.31	3.51
Contig N50 length, kb	67.07	19.96
Scaffold N50 length, Mb	1.39	0.45
Transposons		
TE quantity, Gb (% of genome)	1.86 (64)	1.75 (58)
BUSCO		
Missing core genes (%)	6 (2.0)	16 (5.2)
Genome annotation		
Protein-coding genes	33,932	36,951
Average gene length, bp	4,053	3,549
Average exon length, bp	259	237

Transposable elements (TEs) are the chief mechanistic drivers of genome evolution (17). We found that the CSS genome harbored a total of 1.86 Gb of TEs, representing at least 64% of the assembly (excluding undefined base Ns). This number is larger than that identified in the CSA genome (~58%), suggesting that PacBio data facilitated the construction of repetitive regions in the CSS genome. Long terminal repeat (LTR) retrotransposons encompass ~58% of the CSS assembly. Of them, the *Gypsy* and *Copia* superfamilies made up ~46 and ~8% of the tea genome, respectively (Fig. 1*A* and Datasets S8 and S9).

To annotate gene models and quantify their expression, we generated an average of 11.8 Gb of clean RNA-seq (sequencing) data for each of eight primary tissue types (Dataset S11). We also obtained 80,217 transcripts with an average length of 1,781 bp using the SMRT sequencing platform (Datasets S12 and S13). Integrative gene-finding algorithms predicted 33,932 high-confidence gene models (Fig. 1*A* and Datasets S14

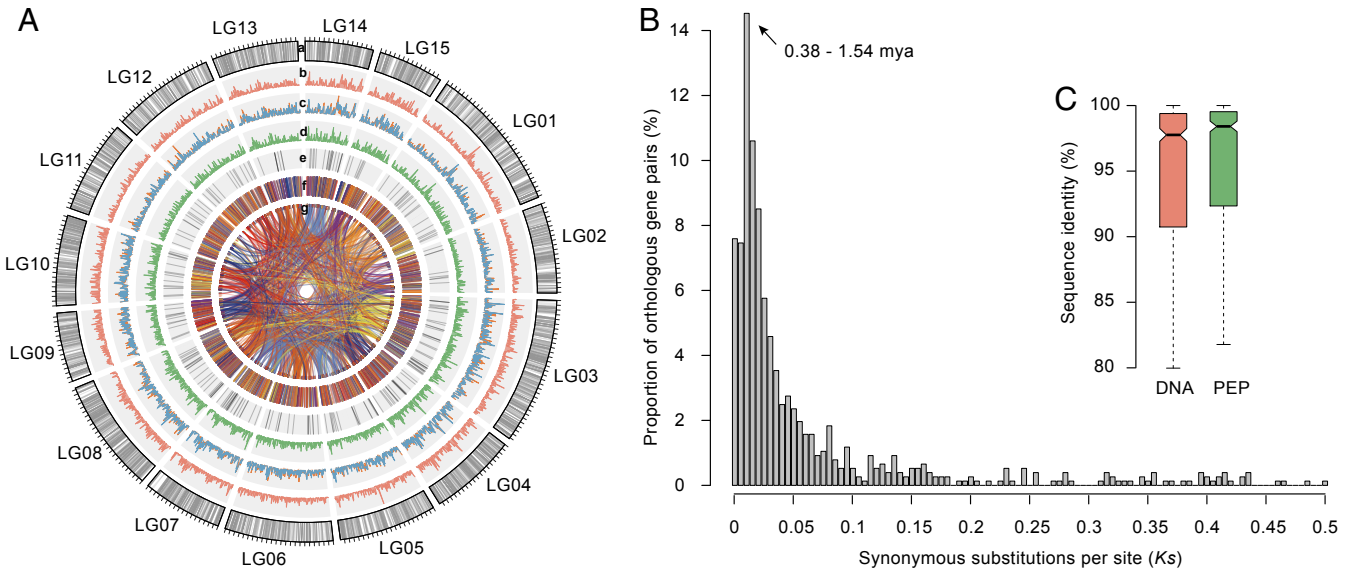

Fig. 1. Landscape of the tea plant genome. (*A*) Global view. SNP markers in the genetic map (gray; a); simple sequence repeat density (orange; b); TE density (c; *Copia*, orange; *Gypsy*, blue); gene density (green; d); transcription factors (black; e); genes in syntenic blocks between tea and grape (each color for syntenic genes from each grape chromosome; f); and oriented paralogous genes (g). A 1-Mb sliding window was used to calculate the density of different elements. (*B*) Estimation of divergence time between CSS and CSA using orthologous gene pairs within collinear blocks. (*C*) DNA and protein sequence similarity of orthologous genes between CSS and CSA. The error bar indicates the maximum and minimum sequence similarity values of orthologous genes.

and S15). More than 92% of the genes were assigned to a suite of function databases (Dataset S16). Of them, 2,486 transcription factor (TF) genes were predicted and classified into 86 different families (Dataset S17).

A total of 121 syntenic blocks that contain 1,543 collinear genes were identified between the CSS and CSA genomes. Calculation of the synonymous substitution rate (K_s) distribution of these gene pairs peaked at 0.005 to 0.02. This suggests that the sequenced haplotypes of CSA and CSS diverged from their common ancestor 0.38 to 1.54 Mya based on a universal substitution rate of 6.5×10^{-9} mutations per site per year (18) (Fig. 1*B*). The average sequence similarity of orthologous genes at DNA and protein levels was 92.4% (median 97.8%) and 93.9% (median 98.4%), respectively (Fig. 1*C*).

Evolution of the Tea Genome and Secondary Metabolite-Associated Genes. The evolutionary dynamics of gene families were analyzed by comparing the tea plant genome with those of 10 representative plant species. A total of 15,224 candidate gene families were identified in the tea genome, of which 429 were tea-specific and 11,128 were shared among all 11 species (*SI Appendix*, Fig. S3). New gene copies were gained within 1,810 families, whereas 1,001 families contracted in the tea genome (*SI Appendix*, Fig. S4). Functional exploration of the tea-specific gene families indicated that domains such as cytochrome P450, NB-ARC, and TFs were markedly amplified in the tea genome [false discovery rate (FDR) < 0.05]. Prominent volatile compounds, crucial for tea aroma and flavor, are derived either by oxidation of lipids and carotenoids or from the terpenoid and shikimate pathways (19). Interestingly, we found that gene families encoding enzymes associated with the biosynthesis of these key metabolites, such as 1-deoxy-D-xylulose-5-phosphate synthase, (−)-germacrene D synthase, α-farnesene synthase, isoprene synthase, geranyl linalool synthase, and (−)-α-terpineol synthase, were also expanded, often in a species-specific manner, in the tea genome (FDR < 0.05) (Datasets S23 and S24).

It is well-documented that whole-genome duplication (WGD) events have been frequent in the evolutionary history of flowering plants and generally shaped the evolutionary trajectory of genomes and genes, in particular those genes associated with agronomic and/or plant-specialized phenotypic traits. With the alignment of more than 32,000 gene models on large scaffolds (>10 kb) to grape gene models, we identified 2,706 grape–tea syntenic gene blocks (Dataset S25), containing >15,894 tea genes. The gene collinearity suggested that CSS carried two duplications compared with the diploid grape genome, which has had no WGD since the γ-event ∼140 Mya (*SI Appendix*, Fig. S5). This indicates that two additional rounds of WGD events occurred in the CSS lineage. We found that ∼14% of the total collinear blocks are from an ancient event (∼90 to 100 Mya; *SI Appendix*, Figs. S5 and S6) that is shared with kiwifruit (20). The long retained homeologous blocks from this event are mostly located in tea genome regions orthologous to grape chromosomes G14, G18, and G19 (*SI Appendix*, Fig. S5). The recent WGD event of tea at ∼30 to 40 Mya occurred after the divergence of tea plant and kiwifruit lineages (∼80 Mya) (*SI Appendix*, Figs. S6 and S7). This most recent WGD was further evidenced by the distribution of distance–transversion rates at fourfold degenerate sites (4dTv) of tea plant syntenic genes (*SI Appendix*, Fig. S8).

To understand the role of WGD events in the evolution of tea plant genes associated with biosynthesis of characteristic secondary metabolites, we systematically depicted the evolutionary landscape of secondary metabolite (SM) genes between tea and six other representative plant species, including kiwifruit, coffee, cacao, *Arabidopsis*, poplar, and grape. Based on the age distribution of K_s of these duplicated SM gene pairs, the relative contributions of WGD and paralogous duplication responsible for these gene family expansions could be determined (*SI Appendix*, Fig. S9 and Dataset S26). We found that, although the SM genes initially duplicated in the common ancestor of these investigated plant species, most amplification of the plant SM genes occurred after their lineages diverged (*SI Appendix*, Fig. S9).

Evolution of the Catechins Biosynthesis Pathway. Catechins are derived from the phenylpropanoid pathway. The key gene families involved in catechins biosynthesis, such as chalcone synthase (CHS), chalcone isomerase (CHI), dihydroflavonol reductase

(DFR), leucoanthocyanidin reductase (LAR), and anthocyanidin reductase (ANR) (21, 22), were identified in our assemblies (Dataset S27). In the tea genome, WGDs have considerably impacted the genes involved in catechins biosynthesis pathways (*SI Appendix*, Fig. S9). An acyltransferase gene family belonging to subclade 1A of serine carboxypeptidase-like (SCPL) acyltransferases that are widely involved in plant secondary metabolism has been found extensively expanded in the tea genome. Of 22 *SCPL* genes found in the tea genome, 4 were generated through the WGD event that occurred ~30 to 40 Mya, and 15 were amplified via tandem duplication that was mostly recent and species-specific (Dataset S26). As expected, given the high rate of paralogue generation in many gene families, we found duplications of upstream genes involved in phenylpropanoid synthesis and downstream genes specifically involved in the catechins biosynthesis pathway that were in addition to those generated by the WGD events (*SI Appendix*, Fig. S9). One pair of *DFR*, two pairs of *CHS*, and one pair of *SCPL* genes exhibit divergences indicating that they were duplicated within the last 2 to 25 My. Relatively few duplications in these gene families can be traced to the most recent WGD, suggesting that the ancient WGD and recent paralogous duplication were most important to generating the diversity leading to the evolution of the catechins pathway in the tea lineage.

Acyltransferase Family Specially Evolved for the Biosynthesis of Galloylated Catechins. Metabolite profiling and transcriptome analysis were conducted to help further understand the evolution of catechins biosynthesis in tea plant tissues. The *cis*-flavan-3-ols (ECs) were found to accumulate in most of the examined tissues. Galloylated catechins (ECG, EGCG) account for up to 80% of total catechins in apical buds, and thus have a major impact on tea quality (Dataset S30). Although the genetic basis for biosynthesis of high levels of the galloylated catechins is not fully understood, they have been proposed to be synthesized from flavan-3-ols via 1-*O*-glucose ester-dependent reactions catalyzed by galloyl-1-*O*-β-D-glucosyltransferase (UGGT) and epicatechin:1-*O*-galloyl-β-D-glucose *O*-galloyltransferase (ECGT) (23, 24). ECGT, an enzyme that belongs to subclade 1A of SCPL acyltransferases, has been shown to play critical roles in galloylation of flavan-3-ols (24). Comparative genomic analysis showed that the tea plant harbors 22 *SCPL* genes, compared with 11 in grape, 19 in *Arabidopsis thaliana*, and 8 in poplar (Fig. 2*A* and Dataset S27). Of the WGDs and extensive gene duplications in the type 1A *SCPL* gene family that are tea-specific, ~68% (15) were generated through tandem duplications, and clustered with three *SCPLs* from grape, a species that also produces galloylated catechins. The other seven *SCPL* genes cluster with *SCPLs* from other plant species and/or are highly expressed in such tissues as roots and flowers that contain low levels of galloylated catechins (Fig. 2*A* and Dataset S28). Expression profiles of *SCPL*s showed that most *SCPL*s were highly expressed in apical buds and young leaves, where most galloylated flavan-3-ols accumulate (Fig. 2*B*). Indeed, transcriptome and metabolite correlation analysis showed that the tissue expression patterns for these 14 tea-specific *SCPL* genes are highly correlated with the accumulation of EGCG and ECG (Pearson's correlation test, $P < 0.05$; Fig. 3 *A* and *B* and Dataset S29). The expression patterns for specific members of the *DFR*, *LAR*, and *ANR* gene families also correlated with the accumulation of *cis*-flaval-3-ols in different tea tissues (Fig. 3*B* and Dataset S29). These observations indicate that gene duplications and expression dynamics of these secondary product-related genes may be responsible for the uniquely high level of production of galloylated catechins in tea, and thus could play a primary role in the determination of tea palatability. Functional experimentation is needed to precisely determine whether some or all of these *SCPL* genes are involved in galloylation.

Particularly interesting is the fact that these galloylated catechins that accumulate to high levels in young tea leaves are present mostly

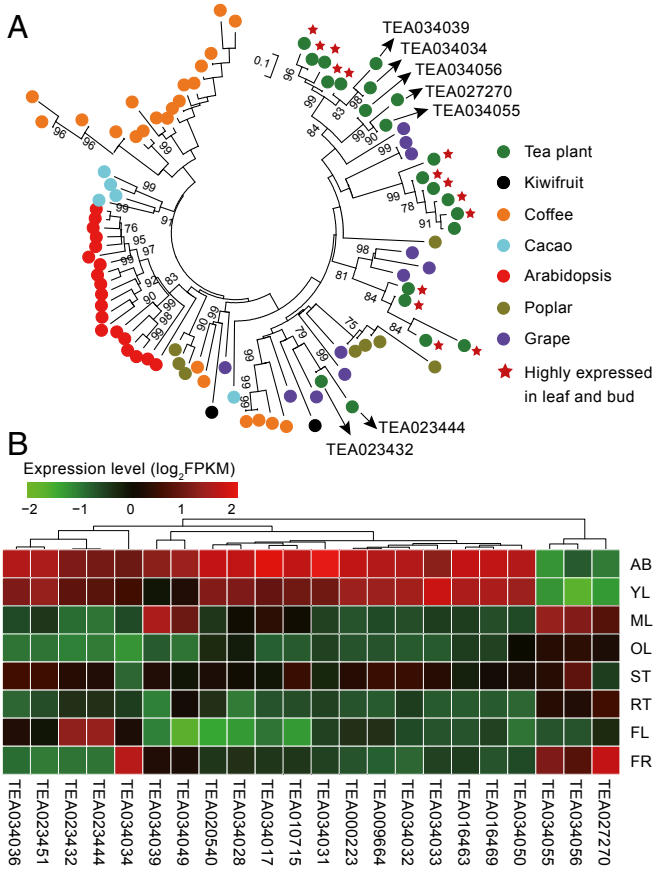

Fig. 2. Evolution of the genes encoding subclade 1A of serine carboxypeptidase-like acyltransferases (SCPL1A) in tea and six other plant species. (*A*) Neighbor-joining phylogenetic tree of SCPL1A-encoding genes from seven plant species, including tea, kiwifruit, coffee, cacao, *Arabidopsis*, poplar, and grape. (*B*) Expression profiles of the 22 tea plant SCPL1A genes (column) in eight different tissues (rows): apical buds (AB), young leaves (YL), mature leaves (ML), old leaves (OL), young stems (ST), tender roots (RT), flowers (FL), and young fruits (FR). Gene expression level was evaluated using FPKM (fragments per kilobase per million reads mapped).

in the monomeric form rather than the condensed polymers usually found in other plants (Dataset S30). We observed that such tea plant tissues as fruits, flowers, and roots, that contain more condensed polymer proanthocyanidins (PAs), accumulate much less galloylated catechins than tissues such as young buds and leaves that express higher levels of *LAR*, *ANR*, and *SCPL* but contain more monomeric galloylated catechins (Datasets S28–S30). Little is known about the enzymology of PA polymerization in plants, but a recent breakthrough in *Medicago* showed that *Medicago* LAR negatively affected the degree of PA oligomerization by converting the extension unit 4β-(*S*-cysteinyl)-epicatechin, whose availability is critical to PA polymerization, back to the starter unit epicatechin (25). Consistent with the *Medicago* results, tea *LAR* genes are highly expressed in buds and young leaves, where high levels of monomeric catechins are detected (22). The lower expression of *LAR* genes in roots and flowers is accompanied by the accumulation of larger amounts of polymerized PAs (Datasets S28–S30).

Multilevel Regulation of Catechins Biosynthesis. Developmental and environmental cues significantly affect accumulation of catechins and other flavonoids in tea plants, and thereby influence how tea collection and processing are pursued to generate the various tea types. When gene expression levels vary, transcription factor expression is likely to be critical to gene regulatory processes. The

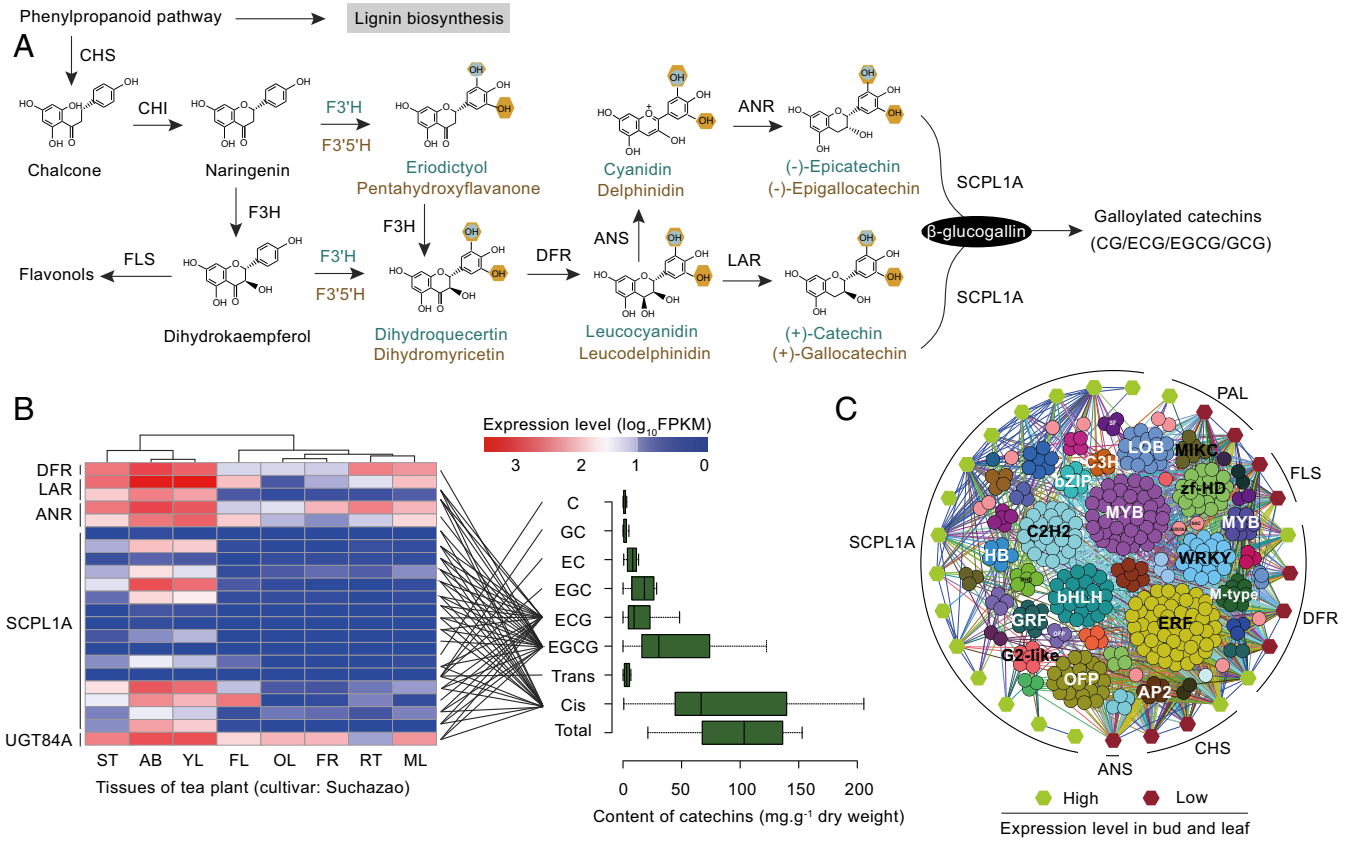

Fig. 3. Evolution and expression of key genes involved in catechins biosynthesis. (*A*) Biosynthetic pathway of the principal catechins. *CHS*, *CHI*, *F3H*, *F3′H*, *F3′5′H*, *DFR*, *ANS*, *LAR*, *ANR*, and *SCPL* represent genes encoding chalcone synthase, chalcone isomerase, flavanone 3-hydroxylase, flavonoid 3′-hydroxylase, flavonoid 3′,5′-hydroxylase, dihydroflavonol 4-reductase, anthocyanidin synthase, leucoanthocyanidin reductase, anthocyanidin reductase, and type 1A serine carboxypeptidase-like acyltransferases, respectively. (*B*) Expression profiles of key genes in different tissues of the tea plant in relation to their contents of different catechins. (*B*, *Left*) Expression levels of key genes associated with catechins biosynthesis in eight tea plant tissues: apical buds, young leaves, mature leaves, old leaves, young stems, flowers, young fruits, and tender roots. Expression data are plotted as log10 values. The horizontal axis of the boxplot (*Right*) shows statistics of catechins contents from different tissues, and the vertical axis exhibits different forms of catechins. "Cis" represents the contents of *cis*-flavan-3-ols, and "trans" represents the contents of *trans*-flavan-3-ols. The significant correlations of gene expression with the contents of ECG, EGCG, and *cis*-flavan-3-ols are indicated by black lines (Pearson's correlation test, $P < 0.05$). The error bar represents the maximum and minimum catechins content in eight different tea plant tissues. (*C*) Transcriptional regulation of catechins biosynthetic genes. A coexpression network connecting structural genes in catechins biosynthesis with transcription factors represents the regulation of catechins biosynthetic genes. The color-filled hexagons represent the structural genes associated with catechins biosynthesis that was highly (green) or lowly (red) expressed in bud and leaf. Expression correlations between TFs (colored solid circles) and catechins-related genes (colored solid hexagons) are shown with colored lines (Pearson's correlation test, $P \leq$ 1e-6).

expression patterns of *CHS*, *DFR*, *ANS*, *ANR*, and *SCPL* gene family members were found to be closely related to the expression patterns of numerous TFs, including MADS box, R2R3-MYB, and bHLH TFs (Fig. 3*C* and Dataset S34). It is known that the multiple MYB activators or repressors, together with WD40 and bHLH TFs, form ternary complexes that positively or negatively regulate flavonoid structural genes (26, 27). However, our comprehensive genome, transcriptome, and metabolome data in the tea plant indicate a much more complex systemology for synthesis of these secondary products, with many types of TFs involved in regulation of key catechins synthesis genes in leaves, including some of those that encode SCPLs, PALs, DFRs, and CHSs (Dataset S34). Several TFs that are associated with biotic and/or abiotic stress responses, such as WRKY, C2H2, C3H, NAC, and ERF family members, also showed strong association with the expression of catechins biosynthesis genes, suggesting that tea plants employ flavan-3-ols for adaptation (*SI Appendix*, Fig. S10), in agreement with previous reports showing antiinsect and antimicrobial activities of these natural products (28, 29).

Identification of a Theanine Synthetase Gene. Theanine (γ-glutamylethylamide) is the most abundant free amino acid in tea and is ubiquitous in tea plant tissues (30). Theanine was proposed to be a major tea plant nitrogen reservoir derived from glutamate and ethylamine (EA) by the action of theanine synthetase (TS) (31). Of genes key to nitrogen assimilation, storage, and accumulation in many plants (32–34), five glutamine synthetase (GS), two glutamate synthase (GOGAT), and four glutamate dehydrogenase (GDH) gene homologs were predicted in the tea genome (Fig. 4*A*). It remains unclear how TS or GS makes use of glutamate and EA to synthesize theanine and glutamine. Although, in vitro, TS and GS activity in crude extracts has been detected (31), no *TS* gene has been identified from any plant. Phylogenetic analysis revealed two distinct clades (GSI and GSII) for plant *GS* genes (Fig. 4*B*). GSI-type genes have also been found in some prokaryotes and other eukaryotes (35, 36), while most GSII-type genes are only found in plants (37). Interestingly, we found the predicted tea GSI-type protein CsGSI shares high homology with PtGS from the bacterium *Pseudomonas taetrolens*. PtGS has been engineered for the high-level production of theanine from the substrates glutamic acid and ethylamine in this natural theanine producer strain (38). Because of this high homology to *PtGS*, we view *CsGSI* as the best *TS* gene candidate and have renamed it *CsTSI*. *CsTSI* encodes an

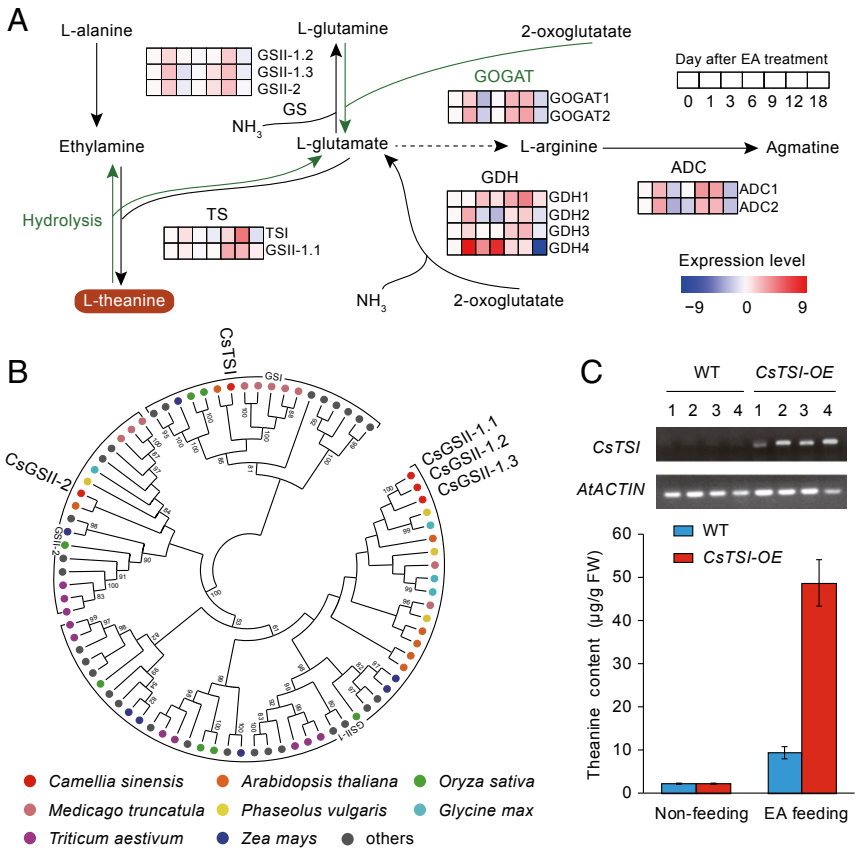

Fig. 4. Key genes involved in the theanine biosynthesis pathway. (*A*) The proposed pathway for theanine biosynthesis and expression of key genes upon precursor ethylamine feeding. *TS*, *GS*, *GOGAT*, *GDH*, and *ADC* represent genes encoding theanine synthetase, glutamine synthetase, glutamate synthetase, glutamate dehydrogenase, and arginine decarboxylase, respectively. Tea seedlings grown hydroponically were fed ethylamine chloride for different numbers of days before being sampled for amino acid profiling and transcriptome analyses. (*B*) Phylogenetic tree of tea TS and GS candidate genes and the available GS genes from prokaryotes, fungi, and plants. The tea TS candidate gene (*CsTSI*) shows high similarity to known *GSI*-type genes, and other *GS* candidate genes exhibit high homology with previously reported *GSII*-type genes in plants. (*C*) Assay of theanine synthesis activity of CsTSI in *Arabidopsis* seedlings. The candidate tea TS gene (*CsTSI*) that shows high similarity to known GSI-type genes was cloned into a binary vector and overexpressed in *Arabidopsis* driven by a 35S promoter. *CsTSI-OE* indicates *CsTSI*-overexpression lines, while WT represents wild type (control). Seedlings were fed with or without 10 mM EA chloride solution (with water as control) for 3 d. Theanine synthesized by the seedlings was extracted and measured. Data are expressed as means ± SD from at least three independent transgenic lines with replicate experiments. FW, fresh weight.

enzyme with a GLN-SYN motif in the C terminus and an amidohydrolyase domain in the N terminus; interestingly, the latter was also identified in MtN6 and GmN6L nodulins (*SI Appendix*, Fig. S11). Sequence alignment and structural analyses show that CsTSI possesses some unshared amino acid residues at the ammonia-binding domain, suggesting that CsTSI has a binding preference for ethylamine over ammonia (*SI Appendix*, Fig. S11).

Metabolite profiling indicated that tea plant roots accumulate the highest level of theanine, followed by young buds and tender leaves (*SI Appendix*, Fig. S12). *CsTSI* was highly expressed in roots, while one *GSII-1.1* gene was highly expressed in young buds (Dataset S38). Overall, the expression patterns of *CsTSI* and *GSII-1.1* are significantly correlated with theanine content across tissues (Pearson's correlation test, $P < 0.05$; Dataset S39). Tracer experiments using a 1-^{14}C-labeled EA precursor indicated that theanine is primarily synthesized in the root and then transported to the shoot (39). When feeding hydroponically grown tea seedlings with EA aqueous solution, the expression of *CsTSI* in roots was elevated to its highest level after 12 d of treatment, concurrent with theanine accumulation reaching a peak (Fig. 4*A*). Amino acid profiling suggests that alanine, arginine, and glutamate shared similar but less dramatic accumulation patterns with theanine over the time of the EA treatment (Dataset S41). Transcripts of *GS*/*TS*, *GOGAT*, and *GDH* genes also increased with EA treatment, consistent with their proposed involvement in theanine accumulation (*SI Appendix*, Fig. S14). The simultaneous increases in glutamate, theanine, and arginine levels and arginine decarboxylase (*ADC*) expression strongly support the idea that theanine biosynthesis is involved in nitrogen assimilation, storage, and recycling (Dataset S41). To further investigate the theanine synthetase activity of CsTSI, we generated *CsTSI*-overexpressing *Arabidopsis* plants for an in planta theanine biosynthesis assay (Fig. 4*C*). Feeding 10 mM EA solutions to T3 *CsTSI* overexpression (*CsTSI-OE*) seedlings and wild-type seedlings led to a great increase in theanine production only in *CsTSI-OE* lines (Fig. 4*C*). This result proves that CsTSI has theanine synthetase activity. Taken together, the expression and homology data strongly support the idea that *CsTSI* is the tea theanine synthase gene, a model that can be further tested by genetic and reverse-genetic experiments.

Convergent Evolution of the Caffeine Synthesis Pathway. Caffeine is a vital metabolite for tea quality, being largely responsible for the energizing briskness of most tea products. Earlier studies have shown that caffeine synthesis independently evolved in cacao and coffee lineages, with the prediction that this would also be true in tea (40). Xia et al. (15) made this conclusion from their analysis of the CSA genome, and our results on CSS also agree. However,

our more complete dataset also indicates that most of the amplification of these responsible *N*-methyltransferase genes was expanded through the most recent gene duplication events (*SI Appendix*, Fig. S16).

Conclusions

We report a high-quality draft genome sequence of CSS, the most widely cultivated type of tea plant in China and the world. We confirm and date a WGD in the tea lineage that occurred 30 to 40 Mya and an earlier WGD that occurred 90 to 100 Mya. We conclude that *C. sinensis* diverged from a shared lineage with kiwifruit about 80 Mya, while the two major varieties CSS and CSA diverged from a common ancestor ~0.38 to 1.54 Mya. The recent WGD was a particularly important event for generating many duplications of genes that contribute to catechins biosynthesis, but most of these extra copies arose from paralogous duplications that occurred long after the tea plant ancestor diverged from the kiwifruit lineage. Correlation analyses of transcriptome and catechins profiles support correlations between specific catechins biosynthetic gene family members, in such families as *CHS*, *DFR*, *ANS*, *LAR*, *ANR*, and *SCPL*, and the accumulation of catechins, particularly galloylated *cis*-flaval-3-ols. These analyses also identified the likely transcription factor networks that regulate synthesis of these key secondary metabolites that determine tea quality. A well-supported candidate for the key gene in the synthesis of the amino acid theanine, providing essential health and flavor characteristics to tea, was also identified. The *C. sinensis* genome sequence can serve as a vital resource for studying the genetic bases of these major plant metabolic pathways and for germplasm utilization to breed improved tea cultivars.

Methods

Plant Materials and Sequencing. Cultivar Shuchazao of *C. sinensis* (accession no. GS2002008) was found to exhibit a relatively low level of heterozygosity by RAD-seq technology and, thus, was selected for genome sequencing. Young tender leaves of an individual plant were used for DNA extraction. A total of 10 paired-end libraries were constructed using paired-end kits (Illumina) with average insert sizes of ~170, 250, 500, and 800 bp. Ten mate-pair libraries were prepared by mate-pair kits (Illumina) with average insert sizes of 2, 5, 10, 20, and 40 kb. Sequencing was performed on the Illumina HiSeq 2500 platform. Ten- and 20-kb libraries were constructed for PacBio sequencing on a PacBio RS II sequencer (Pacific Biosciences). Both the Illumina next-generation sequencing and PacBio SMRT sequencing data were used for the genome assembly.

De Novo Assembly, Annotation, and Quality Assessment. All reads from libraries of 170 and 250 bp were split into 119-mers (118-bp overlap with one overhang) and applied to construct a de Bruijn graph. The contigs were then constructed and assembled into scaffolds with further gap filling using methods as described in *SI Appendix*, section 1.4. We searched transposable elements in the tea plant genome by integrating de novo methods and homology-based methods with the prediction programs RepeatModeler (www.repeatmasker.org/RepeatModeler/), LTR_FINDER (41), RepeatMasker (www.repeatmasker.org), and RepeatProteinMask (version 3.3.0; www.repeatmasker.org/RepeatProteinMask.html). Based on the repeat-masked genome, we combined three pieces of evidence from ab initio gene prediction, homolog searching, and EST/UniGene-based prediction to predict nonredundant protein-encoding gene models in the tea plant (*SI Appendix*, section 2.2). Noncoding RNA genes for miRNA, tRNA, rRNA, small nucleolar RNA, and snRNA were predicted using de novo and/or homology search methods (15).

To investigate the quality of the genome assembly, 18 BACs were randomly selected and sequenced to align against our genome assembly (42). The accuracy and completeness of the gene prediction were evaluated using three methods, including DNA alignment, EST alignment, and BUSCO datasets from the plant lineage.

Genome Evolution and Expansion Analyses. To identify orthologous genes among 11 representative plant genomes, full genome sequences were retrieved from the appropriate websites. Pseudogenes and TE-derived genes were removed, and only the longest ORF was selected to represent each gene. We used the OrthoMCL (43) pipeline to identify gene families, and the generated high-quality single-copy genes were used to construct the phylogenetic tree among them. Divergence times among these 12 plant species were estimated using the MrBayes and MCMC tree programs (44). The expansion or contraction events of gene families were computationally identified by the comparisons of family (cluster) size differences between the most recent common ancestor and each of the current plant species with a significant *P* value of 0.05.

To investigate gene content expansion in CSS, whole-genome duplication events were analyzed. Orthologous genes in tea and grape were identified and mapped onto the grape genome. The homologous syntenic blocks were identified according to the method described previously (45). Syntenic gene blocks in tea scaffolds were determined with the number of the genes in one syntenic block ≥3 and the number of nonsyntenic tea genes between two adjacent syntenic genes <5. The 4dTv of the identified homologous blocks in tea–tea, tea–grape, and tea–cocoa were calculated with the HKY substitution model.

Evolution of Catechins Biosynthesis Genes. To elucidate the evolution of the catechins biosynthetic pathway in tea, we constructed a phylogenetic tree and evaluated the times of gene duplications for each structural gene in the pathway. To detect tandem duplications that may have occurred in each gene family in the pathway, we investigated the locations of all identified genes in the assembly. Timing of divergence of duplicates in each catechins biosynthesis gene was estimated based on the rooted phylogenetic tree with *Arabidopsis* genes as outgroups. Paralogous pairs of each gene were selected and subjected to calculation of the synonymous substitution rate based on the Nei and Gojoberi method of Yang implemented in the PAML program (version 4.9b) (46).

Metabolic Profiling and Transcript Expression Correlations for Catechins Biosynthesis Genes. Metabolites were analyzed by using both HPLC and LC-Q-TOF-MS. Investigation of whether there were significant correlations between expression levels of catechins pathway genes and the contents of catechins in tea plant tissues was performed using the Pearson's correlation test.

Verification of an in Vivo Theanine Synthesis Function of *CsTSI*. The ORF of *CsTSI* was cloned from cDNAs prepared from tea root tissues. After sequencing confirmation, the *CsTSI* ORF was cloned into the pDONR221 vector and then recombined into the pB2GW7 binary vector using the Gateway cloning system (Invitrogen, Life Technologies). The resulting construct, harboring 35S::*CsTSI*, was transformed into *Agrobacterium tumefaciens* GV3101 by electroporation. The positive *A. tumefaciens* GV3101 clones were selected for *A. thaliana* (Col-0) transformation using the standard flower dipping method. At least 10 independent transformants were screened and selected by using BASTA. Homozygous T3 transgenic lines were verified by RT-PCR and used for theanine biosynthesis experiments. qRT-PCR was conducted with standard protocols for RNA extraction, cDNA preparation, PCR with a pair of *CsTSI*-specific primers, and calculation.

Seeds of *CsTSI-OE* lines and wild-type (Col-0) were surface-sterilized and germinated on a half-strength MS medium agar plate supplemented with or without 10 mM ethylamine chloride. The plates were incubated in a 22 °C growth chamber under a light period of 16/8 h for germination and seedling development. After 22 d of growth, four-leaf *Arabidopsis* seedlings were harvested for theanine analysis, according to the method described above. At least three independent transgenic lines were used in each experiment, and three independent experiments were conducted.

Further detailed methods for data analyses are fully provided in *SI Appendix*.

Data Deposition. Clean data from Illumina sequencing reads, transcriptome reads, and PacBio data of tea plant (CSS), genome assembly, gene prediction, and gene functional annotations may be accessed from pcsb.ahau.edu.cn:8080/CSS/ (data downloading with user name 20170705F16HTSCCKF2479, password rwn160912abc).

ACKNOWLEDGMENTS. We thank Dr. Deyu Xie for critical reading and assistance on the manuscript. We thank the 916 Tea Plantation in Shucheng, Anhui Province, the Tea Research Institute, Yunnan Academy of Agricultural Sciences, and the Tea Research Institute, Fujian Academy of Agricultural Sciences for providing us with samples of tea plants. This work was also supported by the People's Government of Anhui Province. This work was mainly supported by the Vitalizing Plan of Tea Industry in Anhui Province (2012–2015), National Tea Industry Technology System from the Ministry of Agriculture of China (nycytx-26), Major Project of the Chinese National Program for Fundamental Research and Development (2012CB722903), Science

and Technology Project of Anhui Province (13Z03012), National Natural Science Foundation of China (30972400, 31170283, 31171608, 31170282, 31300578, 31300576, and 31470689), Doctoral Scientific Fund Project of the Ministry of Education of China (20113418130001), Special Innovative Province Construction in Anhui Province (15czs08032), and Changjiang Scholars and Innovative Research Team in University (IRT1101).

1. Wheeler DS, Wheeler WJ (2004) The medicinal chemistry of tea. *Drug Dev Res* 61: 45–65.
2. Yang CS, Hong J (2013) Prevention of chronic diseases by tea: Possible mechanisms and human relevance. *Annu Rev Nutr* 33:161–181.
3. Kingdom-Ward F (1950) Does wild tea exist? *Nature* 165:297–299.
4. Taniguchi F, et al. (2014) Worldwide core collections of tea (*Camellia sinensis*) based on SSR markers. *Tree Genet Genomes* 10:1555–1565.
5. Kaundun SS, Matsumoto S (2003) Development of CAPS markers based on three key genes of the phenylpropanoid pathway in tea, *Camellia sinensis* (L.) O. Kuntze, and differentiation between *assamica* and *sinensis* varieties. *Theor Appl Genet* 106: 375–383.
6. Ming T, Bartholomew B (2007) Theaceae. *Flora China* 12:366–478.
7. Willson KC, Clifford MN (2012) *Tea: Cultivation to Consumption* (Springer Science & Business Media, Berlin).
8. Yang Y, Liang Y (2014) *Tea Plant Clonal Varieties in China* (Shanghai Scientific & Technical, Shanghai).
9. Namita P, Mukesh R, Vijay KJ (2012) *Camellia sinensis* (green tea): A review. *Glob J Pharmacol* 6:52–59.
10. Asakawa T, Hamashima Y, Kan T (2013) Chemical synthesis of tea polyphenols and related compounds. *Curr Pharm Des* 19:6207–6217.
11. Li S, Lo C-Y, Pan M-H, Lai C-S, Ho C-T (2013) Black tea: Chemical analysis and stability. *Food Funct* 4:10–18.
12. Narukawa M, Morita K, Hayashi Y (2008) L-Theanine elicits an umami taste with inosine 5′-monophosphate. *Biosci Biotechnol Biochem* 72:3015–3017.
13. Li C-F, et al. (2015) Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (*Camellia sinensis*). *BMC Genomics* 16:560.
14. Shi C-Y, et al. (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics* 12:131.
15. Xia EH, et al. (2017) The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol Plant* 10:866–877.
16. Yang H, et al. (2016) Genetic divergence between *Camellia sinensis* and its wild relatives revealed via genome-wide SNPs from RAD sequencing. *PLoS One* 11:e0151424.
17. Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol* 65:505–530.
18. Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc Natl Acad Sci USA* 93: 10274–10279.
19. Yang Z, Baldermann S, Watanabe N (2013) Recent studies of the volatile compounds in tea. *Food Res Int* 53:585–599.
20. Huang S, et al. (2013) Draft genome of the kiwifruit *Actinidia chinensis*. *Nat Commun* 4:2640.
21. Pang Y, et al. (2013) Functional characterization of proanthocyanidin pathway enzymes from tea and their application for metabolic engineering. *Plant Physiol* 161: 1103–1116.
22. Wang P, et al. (2018) Evolutionary and functional characterization of leucoanthocyanidin reductases from *Camellia sinensis*. *Planta* 247:139–154.
23. Cui L, 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.
24. Liu Y, et al. (2012) Purification and characterization of a novel galloyltransferase involved in catechin galloylation in the tea plant (*Camellia sinensis*). *J Biol Chem* 287: 44406–44417.
25. Liu C, Wang X, Shulaev V, Dixon RA (2016) A role for leucoanthocyanidin reductase in the extension of proanthocyanidins. *Nat Plants* 2:16182.
26. Sun B, et al. (2016) Purple foliage coloration in tea (*Camellia sinensis* L.) arises from activation of the R2R3-MYB transcription factor CsAN1. *Sci Rep* 6:32534.
27. Xu W, Dubos C, Lepiniec L (2015) Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. *Trends Plant Sci* 20:176–185.
28. Hahlbrock K, et al. (2003) Non-self recognition, transcriptional reprogramming, and secondary metabolite accumulation during plant/pathogen interactions. *Proc Natl Acad Sci USA* 100:14569–14576.
29. Nathanson JA (1984) Caffeine and related methylxanthines: Possible naturally occurring pesticides. *Science* 226:184–187.
30. Nobre AC, Rao A, Owen GN (2008) L-Theanine, a natural constituent in tea, and its effect on mental state. *Asia Pac J Clin Nutr* 17:167–168.
31. Ashihara H (2015) Occurrence, biosynthesis and metabolism of theanine (γ-glutamyl-L-ethylamide) in plants: A comprehensive review. *Nat Prod Commun* 10:803–810.
32. Bernard SM, Habash DZ (2009) The importance of cytosolic glutamine synthetase in nitrogen assimilation and recycling. *New Phytol* 182:608–620.
33. Gregerson RG, Miller SS, Twary SN, Gantt JS, Vance CP (1993) Molecular characterization of NADH-dependent glutamate synthase from alfalfa nodules. *Plant Cell* 5: 215–226.
34. Melo-Oliveira R, Oliveira IC, Coruzzi GM (1996) *Arabidopsis* mutant analysis and gene regulation define a nonredundant role for glutamate dehydrogenase in nitrogen assimilation. *Proc Natl Acad Sci USA* 93:4718–4723.
35. Doskočilová A, et al. (2011) A nodulin/glutamine synthetase-like fusion protein is implicated in the regulation of root morphogenesis and in signalling triggered by flagellin. *Planta* 234:459–476.
36. Silva LS, Seabra AR, Leitão JN, Carvalho HG (2015) Possible role of glutamine synthetase of the prokaryotic type (GSI-like) in nitrogen signaling in *Medicago truncatula*. *Plant Sci* 240:98–108.
37. Martin A, et al. (2006) Two cytosolic glutamine synthetase isoforms of maize are specifically involved in the control of grain production. *Plant Cell* 18:3252–3274.
38. Yamamoto S, Wakayama M, Tachiki T (2006) Cloning and expression of *Pseudomonas taetrolens* Y-30 gene encoding glutamine synthetase: An enzyme available for theanine production by coupled fermentation with energy transfer. *Biosci Biotechnol Biochem* 70:500–507.
39. Deng W-W, Ogita S, Ashihara H (2009) Ethylamine content and theanine biosynthesis in different organs of *Camellia sinensis* seedlings. *Z Naturforsch C* 64:387–390.
40. Denoeud F, et al. (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345:1181–1184.
41. Xu Z, Wang H (2007) LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35:W265–W268.
42. Tai Y, et al. (2017) Construction and characterization of a bacterial artificial chromosome library for *Camellia sinensis*. *Tree Genet Genomes* 13:89.
43. Li L, Stoeckert CJ, Jr, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189.
44. Arvestad L, Berglund AC, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19(Suppl 1):i7–i15.
45. Peng Z, et al. (2013) The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nat Genet* 45:456–461, 461e1–461e2.
46. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.