frontiers in Genetics

ORIGINAL RESEARCH
published: 10 November 2021
doi: 10.3389/fgene.2021.770570

Genome-Wide Identification, Characterization and Function Analysis of Lineage-Specific Genes in the Tea Plant *Camellia sinensis*

*Zhizhu Zhao and Dongna Ma**

Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, China

OPEN ACCESS

Edited by:
Zefeng Yang,
Yangzhou University, China

Reviewed by:
Swarup Roy Choudhury,
Indian Institute of Science Education and Research, Tirupati, India
Bourlaye Fofana,
Charlottetown Research and Development Centre, Canada

***Correspondence:**
Dongna Ma
m_dongna@163.com

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *04 September 2021*
Accepted: *14 October 2021*
Published: *10 November 2021*

Citation:
Zhao Z and Ma D (2021) Genome-Wide Identification, Characterization and Function Analysis of Lineage-Specific Genes in the Tea Plant Camellia sinensis. *Front. Genet. 12:770570. doi: 10.3389/fgene.2021.770570*

Genes that have no homologous sequences with other species are called lineage-specific genes (LSGs), are common in living organisms, and have an important role in the generation of new functions, adaptive evolution and phenotypic alteration of species. *Camellia sinensis* var. *sinensis* (CSS) is one of the most widely distributed cultivars for quality green tea production. The rich catechins in tea have antioxidant, free radical elimination, fat loss and cancer prevention potential. To further understand the evolution and utilize the function of LSGs in tea, we performed a comparative genomics approach to identify *Camellia*-specific genes (CSGs). Our result reveals that 1701 CSGs were identified specific to CSS, accounting for 3.37% of all protein-coding genes. The majority of CSGs (57.08%) were generated by gene duplication, and the time of duplication occurrence coincide with the time of two genome-wide replication (WGD) events that happened in CSS genome. Gene structure analysis revealed that CSGs have shorter gene lengths, fewer exons, higher GC content and higher isoelectric point. Gene expression analysis showed that CSG had more tissue-specific expression compared to evolutionary conserved genes (ECs). Weighted gene co-expression network analysis (WGCNA) showed that 18 CSGs are mainly associated with catechin synthesis-related pathways, including phenylalanine biosynthesis, biosynthesis of amino acids, pentose phosphate pathway, photosynthesis and carbon metabolism. Besides, we found that the expression of three CSGs (*CSS0030246*, *CSS0002298,* and *CSS0030939*) was significantly down-regulated in response to both types of stresses (salt and drought). Our study first systematically identified LSGs in CSS, and comprehensively analyzed the features and potential functions of CSGs. We also identified key candidate genes, which will provide valuable assistance for further studies on catechin synthesis and provide a molecular basis for the excavation of excellent germplasm resources.

Keywords: tea plant, lineage-specific genes, gene duplication, transcriptome, *Camellia*

INTRODUCTION

Genes that have no homologous sequences with other species are called lineage-specific genes (LSGs), sometimes are also called orphan genes (Fischer and Eisenberg, 1999; Tautz and Domazet-Loso, 2011). LSGs were first found in *Saccharomyces cerevisiae* in 1996, that is, a large number of genes in the genome showed no similarity to the database sequence, accounting for about 26% of the genome (Dujon, 1996). As more and more complete genomes and transcriptomes from different species have been sequenced, LSGs have also been more and more widely studied, from microorganisms to plants, such as legumes (Graham et al., 2004), *Triticeae* (Ma et al., 2020), *Oryza sativa* (Yang et al., 2009), *Arabidopsis* (Lin et al., 2010), Poaceae (Campbell et al., 2007), *Populus* (Yang et al., 2009) and sweet orange(Xu et al., 2015). The proportion of LSGs in different genomes is also different, and it has been found that the average proportion of LSGs in plants is higher than the average proportion in animals (Yang et al., 2013). The significance of the presence of most LSGs remains unknown, but is often associated with the unique features the species have and stress tolerance, which is of important implications for elucidating the evolutionary of species (Khalturin et al., 2009).

Although we cannot analyze the biological functions of LSGs using homology-based functional classification, the structural traits of their sequences may provide some initial clues for LSGs exploration. Compared with evolutionary conserved genes (ECs), LSGs have some differences in gene length, number of introns and exons, GC content and chromosome distribution preference, owing to the shorter generation time. LSGs are normally characterized by shorter gene length and fewer exons in eukaryotes (Domazet-Loso, 2003; Campbell et al., 2007; Toll-Riera et al., 2009; Yan et al., 2014). GC content of LSGs to most species is lower than that of ECs, a characteristic that is not universal. For example, the GC content of LSGs in zebrafish is higher than conserved genes, this characteristic is similar to the LSGs in rice (Yang et al., 2013). The distribution characteristics of LSGs on chromosomes are also different, like zebrafish have uneven distribution of LSGs on chromosomes, with some chromosomes having a high proportion of gene and others having no LSGs, which may be related to the length of chromosome (Yang et al., 2013). Nevertheless, the distribution of LSGs in *Arabidopsis* and ant has no chromosomal preference and LSGs are evenly distributed among non-LSGs throughout the genome (Donoghue et al., 2011; Wissler et al., 2013). In addition to sequence traits, some LSGs show a high degree of tissue-specific expression (Lemos et al., 2005). LSGs were more expressed in callus in sweet orange, a stem-cell like tissue (Xu et al., 2015) and most LSGs in wheat were expressed in sexual tissues (Ma et al., 2020).

Studies found that the expression of some LSGs responded to a wide range of stress conditions, suggesting that these LSGs may enable the species to better adapt to the environment, thus LSGs become important genes during evolution (Khalturin et al., 2009; Donoghue et al., 2011). LSGs in mangrove *Aegiceras corniculatum* are involved in pathways like flavonoid biosynthesis, which play a role in oxidative toxicity mitigating in mangrove plants under high saline environments (Ma et al., 2021). The LSG in *Arabidopsis thaliana QQS* was involved in regulating the partitioning of carbon and nitrogen among proteins and carbohydrates in leaves (Li et al., 2015). The rice orphan gene *OsDR10* was reported to enhance disease resistance by increasing endogenous salicylic acid (SA) levels and suppressing the production of endogenous jasmonic acid (JA) (Xiao et al., 2009). The overexpression of new gene *GS9* in rice results in round grains, which can be used to breed rice varieties with optimized grain shape (Zhao et al., 2018). A study conducted in six orphan genes in *Drosophila* showed that arbitrary suppression of four of these six orphan genes via RNAi caused lethality (Reinhardt et al., 2013). Obviously, LSGs are involved in different metabolic pathways and have diverse functions that affect various aspects of organisms, and the importance of LSGs is just showing up.

Tea is one of the most well known and most consumed beverages in the world, which provides both health benefits and economic value (Song et al., 2012; Lee et al., 2014; Zhu et al., 2020). Due to the benefits tea brings to health, the exploration of tea has increased at molecular level. Tea belongs to the Theaceae family, and is a quite important economically crop worldwide whose leaves can be used to produce various tea. Because of the variation of gene, the difference in growing conditions and the difference in processing modes, tea always has diverse palatability, like bitter, astringent, and sweet flavors (Song et al., 2012). *Camellia sinensis* var. *sinensis* (CSS) is one of the most widely distributed cultivars for quality green tea production (Song et al., 2012). Currently, 67% elite tea plant cultivars belong to CSS. During rapid evolution, LSGs are endowed with new biological functions when subjected to external environmental pressures, which allow species to better adapt to the external environment (Long et al., 2003; Kaessmann, 2010). In addition, some LSGs participate and play important roles in metabolic networks and pathways affecting various aspects of the organism soon after their origin (Chen et al., 2012), and thus making LSGs important genes during evolution. To further understand the evolution and utilize the function of LSGs in tea species CSS, we used a comparative genomics approach to identify *Camellia*-specific genes (CSGs) in the CSS genome for analyzing the origin models, structural properties and subcellular localization of the CSGs. Furthermore, we constructed weighted gene co-expression network analysis (WGCNA) to predict the function of LSGs in CSS. Collectively, these results will provide important information for understanding the role played by CSGs in the evolution of lineage specific phenotypes and adaptive innovation in CSS.

MATERIALS AND METHODS

Data Collection

The predicted CSS high-quality genome annotation and the expression level genes from eight different tissues (apical bud, young leaf, mature leaf, old leaf, stem, flower, fruit and stem) were downloaded from http://tpdb.shengxin.ren/. We

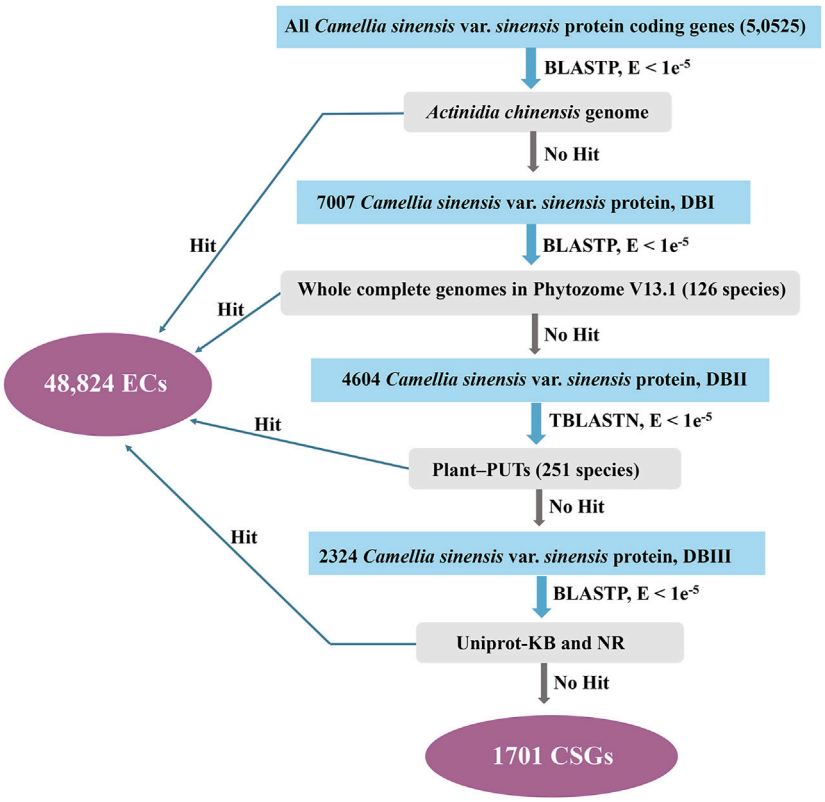

FIGURE 1 | Procedure for identifying lineage-specific genes in *Camellia sinensis* var. *sinensis*.

collected *A. chinensis* genomes from public databases to identify LSGs in CSS (ftp://bioinfo.bti.cornell.edu/pub/kiwifruit/). Other 126 plant genome sequences were downloaded from Phytozome V13.1 (http://phytozome.jgi.doe.gov/pz/ portal.html) (**Supplementary Table S1**),the assembled unique transcripts (PUT) sequences of the plants were downloaded from PlantGDB (http://www.plantgdb.org/prj/ESTCluster/progress.php) (**Supplementary Table S2**), Uniprot-KB were downloaded from Uniprot (ftp://ftp.ebi.ac.uk/pub/ databases/uniprot/knowledgebase/) and NR databases were acquired from NCBI, respectively.

Identification of CSGs

The study on the origin and evolution of CSGs has been improved due to the development of comparative genomics. Based on a homolog search, CSGs within CSS were identified in a pipeline (**Figure 1**). Firstly, CSS protein sequences were searched against *A. chinensis* proteome data using BLASTP. The CSS protein sequence was discarded once it has BLASTP hit with an E-value cutoff of 1e-5. We then performed homology searches with genomes of other plants, Plant-PUTs database, Uniprot-KB database and NR database in turn with an E-value cutoff of 1e-5. Finally, the genes having no homolog to any databases are the CSGs (Zhang et al., 2007; Lin et al., 2010), while the others which are homologous are evolutionarily conserved genes (ECs).

Genic Features

We used the whole genome information of CSS to observe the structural characteristics of the CSGs. The isoelectric point of CSGs and ECs was measured using DAMBE7 software (Xuhua, 2018). Differences between CSGs and ECs, including gene size, length of protein, size of the exons and introns, number of the exons and content of GC were calculated using the in-house python scripts. The significant difference between different groups of CSGs and ECs was then determined with Mann-Whitney U test. We extracted the information of chromosome localization from chromosome sequences and mapped it with MapGene2Chrom (http://mg2c.iask.in/mg2c_v2.0/), and finally predicted CSGs subcellular localizations using BUSCA (Bologna Unified Subcellular Component Annotator) (Savojardo et al., 2018).

Gene Duplication Analysis

There are multiple models explaining for the origin of LSGs according to previous studies (Wu et al., 2011; Wissler et al., 2013), among which gene duplication has long been thought as the primary mechanism of the emergence of LSGs (Zhang, 2003). We first searched for homologous genes with an E-value cutoff of 1e-8 using BLASTP, and then used DupGen_finder.pl to determine the different types of gene duplication, which is able to identify WGD, tandem duplication, proximal duplication, transposon duplication and dispersed duplication

(Qiao et al., 2019). The *Ks* of the duplication paralogous gene pairs were computed with the python script synonymous_calc.py (https://github.com/tanghaibao) with the method of Nei-Gojobori. We finally estimated the time of gene duplication of CSGs with the universal mutation rate of 6.5×10^{-9} (Gaut et al., 1996).

Gene Expression Analysis

To analyze the environmental adaption of CSGs, we downloaded the transcriptome data of CCS from the European Nucleotide Archive database (ENA; http://www.ebi.ac.uk/ena) under project number accession PRJEB11522. Plants in this experiment were divided into three groups, the treatment of the first group was 25% percent polyethylene glycol to simulate drought stress conditions, the second group was 200 mM NaCl, and the last group was a blank control with sampling times at 0, 24, 48, and 72 h (Zhang et al., 2017). We then used Trimmomatic program to filter the raw RNA-seq data (Bolger et al., 2014). In order to identify differentially expressed genes (DEGs) among different treatments, the abundance_estimates_to_matrix.pl, run_DE_analysis.pl (edgeR) and analyze_diff_expr.pl modules of the Trinity package with default settings were used. The $|\log2FC| \geq 1$ and a false discovery rate (FDR) < 0.05 as the threshold were implied to determine the significant differences in gene expression, and RSEM implemented in Trinity package was applied to compute FPKM (fragments per kilobase of exon per million fragments mapped) (Grabherr et al., 2011). Based on RNA-seq data, cluster analysis was performed with R software, and the specific expression of the genes were selected for follow-up functional validation. Genes with FPKM value >0.02 were assumed to have been expressed (Ma et al., 2020). In addition, the genes specifically expressed in certain tissue were identified using PaGeFinder software with specificity measure (SPM) (Pan et al., 2012), and it was identified as a specific gene in this tissue once the SPM value was ≥0.9.

Weighted Gene Co-Expression Network Analysis and Function Annotation

After discarding the genes with FPKM <1, we then constructed WGCNA and divided these genes into modules with the help of WGCNA package in R software (Langfelder and Horvath, 2008). The network was built with default parameters using the automatic network builder function block wise Modules. We then calculated the eigengene value for each module in each tissue, and selected the module owing highest correlation coefficient while satisfying *p*-value < 0.05 as the tissue-specific module for further analysis. The most representative gene in each module was considered to be the module eigengene (ME). Module membership (MM) and gene significance (GS) of each ME were calculated in each tissue-specific module, and once MM > 0.95 and GS > 0.85 were satisfied, this gene was considered as a hub gene of this module. KEGG enrichment analysis was performed on an online platform, OmicShare (https://www.omicshare.com/).

RESULTS

Identification of CSGs

Using the database resources released recently, CSGs in CSS were identified based on methods used in previous studies (**Figure 1**) (Toll-Riera et al., 2009; Yang et al., 2009; Lin et al., 2010; Tautz and Domazet-Loso, 2011). In this study, there were 50,525 annotated protein-coding genes within CSS genome in all, the they were used to perform BLASTP with Theaceae family genome (*Actinidia chinensis*) that had already been published. In this step, a total of 43,518 CSS genes had significant similarity (E-value < 1 e-5) and 7007 genes (DBI) were retained for subsequent analysis. We removed the ECs showing homology and further searched the remaining genes against 126 plant genomes from Phytozome v13.1, resulting in 4604 genes retained for the next step of searches (DBII). In the following comparison of these 4604 genes with 251 PlantGDB-assembled Unique Transcripts (PUTs) sequences, 2324 genes could not find any homologs (DBIII). Finally, to further eliminate the effect of false positives on the analysis, the remained genes were analyzed against the UniProt-KB and NR databases, a step that ultimately left 1701 genes. We termed these last remaining 1701 genes as CSGs in the CSS genome, making up 3.37% of the whole CSS genome (**Supplementary Table S3**), while these remaining 48,824 genes with similarities in the database were defined as ECs.

High Proportion of CSGs Generated *via* Gene Duplications

There are several mechanisms regarding how LSGs were created, among which gene duplication has been long considered to be a major way for the origin of LSGs, and the creation of a new gene in the gene duplication model originates mainly through the differentiation after duplication. In this experiment, of the 1701 CSGs we identified from the genome of CSS, 971 CSGs originated from gene duplication, representing 57.08% of all CSGs (**Supplementary Table S4**) and evenly distributed on each chromosome (**Figure 2A**). Among the CSGs originating from gene duplication, a total of 54 CSGs were detected to create during WGD duplication. Besides, the number of CSGs generated by tandem duplication, transposed duplication, proximal duplication and dispersed duplication were 44, 53, 54, and 766 (**Figure 2B**), respectively. Obviously, CSGs were mainly produced by gene duplication. We used synonymous substitution rates (*Ks*) to assess the timeline for the gene duplication to occur in CSGs. As a result, there were two peaks, one with *Ks* = 0.5—0.7 and a second with *Ks* = 1.2—1.3 (**Figure 2C**), corresponding to the duplication time of 38–54 and 92–100 million years ago (MYA).

Features of CSGs

To clarify whether there were significant differences between CSGs and ECs, we focused on analyzing and comparing the sequence structural features between the 1701 CSGs and 48,824 ECs identified in this study. As a result, both gene size (**Figure 3A**) and protein length (**Figure 3B**) of CSGs were

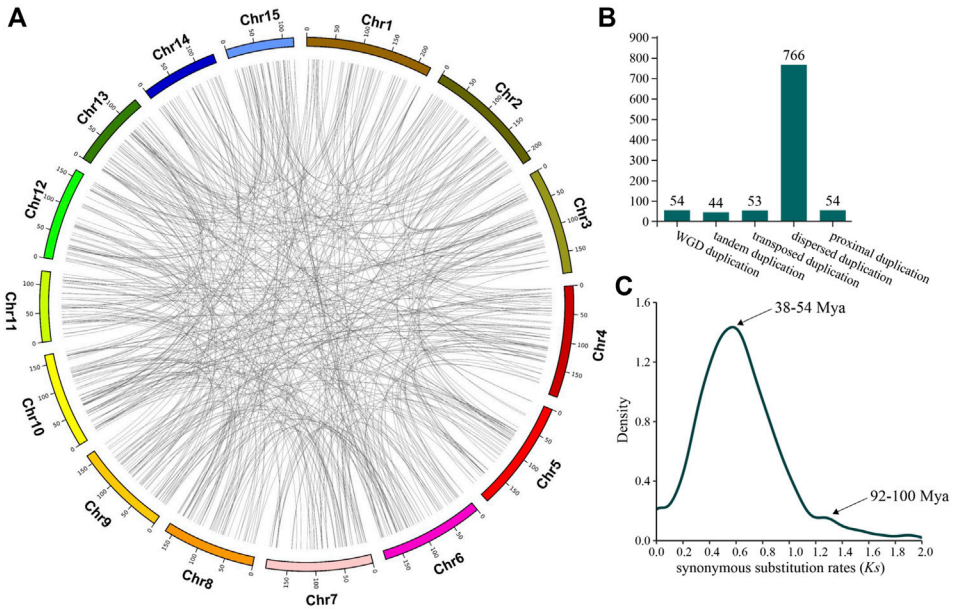

FIGURE 2 | *Camellia*-specific genes (CSGs) originating from gene duplication. **(A)** Distribution of CSGs on chromosomes by five mechanisms of origin. **(B)** The CSGs number of different duplication types. **(C)** Density distribution of synonymous substitution rates (*Ks*) values between CSGs and paralogous genes.

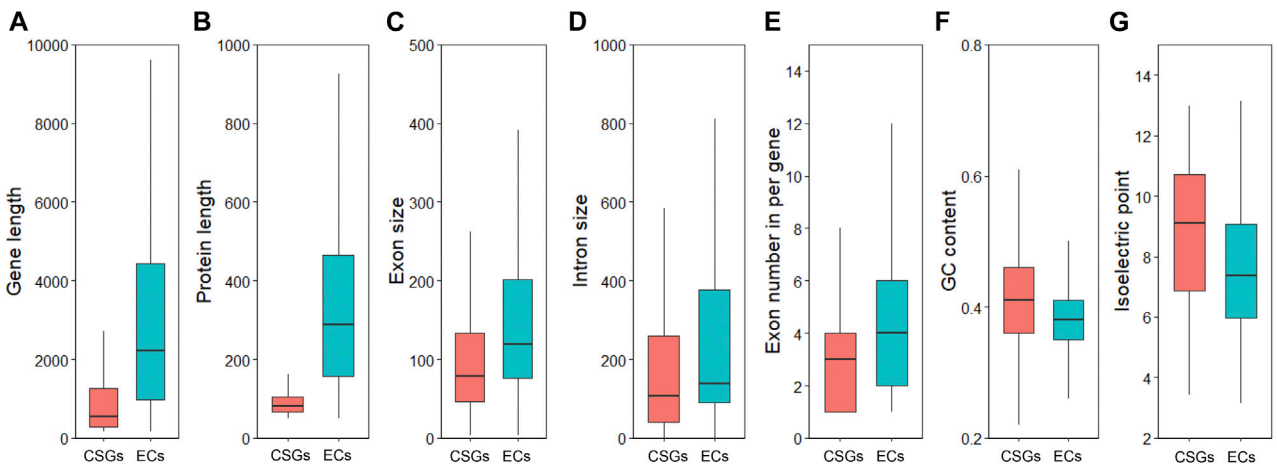

FIGURE 3 | Analyze and compare the structural characteristics of *Camellia*-specific genes (CSGs) and evolutionarily conserved genes (ECs). **(A)** Box-plot comparisons of gene length. **(B)** protein length. **(C)** exon size. **(D)** intron size. **(E)** exon number in per gene. **(F)** GCs content. **(G)** isoelectric point.

significantly smaller compared to ECs (**Table 1**), with 1484.21 bp for CSGs gene size and 92.19 amino acids (aa) for CSGs protein length, 5367.51 bp for ECs gene size and 370.21 aa for ECs protein length (**Table 1**). The exon size (**Figure 3C**) and intron size (**Figure 3D**) of CSGs were both smaller than those of ECs, and the number of exons per gene of CSGs was also significantly less than ECs (**Figure 3E**). GC content in gene (**Figure 3F**), CDS and exon of CSGs were all significantly higher (**Table 1**). In addition, the isoelectric point was 8.65 for CSGs and 7.51 for ECs, which was obviously higher for CSGs. (**Figure 3G**; **Table 1**). Overall, the result indicated that there were obvious differences between CSGs and ECs in genetic features.

To analyze the genomic distribution of CSGs, we mapped the CSGs on the chromosomes of CSS (**Figure 4A**) according to the information annotated in the genome (**Supplementary Table S3**). In total, there were 1429 CSGs distributed on 15 chromosomes. The highest number of CSGs on each chromosome was Chr2 (127), Chr7 (112) and Chr10 (118) in that order, while the highest percentage of CSGs on each chromosome was Chr10 (4.74%) and Chr8 (4.34%). It was clear that CSGs showed a preferential distribution on some chromosomes compared to ECs. In addition, the distribution of CSGs was more balanced on chromosomes except for the regions close to the telomeres of chromosome Chr5, Chr6, and

TABLE 1 | Genic features of *camellia*-specific genes (CSGs) compared with evolutionarily conserved genes (ECs).

Items	CSGs		ECs		Mann-whitney U test Probability
	Mean (SE)	Median	Mean (SE)	Median	
Gene size (bp)	1484.21 (333.02)	573	5367.51 (7663.03)	2825	<0.0005
Protein length (aa)	92.19 (39.37)	80	370.21 (290.56)	301	<0.0005
Exons per gene	2.77 (1.68)	3	5.18 (4.52)	4	<0.0005
Exon size (bp)	106.77 (148.83)	79	247.87 (346.23)	134	<0.0005
Intron size (bp)	584.06 (1782.77)	140	942.51 (2198.02)	229	<0.0005
Gene GC content (%)	41.60 (0.67)	40.74	38.78 (5.17)	37.56	<0.0005
CDS GC content (%)	45.96 (0.55)	45.56	44.78 (4.30)	44.13	<0.0005
Exon GC content (%)	44.51 (0.88)	44.64	43.22 (0.6)	42.92	<0.0005
Isoelectric point	8.65 (2.44)	8.65	7.51(1.96)	7.36	<0.0005

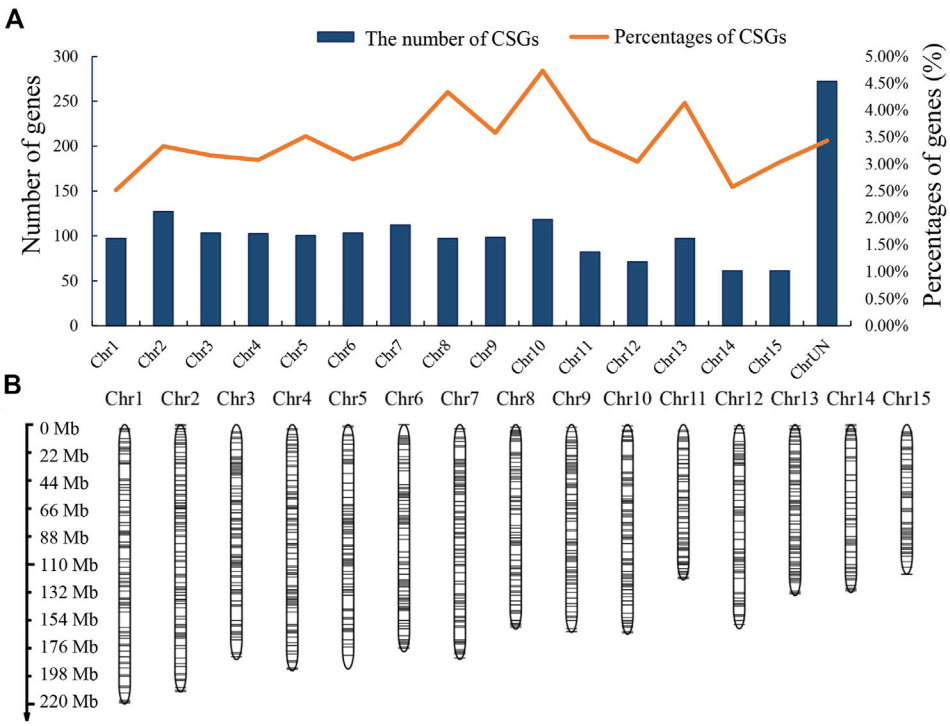

FIGURE 4 | *Camellia*-specific genes (CSGs) distribution on chromosomes. **(A)** The numbers of CSGs on each chromosome of *Camellia sinensis* var. *sinensis*. Both numbers and percentages are shown. **(B)** Chromosomal distribution of the identified CSGs. Black horizontal lines represent CSGs.

Chr15 where the distribution of CSGs was sparse (**Figure 4B**). Overall, CSGs were relatively evenly distributed on these 15 chromosomes.

Subcellular Localization

The function of proteins can usually be inferred to some extent based on their subcellular localization. Of the 1701 CSGs identified in this study, 644 were localized in the nucleus, 547 in extracellular space, 373 in chloroplast, 50 in endomembrane system, 35 on organelle membrane, 21 on mitochondria, 22 on the plasma membrane, eight on chloroplast thylakoid lumen, and only one on cytoplasm (**Figure 5**).

Expression Profiles of CSGs

The expression pattern of a gene in different tissues is crucial to elucidate whether this CSG has a corresponding biological function. We downloaded RNA-seq data from eight tissues of CSS. The transcriptional data contained 400 (23.52%) CSGs and 40,897 (83.76%) ECs with FPKM >0.02. Among them, 194 CSGs were found to be expressed in all eight tissues (FPKM >2 in a minimum of one tissue), and 16 CSGs were shown to be highly expressed in all eight tissues (FPKM >2 in all of them) (**Table 2**). Based on the expression abundance in each tissue, it can be seen that most CSGs are expressed with tissue preference (**Figure 6**). Further studies found, 212 CSGs showed specific expression in

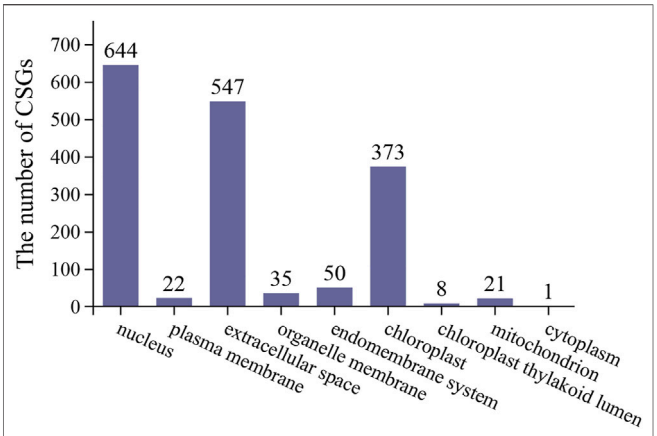

FIGURE 5 | *Camellia*-specific genes assigned to different subcellular locations.

eight tissues, of which 17 were specifically expressed in apical bud, 21 in young leaf, 18 in mature leaf, 28 in old leaf, 25 in stem, 32 in flower, 29 in fruit and 42 in root (**Table 2**; **Supplementary Table S5**), these genes might play unique roles in the corresponding tissues. Besides, a total of 6589 ECs was identified, of which 487, 416, 505, 33, 582, 1855, 525, and 2186 were specifically expressed in apical bud, young leaf, mature leaf, old leaf, stem, flower, fruit and root, respectively (**Table 2**). It was clear that CSGs (53%) were more likely to express in specific tissues than ECs (16.11%).

To explore the potential relationship between CSG and environmental adaptation, we analyzed the expression of CSGs under salt and drought stress. A total of 12 CSGs were found to be stimulated by environmental stress compared to CK treatment under the criteria of ≥1.5-fold expression differential and the false discovery rate (FDR) < 0.01. There were 5 and 4 CSGs responded to salt and drought, respectively (**Supplementary Table S6**). Surprisingly, among these genes, *CSS0030246*, *CSS0002298,* and *CSS0030939* responded to both types of stresses (**Supplementary Table S6**), suggesting that these three genes probably function importantly roles in stress tolerance.

CSGs Function Prediction

Since it was impossible to infer the function of CSGs through homologous genes, but CSGs were specifically expressed in different tissues (**Table 2**). We used WGCNA, a method to identify synergistic gene modules, to further analyze the potential functions of CSGs in different tissues. We identified 17 modules. Treating different tissues as traits, we screened the most optimally correlated modules for the characteristic vector genes and phenotypes and plotted a heat map of module-trait relationships. We finally identified seven modules with extremely strong positive correlation with the trait, and the correlation coefficient (CC) between MEcyan module and apical bud reaches 0.72 (p-value = 0.04), MEbisque4 and mature leaf (CC = 0.94, p-value = 5×10^{-4}), MEred and old leaf (CC = 0.9, p-value = 0.002), MEdarkorange2 and stem (CC = 0.88, p-value = 0.004), MEblue and flower (CC = 1, p-value = 3×10^{-8}), MEdarkmagenta and fruit (CC = 0.91, p-value = 0.001) and MEturquoise and root (CC = 1, p-value = 4×10^{-8}), respectively (**Figure 7**). The Pearson correlation coefficients (PCC) were calculated to derive seven tissue-specific modules (**Figures 8A–G**). We then identified 3187 hub genes in seven modules after screening (**Supplementary Table S7**), among them, including 18 CSGs. In MEbisque4 (mature leaf), there were 148 hub genes, including one CSGs. In MEred model (old leaf), there were 210 hub genes, including 2 CSGs. In MEblue model (flower), there were 1,108 hub genes, including 9 CSGs. In MEturquoise model (root), there were 140 hub genes, including 6 CSGs (**Supplementary Table S8**). These four modules were immediately subjected to KEGG enrichment analysis. In MEbisque4 (mature leaf), it is mainly enriched in biosynthesis of amino acids (ko01230), photosynthesis (ko00195), phenylalanine, tyrosine and tryptophan biosynthesis (ko00400), pentose phosphate pathway (ko00030) and carbon fixation in photosynthetic organisms (ko00710) (**Figure 9A**). In MEred model (old leaf), it is primarily affluent in pentose phosphate pathway (ko00030), carbon fixation in photosynthetic organisms (ko00710), folate biosynthesis (ko00790), carbon metabolism (ko01200), galactose metabolism (ko00052) and carotenoid biosynthesis (ko00906) (**Figure 9B**). In MEblue model (flower), pentose and glucuronate interconversions(ko00040), starch and sucrose metabolism (ko00500), galactose metabolism (ko00052), plant hormone signal transduction (ko04075) and alpha-Linolenic acid metabolism (ko00592) were enriched (**Figure 9C**). In MEturquoise model (root), it is mainly enriched phenylpropanoid biosynthesis (ko00940), plant-pathogen interaction (ko04626), glutathione metabolism (ko00480), pentose and glucuronate interconversions (ko00040) (**Figure 9D**).

TABLE 2 | Tissue expression pattern of *camellia*-specific genes (CSGs) and evolutionary conserved genes (ECs).

Items	Apical bud	Flower	Fruit	Young leaf	Mature leaf	Old leaf	Root	Stem	Total
With tissue-specific expression									
Number of ASGs (%)	17 (8.02)	32 (15.09)	29 (13.68)	21 (9.91)	18 (8.49)	28 (13.21)	42 (19.81)	25 (11.79)	212 (100)
Number of ECs (%)	487 (7.39)	1855 (28.15)	525 (7.97)	416 (6.31)	505 (7.66)	33 (0.5)	2186 (33.18)	582 (8.83)	6589 (100)
With high expression abundance (FPKM >2)									
Number of ASGs (%)	64(10.96)	78(13.36)	72(12.33)	69(11.82)	75(12.84)	65(11.13)	84(14.38)	77(13.18)	584(100)
Number of ECs (%)	24742(13.02)	22251(11.71)	23643(12.44)	24265(12.77)	23830(12.54)	21315(11.22)	24370(12.82)	25639(13.49)	190055(100)

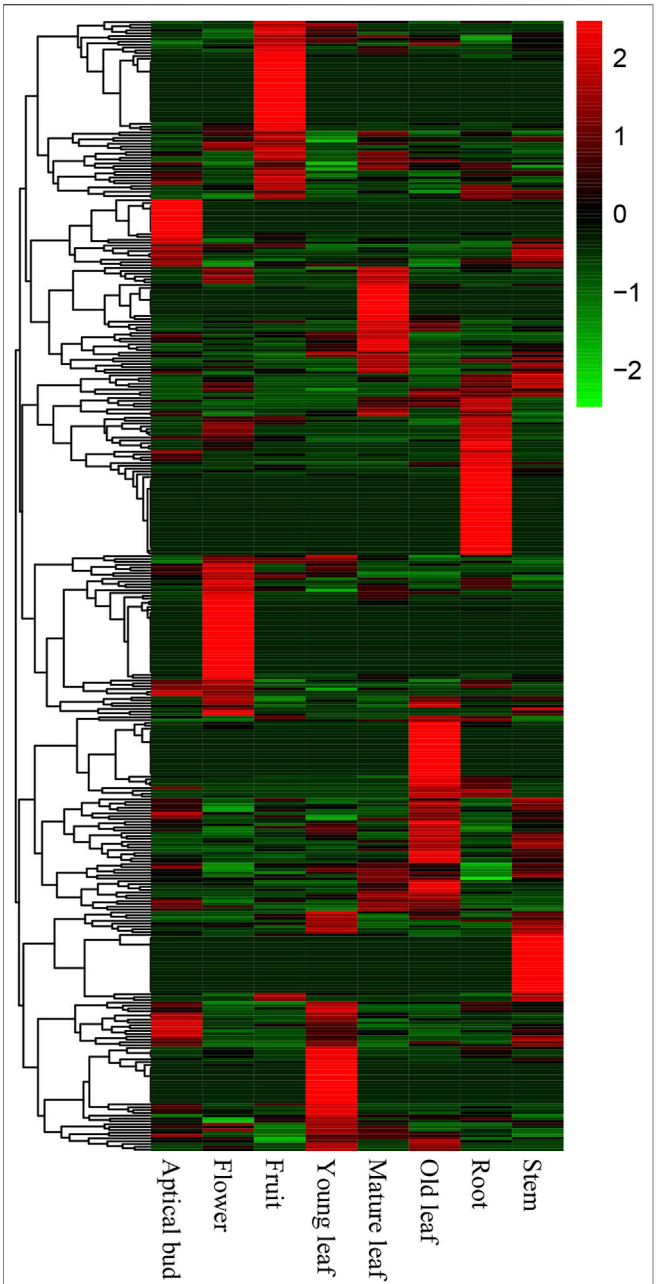

FIGURE 6 | Expression pattern of *Camellia*-specific genes (CSGs) in different tissues includes apical bud, young leaf, mature leaf, old leaf, stem, flower, fruit, and root.

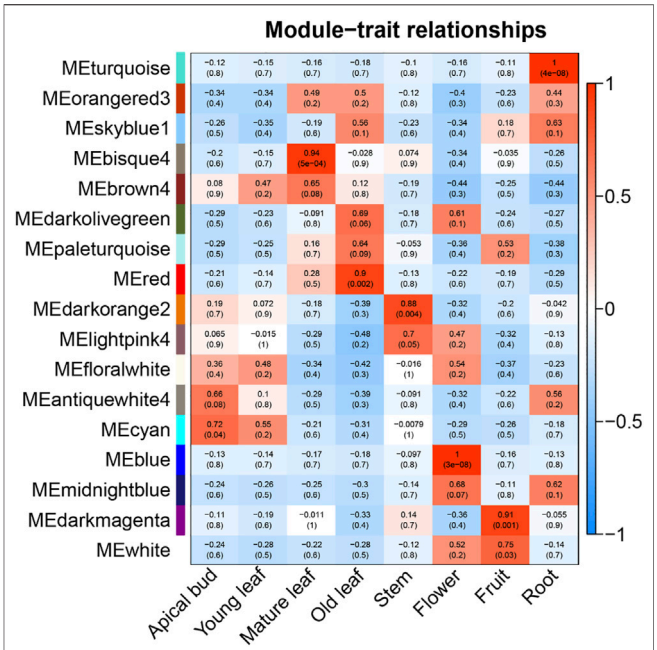

FIGURE 7 | Heat map of module-tissue relationship.

DISCUSSION

With the combination of genome sequencing with comparative analysis, enormous LSGs with potentially important functions have been identified in different species (Wilson et al., 2005; Zhang et al., 2007; Lin et al., 2010; Tsutsui, 2011), which motivated our genome wide exploration of LSGs within tea plant CSS. Before further analysis of LSGs, we need to identify LSGs first. Lin et al. identified 1324 LSGs in *A. thaliana* genome (Lin et al., 2010) and Ma et al. identified 3812 LSGs in wheat genome (Ma et al., 2020). Among this research, a grand sum of 1701 CSGs in the genome of CSS were identified, representing approximately 3.37% of the entire genome. This CSGs percentage was similar to the 4.9% found in *A. thaliana* (Lin et al., 2010) and 3.2% in rice (Yang et al., 2009). Since we used the genomes of published homologous species to identify LSGs, the more abundant the genome data of reference species available, the more information we could annotate and the less the false positives would be, though the number of LSGs might decrease. Although there are still shortcomings in our currently available identification tools such as pseudogene exclusion, our study remains a vital step in exploring new genes in CSS genome, and the identification of CSGs will become more accurate.

Accumulating researches have showed that some characteristics of LSGs may be somewhat different compared to ECs in all species, such as gene size, length of protein, GC content and number of exons, mainly related to the mechanism of origin and evolutionary time of LSGs. To reveal whether these differences in genic characteristics exist between CSGs and ECs, the sequence structure of CSGs and ECs were compared and analyzed. The average size of LSGs is normally smaller than ECs (Campbell et al., 2007; Zhang et al., 2007; Cai and Petrov, 2010; Lin et al., 2010; Yang et al., 2013), our result in CSS also comply with this conclusion (**Table 1**). This phenomenon may be related to the fact that each CSG has fewer exons (**Table 1**). Besides, CSGs have shorter protein lengths (**Table 1**), consistent with the LSGs of other eukaryotes (Domazet-Loso, 2003; Campbell et al., 2007; Toll-Riera et al., 2009; Donoghue et al., 2011). One reason for such differences between CSGs and ECs may be that intronless genes can be created by retroposition, which has

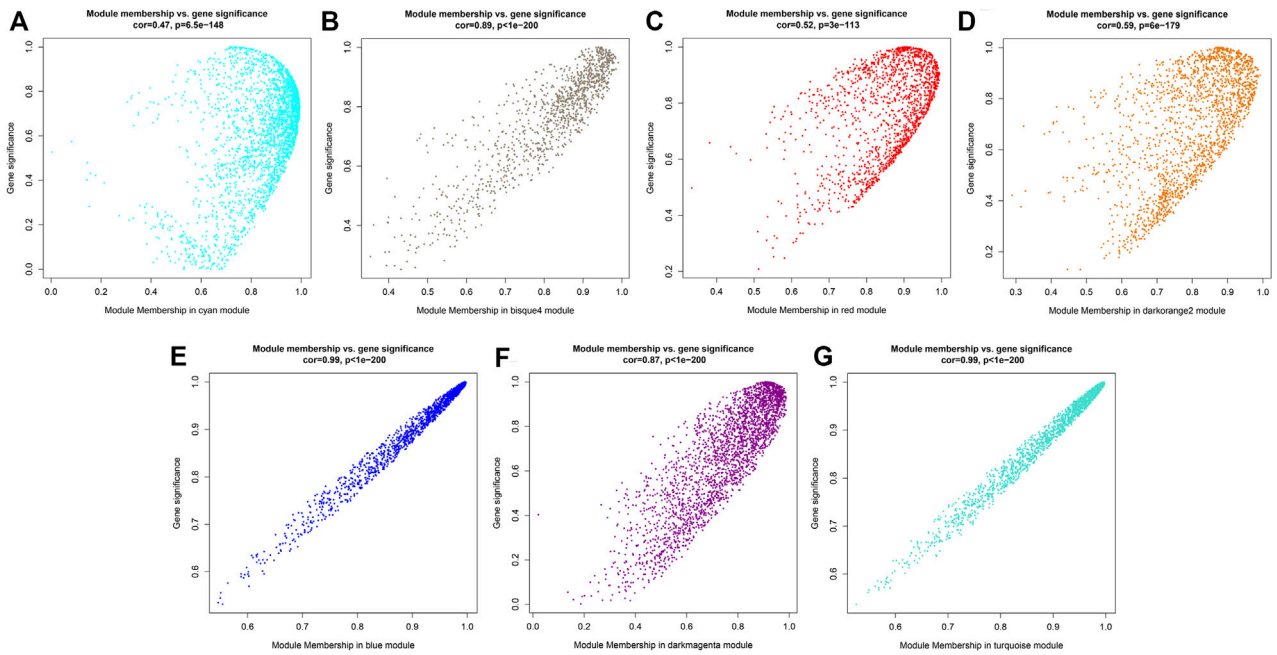

FIGURE 8 | Gene significance map. **(A)** Gene significance map for members of module MEcyan. **(B)** Gene significance map for members of module MEbisque4. **(C)** Gene significance map for members of module MEred. **(D)** Gene significance map for members of module MEdarkorange2. **(E)** Gene significance map for members of module MEblue. **(F)** Gene significance map for members of module MEdarkmagenta. **(G)** Gene significance map for members of module MEturquoise.

been shown to create a large number of LSGs in the zebrafish genome (Fu et al., 2010). Alternatively, this phenomenon may be a result of the "introns late" hypothesis, which suggests that the accretion of intron into the protein-coding genes is continuous during the evolution of eukaryotes (Koonin, 2006). As a result, younger genes have fewer exons. Furthermore, since LSGs are species specific, they generally have emerged in relatively recent years. In summary, these reasons may partly explain why LSGs have fewer numbers of exons per gene and why LSGs are shorter than ECs. On the other hand, CSGs has significantly higher GC content than ECs in CSS, consistent with the results in *Bombyx mori* (Sun et al., 2015) and zebrafish (Yang et al., 2013). This is consistent with the observation in previous studies that the enrichment of high GC content class usually occurs in genes lacking introns (Carels and Bernardi, 2000; Alexandrov et al., 2009). However, this property is not universal, the GC content of LSG in other species like *Triticeae* is lower than that of ECs (Ma et al., 2020). Differences in GC content are the result of a combination of factors such as the external environment and habits of organisms, and the possible mechanisms responsible for these significant differences still need further study (Carels and Bernardi, 2000; Galtier et al., 2001; Halder et al., 2017). The isoelectric point has been considered to alter the protein function and indirectly reflect the species-specific adaption made in response to the variable environment (Nandi et al., 2005). In this study, the isoelectric points of CSGs were found higher than those of ECs and the difference can indirectly reflect the species-specific applicability of CSS to the environment.

The mechanisms of the origin of LSGs are vital for explaining the origin and evolution of new phenotypes and ultimately the genetic basis of biodiversity. There are four main mechanisms explaining for the origin of LSGs including gene duplication, transposon pattern, gene overlap, and *de novo* origin (Kaessmann, 2010; Long et al., 2013), among which gene duplication was considered to be most predominant (Long et al., 2003; Kaessmann, 2010; Tautz and Domazet-Loso, 2011; Wissler et al., 2013). Tautz believed that LSGs are formed by sequence variation after gene replication, and because of the acceleration of evolution, this gene loses its sequence similarity with other species genes, and thus LSGs appear (Opazo and Storz, 2008; Tautz and Domazet-Loso, 2011; Kondrashov, 2012). In this study, we found that 971 CSGs in CSS were derived from gene duplication, occupying 57.08% of the total CSGs. We evaluated the duplication time of CSGs using *Ks* peaks, and the result showed concordance with the synchronization of the two WGD events in the CSS genome (Wei et al., 2018). In CSS genome, gene duplication had brought large impact on the evolution of genes associated with the biosynthesis of secondary metabolites that are essential for tea aroma and flavor, such as genes involved in the catechin biosynthesis pathway were mostly generated by gene duplication.

Due to the rapid development of sequencing technology, the study of LSGs is now no longer limited to sequence structure but exploring gene function. RNA-Seq is an effective way to characterize the expression schemas of CSGs among various tissues (Wang et al., 2009). Studies have shown that there is difference in the expression of LSGs in different tissues, usually with higher expression in the reproductive system in animals (Begun et al., 2007; Chen et al., 2020) and also in plant tissues such as mature pollen (Wu et al., 2014) and callus (Xu et al.,

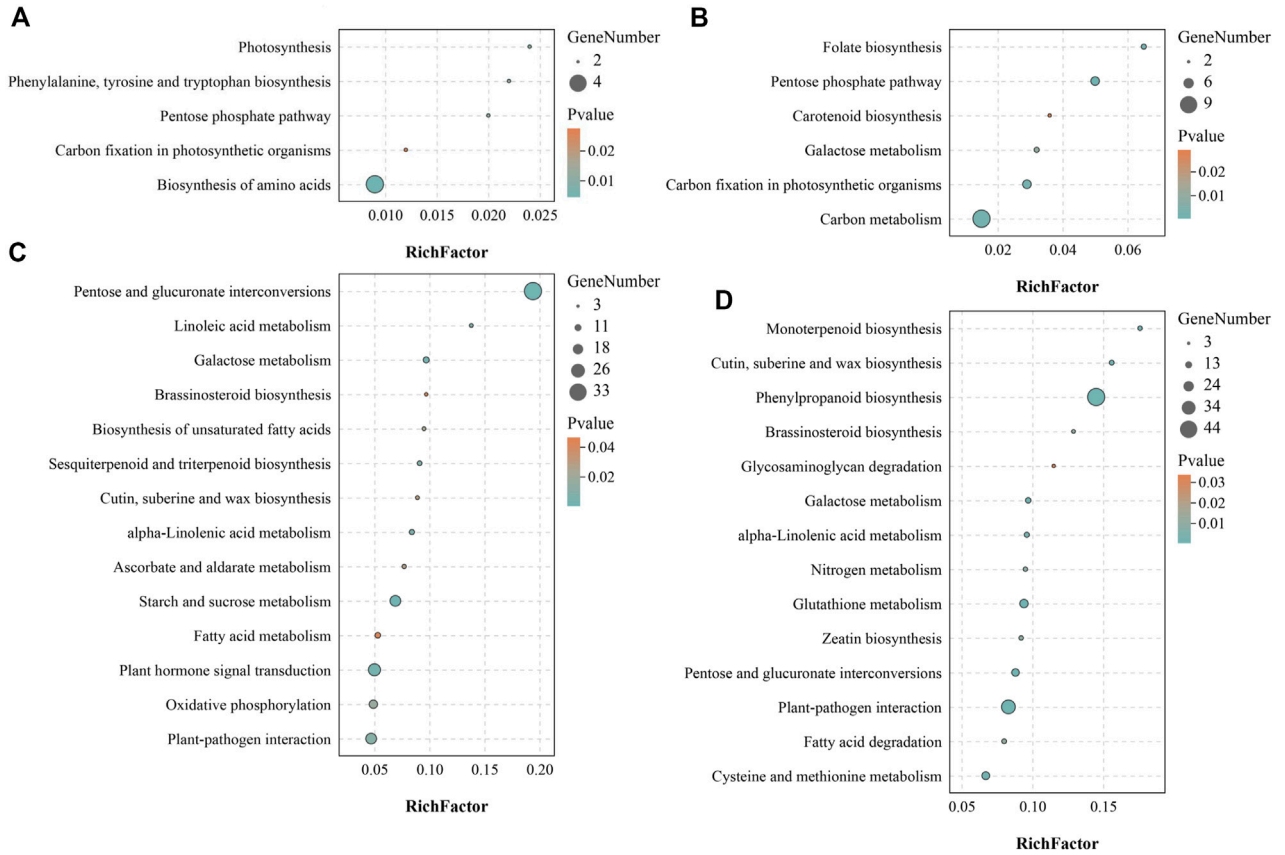

FIGURE 9 | KEGG enrichment analysis of four tissue-specific modules. **(A)** KEGG enrichment analysis result of MEbisque4 module genes. **(B)** KEGG enrichment analysis result of MEred module genes. **(C)** KEGG enrichment analysis result of MEblue module genes. **(D)** KEGG enrichment analysis result of MEturquoise module genes.

2015). In this study, 212 genes were found to have significant tissue specific expression. There were 17, 32 and 29 CSGs expressed only in reproductive organs including apical bud, flowers and fruits, respectively, and 21, 18, 28, 25, and 42 genes in trophic organs including young leaves, mature leaves, old leaves, stems and roots, respectively (**Table 2**), implying that most CSGs play an important role in reproductive development. Besides, some LSGs have been reported to be important for tackling with extreme environmental conditions like cold, drought, heat and salt stress according to previous studies (Bosch et al., 2009; Yang et al., 2009; Donoghue et al., 2011). For CSS, both salinity and drought constitute severe challenges that significantly affect the production and qualities of CSS. We here checked the expression of CSGs under salt and drought stress, and observed that 12 genes had been stimulated, indicating that these 12 stress-responsive CSGs may be related to adaption to the extreme environmental conditions (**Supplementary Table S6**). CSS0030246, CSS0002298, CSS0018115, CSS0048226, and CSS0006611 were down regulated under 24 h salt stress. CSS0030246 was down regulated under 48 h salt stress. CSS0030939 and CSS0038744 were down regulated under 72 h salt stress. CSS0040193 was up regulated under 72 h salt stress (**Supplementary Table S6**). Over-expression of CSS0040193 may be associated with the tolerance of CSS to salinity. At the same time, CSS0002298, CSS0023764, CSS0046868, CSS0005736 and CSS0027450 were down regulated under 24 h drought stress. CSS0030246 and CSS0030939 were down regulated under 48 and 72 h drought stress, respectively (**Supplementary Table S6**). Interestingly, CSS0030246, CSS0002298 and CSS0030939 responded to both salt and drought stress, which may be candidates for further studying environmental adaptation in CSS.

Since having no homologous genes related in other species, the possible expression characteristics and functions of CSGs cannot be inferred by homology comparisons. However, we can infer the possible biological processes involved in CSGs by means of the co-expressed gene modules. In this study, we identified 18 CSGs in 4 tissue-specific modules with WGCNA (**Supplementary Table S8**), and identified that these co-expression gene modules were predominantly involved in phenylalanine biosynthesis, biosynthesis of amino acids, pentose phosphate pathway, photosynthesis and carbon fixation in photosynthetic organisms with KEGG analysis (**Figure 9**). Catechins are the main components of polyphenolic substances in tea leaves, which determine the unique aroma and flavor of tea. At the same time, the biological activity of catechins is of great significance to the prevention of various diseases and human health, such as the

suppression of postprandial hypertriacylglycerolemia (Ikeda et al., 2005) and the prevention and therapy of cancer (Yiannakopoulou, 2014). Catechins are synthesized by a series of complex metabolic pathways, notably the flavonoids synthesis pathway, the pentose phosphate pathway and the shikimic acid pathway (Eungwanichayapant and Popluechai, 2009). The phenylpropane pathway is the starting pathway for flavonoids metabolism in plants, and the phenylpropane pathway uses phenylalanine as starting substrate(Lepiniec et al., 2006), indicating that the CSGs we identified play a crucial role in the synthesis of catechins and the formation of the specific flavor of CSS. In addition, photosynthesis not only determines the growth and productivity of plants, but also has a great impact on secondary metabolic pathways. Studies have shown that the biosynthesis of catechins in tea plants is regulated by light and its content is negatively correlated with chlorophyll concentration (Li et al., 2016; Hwa et al., 2019). The higher the chlorophyll content under low light, the lower the catechin content (Li et al., 2016; Hwa et al., 2019). In conclusion, CSGs involved in photosynthesis and carbon fixation are closely related to the productivity and quality of CSS.

CONCLUSION

In this study, we identified 1701 CSGs from the CSS genome, accounting for 3.37% of the genome. Through structural characterizations analysis, we found that CSGs had shorter protein length, higher GC content and isoelectric point compared to ECs. Analysis of the origin of 1701 CSGs showed that 971 CSGs were derived from gene duplication, making up 57.08% of total CSGs. Besides, most CSGs were found mainly localized in the nucleus, extracellular space and chloroplasts. Gene expression analysis revealed tissue-specific expression of CSGs. The results of WGCNA showed that CSGs were mainly involved in pathways such as phenylalanine, tyrosine and tryptophan biosynthesis, pentose phosphate pathway, biosynthesis of amino acids, photosynthesis and carbon fixation. In addition, the expression of some CSGs was associated with stress tolerance. In conclusion, this study has provided a basis for studying the specific genetic resources of tea and provides some clues for future interpretation of the role played by tea LSGs in tea-specific features.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

DM designed and supervised the study. DM and ZZ analyzed the data and drafted the manuscript. DM and ZZ critically revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We appreciate State Key Laboratory of Tea Plant Biology and Utilization Anhui Agricultural University for providing their valuable databases in public.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.770570/full#supplementary-material

REFERENCES

Alexandrov, N. N., Brover, V. V., Freidin, S., Troukhan, M. E., Tatarinova, T. V., Zhang, H., et al. (2009). Insights into Corn Genes Derived from Large-Scale cDNA Sequencing. *Plant Mol. Biol.* 69 (1-2), 179–194. doi:10.1007/s11103-008-9415-4

Begun, D. J., Lindfors, H. A., Kern, A. D., and Jones, C. D. (2007). Evidence for De Novo Evolution of Testis-Expressed Genes in the *Drosophila yakuba/Drosophila erecta* Clade. *Genetics* 176 (2), 1131–1137. doi:10.1534/genetics.106.069245

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* 30 (15), 2114–2120. doi:10.1093/bioinformatics/btu170

Bosch, T. C. G., Augustin, R., Anton-Erxleben, F., Fraune, S., Hemmrich, G., Zill, H., et al. (2009). Uncovering the Evolutionary History of Innate Immunity: The Simple Metazoan Hydra Uses Epithelial Cells for Host Defence. *Developmental Comp. Immunol.* 33 (4), 559–569. doi:10.1016/j.dci.2008.10.004

Cai, J. J., and Petrov, D. A. (20102010). Relaxed Purifying Selection and Possibly High Rate of Adaptation in Primate Lineage-specific Genes. *Genome Biol. Evol.* 22 (1), 2016393–2016409. doi:10.1093/gbe/evq019

Campbell, M. A., Zhu, W., Jiang, N., Lin, H., Ouyang, S., Childs, K. L., et al. (2007). Identification and Characterization of Lineage-specific Genes within the Poaceae. *Plant Physiol.* 145 (4), 1311–1322. doi:10.1104/pp.107.104513

Carels, N., and Bernardi, G. (2000). Two Classes of Genes in Plants. *Genetics* 154 (4), 1819–1825. doi:10.1093/genetics/154.4.1819

Chen, K., Tian, Z., Chen, P., He, H., Jiang, F., and Long, C.-a. (2020). Genome-wide Identification, Characterization and Expression Analysis of Lineage-specific Genes within Hanseniaspora Yeasts. *FEMS Microbiol. Lett.* 367, fnaa077. doi:10.1093/femsle/fnaa077

Chen, S., Ni, X., Krinsky, B. H., Zhang, Y. E., Vibranovski, M. D., White, K. P., et al. (2012). Reshaping of Global Gene Expression Networks and Sex-Biased Gene Expression by Integration of a Young Gene. *Embo J.* 31 (12), 2798–2809. doi:10.1038/emboj.2012.108

Domazet-Loso, T. (2003). An Evolutionary Analysis of Orphan Genes in *Drosophila*. *Genome Res.* 13 (10), 2213–2219. doi:10.1101/gr.1311003

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., and Spillane, C. (2011). Evolutionary Origins of Brassicaceae Specific Genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* 11, 47. doi:10.1186/1471-2148-11-47

Dujon, B. (1996). The Yeast Genome Project: what Did We Learn? *Trends Genet.* 12 (7), 263–270. doi:10.1016/0168-9525(96)10027-5

Eungwanichayapant, P. D., and Popluechai, S. (2009). Accumulation of Catechins in tea in Relation to Accumulation of mRNA from Genes Involved in Catechin Biosynthesis. *Plant Physiol. Biochem.* 47 (2), 94–97. doi:10.1016/j.plaphy.2008.11.002

Fischer, D., and Eisenberg, D. (1999). Finding Families for Genomic ORFans. *Bioinformatics* 15 (9), 759–762. doi:10.1093/bioinformatics/15.9.759

Fu, B., Chen, M., Zou, M., Long, M., and He, S. (2010). The Rapid Generation of Chimerical Genes Expanding Protein Diversity in Zebrafish. *Bmc Genomics* 11. doi:10.1186/1471-2164-11-657

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content Evolution in Mammalian Genomes: the Biased Gene Conversion Hypothesis. *Genetics* 159 (2), 907–911. doi:10.1093/genetics/159.2.907

Gaut, B. S., Morton, B. R., McCaig, B. C., and Clegg, M. T. (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.* 93 (19), 10274–10279. doi:10.1073/pnas.93.19.10274

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length Transcriptome Assembly from RNA-Seq Data without a Reference Genome. *Nat. Biotechnol.* 29 (7), 644–652. doi:10.1038/nbt.1883

Graham, M. A., Silverstein, K. A. T., Cannon, S. B., and VandenBosch, K. A. (2004). Computational Identification and Characterization of Novel Genes from Legumes. *Plant Physiol.* 135 (3), 1179–1197. doi:10.1104/pp.104.037531

Halder, B., Malakar, A. K., Malakar, A. K., and Chakraborty, S. (2017). Nucleotide Composition Determines the Role of Translational Efficiency in Human Genes. *Bioinformation* 13 (2), 46–53. doi:10.6026/97320630013046

Ikeda, I., Tsuda, K., Suzuki, Y., Kobayashi, M., Unno, T., Tomoyori, H., et al. (2005). Tea Catechins with a Galloyl Moiety Suppress Postprandial Hypertriacylglycerolemia by Delaying Lymphatic Transport of Dietary Fat in Rats. *J. Nutr.* 135 (2), 155–159. doi:10.1093/jn/135.2.155

Kaessmann, H. (2010). Origins, Evolution, and Phenotypic Impact of New Genes. *Genome Res.* 20 (10), 1313–1326. doi:10.1101/gr.101386.109

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T. C. G. (2009). More Than Just Orphans: Are Taxonomically-Restricted Genes Important in Evolution? *Trends Genet.* 25 (9), 404–413. doi:10.1016/j.tig.2009.07.006

Kondrashov, F. A. (2012). Gene Duplication as a Mechanism of Genomic Adaptation to a Changing Environment. *Proc. R. Soc. B.* 279 (1749), 5048–5057. doi:10.1098/rspb.2012.1108

Koonin, E. V. (2006). The Origin of Introns and Their Role in Eukaryogenesis: a Compromise Solution to the Introns-Early versus Introns-Late Debate? *Biol. Direct* 1, 22. doi:10.1186/1745-6150-1-22

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Lee, L.-S., Kim, S.-H., Kim, Y.-B., and Kim, Y.-C. (2014). Quantitative Analysis of Major Constituents in Green Tea with Different Plucking Periods and Their Antioxidant Activity. *Molecules* 19 (7), 9173–9186. doi:10.3390/molecules19079173

Lemos, B., Bettencourt, B. R., Meiklejohn, C. D., and Hartl, D. L. (2005). Evolution of Proteins and Gene Expression Levels Are Coupled in *Drosophila* and Are Independently Associated with mRNA Abundance, Protein Length, and Number of Protein-Protein Interactions. *Mol. Biol. Evol.* 22 (5), 1345–1354. doi:10.1093/molbev/msi122

Lepiniec, L., Debeaujon, I., Routaboul, J.-M., Baudry, A., Pourcel, L., Nesi, N., et al. (2006). Genetics and Biochemistry of Seed Flavonoids. *Annu. Rev. Plant Biol.* 57, 405–430. doi:10.1146/annurev.arplant.57.032905.105252

Li, L., Zheng, W., Zhu, Y., Ye, H., Tang, B., Arendsee, Z. W., et al. (2015). *QQS* Orphan Gene Regulates Carbon and Nitrogen Partitioning across Species via NF-YC Interactions. *Proc. Natl. Acad. Sci. USA* 112 (47), 14734–14739. doi:10.1073/pnas.1514670112

Li, Z.-X., Yang, W.-J., Ahammed, G. J., Shen, C., Yan, P., Li, X., et al. (2016). Developmental Changes in Carbon and Nitrogen Metabolism Affect tea Quality in Different Leaf Position. *Plant Physiol. Biochem.* 106, 327–335. doi:10.1016/j.plaphy.2016.06.027

Lin, H., Moghe, G., Ouyang, S., Iezzoni, A., Shiu, S. H., Gu, X., et al. (2010). Comparative Analyses Reveal Distinct Sets of Lineage-specific Genes within *Arabidopsis thaliana*. *BMC Evol. Biol.* 10, 41. doi:10.1186/1471-2148-10-41

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The Origin of New Genes: Glimpses from the Young and Old. *Nat. Rev. Genet.* 4 (11), 865–875. doi:10.1038/nrg1204

Long, M., Vankuren, N. W., Chen, S., and Vibranovski, M. D. (2013). New Gene Evolution: Little Did We Know. *Annu. Rev. Genet.* 47 (1), 307–333. doi:10.1146/annurev-genet-111212-133301

Ma, D., Guo, Z., Ding, Q., Zhao, Z., Shen, Z., Wei, M., et al. (2021). Chromosome-level Assembly of the Mangrove Plant *Aegiceras corniculatum* Genome Generated through Illumina, PacBio and Hi-C Sequencing Technologies. *Mol. Ecol. Resour.* 21, 1593–1607. doi:10.1111/1755-0998.13347

Ma, S., Yuan, Y., Tao, Y., Jia, H., and Ma, Z. (2020). Identification, Characterization and Expression Analysis of Lineage-specific Genes within Triticeae. *Genomics* 112 (2), 1343–1350. doi:10.1016/j.ygeno.2019.08.003

Nandi, S., Mehra, N., Lynn, A. M., and Bhattacharya, A. (2005). Comparison of Theoretical Proteomes: Identification of COGs with Conserved and Variable pI within the Multimodal pI Distribution. *BMC Genomics* 6 (1), 116–128. doi:10.1186/1471-2164-6-116

Opazo, F. G. H. J. C., and Storz, J. F. (2008). Rapid Rates of Lineage-specific Gene Duplication and Deletion in the α-Globin Gene Family. *Mol. Biol. Evol.* 25 (3), 591–602.

Pan, J.-B., Hu, S.-C., Wang, H., Zou, Q., and Ji, Z.-L. (2012). PaGeFinder: Quantitative Identification of Spatiotemporal Pattern Genes. *Bioinformatics* 28 (11), 1544–1545. doi:10.1093/bioinformatics/bts169

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene Duplication and Evolution in Recurring Polyploidization-Diploidization Cycles in Plants. *Genome Biol.* 20 (1), 38. doi:10.1186/s13059-019-1650-2

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., and Jones, C. D. (2013). De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. *Plos Genet.* 9 (10), e1003860. doi:10.1371/journal.pgen.1003860

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018). BUSCA: an Integrative Web Server to Predict Subcellular Localization of Proteins. *Nucleic Acids Res.* 46 (W1), W459–W466. doi:10.1093/nar/gky320

Song, K. E., Jeon, S. H., Shim, D. B., Jun, W. J., Chung, J. W., and Shim, S. (2019). Strong Solar Irradiance Reduces Growth and Alters Catechins Concentration in tea Plants over winter. *J. Crop Sci. Biotechnol.* 22 (5), 475–480. doi:10.1007/s12892-019-0215-0

Song, R., Kelman, D., Johns, K. L., and Wright, A. D. (2012). Correlation between Leaf Age, Shade Levels, and Characteristic Beneficial Natural Constituents of tea (Camellia Sinensis) Grown in Hawaii. *Food Chem.* 133 (3), 707–714. doi:10.1016/j.foodchem.2012.01.078

Sun, W., Zhao, X.-W., and Zhang, Z. (2015). Identification and Evolution of the Orphan Genes in the Domestic Silkworm,*Bombyx mori*. *FEBS Lett.* 589 (19), 2731–2738. doi:10.1016/j.febslet.2015.08.008

Tautz, D., and Domazet-Lošo, T. (2011). The Evolutionary Origin of Orphan Genes. *Nat. Rev. Genet.* 12 (10), 692–702. doi:10.1038/nrg3053

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., et al. (2009). Origin of Primate Orphan Genes: a Comparative Genomics Approach. *Mol. Biol. Evol.* 26 (3), 603–612. doi:10.1093/molbev/msn281

Tsutsui, J. (2011). Taxonomically Restricted Genes Are Associated with the Evolution of Sociality in the Honey Bee. *BMC Genomics* 12, 164. doi:10.1186/1471-2164-12-164

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* 10 (1), 57–63. doi:10.1038/nrg2484

Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., et al. (2018). Draft Genome Sequence of *Camellia Sinensis* Var. *Sinensis* Provides Insights into the Evolution of the tea Genome and tea Quality. *Proc. Natl. Acad. Sci. U S A.* 115 (18), E4151. doi:10.1073/pnas.1719622115

Wilson, G. A., Bertrand, N., Patel, Y., Hughes, J. B., Feil, E. J., and Field, D. (2005). Orphans as Taxonomically Restricted and Ecologically Important Genes. *Microbiology (Reading)* 151 (Pt 8), 2499–2501. doi:10.1099/mic.0.28146-0

Wissler, L., Gadau, J., Simola, D. F., Helmkampf, M., and Bornberg-Bauer, E. (2013). Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes. *Genome Biol. Evol.* 5 (2), 439–455. doi:10.1093/gbe/evt009

Wu, D.-D., Irwin, D. M., and Zhang, Y.-P. (2011). *De Novo* origin of Human Protein-Coding Genes. *Plos Genet.* 7 (11), e1002379. doi:10.1371/journal.pgen.1002379

Wu, D.-D., Wang, X., Li, Y., Zeng, L., Irwin, D. M., and Zhang, Y.-P. (2014). "Out of Pollen" Hypothesis for Origin of New Genes in Flowering Plants: Study from *Arabidopsis thaliana*. *Genome Biol. Evol.* 6 (10), 2822–2829. doi:10.1093/gbe/evu206

Xia, X. (2018). DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Mol. Biol. Evol.* 35 (6), 1550–1552. doi:10.1093/molbev/msy073

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., et al. (2009). A Rice Gene of *De Novo* Origin Negatively Regulates Pathogen-Induced Defense Response. *Plos One* 4 (2), e4603. doi:10.1371/journal.pone.0004603

Xu, Y., Wu, G., Hao, B., Chen, L., Deng, X., and Xu, Q. (2015). Identification, Characterization and Expression Analysis of Lineage-specific Genes within Sweet orange (*Citrus Sinensis*). *Bmc Genomics* 16, 995. doi:10.1186/s12864-015-2211-z

Yan, H., Zhang, W., Lin, Y., Dong, Q., Peng, X., Jiang, H., et al. (2014). Different Evolutionary Patterns Among Intronless Genes in maize Genome. *Biochem. Biophysical Res. Commun.* 449 (1), 146–150. doi:10.1016/j.bbrc.2014.05.008

Yang, L., Zou, M., Fu, B., and He, S. (2013). Genome-wide Identification, Characterization, and Expression Analysis of Lineage-specific Genes within Zebrafish. *BMC Genomics* 14 (1), 65. doi:10.1186/1471-2164-14-65

Yang, X., Jawdy, S., Tschaplinski, T. J., and Tuskan, G. A. (2009). Genome-wide Identification of Lineage-specific Genes in *Arabidopsis*, Oryza and Populus. *Genomics* 93 (5), 473–480. doi:10.1016/j.ygeno.2009.01.002

Yiannakopoulou, E. C. (2014). Effect of green tea Catechins on Breast Carcinogenesis. *Eur. J. Cancer Prev.* 23 (2), 84–89. doi:10.1097/CEJ.0b013e328364f23e

Zhang, G., Wang, H., Shi, J., Wang, X., Zheng, H., Wong, G. K.-S., et al. (2007). Identification and Characterization of Insect-specific Proteins by Genome Data Analysis. *BMC Genomics* 8 (1), 93. doi:10.1186/1471-2164-8-93

Zhang, J. (2003). Evolution by Gene Duplication: an Update. *Trends Ecol. Evol.* 18 (6), 292–298. doi:10.1016/s0169-5347(03)00033-8

Zhang, Q., Cai, M., Yu, X., Wang, L., Guo, C., Ming, R., et al. (2017). Transcriptome Dynamics of *Camellia Sinensis* in Response to Continuous Salinity and Drought Stress. *Tree Genet. Genomes* 13 (4), 78. doi:10.1007/s11295-017-1161-9

Zhao, D.-S., Li, Q.-F., Zhang, C.-Q., Zhang, C., Yang, Q.-Q., Pan, L.-X., et al. (2018). *GS9* Acts as a Transcriptional Activator to Regulate rice Grain Shape and Appearance Quality. *Nat. Commun.* 9, 1240. doi:10.1038/s41467-018-03616-y

Zhu, M.-z., Lu, D.-m., Ouyang, J., Zhou, F., Huang, P.-f., Gu, B.-z., et al. (2020). Tea Consumption and Colorectal Cancer Risk: a Meta-Analysis of Prospective Cohort Studies. *Eur. J. Nutr.* 59 (8), 3603–3615. doi:10.1007/s00394-020-02195-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.